MDL-BASED IDENTIFICATION OF RELEVANT TEMPORAL SCALES IN TIME SERIES

Ugo Vespier[1]*, Arno Knobbe*[1]*, Siegfried Nijssen*[2] *and Joaquin Vanschoren*[1]

[1]LIACS, Leiden University, the Netherlands
[2]Katholieke Universiteit Leuven, Belgium
uvespier@liacs.nl

ABSTRACT

The behavior of many complex physical systems is affected by a variety of phenomena occurring at different temporal scales. Time series data produced by measuring properties of such systems often mirrors this fact by appearing as a composition of signals across different time scales. When the final goal of the analysis is to model the individual phenomena affecting a system, it is crucial to be able to recognize the right temporal scales and to separate the individual components of the data. We introduce a solution to this challenge based on a combination of the Minimum Description Length (MDL) principle, feature selection strategies, and convolution techniques from the signal processing field. As a result, we show that our algorithm produces a good decomposition of a given time series and, as a side effect, builds a compact representation of its identified components.

1. INTRODUCTION

Our work [5] is concerned with the analysis of sensor data. When monitoring complex physical systems over time, one often finds multiple phenomena in the data that work on different time scales. If one is interested in analyzing and modeling these individual phenomena, it is crucial to recognize these different scales and separate the data into its underlying components. Here, we present a method for extracting the time scales of various phenomena present in large time series.

The need for analyzing time series data at multiple time scales is nicely demonstrated by a large monitoring project in the Netherlands, called *InfraWatch* [4]. In this project, we employ a range of sensors to measure the dynamic response of a large Dutch highway bridge to varying traffic and weather conditions. When viewing this data (see Fig. 1, upper plot), one can easily distinguish various *transient events* in the signal that occur on different time scales. Most notable are the gradual change in strain over the course of the day (as a function of the outside temperature, which influences stiffness parameters of the concrete), a prolonged increase in strain caused by rush hour traffic congestion, and individual bumps in the signal due to cars and trucks traveling over the bridge. In order to understand the various changes in the sensor signal, one would benefit substantially from separating out the events at various scales. The main goal of the work described here is to do just that: we consider the temporal data as a series of superimposed effects at different time scales, establish at which scales events most often occur, and from this we extract the underlying signal components.

We approach the scale selection problem from a Minimum Description Length [1] (MDL) perspective. The motivation for this is that we need a framework in which we can deal with a wide variety of representations for scale components. Our main assumption is that separating the original signal into components at different time scales will simplify the shape of the individual components, making it easier to model them separately. Our results show that, indeed, these multiple models outperform (in terms of MDL score) a single model derived from the original signal. While introducing multiple models incurs the penalty of having to describe them, there are much fewer 'exceptions' to be described compared to the single model, yielding a lower overall description length.

The analysis of time scales in time series data is often approached from a *scale-space* perspective, which involves convolution of the original signal with Gaussian kernels of increasing size [6] to remove information at smaller scales. By subtracting carefully selected components of the scale-space, we can effectively cut up the scale space into k ranges. In other words, signal processing offers methods for producing a large collection of derived features, and the challenge we face in this paper is how to select a subset of k features, such that the original signal is decomposed into a set of meaningful components at different scales.

Our approach applies the MDL philosophy to various aspects of modeling: choosing the appropriate scales at which to model the components, determining the optimal number of components (while avoiding overfitting on overly specific details of the data), and deciding which class of models to apply to each individual component. For this last decision, we propose two classes of models representing the components respectively on the basis of a discretization and a segmentation scheme. For this last scheme, we allow three levels of complexity to approximate the segments: piecewise constant approximations, piecewise linear approximations, as well as quadratic ones. These options result in different trade-offs between model cost and accuracy, depending on the type of signal we are dealing with.

A useful side product of our approach is that it identifies a concise representation of the original signal. This representation is useful in itself: queries run on the decomposed signal may be answered more quickly than when run on the original data. Furthermore, the parameters of the encoding may indicate useful properties of the data as well.

2. PRELIMINARIES

We deal with finite sequences of numerical measurements (samples), collected by observing some property of a system with a sensor, and represented in the form of time series as defined below.

