entropy

MDPI

Article

Data Discovery and Anomaly Detection Using Atypicality for Real-Valued Data

Elyas Sabeti [1] **and Anders Høst-Madsen** [2,3,*]

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, NCRC 10-A108, 2800 Plymouth Rd, Ann Arbor, MI 48109-2800, USA; Sabeti@umich.edu
[2] Department of Electrical Engineering, University of Hawaii at Manoa, Honolulu, HI 96822, USA
[3] Shenzhen Research Institute of Big Data, Shenzhen 518172, China
* Correspondence: Sabeti@umich.edu

Received: 31 October 2018; Accepted: 21 February 2019; Published: 26 February 2019

Abstract: The aim of using atypicality is to extract small, rare, unusual and interesting pieces out of big data. This complements statistics about typical data to give insight into data. In order to find such "interesting" parts of data, universal approaches are required, since it is not known in advance what we are looking for. We therefore base the atypicality criterion on codelength. In a prior paper we developed the methodology for discrete-valued data, and the current paper extends this to real-valued data. This is done by using minimum description length (MDL). We develop the information-theoretic methodology for a number of "universal" signal processing models, and finally apply them to recorded hydrophone data and heart rate variability (HRV) signal.

Keywords: atypicality; minimum description length; big data; codelength

1. Introduction

A central question in the era of "big data" is what to do with the enormous amount of information. One possibility is to characterize it through statistics, e.g., averages, or classify it using machine learning, in order to understand the general structure of the overall data. The perspective in this paper is the opposite, namely that most of the value in the information—in some applications—is in the parts that deviate from the average, that are unusual, atypical. Think of art: The valuable paintings or writings are those that deviate from the norms and brake the rules, that are atypical. Or groundbreaking scientific discoveries, which find new structure in data. Finding such unusual data is often done by painstaking human evaluation of data. The goal of our work is to find practical, automatic methods for localizing atypical parts of data.

When searching for atypical data, a key characteristic is that we do not know what we are looking for, we are looking for the "unknown unknowns". We therefore need universal methods. In the paper [1] we developed a methodology, atypicality, that can be used to discover such data. The basic idea is that if some data can be encoded with a shorter codelength in itself, i.e., with a universal source coder, rather than using the optimum coder for typical data, then it is atypical. The purpose of the current paper is to generalize this to real-valued data. Lossless source coding does not generalize directly to real-valued data. Instead we can use minimum description length (MDL). In the current paper we develop an approach to atypicality based on MDL, and show its usefulness on a real dataset.

In this section before an extensive literature review of detection problems, we first describe the concepts of atypicality and how this framework can be used for data discovery. This arrangement is essential in order to compare the atypicality with the state of the art methods.

1.1. Anomaly Detection and Data Discovery Based on Description Length

A common way to define an outlier or anomaly in data is a sample that does not fit the statistics of typical data [2], e.g., if typical data is described by a pdf $f_T(x)$, and if $f_T(x) < \tau$ for some threshold τ then x is an outlier. In this paper we approach the problem of anomaly detection, and in particular data discovery, from a different point of view. We consider sequences of data x^l, and say that a sequence of data x^l is atypical if there is some alternative model that 'fits' the data better than the typical model. This point of view has been considered before in anomaly detection, e.g., [3]. Given a typical probability distribution, data that is unlikely could simply be, well, outliers, e.g., faulty measurements, and not of much interest in itself. Requiring data to fit an alternative model gives an indication that there is some interesting, new relationship in the data. We therefore think of this approach going beyond simply finding anomalous data, to finding interesting data, i.e., data discovery.

In our paper [1] we used universal source coding for anomaly detection; in [3–5] the authors used a type of universal empirical histogram. This kind of methodology is feasible when data is discrete. However, real-valued data is too rich for such universal descriptions. Models for real-valued data is almost always given as parametric models, either directly or indirectly. Our approach to atypicality for real-valued data, in the absence of universal coders, is to consider multiple 'universal' real-valued models given by parametric models. For example, it is well-known [6] that by the Wold decomposition (almost) all Gaussian stationary processes can be described in terms of a linear prediction model. Wavelets are also good for compressing (lossily) many signals and images. One can therefore expect these will also work well as alternative models. Most modeling and compression are based on a second order analysis, and therefore fit with Gaussian models. One could be interested in also finding atypical data that does not fit a Gaussian model; however, apart from iid (independently, identically distributed) models (similar to [3]), this is difficult to do, so the richness of non-Gaussian models is limited. We will therefore focus on Gaussian models in this paper; notice, however, this is not a limitation of atypicality, we have considered non-Gaussian models in [7].

Consider an atypicality setup where the typical model is given by a probability density function (pdf) $f_T(x^l)$ and the atypical model is given by $f(\mathbf{x}|\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ unknown. Asking if the atypical model is better can be thought of simply as a generalized likelihood ratio test (GLRT) hypothesis test [8].

$$\frac{\min_{\boldsymbol{\theta}} f(x^l|\boldsymbol{\theta})}{f_T(x^l)} \geq \tau.$$

However, in atypicality we would like to test the sequence with respect to a large class of alternative hypotheses—even the class of linear prediction models is infinite. So, assume we have a finite or countable infinite set of model classes $\mathcal{M}_i$ with corresponding pdfs $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$. A test could then be

$$\frac{\min_i \min_{\boldsymbol{\theta}_i} f_i(x^l|\boldsymbol{\theta}_i)}{f_T(x^l)} \geq \tau. \tag{1}$$

However, this is clearly not very useful. More and more complex model will fit data better and better [9], so that the false alarm probability will be very large—model complexity has to be taken into account. One way to do this through Bayesian statistics assigning prior probabilities to both models and parameters, ending up in the test

$$\frac{P_A \sum_i P(\mathcal{M}_i) \int f_i(x^l|\boldsymbol{\theta}_i) f_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{(1-P_A) f_T(x^l)} \geq 1 \tag{2}$$

where P_A is the probability of a sequence being atypical and $P(\mathcal{M}_i)$ the probability of an alternative model $\mathcal{M}_i$. The issue is that using (2) requires choosing a lot of prior distributions and being able to calculate marginal distributions $\int f_i(x^l|\boldsymbol{\theta}_i) f_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$. As explained in for example (3.4–3.5 [9]), these are not easy problems to tackle. Priors are often dictated by the need for the integral to be calculable, rather than actual prior information, and it still leaves parameters unknown ('hyperparameters'). In addition,

choosing prior distributions is anathema to the central idea of looking for unknown data in big data. The whole point is that we know very little about the data we are looking for.

This is where we can use description length. Suppose at first that data is discrete-valued. To each sequence x^l we assign a codeword $c(x^l)$ with length $L(x^l)$. The codewords have to be prefix free and the lengths therefore have to satisfy the Kraft inequality [10]: $\sum_{x^l} 2^{-L(x^l)} \leq 1$. If we let $p(x^l) = 2^{-L(x^l)}$ this defines a (sub)probability on the data, which can be used in a hypothesis test. One can think of description length and coding as a method to find probabilities. There is a key advantage in using description length, as explained in the following. In decoding, a decoder reads a sequence of bits sequentially and turns this into a copy of the source sequence; the codes must be prefix-free. Key here is that in the current step the decoder can only use what is decoded in prior steps. Therefore, when the source sequence is encoded, the encoder cannot use future samples to encode the current sample. We call this 'the principle of sequentiality'. It is the Kraft inequality in reverse: In one direction, as above, we can use the Kraft inequality to verify that a set of codelengths gives valid codes. In the other direction, when codes are decodable (in the pre-fix free sense), they must satisfy the Kraft inequality, and the corresponding probabilities must therefore be valid. An example is Lempel-Ziv coding [10–12], which does not explicitly rely on probabilities. It gives valid codewords because the coding is decodable with a sequential decoder.

To generalize the coding approach to real-valued data, lossless coding is needed. One can notice that lossless coding of real-valued data is used in many applications, for example lossless audio coding [13]. However, direct encoding of the reals represented as binary numbers, such as done in lossless audio coding, makes the methods too dependent on data representation rather than the underlying data. Instead we will use a more abstract model of (finite-precision) reals. We will assume a fixed point representation with a (large) finite number, r, bits after the period, and an unlimited number of bits prior to the period as in [14]. Assume that the actual data is distributed according to a pdf $f(x)$. Then the number of bits required to represent x is given by

$$\begin{aligned} L(x) &= -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log(f(x)2^{-r}) \\ &= -\log(f(x)) + r. \end{aligned} \tag{3}$$

As we are only interested in comparing codelengths the dependency on r cancels out. Suppose we want to decide between two models $f_1(x)$ and $f_2(x)$ for data. Then we decide $f_1(x)$ if $\lim_{r\to\infty} -\log \int_x^{x+2^{-r}} f_1(t)dt + \log \int_x^{x+2^{-r}} f_2(t)dt > 0$, which is $-\log f_1(x) > -\log f_2(x)$. Thus, for the typical codelength we can simply use $L_T(x) = -\log f_T(x)$. One can also argue for this codelength more fundamentally from finite blocklength rate-distortion in the limit of low distortion [15], which makes it more theoretically well-founded. Notice that this codelength is not scaling invariant:

$$\begin{aligned} y &= ax + b \\ L_t(y) &= -\log f(x) + \log |a| \end{aligned} \tag{4}$$

which means care has to be taken when transforms of data are considered. To code the atypical distributions, as the decoder does not know the values of the parameters, both data and parameters in parametric models have to be encoded for a decoder to be able to decode; this was the starting point in the original paper on MDL [14]. One could also use a Bayesian distribution $\int f_i(x^l|\boldsymbol{\theta}_i) f_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ from (2), which does not solve the issues with using Bayes. Instead we can use the principle of sequentiality of coding as follows. We replace $\int f_i(x^l|\boldsymbol{\theta}_i) f_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ in (2) with a codelength based on Rissanen's predictive MDL [16].

$$L_i(x^l) = -\sum_{n=0}^{l-1} \log f_i(x_{n+1}|\hat{\boldsymbol{\theta}}_i(x^n)) \tag{5}$$

where $\hat{\boldsymbol{\theta}}(x^i)$ is the maximum likelihood estimate of the parameter. Since this is sequentially decodable, it gives a valid codelength, and hence probability, without any prior distribution on $\boldsymbol{\theta}$. It does not work for the first sample, as there is no estimate. Instead we encode x_1 with a default distribution. In general application of MDL choice of the default distribution can be tricky, but for atypicality we have a good default distribution: The typical distribution, giving the codelength

$$L_i(x^l) = -\sum_{n=1}^{l-1} \log f_i(x_{n+1}|\hat{\boldsymbol{\theta}}_i(x^n)) - \log f_T(x_1). \tag{6}$$

Notice that default distribution is the same for all models $\mathcal{M}_i$: we do not have to choose a prior for each model. There are no prior assumptions involved, since we use the typical distribution. We still need the probabilities $P(\mathcal{M}_i)$; here we can use Rissanen's universal coder of the integers [14]. The codelength for an integer i is $c + \log^* i$ where c is a normalization constant in the Kraft inequality [14] and $\log^* l = \log l + \log\log l + \cdots$ with the sum continuing as long as the log is defined. We order the models according to complexity and encode the ordinal of a model. The description length test for the sequence x^l to be atypical then becomes

$$\begin{aligned} &-\log\left(\sum 2^{-L_i(x^l)-\log^* i - c}\right) - \log P_A \\ &\quad \leq -\log f_T(x^l) - \log(1 - P_A). \end{aligned} \tag{7}$$

The appeal of coding becomes even more clear when we search for atypical subsequences of long sequences. Using coding this can be done as follows. The coder uses a special header, a codeword not a prefix of any codeword used for the actual data, to denote the start of a subsequence—the decoder will now know it needs to use the atypical decoder. It also encodes the length of the atypical subsequence using Rissanen's universal coder for the integers [14], adding $c + \log^* l$ to the code length, so that the decoder knows when to switch back to the typical coder. The whole sequence is sequentially decodable, thus has a valid probability, and we know from [1] that this gives a valid criterion, at least for iid sequences, in the sense that not the whole sequence will be classified as atypical; the key is the insistence on decodability. It would be difficult to do this directly using Bayesian analysis, as we would have to develop probability distributions for the total sequence for every combination of atypical subsequences in the long sequence. To be precise, for every set of potential subsequences $\mathcal{S} = \{x_{s_1}^{e_1}, x_{s_2}^{e_2}, \ldots\}$ we would have to calculate $p(x^l|\mathcal{S})p(\mathcal{S})$, and then choose the $\mathcal{S}$ giving the largest probability, i.e., MAP.

To understand how (7) avoids the problem overfitting of (1), we notice that asymptotically for large l by [16]

$$L_i(x^l) \approx f(x^l|\hat{\boldsymbol{\theta}}(x^l)) + \frac{k}{2}\log l \tag{8}$$

where k is the number of parameters in $\boldsymbol{\theta}$; this is true for many MDL and Bayesian methods, including Rissanen's original approach [14]. Because (8) penalizes models with many parameters, overfitting is avoided even if we consider an infinite collection of models. While (8) is often used for model selection, it is not accurate enough for our purposes, and we use (5) directly. However, (8) is useful for discussion and analysis.

The above approach can be seen as a generalization to real-valued data of the approach in [1]

Definition 1. *A sequence is atypical if it can be described (coded) with fewer bits in itself rather than using the (optimum) code for typical sequences.*

There is a further difference from Bayes (2), which is more philosophical than computational and practical. When we describe the problem as a hypothesis test problem as in (2), we are asking which hypothesis is correct (which is also the basis of Bayesian model selection [9]). However, in stating the problem as a description length problem, we are just asking if we can find a shorter description length, not if a model is correct. By considering a very large class of alternative models (most pronounced

when we use universal source coding), none might fit very well, none might be even close to the actual model, but we might find one that fits better than the typical model, and that is sufficient for a sequence to be atypical. We have no idea how atypical data might look like, so we cast a very wide net.

1.2. Alternative Approaches

Atypicality has many applications: Anomaly detection, outlier detection, data discovery, novelty detection, transient detection, search for 'interesting' data etc. What all of these applications have in common is that we seek data that is unusual in some way, and atypicality is a general method for finding such data. Each of these applications have specific alternative methods, and we will discuss atypicality compared to other approaches in some of these applications.

There is a very large existing literature on anomaly detection [17–26]; The paper [17] gives an overview until 2009. What is characteristic of all methods, as far as we know, is that they look for data that do not fit the characteristics of normal data, either statistically or according to some other measure. From [17]: "At an abstract level, an anomaly is defined as a pattern that does not conform to expected normal behavior." Atypicality takes a different approach. Atypicality looks for data where an alternative model fits the the data better. Atypicality will still find the first type of anomalies according to [17], but it will also be able to find a wider, more subtle class of anomalies. As a simple example, suppose the normal data is iid Gaussian with zero mean and variance σ^2. The anomalous data is also Gaussian with zero mean and variance σ^2, but the noise is colored. This is in no way anomalous according to the definition in [17]. However, by coding data with a linear predictive coder (see Section 3.2 later) atypicality will detect the anomalous sequence. In [27] we in fact prove that atypicality is exactly optimum for discrete data in the class of finite state machines. While we do not have a similar theorem for real-valued data, this indicates the advantages of atypicality for anomaly detection.

Another advantage of atypicality is that it can straightforwardly be applied to data of unknown/variable length, as discussed in Section 1.1. All existing anomaly detection algorithms we know of use fixed windows, so they cannot make decisions between long, slightly unusual sequences, and short, very unusual sequences; atypicality can. On the other hand, atypicality cannot find single, anomalous samples—outliers: To be able to find a new model for anomalous data, it needs a collection of samples. For this kind of application, more traditional methods must be used.

A type of detection problem closely related to anomaly detection is transient detection [28–34]. In many signal processing applications, it is of interest to detect short-duration statistical changes in observed data. For a parametric class of probability distribution $\{f(x|\theta) : \theta \in \Theta\}$ and for an unknown n_s and n_d the following two hypotheses are considered:

$$\begin{aligned} H_0 &: x_1^l \sim f(x|\theta_0) \\ H_1 &: x_1^{n_s-1} \sim f(x|\theta_0)\,,\ x_{n_s}^{n_d-1} \sim f(x|\theta_1)\,,\ x_{n_d}^{l} \sim f(x|\theta_0)\,. \end{aligned}$$

If θ_0 and θ_1 are known, the Page test is optimal for this in the sense that by using a GLRT; given an average wait between false alarms, it minimizes the worst-case average delay to detection [31]. However in many applications, there is either no information about θ_1 or it varies from one transient signal to another. In this case, it is shown that Variable Threshold Page (VTP) gives a reliable result [29,31]. There are also other approaches of transient detection based on Nuttall's power-law detector that are often used in the literature [29,30]. Other methods are [32–34]. In general atypicality will outperform this methods since it not only allows a more comprehensive class of models, but also it can take advantage of various powerful signal processing methods such as filterbanks and linear prediction to find transient signals with various statistics.

Finally, we will mention change point detection and quickest change detection [35–42]. The goal here is to find a point in time where the distribution of data changes from one to another. The difference from atypicality is that in atypicality, subsequences have both a start and end point. In principle one

could use atypicality for change point detection, but since it is not optimized for this application, the comparison is not that relevant, and atypicality might not perform well. We refer to [35,36] for how to use MDL for change point detection.

2. Minimum Description Length Methods

Above we have argued for using (5) as a codelength. The issue with this method is how to initialize the recursion. In (6) this is solved by using the typical distribution for the first sample, but in general, with more than one parameter, $\hat{\boldsymbol{\theta}}_i(x^i)$ may not be defined until i becomes larger than 1. The further issue is that even when $\hat{\boldsymbol{\theta}}(x^i)$ *is* defined, the estimate might be poor for small i, and using this in (5) can give very long codelengths, see Figure 1 below.

Our solution to the first issue is to encode with increasingly complex models as i increases; we therefore only have to use the default distribution for the very first sample. Since we are not interested in finding a specific model, this is not problematic in atypicality. Our solution to the second issue is rather than using the ML estimate for encoding as though it is the actual parameter value, we use it as an uncertain estimate of $\boldsymbol{\theta}$. We then take this uncertainty into account in the codelength. This is similar to the idea of using confidence intervals in statistical estimates [43]. Below we introduce two methods using this general principle. This is different to the sequentially normalized maximum likelihood method [44], which modifies the encoder itself.

2.1. Sufficient Statistic Method (SSM)

As explained above, our approach to predictive MDL is to introduce uncertainty in the estimate of $\boldsymbol{\theta}$. Our first methodology is best explained through a simple example. Suppose our model is $\mathcal{N}(\mu, \sigma^2)$, with σ known. The average $\bar{x}_n$ is the ML estimate of μ at time n. We know that

$$\bar{x}_n = \mu + z, \quad z \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right).$$

We can re-arrange this as

$$\mu = \bar{x}_n - z.$$

Thus, *given* $\bar{x}_n$, we can think of μ as random $\mathcal{N}\left(\bar{x}_n, \frac{\sigma^2}{n}\right)$. Now

$$x_{n+1} = \mu + z_{n+1} \sim \mathcal{N}\left(\bar{x}_n, \sigma^2 + \frac{\sigma^2}{n}\right)$$

which we can use as a coding distribution for x_{n+1}. This compares to $\mathcal{N}\left(\bar{x}_n, \sigma^2\right)$ that we would use in traditional predictive MDL. Thus, we have taken into account that the estimate of μ is uncertain for n small. The idea of thinking of the non-random parameter μ as random is very similar to the philosophical argument for confidence intervals [43].

In order to generalize this example to more complex models, we take the following approach. Suppose $\mathbf{t}(x^n)$ is a k-dimensional sufficient statistic for the k-dimensional $\boldsymbol{\theta} \in \Theta$. Also suppose there exists some function $\mathbf{s}$ and a k-dimensional (vector) random variable $\mathbf{Y}$ independent of $\boldsymbol{\theta}$ so that

$$\mathbf{t}(x^n) = \mathbf{s}(\mathbf{Y}, \boldsymbol{\theta}). \tag{9}$$

We now assume that for every $(\mathbf{t}, \mathbf{Y})$ in their respective support, (9) has a solution for $\boldsymbol{\theta} \in \Theta$ so that we can write

$$\boldsymbol{\theta} = \mathbf{r}(\mathbf{Y}, \mathbf{t}(x^n)). \tag{10}$$

The parameter $\boldsymbol{\theta}$ is now a random variable (assuming $\mathbf{r}$ is measurable, clearly) with a pdf $f_{x^n}(\boldsymbol{\theta})$ This then gives a distribution on x_{n+1}, i.e.,

$$f(x_{n+1}|x^n) = \int f(x_{n+1}|\boldsymbol{\theta}) f_{x^n}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{11}$$

The method has the following property:

Theorem 1. *The distribution of x_{n+1} is invariant to arbitrary parameter transformations.*

This is a simple observation from the fact that (11) is an expectation, and that when $\boldsymbol{\theta}$ is transformed, the distribution according to (10) is also transformed with the same function.

One concern is the way the method is described. Perhaps we could use different functions $\mathbf{s}$ and $\mathbf{r}$ and get a different result? In the following we will prove that the distribution of $\boldsymbol{\theta}$ is independent of which $\mathbf{s}$ and $\mathbf{r}$ are used.

It is well-known [6,10] that if the random variable X has CDF F, then $U = F(X)$ has a uniform distribution (on $[0,1]$). Equivalently, $X = F^{-1}(U)$ for some uniform random variable U. We need to generalize this to n dimensions. Recall that for a continuous random variable [6]

$$\begin{aligned} F_{i|i-1,\ldots,1}(x_i|x_{i-1},\ldots x_1) &= \int_{-\infty}^{x_i} f(t|x_{i-1},\ldots,x_1)dt \\ &= \frac{1}{f(x_{i-1},\ldots,x_1)} \int_{-\infty}^{x_i} f(t,x_{i-1},\ldots,x_1)dt \end{aligned}$$

whenever $f(x_{i-1},\ldots,x_1) \neq 0$. As an example, let $n = 2$. Then the map $(X_1, X_2) \mapsto (F_1(X_1), F_{2|1}(X_2, X_1))$ is a map from $\mathbb{R}^2$ onto $[0,1]^2$, and $(F_1(X_1), F_{2|1}(X_2, X_1))$ has uniform distribution on $[0,1]^2$. Here $F_1(X_1)$ is continuous in X_1 and $F_{2|1}(X_2, X_1)$ is continuous in X_2

We can write $X_1 = F_1^{-1}(U_1)$. For fixed x_1 we can also write $X_2 = F_{2|1}^{-1}(U_2|x_1)$ for those x_1 where $F_{2|1}$ is defined, and where the inverse function is only with respect to the parameter before $|$. Then

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} F_1^{-1}(U_1) \\ F_{2|1}^{-1}(U_2|F_1^{-1}(U_1)) \end{bmatrix} \triangleq \check{\mathbf{F}}^{-1}(U_1, U_2).$$

This gives the correct joint distribution on (X_1, X_2): The marginal distribution on X_1 is correct, and the conditional distribution of X_2 given X_1 is also correct, and this is sufficient. Clearly $\check{\mathbf{F}}^{-1}$ is not defined for all U_1, U_2; the relationship should be understood as being valid for almost all (X_1, X_2) and (U_1, U_2). We can now continue like this for $X_3, X_4, \ldots, X_n$. We will state this result as a lemma

Lemma 2. *For any continuous random variable $\mathbf{X}$ there exists an n-dimensional uniform random variable $\mathbf{U}$, so that $\mathbf{X} = \check{\mathbf{F}}^{-1}(\mathbf{U})$.*

Theorem 2. *Consider a model $\mathbf{t} = \mathbf{s}_1(\mathbf{Y}_1; \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \mathbf{r}_1(\mathbf{Y}_1; \mathbf{t})$ and an alternative model $\mathbf{t} = \mathbf{s}_2(\mathbf{Y}_2; \boldsymbol{\theta})$, with $\boldsymbol{\theta} = \mathbf{r}_2(\mathbf{Y}_2; \mathbf{t})$. We make the following assumptions:*

1. *The support of $\mathbf{t}$ is independent of $\boldsymbol{\theta}$ and its interior is connected.*
2. *The extended CDF $\check{\mathbf{F}}_i$ of $\mathbf{Y}_i$ is continuous and differentiable.*
3. *The function $\mathbf{Y}_i \mapsto \mathbf{s}_i(\mathbf{Y}_i; \boldsymbol{\theta})$ is one-to-one, continuous, and differentiable for fixed $\boldsymbol{\theta}$.*

Then the distributions of $\boldsymbol{\theta}$ given by $\mathbf{r}_1$ and $\mathbf{r}_2$ are identical.