Definition 1. *A **time series** of length n is a finite sequence of values $\boldsymbol{x} = x[1], \ldots, x[n]$ of finite precision.*[1] *A subsequence $\boldsymbol{x}[a:b]$ of $\boldsymbol{x}$ is defined as follows:*

$$\boldsymbol{x}[a:b] = (\boldsymbol{x}[a], \boldsymbol{x}[a+1], \ldots, \boldsymbol{x}[b]),\ a < b$$

We also assume that all the considered time series have no missing values and that their sampling rate is constant.

2.1. The Scale-Space Image

The *scale-space image* [6] is a scale parametrization technique for one-dimensional signals[2] based on the operation of convolution.

Definition 2. *Given a signal $\boldsymbol{x}$ of length n and a response function (kernel) $\boldsymbol{h}$ of length m, the result of the **convolution** $\boldsymbol{x} * \boldsymbol{h}$ is the signal $\boldsymbol{y}$ of length n, defined as:*

$$\boldsymbol{y}[t] = \sum_{j=-m/2+1}^{m/2} \boldsymbol{x}[t-j]\,\boldsymbol{h}[j]$$

In this paper, $\mathbf{h}$ is a Gaussian kernel with mean $\mu = 0$, standard deviation σ, area under the curve equal to 1, discretized into m values.[3]

Given a signal $\mathbf{x}$, the family of σ-smoothed signals $\Phi_{\mathbf{x}}$ over scale parameter σ is defined as follows:

$$\Phi_{\mathbf{x}}(\sigma) = \mathbf{x} * \mathbf{g}_{\sigma}\ ,\ \sigma > 0$$

where $\mathbf{g}_{\sigma}$ is a Gaussian kernel having standard deviation σ, and $\Phi_{\mathbf{x}}(0) = \mathbf{x}$.

The signals in $\Phi_{\mathbf{x}}$ define a surface in the time-scale plane (t, σ) known in the literature as the *scale-space image* [3, 6]. This visualization gives a complete description of the scale properties of a signal in terms of Gaussian smoothing. For practical purposes, the scale-space image is quantized across the scale dimension by computing the convolutions only for a finite number of scale parameters. More formally, for a given signal $\mathbf{x}$, we fix a set of scale parameters $S = \{2^i \mid 0 \leq i \leq \sigma_{max}\ \wedge i \in \mathbb{N}\}$ and we compute $\Phi_{\mathbf{x}}(\sigma)$ only for $\sigma \in S$ where σ_{max} is such that $\Phi_{\mathbf{x}}(\sigma)$ is approximately equal to the mean signal of $\mathbf{x}$.

[1] 32-bit floating point values in our experiments.

[2] From now on, we will use the term signal and time series interchangeably.

[3] To capture almost all non-zero values, we define $m = \lfloor 6\sigma \rfloor$.

2.2. Scale-Space Decomposition

We define a decomposition scheme of a signal $\mathbf{x}$ by considering adjacent ranges of scales of the signal scale-space image as below.

Definition 3. *Given a signal $\boldsymbol{x}$ and a set of $k-1$ scale parameters $C = \{\sigma_1, \ldots, \sigma_{k-1}\}$ (called the cut-point set) such that $\sigma_1 < \ldots < \sigma_{k-1}$, the **scale decomposition** of $\boldsymbol{x}$ is given by the set of component signals $D_{\boldsymbol{x}}(C) = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, defined as follows:*

$$\boldsymbol{x}_i = \begin{cases} \Phi_{\boldsymbol{x}}(0) - \Phi_{\boldsymbol{x}}(\sigma_1) & \text{if } i = 1 \\ \Phi_{\boldsymbol{x}}(\sigma_{i-1}) - \Phi_{\boldsymbol{x}}(\sigma_i) & \text{if } 1 < i < k \\ \Phi_{\boldsymbol{x}}(\sigma_{k-1}) & \text{if } i = k \end{cases}$$

Note that for k components we require $k-1$ cut-points.