Proof. By Lemma 2 write $\mathbf{Y}_1 = \mathbf{F}_1^{-1}(\mathbf{U}_1)$, $\mathbf{Y}_2 = \mathbf{F}_2^{-1}(\mathbf{U}_2)$. Let u be the k-dimensional uniform pdf, i.e., $u(\mathbf{x}) = 1$ for $\mathbf{x} \in [0,1]^k$ and 0 otherwise, and let $\mathbf{Y}_i = \mathbf{s}_i^{-1}(\mathbf{t}; \boldsymbol{\theta})$ denote the solution of $\mathbf{t} = \mathbf{s}_i(\mathbf{Y}_i; \boldsymbol{\theta})$ with

respect to $\mathbf{Y}_i$, which is a well-defined due to Assumption 3. We can then write the distribution of $\mathbf{t}$ in two ways as follows ([6]), due to the differentiability assumptions

$$f(\mathbf{t};\boldsymbol{\theta}) = u(\mathbf{F}_1(\mathbf{s}_1^{-1}(\mathbf{t};\boldsymbol{\theta}))\left|\frac{\partial \mathbf{F}_1(\mathbf{s}_1^{-1}(\mathbf{t};\boldsymbol{\theta})}{\partial \mathbf{t}}\right| = u(\mathbf{F}_2(\mathbf{s}_2^{-1}(\mathbf{t};\boldsymbol{\theta}))\left|\frac{\partial \mathbf{F}_2(\mathbf{s}_2^{-1}(\mathbf{t};\boldsymbol{\theta})}{\partial \mathbf{t}}\right|.$$

Due to Assumption 1 we can then that conclude $\frac{\partial \mathbf{F}_1(\mathbf{s}_1^{-1}(\mathbf{t};\boldsymbol{\theta})}{\partial \mathbf{t}} = \frac{\partial \mathbf{F}_2(\mathbf{s}_2^{-1}(\mathbf{t};\boldsymbol{\theta})}{\partial \mathbf{t}}$, or

$$\mathbf{F}_1(\mathbf{s}_1^{-1}(\mathbf{t};\boldsymbol{\theta}) = \mathbf{F}_2(\mathbf{s}_2^{-1}(\mathbf{t};\boldsymbol{\theta})) + \mathbf{k}(\boldsymbol{\theta}).$$

But both $\mathbf{F}_1$ and $\mathbf{F}_2$ have range $[0,1]^k$, and it follows that $\mathbf{k}(\boldsymbol{\theta}) = \mathbf{0}$. Therefore

$$\mathbf{t} = \mathbf{s}_1(\mathbf{F}_1^{-1}(\mathbf{U});\boldsymbol{\theta}) = \mathbf{s}_2(\mathbf{F}_2^{-1}(\mathbf{U});\boldsymbol{\theta}).$$

If we then solve either for $\boldsymbol{\theta}$ as a function of $\mathbf{U}$ (for fixed $\mathbf{t}$), we therefore get exactly the same result, and therefore the same distribution. □

The assumptions of Theorem 2 are very restrictive, but we believe they are far from necessary. In [45] we proved uniqueness in the one-dimensional case under much weaker assumptions (e.g., no differentiability assumptions), but that proof is not easy to generalize to higher dimensions.

Corollary 3. *Let $\mathbf{t}_1(x^n)$ and $\mathbf{t}_2(x^n)$ be equivalent sufficient statistic for $\boldsymbol{\theta}$, and assume the equivalence map is a diffeomorphism. Then the distribution on $\boldsymbol{\theta}$ given by the sufficient statistic approach is the same for $\mathbf{t}_1$ and $\mathbf{t}_2$.*

Proof. We have $\mathbf{t}_1 = \mathbf{s}_1(\mathbf{Y}_1,\boldsymbol{\theta})$ and $\mathbf{t}_2 = \mathbf{s}_2(\mathbf{Y}_2,\boldsymbol{\theta})$. By assumption, there exists a one-to-one map a so that $\mathbf{t}_1 = a(\mathbf{t}_2)$, thus $\mathbf{t}_1 = a(\mathbf{s}_2(\mathbf{Y}_2,\boldsymbol{\theta}))$. Since the distribution of $\boldsymbol{\theta}$ is independent of how the problem is stated, $\mathbf{t}_1$ and $\mathbf{t}_2$ gives the same distribution on $\boldsymbol{\theta}$. □

2.2. Normalized Likelihood Method (NLM)

The issue with the sufficient statistic method is that a sufficient statistic of the same dimension of the parameter vector can be impossible to find. We will therefore introduce a simpler method. Let the likelihood function of the model be $f(x^l|\boldsymbol{\theta})$. For a fixed x^l we can consider this as a 'distribution' on $\boldsymbol{\theta}$; the ML estimate is of course the most likely value of this distribution. To account for uncertainty in the estimate, we can instead try use the total $f(x^l|\boldsymbol{\theta})$ to give a distribution on $\boldsymbol{\theta}$, and then use this for prediction. In general $f(x^l|\boldsymbol{\theta})$ is not a probability distribution as it does not integrate to 1 in $\boldsymbol{\theta}$. We can therefore normalize it to get a probability distribution

$$f_{x^l}(\boldsymbol{\theta}) = \frac{f(x^l|\boldsymbol{\theta})}{C(x^l)}; \quad C(x^l) = \int f(x^l|\boldsymbol{\theta})d\boldsymbol{\theta} \tag{12}$$

if $\int f(x^l;\boldsymbol{\theta})d\boldsymbol{\theta}$ is finite. For comparison, the Bayes posteriori distribution is

$$f(\boldsymbol{\theta}|x^l) = \frac{f(x^l|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(x^l|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

If the support Θ of $\boldsymbol{\theta}$ has finite area, (12) is just the Bayes predictor with uniform prior. If the support Θ of $\boldsymbol{\theta}$ does not have finite area, we can get (12) as a limiting case when we take the limit of uniform distributions on finite Θ_n that converge towards Θ. This is the same way the ML estimator can be seen as a MAP estimator with uniform prior [46]. One can reasonably argue that if we have no further information about $\boldsymbol{\theta}$, a uniform distribution seems reasonable, and has indeed been used

for MDL [47] as well as universal source coding ([10], Section 13.2). What the Normalized Likelihood Method does is simply extend this to the case when there is no proper uniform prior for $\boldsymbol{\theta}$.

The method was actually implicitly mentioned as a remark by Rissanen in ([48], Section 3.2), but to our knowledge was never further developed; the main contribution in this paper is to introduce the method as a practical method. From Rissanen we also know the coding distribution for x_n:

$$f(x_{n+1}|x^n) = \int f(x_{n+1}|\boldsymbol{\theta}) f_{x^n}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{C\left(x^{n+1}\right)}{C\left(x^n\right)}. \tag{13}$$

Let us assume $C(x^n)$ becomes finite for $n > 1$ (this is not always the case, often n needs to be larger). The total codelength can then be written as

$$\begin{aligned} L(x^l) &= \sum_{i=1}^{l-1} -\log f(x_{i+1}|x^i) - \log f_d(x_1) \\ &= -\log C(x^l) + \log C(x^2) - \log f_d(x_1), \end{aligned} \tag{14}$$

where $f_d(x)$ is the default distribution, which for application in atypicality can be taken as the typical distribution. One might see this simply as a (generalized) Bayesian method. However, in general $C(x^n)$ is not a valid probability, and as mentioned in ([9], Section 3.4) an improper prior cannot be used for Bayesian model selection. But when implemented sequentially, as indicated in (14) it does give a valid codelength, because of the principle of sequentiality, central to coding.

2.3. Examples

We will compare the different methods for a simple model. Assume our model is $\mathcal{N}(0, \sigma^2)$ with σ unknown. The likelihood function is $f(x^n|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)$. For $n = 1$ we have $\int_0^\infty f(x^n|\sigma^2)d\sigma^2 = \infty$, but for $n \geq 2$

$$C\left(x^n\right) = \int f(x^n|\sigma^2)d\sigma^2 = \frac{1}{\pi^{\frac{n}{2}}2} \frac{\Gamma\left(\frac{n-2}{2}\right)}{\left[n\widehat{\sigma^2}_n\right]^{\frac{n-2}{2}}}$$

then

$$f_{\text{nlm}}(x_{n+1}|x^n) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)} \frac{\left[n\widehat{\sigma^2}_n\right]^{\frac{n-2}{2}}}{\left[(n+1)\,\widehat{\sigma^2}_{n+1}\right]^{\frac{n-1}{2}}}$$

where $\widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^n x_i^2$. Thus, for coding, the two first samples would be encoded with the default distribution, and after that the above distribution is used. For the SSM, we note that $\widehat{\sigma^2}_n$ is a sufficient statistic for σ^2 and that $z = \frac{n}{\sigma^2}\widehat{\sigma^2}_n \sim \chi^2_{(n)}$, i.e., $\widehat{\sigma^2}_n = s(z, \sigma^2) = \frac{\sigma^2}{n}z$, which we can be solved as $\sigma^2 = r(z, \widehat{\sigma^2}_n) = \frac{n}{z}\widehat{\sigma^2}_n$, in the notation of (9)–(10). This is a transformation of the $\chi^2_{(n)}$ distribution which can be easily found as [6]

$$f_{x^n}(\sigma^2) = \frac{\left[n\widehat{\sigma^2}_n\right]^{\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)(\sigma^2)^{\frac{n+2}{2}}} \exp\left\{-\frac{n}{2\sigma^2}\widehat{\sigma^2}_n\right\}.$$

Now we have

$$f_{\text{ssm}}(x_{n+1}|x^n) = \int f(x_{n+1}|\sigma^2) f_{x^n}(\sigma^2) d\sigma^2$$
$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \frac{\left[n\widehat{\sigma^2}_n\right]^{\frac{n}{2}}}{\left[(n+1)\,\widehat{\sigma^2}_{n+1}\right]^{\frac{n+1}{2}}}. \tag{15}$$

For comparison, the ordinary predictive MDL is

$$f(x_{n+1}|x^n) = \frac{1}{\sqrt{2\pi\widehat{\sigma^2}_n}} \exp\left(-\frac{1}{2\widehat{\sigma^2}_n} x_{n+1}^2\right) \tag{16}$$

which is of a completely different form. To understand the difference, consider the codelength for x_2:

$$\begin{aligned} L(x_2) &= \log\left(\tfrac{x_1^2+x_2^2}{|x_1|}\right) + \log\left(\tfrac{\sqrt{\pi}\Gamma(\frac{1}{2})}{\Gamma(1)}\right) && \text{SSM,} \\ L(x_2) &= \tfrac{1}{2}\log\left(2\pi x_1^2\right) + \tfrac{x_2^2}{x_1^2} && \text{predictive MDL.} \end{aligned}$$

It can be seen that if x_1 is small and x_2 is large, the codelength for x_2 is going to be large. But in the sufficient statistic method this is strongly attenuated due to the log in front of the ratio. Figure 1 shows this quantitatively in the redundancy sense. The redundancy is the difference between the codelength using true and estimated distributions. As can be seen, the CDF of the ordinary predictive MDL redundancy has a long tail, and this is taken care of by SSM.

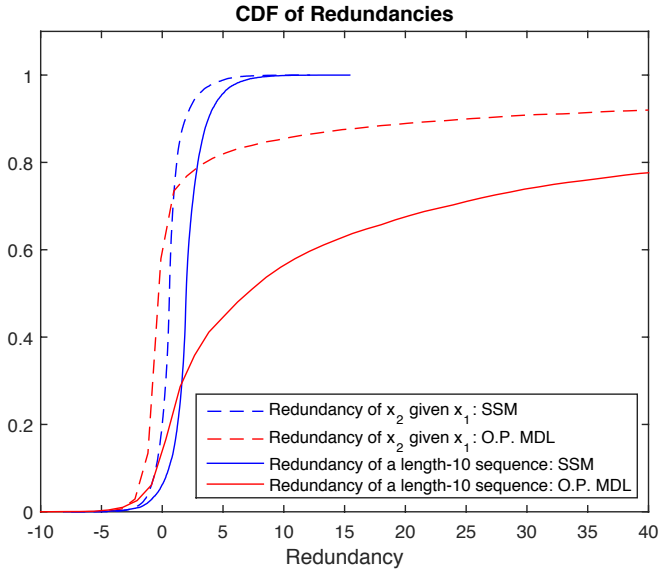

Figure 1. Redundancy comparison between ordinary predictive minimum description length (O.P. MDL) and our proposed sufficient statistic method for $\mu = 0$ and $\sigma^2 = 4$.

3. Scalar Signal Processing Methods

In the following we will derive MDL for various scalar signal processing methods. We can take inspiration from signal processing methods generally used for source coding, such as linear prediction and wavelets; however, the methods have to be modified for MDL, as we use lossless coding, not lossy coding. As often in signal processing, the models are a (deterministic) signal in Gaussian noise. In a previous paper we have also considered non-Gaussian models [7]. All proofs are in Appendices.

3.1. Iid Gaussian Case

A natural extension of the examples considered in Section 2.1 is $x_n \sim \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown. Define $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu}_n)^2$. Then the sufficient statistic method is

$$f(x_{n+1}|x^n) = \sqrt{\frac{n}{\pi\,(n+1)}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \times \frac{\left[(n-1)\,S_n^2\right]^{\frac{n-1}{2}}}{\left[nS_{n+1}^2\right]^{\frac{n}{2}}}. \tag{17}$$

This is a special case of the vector Gaussian model considered later, so we will not provide a proof.

3.1.1. Linear Transformations

The iid Gaussian case is a fundamental building block for other MDL methods. The idea is to find a linear transformation so that we can model the result as iid, and then use the iid Gaussian MDL. For example, in the vector case, suppose $\mathbf{x}_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is (temporally) iid, and let $\mathbf{y}_n = \mathbf{A}\mathbf{x}_n \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$. If we then assume that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ is diagonal, we can use the iid Gaussian MDL on each component. Similarly, in the scalar case, we can use a filter instead of a matrix. Because of (4) we need to require $\mathbf{A}$ to be orthonormal: For any input we then have $\mathbf{y}_n^T\mathbf{y}_n = \mathbf{x}_n^T\mathbf{A}^T\mathbf{A}\mathbf{x}_n = \mathbf{x}_n^T\mathbf{x}_n$, and in particular $E[\mathbf{y}_n^T\mathbf{y}_n] = E[\mathbf{x}_n^T\mathbf{x}_n]$ independent of the actual $\boldsymbol{\Sigma}$. We will see this approach in several cases in the following.

3.2. Linear Prediction

Linear prediction is a fundamental to random processes. Write

$$\hat{x}_{n+1|x^n} = \sum_{k=0}^{\infty} w_k x_{n-k}$$
$$e_{n+1} = x_{n+1} - \hat{x}_{n+1|x^n}.$$

Then for most stationary random processes the resulting random process $\{e_n\}$ is uncorrelated, and hence in the Gaussian case, iid, by the Wold decomposition [6]. It is therefore a widely used method for source coding, e.g., [13]. In practical coding, a finite prediction order M is used,

$$\hat{x}_{n+1|x^n} = \sum_{k=1}^{M} w_k x_{n-k+1}, \quad n \geq M$$

Denote by τ the power of $\{e_n\}$. Consider the simplest case with $M = 1$: There are two unknown parameters (w_1, τ). However, the minimal sufficient statistic has dimension three [49]: $\left(\sum_{k=1}^n x_k^2, \sum_{k=1}^{n-1} x_k^2, \sum_{k=2}^n x_k x_{k-1}\right)$. Therefore, we cannot use SSM; and even if we could, the distribution of the sufficient statistic is not known in closed form [49]. We therefore turn to the NLM.

We assume that $e_{n+1} = x_{n+1} - \hat{x}_{n+1|x^n}$ is iid normally distributed with zero mean and variance τ,

$$f(x^n|\tau, \mathbf{w}) = \frac{1}{(2\pi\tau)^{(n-M)/2}} \times \exp\left(-\frac{1}{2\tau}\sum_{i=M+1}^{n}\left[x_i - \sum_{k=1}^{M} w_k x_{i-k}\right]^2\right). \tag{18}$$

Define

$$\hat{r}_{(n)}(k) = \sum_{i=M+1}^{n} x_i x_{i-k}.$$

Then a simple calculation shows that

$$\sum_{i=M+1}^{n} e_i^2 = \hat{r}_{(n)}(0) - 2\mathbf{w}^T \mathbf{p}_{(n)} + \mathbf{w}^T R_{(n)}^{(M)} \mathbf{w}$$

where $\mathbf{w}^T = [w_1\ w_2\ \cdots\ w_M]$, $\mathbf{p}_{(n)}^T = [\hat{r}_{(n)}(1)\ \hat{r}_{(n)}(2)\ \cdots\ \hat{r}_{(n)}(M)]$,

$$R_{(n)}^{(M)} = \sum_{i=M+1}^{n} \mathbf{x}_{i-M}^{i-1} \left(\mathbf{x}_{i-M}^{i-1}\right)^T \tag{19}$$

and $x_{i-M}^{i-1} = [x_{i-1}, x_{i-2}, \ldots, x_{i-M}]$. Thus

$$f(x^n|\tau, \mathbf{w}) = \frac{1}{(2\pi\tau)^{(n-M)/2}} \times \exp\left(-\frac{1}{2\tau}\left[\hat{r}_{(n)}(0) - 2\mathbf{w}^T \mathbf{p}_{(n)} + \mathbf{w}^T R_{(n)}^{(M)} \mathbf{w}\right]\right)$$

giving (see Appendix A)

$$C(x^n) = \frac{1}{2\,(\pi)^{\frac{n-2M}{2}} \sqrt{\det\left(R_{(n)}\right)}} \frac{\Gamma\left(\frac{n-2M-2}{2}\right)}{\left(\hat{\tau}_{(n)}^{(M)}\right)^{\frac{n-2M-2}{2}}}$$

and

$$f_M(x_{n+1}|x^n) = \sqrt{\frac{\det\left(R_{(n)}^{(M)}\right)}{\det\left(R_{(n+1)}^{(M)}\right)} \frac{\Gamma\left(\frac{n-2M-1}{2}\right)}{\Gamma\left(\frac{n-2M-2}{2}\right)}} \times \frac{1}{\sqrt{\pi}} \frac{\left(\hat{\tau}_{(n)}^{(M)}\right)^{\frac{n-2M-2}{2}}}{\left(\hat{\tau}_{(n+1)}^{(M)}\right)^{\frac{n-2M-1}{2}}} \tag{20}$$

with $\hat{\tau}_{(n)}^{(M)} = \hat{r}_{(n)}(0) - \mathbf{p}_{(n)}^T R_{(n)}^{-1} \mathbf{p}_{(n)}$.

The Equation (20) is defined for $n \geq 2M + 2$: The vector x_{i-M}^{i-1} is defined for $i \geq M + 1$, and $R_{(n)}^{(M)}$ defined by (19) becomes full rank when the sum contains M terms. Before the order M linear predictor becomes defined, the data needs to be encoded with other methods. Since in atypicality we are not seeking to determine the model of data, just if a different model than the typical is better, we encode data with lower order linear predictors until the order M linear predictor becomes defined. So, the first sample is encoded with the default pdf. The second and third samples are encoded with the iid unknown variance coder (There is no issue in encoding some samples with SSM and others with NLM) (15). Then the order 1 linear predictor takes over, and so on.

3.3. Filterbanks and Wavelets

A popular approach to source coding is sub-band coding and wavelets [50–52]. The basic idea is to divide the signal into (perhaps overlapping) spectral sub-bands and then allocate different bitrates to each sub-band; the bitrate can be dependent on the power in the sub-band and auditory properties

of the ear in for example audio coding. In MDL we need to do lossless coding, so this approach cannot be directly applied, but we can still use sub-band coding as explained in the following.

As we are doing lossless coding, we will only consider perfect reconstruction filterbanks [50,53]. Furthermore, in light of Section 3.1.1 we also consider only (normalized) orthogonal filterbanks [50,52].

The basic idea is that we split the signal into a variable number of sub-bands by putting the signal through the filterbank and downsampling. Then the output of each downsampled filter is coded with the iid Gaussian coder of Section 3.1 with an unknown mean and variance, which are specific to each sub-band. In order to understand how this works, consider a filterbank with two sub-bands. Assume that the signal is stationary zero mean Gaussian with power σ^2, and let the power at the output of sub-band 1 be σ_1^2 and of sub-band 2 be σ_2^2. Because the filterbank is orthogonal, we have $\sigma^2 = \frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)$. To give some intuition to why a sub-band coder can give shorter codelengh, we use (8) to get the approximate codelengths

$$\begin{aligned} L_{\text{direct}} &= \frac{l}{2}\log\left(\sigma^2\right) + \frac{l}{2}\left(\log 2\pi + \log e\right) + \frac{1}{2}\log l \\ L_{\text{filterbank}} &= \frac{l}{4}\log\left(\sigma_1^2\right) + \frac{l}{4}\left(\log 2\pi + \log e\right) + \frac{1}{2}\log l \\ &\quad + \frac{l}{4}\log\left(\sigma_2^2\right) + \frac{l}{4}\left(\log 2\pi + \log e\right) + \frac{1}{2}\log l \\ &= \frac{l}{2}\log\left(\sqrt{\sigma_1^2\sigma_2^2}\right) + \frac{l}{2}\left(\log 2\pi + \log e\right) + \log l. \end{aligned}$$

Since $\sqrt{\sigma_1^2\sigma_2^2} \leq \sigma^2$ (with equality only if $\sigma_1^2 = \sigma_2^2$), the sub-band coder will result in shorter codelength for sufficiently large l if the signal is non-white.

The above analysis is a stationary analysis for long sequences. However, when considering shorter sequences, we also need to consider the transient. The main issue is that output power will deviate from the stationary value during the transient, and this will affect the estimated power $\widehat{\sigma_n^2}$ used in the sequential MDL. The solution is to transmit to the receiver the input to the filterbank during the transient, and only use the output of the filterbank once the filters have been filled up. It is easy to see that the system is still perfect reconstruction: Using the received input to the filterbank, the receiver puts this through the analysis filterbank. It now has the total sequence produced by the analysis filterbank, and it can then put that through the reconstruction filterbank. When using multilevel filterbanks, this has to be done at each level.