3. MDL SCALE DECOMPOSITION SELECTION

Given an input signal $\mathbf{x}$, the main computational challenge we face is twofold:

- find a good subset of cut-points C such that the resulting k components of the decomposition $D_{\mathbf{x}}(C)$ optimally capture the effect of transient events at different scales,
- select a representation for each component, according to its inherent complexity.

We propose to use the Minimum Description Length (MDL) principle to approach this challenge. The two-part MDL principle states that the best model M to describe the signal $\mathbf{x}$ is the one that minimizes the sum of the description lengths $L(M) + L(\mathbf{x} \mid M)$.
The possible models depend on the scale decomposition $D_{\mathbf{x}}(C)$ considered[4] and on the representations used for its individual components. An ideal set of representations would adapt to the specific features of every single component, resulting in a concise summarization of the decomposition and, thus, of the signal. In order to apply the MDL principle, we need to define a model $M_{D_{\mathbf{x}}(C)}$ for a given scale decomposition $D_{\mathbf{x}}(C)$ and, consequently, how to compute both $L(M_{D_{\mathbf{x}}(C)})$ and $L(\mathbf{x} \mid M_{D_{\mathbf{x}}(C)})$. The latter term is the length in bits of the information lost by the model, i.e., the residual signal $\mathbf{x} - M_{D_{\mathbf{x}}(C)}$.

Note that, in order to employ MDL, we discretize the input signal $\mathbf{x}$. Below, we introduce the proposed representation schemes for the components. We also define the bit complexity of the residual and the model selection procedure.

3.1. Component Representation Schemes

Within our general framework, many different approaches could be used for representing the components of a decomposition. In the next paragraphs we introduce two such methods.

[4] Including the decomposition formed by zero cut-points ($C = \emptyset$), i.e., the signal itself.

3.1.1. Discretization-based representation

As a first representation, we propose to consider more coarse-grained discretizations of the original range of values. By doing so, similar values will be grouped together in the same bin. The resulting sequence of integers is compacted further by performing run-length encoding, resulting in a string of (v, l) pairs, where l represents the number of times value v is repeated consecutively. This string is finally encoded using a Shannon-Fano or Huffman code (see Section 3.2).

3.1.2. Segmentation-based representation

The main assumption on which we base this method is that a clear transient event can be accurately represented by a simple function, such as a polynomial of a bounded degree. Hence, if a signal contains a number of clear transient events, it should be possible to accurately represent this signal with a number of segments, each of which represented by a simple function.

Given a component $\mathbf{x}_i$ of length n, let

$$z(\mathbf{x}_i) = \{t_1, t_2, ..., t_m\}, \quad 1 < t_i \leq n$$

be a set of indexes of the segment boundaries.

Let $\texttt{fit}(\mathbf{x}_i[a:b], d_i)$ be the approximation of $\mathbf{x}_i[a:b]$ obtained by fitting a polynomial of degree d_i. Then, we represent each component $\mathbf{x}_i$ with the approximation $\hat{\mathbf{x}}_i$, such that:

$$\begin{array}{ll} \hat{\mathbf{x}}_i[0:z_1] & = \texttt{fit}(\mathbf{x}_i[0:z_1], d_i) \\ \hat{\mathbf{x}}_i[z_i:z_{i+1}] & = \texttt{fit}(\mathbf{x}_i[z_i:z_{i+1}], d_i), \quad 1 \leq i < m \\ \hat{\mathbf{x}}_i[z_m:n] & = \texttt{fit}(\mathbf{x}_i[z_m:n], d_i) \end{array}$$

Note that approximation $\hat{\mathbf{x}}_i$ is quantized again by reapplying the function Q to each of its values.