We assume the decoder knows which filters are used and the maximum depth D used. In principle the encoder could now search over all trees of level at most D. The issue is that there are an astonishing large number of such trees; for example for $D = 4$ there are 676 such trees. Instead of choosing the best, we can use the idea of the CTW [1,54,55] and weigh in each node: Suppose after passing a signal x^n of an internal node S through low-pass and high-pass filters and downsampler, $x_L^{n/2}$ and $x_H^{n/2}$ are produced in the children nodes of S. The weighted probability of x^n in the internal node S will be

$$f_w\left(x^n\right) = \frac{1}{2}f\left(x^n\right) + \frac{1}{2}f_w\left(x_L^{n/2}\right)f_w\left(x_H^{n/2}\right)$$

which is a good coding distribution for both a memoryless source and a source with memory [54,55].

4. Vector Case

We now assume that a vector sequence $\mathbf{x}^n$, $\mathbf{x}_i \in \mathbb{R}^M$ is observed. The vector case allows for a more rich set of model and more interesting data discovery than the scalar case, for example atypical correlation between multiple sensors. It can also be applied to images [56], and to scalar data by dividing into blocks. That is in particular useful for the DFT, Section 4.4.

A specific concern is initialization. Applying sequential coding verbatim to the vector case means that the first vector $\mathbf{x}_1$ needs to be encoded with the default coder, but this means the default coder

influences the codelength too much. Instead we suggest to encode the first vector as a scalar signal using the scalar Gaussian coder (unknown variance→unknown mean/variance). That way only the first component of the first vector needs to be encoded with the default coder.

4.1. Vector Gaussian Case with Unknown Mean

First assume $\boldsymbol{\mu}$ is unknown but Σ is given. We define $\text{etr}\left(\cdots\right) = \exp\left(\text{trace}\left(\cdots\right)\right)$ and we have

$$f\left(\mathbf{x}^n|\boldsymbol{\mu}\right) = \frac{1}{\sqrt{\left(2\pi\right)^{kn}\det\left(\Sigma\right)^n}} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right\}.$$

We first consider the NLM. By defining $\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\hat{\Sigma}_n = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i$ (note that $\hat{\Sigma}_n$ is not the estimate of Σ) we have

$$\begin{aligned} C\left(\mathbf{x}^n\right) &= \int f\left(\mathbf{x}^n|\boldsymbol{\mu}\right)d\boldsymbol{\mu} \\ &= \frac{1}{\sqrt{\left(2\pi\right)^{kn}\det\left(\Sigma\right)^n}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i\Sigma^{-1}\mathbf{x}_i\right\} \\ &\times \int \exp\left\{-\frac{n}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}+n\hat{\boldsymbol{\mu}}_n^T\Sigma^{-1}\boldsymbol{\mu}\right\}d\boldsymbol{\mu} \\ &= C\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i\Sigma^{-1}\mathbf{x}_i-\hat{\boldsymbol{\mu}}_n^T\Sigma^{-1}\hat{\boldsymbol{\mu}}_n\right)\right\} \\ &= C\text{etr}\left\{-\frac{1}{2}\left(\hat{\Sigma}_n-n\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_n^T\right)\Sigma^{-1}\right\} \end{aligned}$$

where $C = \frac{1}{\sqrt{\left(2\pi\right)^{k(n-1)}n^k\det(\Sigma)^{n-1}}}$, hence we can write

$$\begin{aligned} f\left(\mathbf{x}_{n+1}|\mathbf{x}^n\right) &= \frac{C\left(\mathbf{x}^{n+1}\right)}{C\left(\mathbf{x}^n\right)} \\ &= \sqrt{\left(\frac{n}{n+1}\right)^k}\frac{1}{\sqrt{\left(2\pi\right)^k\det\left(\Sigma\right)}} \\ &\times \frac{\text{etr}\left\{-\frac{1}{2}\left(\hat{\Sigma}_{n+1}-\left(n+1\right)\hat{\boldsymbol{\mu}}_{n+1}\hat{\boldsymbol{\mu}}_{n+1}^T\right)\Sigma^{-1}\right\}}{\text{etr}\left\{-\frac{1}{2}\left(\hat{\Sigma}_n-n\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_n^T\right)\Sigma^{-1}\right\}}. \end{aligned} \tag{21}$$

It turns out that in this case, the SSM gives the same result.

4.2. Vector Gaussian Case with Unknown Σ

Assume $\mathbf{x}_n \sim \mathcal{N}\left(\mathbf{0},\Sigma\right)$ where the covariance matrix is unknown:

$$f\left(\mathbf{x}^n|\Sigma\right) = \frac{1}{\sqrt{\left(2\pi\right)^{kn}\det\left(\Sigma\right)^n}}\text{etr}\left\{-\frac{1}{2}\hat{\Sigma}_n\Sigma^{-1}\right\}$$

where $\hat{\boldsymbol{\Sigma}}_n = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$.

In order to find the MDL using SSM, notice that we can write

$$\mathbf{x}_n = \mathbf{S}\mathbf{z}_n, \quad \mathbf{z}_n \sim \mathcal{N}(0,\mathbf{I})$$

where $\mathbf{S} = \mathbf{\Sigma}^{\frac{1}{2}}$, that is $\mathbf{S}$ is some matrix that satisfies $\mathbf{S}\mathbf{S}^T = \mathbf{\Sigma}$. A sufficient statistic for $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}}_n = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{S} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T \mathbf{S}^T \overset{\text{def}}{=} \mathbf{S}\mathbf{U}\mathbf{S}^T.$$

Let $\hat{\mathbf{S}}_n = \hat{\mathbf{\Sigma}}_n^{\frac{1}{2}} = \mathbf{S}\mathbf{U}^{\frac{1}{2}}$. Then we can solve $\mathbf{S} = \hat{\mathbf{S}}_n \mathbf{U}^{-\frac{1}{2}}$ and $\mathbf{\Sigma} = \hat{\mathbf{S}}_n \mathbf{U}^{-1} \hat{\mathbf{S}}_n^T$. Since $\mathbf{U}^{-1}$ has Inverse-Wishart distribution $\mathbf{U}^{-1} \sim \mathcal{W}_M^{-1}(I, n)$, one can write $\Sigma \sim \mathcal{W}_M^{-1}\left(\hat{\Sigma}_n, n\right)$. Using this distribution we calculate in Appendix B that

$$f\left(\mathbf{x}_{n+1} | \mathbf{x}^n\right) = \frac{1}{\pi^{\frac{M}{2}}} \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n+1}{2}}} \frac{\Gamma_M\left(\frac{n+1}{2}\right)}{\Gamma_M\left(\frac{n}{2}\right)} \tag{22}$$

where Γ_M is the multivariate gamma function [57].

On the other hand, using the normalized likelihood method we have

$$C\left(\mathbf{x}^n\right) = \frac{\Gamma_M\left(\frac{n}{2} - \frac{M+1}{2}\right)}{2^{\frac{M(M+1)}{2}} \pi^{\frac{kn}{2}} \det\left(\hat{\Sigma}_n\right)^{\frac{n}{2} - \frac{M+1}{2}}},$$

from which

$$\begin{aligned} f\left(\mathbf{x}_{n+1} | \mathbf{x}^n\right) &= \frac{C\left(\mathbf{x}^{n+1}\right)}{C\left(\mathbf{x}^n\right)} \\ &= \frac{1}{\pi^{\frac{k}{2}}} \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n}{2} - \frac{M+1}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n}{2} - \frac{M}{2}}} \frac{\Gamma_M\left(\frac{n}{2} - \frac{M}{2}\right)}{\Gamma_M\left(\frac{n}{2} - \frac{M+1}{2}\right)}. \end{aligned} \tag{23}$$

4.3. Vector Gaussian Case with Unknown Mean and Σ

Assume $\mathbf{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where both mean and covariance matrix are unknown:

$$\begin{aligned} f\left(\mathbf{x}^n | \boldsymbol{\mu}, \Sigma\right) = &\frac{1}{\sqrt{(2\pi)^{Mn} \det(\Sigma)^n}} \\ &\times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right) \right\}. \end{aligned}$$

It is well-known [46] that sufficient statistics are $\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ and $\hat{\mathbf{\Sigma}}_n = (n-1) S_n = \sum_{i=1}^{n} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n\right) \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n\right)^T$. Let $\mathbf{S}$ be a square root of $\mathbf{\Sigma}$, i.e., $\mathbf{S}\mathbf{S}^T = \mathbf{\Sigma}$. We can then write

$$\begin{aligned} \hat{\boldsymbol{\mu}}_n &= \boldsymbol{\mu} + \frac{1}{\sqrt{n}} \mathbf{S}\mathbf{z} \\ \hat{\mathbf{\Sigma}}_n &= \mathbf{S}\mathbf{U}\mathbf{S}^T \end{aligned}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathbf{U} \sim \mathcal{W}_M(\mathbf{I}, n-1)$, $\mathbf{z}$ and $\mathbf{U}$ are independent, and $\mathcal{W}_M$ is the Wishart distribution. We solve the second equation with respect to $\mathbf{S}$ as in Section 4.2 and the first with respect to $\boldsymbol{\mu}$, to get

$$\boldsymbol{\Sigma} = \hat{\mathbf{S}}_n \mathbf{U}^{-1} \hat{\mathbf{S}}_n^T \sim \mathcal{W}_M^{-1}\left(\hat{\boldsymbol{\Sigma}}_n, n-1\right)$$
$$\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_n - \frac{1}{\sqrt{n}}\mathbf{S}\mathbf{z} = \hat{\boldsymbol{\mu}}_n - \frac{1}{\sqrt{n}}\hat{\mathbf{S}}_n \mathbf{U}^{-\frac{1}{2}}\mathbf{z} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_n, \frac{1}{n}\Sigma\right)$$

where $\hat{\mathbf{S}}_n$ is a square root of $\hat{\boldsymbol{\Sigma}}_n$. We can explicitly write the distributions as

$$f_{\mathbf{x}^n}(\boldsymbol{\mu}|\Sigma) = \sqrt{\frac{n^M}{(2\pi)^M \det(\Sigma)}} \exp\left\{-\frac{n}{2}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T \Sigma^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)\right\}$$
$$f_{\mathbf{x}^n}(\Sigma) = \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n-1}{2}}}{2^{\frac{M(n-1)}{2}}\Gamma_M\left(\frac{n-1}{2}\right)} \det(\Sigma)^{-\frac{n+M}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_n\Sigma^{-1}\right\}.$$

Using these distributions, in Appendix C we calculate

$$f(\mathbf{x}_{n+1}|\mathbf{x}^n) = \frac{1}{\pi^{\frac{M}{2}}}\sqrt{\left(\frac{n}{n+1}\right)^M} \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n-1}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n}{2}}} \frac{\Gamma_M\left(\frac{n}{2}\right)}{\Gamma_M\left(\frac{n-1}{2}\right)}$$

and for NLM

$$f(\mathbf{x}_{n+1}|\mathbf{x}^n) = \frac{1}{\pi^{\frac{M}{2}}}\sqrt{\left(\frac{n}{n+1}\right)^M} \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n-1}{2}-\frac{M+1}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n}{2}-\frac{M+1}{2}}} \times \frac{\Gamma_M\left(\frac{n-M-1}{2}\right)}{\Gamma_M\left(\frac{n-M-2}{2}\right)}.$$

These are very similar to the case of known mean, Section 4.2. We require one more sample before the distributions become well-defined, and $\boldsymbol{\Sigma}_n$ is defined differently.

4.4. Sparsity and DFT

We can specify a general method as follows. Let $\boldsymbol{\Phi}$ be an orthonormal basis of $\mathbb{R}^M$ and write the signal model as

$$\mathbf{x}_n = \sum_{i=1}^{N}(A_i + s_{i,n})\boldsymbol{\phi}_{j(i)} + \mathbf{w}_n.$$

Here N is the number of basis vectors used, and $j(i), i = 1, \ldots, N$ their indices. The signal $s_{i,n}$ is iid $\mathcal{N}(0, \sigma_i)$, the noise $\mathbf{w}_n$ iid $\mathcal{N}(0, \sigma^2\mathbf{I})$, and $A_i, \sigma_i^2, \sigma^2$ are unknown. If we let $\mathbf{y}_n = \boldsymbol{\Phi}^T\mathbf{x}_n$ and J the indices of the signal components then

$$y_{j(i),n} = A_i + s_{i,n} + w_{j(i),n} = A_i + \tilde{s}_{i,n}, \quad j(i) \in J$$
$$y_{j,n} = w_{j,n}, \quad j \notin J.$$

Thus the $y_{j(i),n}$ can be encoded with the scalar Gaussian encoder of Section 3.1, while the $y_{j,n}$ can be encoded with a vector Gaussian encoder for $\mathcal{N}(0, \sigma^2\mathbf{I}_{M-N})$ using the following equation that is achieved using the SSM:

$$f\left(\mathbf{w}_{n+1}|\mathbf{w}^{n}\right)=\frac{1}{\pi^{\frac{(M-N)}{2}}}\frac{\Gamma\left(\frac{(M-N)(n+1)}{2}\right)}{\Gamma\left(\frac{(M-N)n}{2}\right)}\times\frac{\left[n\widehat{\tau}_{n}\right]^{\frac{(M-N)n}{2}}}{\left[(n+1)\,\widehat{\tau}_{n+1}\right]^{\frac{(M-N)(n+1)}{2}}}$$

where $\widehat{\tau}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i^T\mathbf{w}_i$. Now we need to choose which coefficients $j(i)$ to choose as signal components and inform the decoder. The set J can be communicated to the decoder by sending a sequence of $0,1$ encoded with the universal encoder of ([10], Section 13.2) with $MH\left(\frac{N}{M}\right)+\frac{1}{2}\log M$ bits. The optimum set can in general only be found by trying all sets J and choosing the one with shortest codelength, which is infeasible. A heuristic approach is to find the N components with maximum power when calculated over the whole blocklength l (the decoder does not need to know how J was chosen, only what J is, it is therefore fine to use the power at the end of the block). What still remains is how to choose N. It seems computationally feasible to start with $N=1$ and then increase N by 1 until the codelength no longer decreases, since most of the calculations for N can be reused for $N+1$.

We can apply this in particular when $\mathbf{\Phi}$ is a DFT matrix. In light of Section 3.1.1 we need to use the normalized form of the DFT. The complication is that the output is complex, i.e., the M real inputs result in M complex outputs, or $2M$ real outputs. Therefore, care has to be taken with the symmetry properties of the output. Another option is to use DCT instead, which is well-developed and commonly used for compression.

5. Experimental Results

5.1. Transient Detection Using Hydrophone Recordings

As an example of the application of atypicality, we will consider transient detection [28]. In transient detection, a sensor records a signal that is pure noise most of the time, and the task is to find the sections of the signal that are not noise. In our terminology, the typical signal is noise, and the task is to find the atypical parts.

As data we used hydrophone recordings from a sensor in the Hawaiian waters outside Oahu, the Station ALOHA Cabled Observatory (ACO) [58]. The data used for this paper were collected (with sampling freuquency of 96 kHz which was then downsampled to 8 kHz) during a proof module phase of the project conducted between February 2007 and October 2008. The data was pre-processed by differentiation ($y[n]=x[n]-x[n-1]$) to remove a non-informative mean component.

The principal goal of this two years of data is to locate whale vocalization. Fin (22 m, up to 80 tons) and sei (12–18 m, up to 24.6 tons) whales are known by means of visual and acoustic surveys to be present in the Hawaiian Islands during winter and spring months, but migration patterns in Hawaii are poorly understood [58].

Ground truth has been established by manual detection, which is achieved using visual inspection of spectrogram by a human operator. 24 h of manual detections for both the 20 Hz and the 20–35 Hz variable calls were recorded for each the following dates (randomly chosen): 1 March 2007, 17 November 2007, 29 May 2008, 22 August 2008, 4 September 2008 and 9 February 2008 [58].

In order to analyze the performance of different detectors on such a data, first the measures 'Precision' and 'Recall' are defined as below

$$\text{Recall} = \frac{\text{number of correct detections}}{\text{total number of manual detections}}$$
$$\text{Precision} = \frac{\text{number of correct detections}}{\text{total number of algorithm detections}}$$

where Recall measures the probability of correctly obtained vocalizations over expected number of detections and Precision measures the probability of correctly detected vocalizations obtained by the detector. The Precision versus Recall curve show the detectors ability to obtain vocalizations as well as the accuracy of these detections [58].

In order to compare our atypicality method with alternative approaches in transient detection, we compare its performance with Variable Threshold Page (VTP) which outperforms other similar methods in detection of non-trivial signals [31].

For the atypicality approach, we need a typical and an atypical coder. The typical signal is pure noise, which, however, is not necessarily white: It consists of background noise, wave motion, wind and rain. We therefore used a linear predictive coder. The order of the linear predictive coder was globally set to 10 as a compromise between performance and computational speed. An order above 10 showed no significant decrease in codelength, while increasing computation time. The prediction coefficients were estimated for each 5 min segment of data. It seems unreasonable to expect the prediction coefficients to be globally constant due to for example variations in weather, but over short length segments they can be expected to be constant. Of course, a 5 min segment could contain atypical data and that would result in incorrect typical prediction coefficients. However, for this particular data we know (or assume) that atypical segments are of very short duration, and therefore will affect the estimated coefficients very little. This cannot be used for general data sets, only for data sets where there is a prior knowledge (or assumption) that atypical data are rare and short. Otherwise the typical coder should be trained on data known to be typical as in [1] or by using unsupervised atypicality, which we are developing for a future paper.

For the atypical coder, we implemented all the scalar methods of Section 3 in addition to the DFT, Section 4.4, with optimization over blocklength. Let $\mathcal{X}(n,l) = (x_n, ..., x_{n+l-1})$ be a subsequence of length l to be tested for atypicality, and suppose $L_T(\mathcal{X}(n,l))$ and $L_A(\mathcal{X}(n,l))$ are the typical codelength and atypical codelength of sequence $\mathcal{X}(n,l)$, respectively. Note that $L_A(\mathcal{X}(n,l)) = -\log(f(\mathcal{X}(n,l))) + \log^* l$ where f is any encoder of Sections 3 and 4.4, and $\log^* l$ is the number of bits to tell the decoder the length of the atypical subsequence, as discussed in Section 1.1, see also [1,59]. Then for every sample of data we calculate

$$\Delta L(n) = \max_l \{L_T(\mathcal{X}(n,l)) - L_A(\mathcal{X}(n,l))\} \tag{24}$$

and the atypicality criterion would be $\Delta L(n) > \tau$ for some threshold (which does not need to be chosen prior to running the algorithm, since the larger $\Delta L(n)$ is the more atypical). Please note that the threshold τ can be seen as the length of the header the encoder uses to tell the decoder an atypical sequence is next. Calculating $\Delta L(n)$ requires examining every subsequence (perhaps up to a maximum length). Because the coders (e.g., (5)) are recursive, we can efficiently calculate $L_A(\mathcal{X}(n,l+1))$ from $L_A(\mathcal{X}(n,l))$, so the complexity is not prohibitive. Still, for a large dataset (i.e., big data), direct implementation of atypicality search is too computationally complex; so instead, similar to [59] we propose a tree-structured searching algorithm in which discovery of atypical sequence (in this case, whale vocalizations) can be performed in different stages. First in coarse search, a tree-structured division of data is considered such that at each level i, data is divided into non-overlapping blocks of length 2^i, then for each block typical and atypical codelengths are compared. Obviously due to non-overlapping division some atypical sequences are missed, and the worse case is if an atypical sequence of length l is divided equally into two consecutive non-overlapping blocks of length 2^i.

However, each of these sequences of length $l/2$ might be detected at the level $i-1$. The issue is that the complexity penalty per sample from (8) is about $\frac{\frac{k}{2}\log l}{l}$, which is decreasing in l. Thus, a sequence of length l may be atypical, but each of the length $\frac{l}{2}$ halves may not be. This can be compensated by repeating every block once and encoding this double length block. By experimentation we have found that this gives a very low chance of missing an atypical subsequence. On the other hand, it does give false positives, because an exactly repeated block clearly has a strong (false) pattern. This is not a big issue, as these false positives are eliminated during the next stage.

After the coarse search, the next stage is fine search, in which the blocks flagged by coarse search are expanded and every subsequence of this expanded block is tested in an exhaustive search, which eliminates false positives. The final stage is segmentation, where the exact start and end point of atypical sequences are determined by minimizing the total codelength of the whole sequence of data. Figure 2 shows Precision vs. Recall curve for both atypicality and VTP.

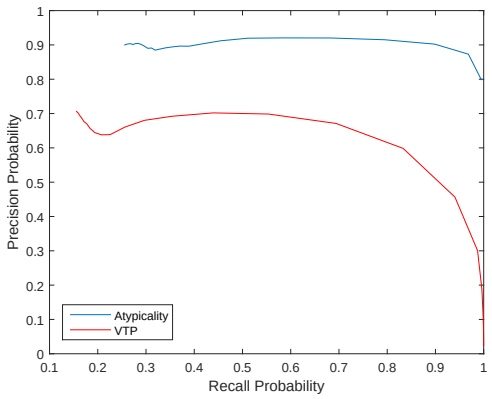

Figure 2. Precision vs. Recall probability for all six days that manual detections are available.

5.2. Anomaly Detection Using Holter Monitoring Data

As another example of atypicality application, we consider an anomaly detection problem. We consider data obtained by Holter Monitoring, i.e., a continuous tape recording of a patient's ECG for 24 h. We use the MIT-BIH Normal Sinus Rhythm Database (nsrdb) which is provided by PhysioNet [60]. Even though the subjects included in this database were found to have had no significant persistent arrhythmias, there still existed arrhythmic beats and patterns to look for [60]. We apply atypicality to find interesting parts of the the dataset.

Since the data is assumed to be 'Normal Sinus Rhythm', a Gaussian model with unknown mean and variance is assumed for the typical data. For atypical encoding, we used the same methodology as in the previous section. As can be seen in the Figure 3, atypicality as an anomaly detector was able to find two major atypical segments, both of which contained multiple supraventricular beats and ventricular contraction (provided by HRV annotation files, PhysioNet [60]). Based on the data annotation these two segments were the only fractions in the data that contained abnormal beats and rhythms, which shows the efficacy of the atypicality framework. For comparison we included VTP as a transient detection method and the pruned exact linear time (PELT) method [37] as a change-point detection algorithm. As can be seen, VTP and PELT detected only one of the anomalous segments, while atypicality detected both.