For a given k-component scale decomposition $D_{\mathbf{x}}(C)$ and a fixed polynomial degree for each of its components, we calculate the complexity in bits of the model $M_{D_{\mathbf{x}}(C)}$, based on this representation scheme, as follows. Each approximated component $\hat{\mathbf{x}}_i$ consists of $|z(\mathbf{x}_i)|+1$ segments. For each segment, we need to represent its length and the $d_i + 1$ coefficients of the fitted polynomial. The length ls_i of the longest segment in $\hat{\mathbf{x}}_i$ is given by

$$ls_i = \max(z_1 \cup \{z_{i+1} - z_i \mid 0 < i \leq m\})$$

We therefore use $\log_2(ls_i)$ bits to represent the segment lengths, while for the coefficients of the polynomials we employ floating point numbers of fixed[5] bit complexity c. The MDL model cost is thus defined, omitting minor terms, as:

$$L(M_{D_{\mathbf{x}}(C)}) = \sum_{i=1}^{k} \left(|z(\mathbf{x}_i)|+1\right)\left(\lceil \log_2(ls_i) \rceil + c\,(d_i+1)\right)$$

So far we assumed to have a set of boundaries $z(\mathbf{x}_i)$, but we did not specify how to compute them. A desirable property for our segmentation would be that a segmentation at a coarser scale does not contain more segments than a segmentation at a finer scale.

The scale space theory assures that there are fewer zero-crossing of the derivatives of a signal at coarser scales [6]. In our segmentation we use the zero-crossings of the first and second derivatives.

3.2. Residual Encoding

Given a model $M_{D_{\mathbf{x}}(C)}$, its residual $\mathbf{r} = \mathbf{x} - \sum_{i=1}^{k} \hat{\mathbf{x}}_i$, computed over the component approximations, represents the information of $\mathbf{x}$ not captured by the model. Having already defined the model cost for the two proposed encoding schemes, we only still need to define $L(\mathbf{x} \mid M_{D_{\mathbf{x}}(C)})$, i.e., a bit complexity $L(\mathbf{r})$ for the residual $\mathbf{r}$.

Here, we exploit the fact that we operate in a quantized space; we encode each bin in the quantized space with a code that uses approximately $-\log(P(x))$ bits, where $P(x)$ is the frequency of the xth bin in our data. The main justification for this encoding is that we expect that the errors are normally distributed around 0. Hence, the bins in the discretization that reflect a low error will have the highest frequency of occurrences; we will give these the shortest codes. In practice, ignoring small details, such codes can be obtained by means of Shannon-Fano coding or Huffman coding; as Hu et al. [2] we use Huffman coding in our experiments.

3.3. Model Selection

We can now define the MDL score that we are optimizing as follows:

Definition 4. *Given a model $M_{D_{\mathbf{x}}(C)}$, its **MDL score** is defined as:*

$$L(M_{D_{\mathbf{x}}(C)}) + L(\boldsymbol{r})$$

In the case of discretization-based encoding, the MDL score is affected by the cardinality used to encode each component. In the case of segmentation-based encoding the MDL score depends on the boundaries of the segments and the degrees of the polynomials in the representation. In both cases, also the cut-points of the considered decomposition affect the final score.

The simplest way to find the model that minimizes this score is to enumerate, encode and compute the MDL score for every possible scale-space decomposition and all possible encoding parameters. This brute-force approach results to be feasible in practice.

4. EXPERIMENTS

In this section, we experimentally evaluate our method actual sensor data from a real-world application. For a complete evaluation of the method, including a more systematic one over artificial data, please refer to [5].

We consider the strain measurements produced by a sensor attached to a large highway bridge in the Netherlands. The considered time series consists of 24 hours of strain measurements sampled at 1 Hz (totaling $86,400$

[5] In our experiments $c = 32$.