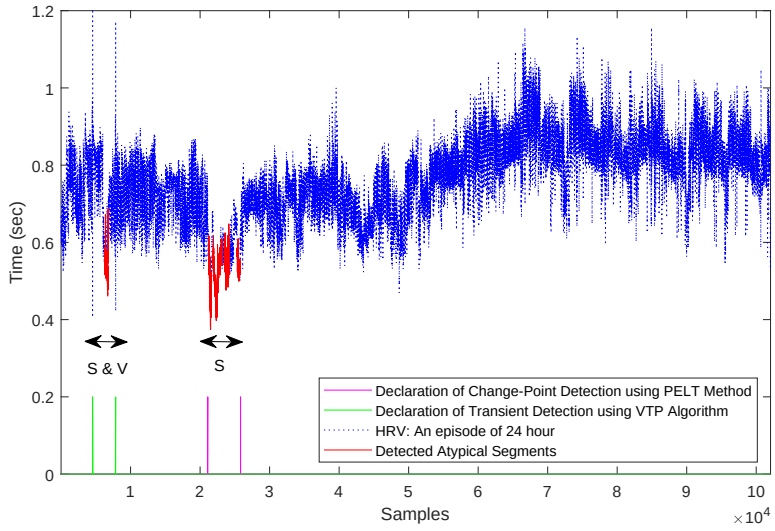

Figure 3. Detected atypical segments of Holter Monitoring heart rate variability (HRV): "S" stands for supraventricular arrhythmia and "V" stands for ventricular contraction based on annotation provided by PhysioNet [60].

6. Conclusions

Atypicality is a method for finding rare, interesting snippets in big data. It can be used for anomaly detection, data mining, transient detection, and knowledge extraction among other things. The current paper extended atypicality to real-valued data. It is important here to notice that discrete-valued and real-valued atypicality is one theory. Atypicality can therefore be used on data that are of mixed type. One advantage of atypicality is that it directly applies to sequences of variable length. Another advantage is that there is only one parameter that regulates atypicality, the single threshold parameter τ, which has the concrete meaning of the logarithm of the frequency of atypical sequences. This contrasts with other methods that have multiple parameters.

Atypicality becomes really interesting in combination with machine learning. First, atypicality can be used to find what is not learned in machine learning. Second, for many data sets, machine learning is needed to find the typical coder. In the experiments in this paper, we did not need machine learning because the typical data was pure noise. But in many other types of data, e.g., ECG (electrocardiogram), 'normal' data is highly complex, and the optimum coder has to be learned with machine learning. This is a topic for future research.

Author Contributions: Conceptualization, E.S. and A.H.-M.; Methodology, E.S. and A.H.-M.; software, E.S.; Validation, E.S.; Formal analysis, E.S. and A.H.-M.; Investigation, E.S.; resources, A.H.-M.; Data curation, E.S.; Writing—original draft preparation, E.S. and A.H.-M.; Writing—review and editing, E.S. and A.H.-M.; Visualization, E.S.; Supervision, A.H.-M.; Project administration, A.H.-M.; Funding acquisition, A.H.-M.

Funding: This work was supported in part by the NSF grants EECS-1546980, CCF-1434600 and the NSF Center for Science of Information (CSoI), and by Shenzhen Peacock Plan under Grant No. KQTD2015033114415450.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Linear Prediction

We showed

$$f(x^n|\tau,\mathbf{w}) = \frac{1}{(2\pi\tau)^{(n-M)/2}} \times \exp\left(-\frac{1}{2\tau}\left[\hat{r}_{(n)}(0) - 2\mathbf{w}^T\mathbf{p}_{(n)} + \mathbf{w}^T R^{(M)}_{(n)}\mathbf{w}\right]\right),$$

therefore using NLM we have

$$\begin{aligned} C(x^n) &= \int\int f(x^n|\tau,\mathbf{w})d\mathbf{w}d\tau \\ &= A\int_\tau \tau^{-\frac{(n-M)}{2}} \exp\left\{-\frac{\hat{r}_{(n)}(0)}{2\tau}\right\} e_1(\tau)\, d\tau \end{aligned}$$

where $A = \frac{1}{(2\pi)^{(n-M)/2}}$ and $e_1(\tau) = \int_{\mathbf{w}} \exp\left\{-\frac{1}{2\tau}\left[\mathbf{w}^T R_{(n)}\mathbf{w} - 2\mathbf{p}^T_{(n)}\mathbf{w}\right]\right\} d\mathbf{w}$. Hence

$$\begin{aligned} C(\mathbf{x}^n) &= B\int_\tau \tau^{-\frac{n-2M}{2}} \exp\left\{-\frac{1}{2\tau}\left[\hat{r}_{(n)}(0) - \mathbf{p}^T_{(n)}R^{-1}_{(n)}\mathbf{p}_{(n)}\right]\right\} d\tau \\ &= B\int_\tau \tau^{-\frac{n-2M}{2}} \exp\left\{-\frac{1}{2\tau}\hat{\tau}^{(M)}_{(n)}\right\} d\tau \\ &= \frac{1}{2\left(\pi\right)^{\frac{n-2M}{2}}\sqrt{\det\left(R_{(n)}\right)}} \frac{\Gamma\left(\frac{n-2M-2}{2}\right)}{\left(\hat{\tau}^{(M)}_{(n)}\right)^{\frac{n-2M-2}{2}}} \end{aligned}$$

where $B = \frac{1}{(2\pi)^{(n-2M)/2}\sqrt{\det\left(R_{(n)}\right)}}$.

Appendix B. Vector Gaussian Case: Unknown Σ

We showed that Σ has Inverse-Wishart distribution $\Sigma \sim \mathcal{W}_M^{-1}\left(\hat{\Sigma}_n, n\right)$ where $\hat{\Sigma}_n = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i$, hence

$$f_{\mathbf{x}^n}(\Sigma) = \frac{\det\left(\hat{\Sigma}_n\right)^{\frac{n}{2}}}{2^{\frac{nM}{2}}\Gamma_M\left(\frac{n}{2}\right)} \det(\Sigma)^{-\frac{n+M+1}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_n\Sigma^{-1}\right\},$$

and since

$$f(\mathbf{x}_{n+1}|\Sigma) = \frac{1}{\sqrt{(2\pi)^M \det(\Sigma)}} \operatorname{etr}\left\{-\frac{1}{2}\left(\hat{\Sigma}_{n+1} - \hat{\Sigma}_n\right)\Sigma^{-1}\right\},$$

therefore we have

$$\begin{aligned} f\left(\mathbf{x}_{n+1}|\mathbf{x}^{n}\right) &= \int_{\Sigma>0} f\left(\mathbf{x}_{n+1}|\Sigma\right) f_{\mathbf{x}^{n}}\left(\Sigma\right) d\Sigma \\ &= C\int_{\Sigma>0} \det\left(\Sigma\right)^{-\frac{n+M+2}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_{n+1}\Sigma^{-1}\right\} d\Sigma \\ &\overset{(A)}{=} C\int_{Y>0} \det\left(Y\right)^{\frac{n}{2}-\frac{M}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_{n+1}Y\right\} dY \\ &= D\int_{V>0} \det\left(V\right)^{\frac{n}{2}-\frac{M}{2}} \operatorname{etr}\left\{-V\right\} dV \\ &\overset{(B)}{=} D\int_{V>0} \det\left(V\right)^{\frac{n+1}{2}-\frac{M+1}{2}} \operatorname{etr}\left\{-V\right\} dV \\ &= \frac{1}{\pi^{\frac{M}{2}}} \frac{\det\left(\hat{\Sigma}_{n}\right)^{\frac{n}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n+1}{2}}} \frac{\Gamma_{M}\left(\frac{n+1}{2}\right)}{\Gamma_{M}\left(\frac{n}{2}\right)} \end{aligned}$$

where $C = \frac{\det\left(\hat{\Sigma}_{n}\right)^{\frac{n}{2}}}{2^{\frac{M(n+1)}{2}}\Gamma_{M}\left(\frac{n}{2}\right)\pi^{\frac{M}{2}}}$ and $D = \frac{\det\left(\hat{\Sigma}_{n}\right)^{\frac{n}{2}}}{\det\left(\hat{\Sigma}_{n+1}\right)^{\frac{n+1}{2}}} \frac{1}{\Gamma_{M}\left(\frac{n}{2}\right)\pi^{\frac{M}{2}}}$, and in equations (A) and (B) we changed the variable $\Sigma = Y^{-1}$ and $Y = 2\hat{\Sigma}_{n}^{-\frac{1}{2}} V \hat{\Sigma}_{n}^{-\frac{1}{2}}$ respectively and $\Gamma_{m}\left(a\right) = \int_{V>0} \det\left(V\right)^{a-\frac{(m+1)}{2}} \operatorname{etr}\left\{-V\right\} dV$ is the multivariate Gamma function.

Appendix C. Vector Gaussian Case: Unknown Mean and Σ

We showed that $\boldsymbol{\mu} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{n}, \frac{1}{n}\Sigma\right)$ and $\Sigma \sim \mathcal{W}_{M}^{-1}\left(\hat{\Sigma}_{n}, n-1\right)$ where $\hat{\boldsymbol{\mu}}_{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i}$ and $\hat{\boldsymbol{\Sigma}}_{n} = \sum_{i=1}^{n}\left(\mathbf{x}_{i}-\hat{\boldsymbol{\mu}}_{n}\right)\left(\mathbf{x}_{i}-\hat{\boldsymbol{\mu}}_{n}\right)^{T}$. Now using Bayes we can write the joint pdf as $f_{\mathbf{x}^{n}}\left(\boldsymbol{\mu},\Sigma\right) = f_{\mathbf{x}^{n}}\left(\boldsymbol{\mu}|\Sigma\right) f_{\mathbf{x}^{n}}\left(\Sigma\right)$. Define $A \overset{\text{def}}{=} f\left(\mathbf{x}_{n+1}|\mathbf{x}^{n}\right) = \int_{\Sigma>0}\int f\left(\mathbf{x}_{n+1}|\boldsymbol{\mu},\Sigma\right) f_{\mathbf{x}^{n}}\left(\boldsymbol{\mu},\Sigma\right) d\boldsymbol{\mu} d\Sigma$,

$$A = B\int_{\Sigma>0} \det\left(\Sigma\right)^{-\frac{n+M+2}{2}} e_{1}\left(\Sigma\right) e_{2}\left(\Sigma\right) d\Sigma$$

where

$$\begin{aligned} e_{1}\left(\Sigma\right) &= \operatorname{etr}\left\{-\frac{1}{2}\left(\hat{\Sigma}_{n} + n\hat{\boldsymbol{\mu}}_{n}\hat{\boldsymbol{\mu}}_{n}^{T} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^{T}\right)\Sigma^{-1}\right\} \\ e_{2}\left(\Sigma\right) &= \int \exp\left\{-\frac{n+1}{2}\left[\boldsymbol{\mu}^{T}\Sigma^{-1}\boldsymbol{\mu} - 2\hat{\boldsymbol{\mu}}_{n+1}\Sigma^{-1}\boldsymbol{\mu}\right]\right\} d\boldsymbol{\mu} \\ &= \sqrt{\frac{\left(2\pi\right)^{M}\det\left(\Sigma\right)}{\left(n+1\right)^{M}}} \exp\left\{\frac{n+1}{2}\hat{\boldsymbol{\mu}}_{n+1}^{T}\Sigma^{-1}\hat{\boldsymbol{\mu}}_{n+1}\right\} \\ B &= \frac{\det\left(\hat{\Sigma}_{n}\right)^{\frac{n-1}{2}}}{\Gamma_{M}\left(\frac{n-1}{2}\right)} \frac{n^{\frac{M}{2}}}{2^{\frac{M(n-1)}{2}}\left(2\pi\right)^{M}}. \end{aligned}$$

Now since $\hat{\Sigma}_{n+1} = \hat{\Sigma}_{n} + n\hat{\boldsymbol{\mu}}_{n}\hat{\boldsymbol{\mu}}_{n}^{T} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^{T} - \left(n+1\right)\hat{\boldsymbol{\mu}}_{n+1}\hat{\boldsymbol{\mu}}_{n+1}^{T}$, by defining $C \overset{\text{def}}{=} B\sqrt{\frac{\left(2\pi\right)^{M}}{\left(n+1\right)^{M}}} = \sqrt{\left(\frac{n}{n+1}\right)^{M}} \frac{\det\left(\hat{\Sigma}_{n}\right)^{\frac{n-1}{2}}}{\Gamma_{M}\left(\frac{n-1}{2}\right)} \frac{1}{2^{\frac{M(n-1)}{2}}\left(2\pi\right)^{\frac{M}{2}}}$ we can write

$$\begin{aligned} A &= C \int_{\Sigma>0} \det(\Sigma)^{-\frac{n+M+1}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_{n+1}\Sigma^{-1}\right\} d\Sigma \\ &= C \int_{Y>0} \det(Y)^{\frac{n}{2}-\frac{M+1}{2}} \operatorname{etr}\left\{-\frac{1}{2}\hat{\Sigma}_{n+1}Y\right\} dY \\ &= C \frac{2^{\frac{Mn}{2}}}{\det(\hat{\Sigma}_{n+1})^{\frac{n}{2}}} \int_{V>0} \det(V)^{\frac{n}{2}-\frac{M+1}{2}} \operatorname{etr}\{-V\} dV \\ &= \frac{1}{\pi^{\frac{M}{2}}} \sqrt{\left(\frac{n}{n+1}\right)^{M}} \frac{\det(\hat{\Sigma}_n)^{\frac{n-1}{2}}}{\det(\hat{\Sigma}_{n+1})^{\frac{n}{2}}} \frac{\Gamma_M\left(\frac{n}{2}\right)}{\Gamma_M\left(\frac{n-1}{2}\right)}. \end{aligned}$$

References

1. Høst-Madsen, A.; Sabeti, E.; Walton, C. Data Discovery and Anomaly Detection Using Atypicality: Theory. *IEEE Trans. Inf. Theory* **2016**, submitted.
2. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection for Discrete Sequences: A Survey. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 823–839. [CrossRef]
3. Li, Y.; Nitinawarat, S.; Veeravalli, V.V. Universal Outlier Hypothesis Testing. *IEEE Trans. Inf. Theory* **2014**, *60*, 4066–4082. [CrossRef]
4. Li, Y.; Nitinawarat, S.; Veeravalli, V.V. Universal Outlier Detection. In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 10–15 February 2013; pp. 1–5.
5. Li, Y.; Nitinawarat, S.; Veeravalli, V.V. Universal Sequential Outlier Hypothesis Testing. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 3205–3209.
6. Grimmett, G.R.; Stirzaker, D.R. *Probability and Random Processes*, 3rd ed.; Oxford University Press: Oxford, UK, 2001.
7. Sabeti, E.; Host-Madsen, A. Atypicality for the Class of Exponential Family. In Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 27–30 September 2016.
8. Kay, S.M. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*; Prentice-Hall: Upper Sadle River, NJ, USA, 1993.
9. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
10. Cover, T.; Thomas, J. *Information Theory*, 2nd ed.; John Wiley: Hoboken, NJ, USA, 2006.
11. Ziv, J.; Lempel, A. A Universal Algorithm for Sequential Data Compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343. [CrossRef]
12. Ziv, J.; Lempel, A. Compression of Individual Sequences via Variable-Rate Coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536. [CrossRef]
13. Ghido, F.; Tabus, I. Sparse Modeling for Lossless Audio Compression. *IEEE Trans. Audio Speech Lang. Proc.* **2013**, *21*, 14–28. [CrossRef]
14. Rissanen, J. A Universal Prior for Integers and Estimation by Minimum Description Length. *Ann. Stat.* **1983**, *11*, 416–431. [CrossRef]
15. Kostina, V. Data Compression With Low Distortion and Finite Blocklength. *IEEE Trans. Inf. Theory* **2017**, *63*, 4268–4285. [CrossRef]
16. Rissanen, J. Stochastic Complexity and Modeling. *Ann. Stat.* **1986**, *14*, 1080–1100. [CrossRef]
17. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv.* **2009**, *41*, 15. [CrossRef]
18. Ranshous, S.; Shen, S.; Koutra, D.; Harenberg, S.; Faloutsos, C.; Samatova, N.F. Anomaly Detection in Dynamic Networks: A Survey. *WIREs Comput. Stat.* **2015**, *7*, 223–247. [CrossRef]
19. Lee, Y.J.; Yeh, Y.R.; Wang, Y.C.F. Anomaly Detection via Online Oversampling Principal Component Analysis. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 1460–1470. [CrossRef]

20. Pimentel, M.A.; Clifton, D.A.; Clifton, L.; Tarassenko, L. A Review of Novelty Detection. *Signal Process.* **2014**, *99*, 215–249. [CrossRef]
21. Esling, P.; Agon, C. Time-Series Data Mining. *ACM Comp. Surv. (CSUR)* **2012**, *45*, 12. [CrossRef]
22. Li, W.; Mahadevan, V.; Vasconcelos, N. Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 18–32. [PubMed]
23. Jia, Z.; Shen, C.; Yi, X.; Chen, Y.; Yu, T.; Guan, X. Big-Data Analysis of Multi-Source Logs for Anomaly Detection on Network-Based System. In Proceedings of the 13th IEEE Conference on Automation Science and Engineering (CASE), Xi'an, China, 20–23 August 2017; pp. 1136–1141.
24. Ahmed, M.; Mahmood, A.N.; Hu, J. A Survey of Network Anomaly Detection Techniques. *J. Netw. Comp. Appl.* **2016**, *60*, 19–31. [CrossRef]
25. Yoon, M.K.; Mohan, S.; Choi, J.; Christodorescu, M.; Sha, L. Learning Execution Contexts from System Call Distribution for Anomaly Detection in Smart Embedded System. In Proceedings of the Second International Conference on Internet-of-Things Design and Implementation, Pittsburgh, PA, USA, 18–21 April 2017; pp. 191–196.
26. Sari, A. A Review of Anomaly Detection Systems in Cloud Networks and Survey of Cloud Security Measures in Cloud Storage Applications. *J. Inf. Secur.* **2015**, *6*, 142. [CrossRef]
27. Høst-Madsen, A.; Sabeti, E.; Walton, C.; Lim, S.J. Universal Data Discovery Using Atypicality. In Proceedings of the 3rd International Workshop on Pattern Mining and Application of Big Data (BigPMA 2016) at the 2016 IEEE International Conference on Big Data (Big Data 2016), Washington, DC, USA, 5–8 December 2016.
28. Han, C.; Willett, P.; Chen, B.; Abraham, D. A Detection Optimal Min-Max Test for Transient Signals. *IEEE Trans. Inf. Theory* **1998**, *44*, 866–869. [CrossRef]
29. Wang, Z.; Willett, P. A Performance Study of Some Transient Detectors. *IEEE Trans. Signal Proc.* **2000**, *48*, 2682–2685. [CrossRef]
30. Wang, Z.; Willett, P.K. All-Purpose and Plug-In Power-Law Detectors for Transient Signals. *Trans. Signal Proc.* **2001**, *49*, 2454–2466. [CrossRef]
31. Wang, Z.J.; Willett, P. A Variable Threshold Page Procedure for Detection of Transient Signals. *IEEE Trans. Signal Proc.* **2005**, *53*, 4397–4402. [CrossRef]
32. Guépié, B.K.; Fillatre, L.; Nikiforov, I. Sequential Detection of Transient Changes. *Seq. Anal.* **2012**, *31*, 528–547. [CrossRef]
33. Egea-Roca, D.; López-Salcedo, J.A.; Seco-Granados, G.; Poor, H.V. Performance Bounds for Finite Moving Average Tests in Transient Change Detection. *IEEE Trans. Signal Proc.* **2018**, *66*, 1594–1606. [CrossRef]
34. Guépié, B.K.; Fillatre, L.; Nikiforov, I. Detecting a Suddenly Arriving Dynamic Profile of Finite Duration. *IEEE Trans. Inf. Theory* **2017**, *63*, 3039–3052.
35. Hirai, S.; Yamanishi, K. Detecting Changes of Clustering Structures Using Normalized Maximum Likelihood Coding. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 343–351.
36. Yamanishi, K.; Miyaguchi, K. Detecting Gradual Changes from Data Stream Using MDL-Change Statistics. In Proceedings of the IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 156–163.
37. Killick, R.; Fearnhead, P.; Eckley, I.A. Optimal Detection of Changepoints with a Linear Computational Cost. *J. Am. Stat. Assoc.* **2012**, *107*, 1590–1598. [CrossRef]
38. Zou, S.; Fellouris, G.; Veeravalli, V.V. Quickest Change Detection under Transient Dynamics: Theory and Asymptotic Analysis. *IEEE Trans. Inf. Theory* **2018**, 1. [CrossRef]
39. Molloy, T.L.; Ford, J.J. Minimax Robust Quickest Change Detection in Systems and Signals with Unknown Transients. *IEEE Trans. Autom. Control* **2018**, 1. [CrossRef]
40. Veeravalli, V.V.; Banerjee, T. Quickest Change Detection. *Acad. Press Library Signal Proc.* **2013**, *3*, 209–256.
41. Fuh, C.D.; Tartakovsky, A.G. Asymptotic Bayesian Theory of Quickest Change Detection for Hidden Markov Models. *IEEE Trans. Inf. Theory* **2019**, *65*, 511–529. [CrossRef]
42. Lavielle, M. Using Penalized Contrasts for the Change-Point Problem. *Signal Proc.* **2005**, *85*, 1501–1510. [CrossRef]
43. Larsen, R.J.; Marx, M. *An Introduction to Mathematical Statistics and Its Applications*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1986; Volume 2.

44. Roos, T.; Rissanen, J. On Sequentially Normalized Maximum Likelihood Models. In Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08), Tampere, Finland, 18 August 2008.
45. Sabeti, E.; Host-Madsen, A. Enhanced MDL with Application to Atypicality. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
46. Scharf, L.L. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*; Addison-Wesley: Boston, MA, USA, 1990.
47. Grunwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
48. Rissanen, J. *Stochastic Complexity in Statistical Inquiry*; World Scientific: Singapore, 1998; Volume 15.
49. Forchini, G. The Density of the Sufficient Statistics for a Gaussian AR(1) Model in Terms of Generalized Functions. *Stat. Probab. Let.* **2000**, *50*, 237–243. [CrossRef]
50. Mallat, S. *A Wavelet Tour of Signal Processing: The Sparse Way*; Academic press: Cambridge, MA, USA, 2008.
51. Vetterli, M.; Kovacevic, J. *Wavelets and Subband Coding*; Prentice Hall: Englewood Cliffs, NJ, USA, 1995; Volume 995.
52. Vetterli, M.; Herley, C. Wavelets and Filter Banks: Theory and Design. *IEEE Trans. Signal Process.* **1992**, *40*, 2207–2232. [CrossRef]
53. Mitra, S.K.; Kuo, Y. *Digital Signal Processing: A Computer-Based Approach*; McGraw-Hill New York: New York, NY, USA, 2006; Volume 2.
54. Willems, F.M.J.; Shtarkov, Y.; Tjalkens, T. The Context-Tree Weighting Method: Basic Properties. *IEEE Trans. Inf. Theory* **1995**, *41*, 653–664. [CrossRef]
55. Willems, F.; Shtarkov, Y.; Tjalkens, T. Reflections on "The Context Tree Weighting Method: Basic properties". *Newslett. IEEE Inf. Theory Soc.* **1997**, *47*, 1.
56. Sabeti, E.; Høst-Madsen, A. How interesting images are: An Atypicality Approach For Social Networks. In Proceedings of the IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016.
57. Muirhead, R.J. *Aspects of Multivariate Statistical Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 197.
58. Silver, K. A Passive Acoustic Automated Detector for Sei and Fin Whale Calls. Master's Thesis, University of Hawaii, Honolulu, HI, USA, 12 November 2014.
59. Host-Madsen, A.; Sabeti, E. Atypical Information Theory for Real-Valued Data. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 666–670.
60. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef] [PubMed]