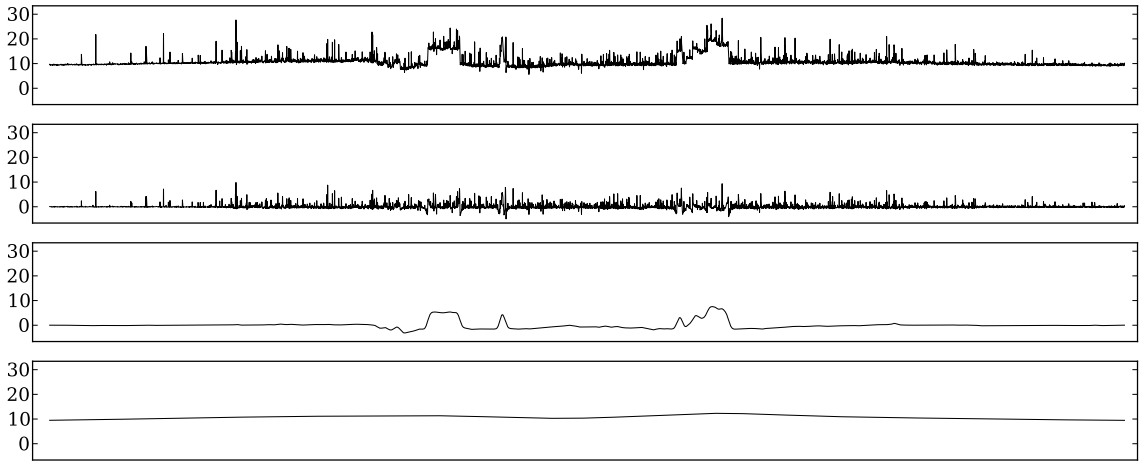

Figure 1: Signal (top) and top-ranked scale decomposition for the InfraWatch data.

data points). A plot of the data is shown in Figure 1 (topmost plot). We evaluated all the possible decompositions up to three components (two cut-points) allowing both the representation schemes we introduced. In the case of the discretization-based representations, we limit the possible cardinalities to 4, 16 and 64. The top-ranked decomposition results in 3 components as shown in the last three plots in Figure 1. The selected cut-points appear at scales $2^6 = 64$ and $2^{11} = 2048$. All three components are represented with the discretization-based scheme, with a cardinality of respectively 4, 16, and 16 symbols. The decomposition has an MDL-score of $344,276$, where $L(M) = 19,457$ and $L(D \mid M) = 324,818$. The found components accurately correspond to physical events on the bridge. The first component, covering scales lower than 2^6, reflects the short-term influence caused by passing vehicles and represented as peaks in the signal. Note that the cardinality selected for this component is the lowest admissible in our setting (4). This is reasonable considering that the relatively simple dynamic behavior occurring at these scales, mostly the presence or not of a peak over a flat baseline, can be cheaply described with 4 or fewer states without incurring a too large error. The middle component, covering scales between 2^6 and 2^{11}, reflects the medium-term effects caused by traffic jams. The first component is slightly influenced by the second one, especially at the start and ending points of a traffic jam. Finally, the third component captures all the scales greater than 2^{11}, here representing the effect of temperature during a whole day. To sum up, the top-ranked decomposition successfully reflects the real physical phenomena affecting the data. The decompositions with rank 8 or less all present similar configurations of cut-points and cardinalities, resulting in comparable components where the conclusions above still hold. The first 2-component decomposition appears at rank 10 with the cut-point placed at scale 2^6, which separates the short-term peaks from all the rest of the signal (traffic jams and baseline mixed together). These facts make the result pretty stable as most of the good decompositions are ranked first.

5. CONCLUSIONS AND FUTURE WORK

We introduced a novel methodology to discover the fundamental scale components in a time series in an unsupervised manner. The methodology is based on building candidate scale decompositions, defined over the scale-space image [6] of the original time series, with an MDL-based selection procedure aimed at choosing the optimal one.

As shown, our approach identifies the relevant scale components in a relevant real-world application, giving meaningful insights about the data.

Future work will experiment with diverse representation schemes and hybrid approaches (such as using combinations of segmentation, discretization and Fourier-based encodings).

6. REFERENCES

[1] P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

[2] B. Hu, T. Rakthanmanon, Y. Hao, S. Evans, S. Lonardi, and E. Keogh. Discovering the intrinsic cardinality and dimensionality of time series using mdl. In *Proceedings of ICDM 2011*, pages 1086–1091, 2011.

[3] T. Lindeberg. Scale-space for discrete signals. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 12(3):234–254, Mar. 1990.

[4] U. Vespier *et al.* Traffic Events Modeling for Structural Health Monitoring. In *Proceedings IDA 2011*, 2011.

[5] U. Vespier *et al.* MDL-based Analysis of Time Series at Multiple Time-Scales. In *Proceedings ECML-PKDD 2012*, 2012.

[6] A. P. Witkin. Scale-space filtering. In *Proceedings IJCAI 1983*, pages 1019–1022, San Francisco, CA, USA, 1983.