LUND UNIVERSITY

Estimating Periodicities in Symbolic Sequences Using Sparse Modelling

S. I. ADALBJÖRNSSON, J. SWÄRD, J. WALLIN, AND A. JAKOBSSON

Published in: IEEE Transactions on Signal Processing
doi:10.1109/TSP.2015.2404314

Lund 2015

Mathematical Statistics
Centre for Mathematical Sciences
Lund University

Estimating Periodicities in Symbolic Sequences Using Sparse Modelling

Stefan I. Adalbjörnsson*, Johan Swärd*, Jonas Wallin†, and Andreas Jakobsson*

***Abstract*—In this work, we propose a method for estimating statistical periodicities in symbolic sequences. Different from other common approaches used for the estimation of periodicities of sequences of arbitrary, finite, symbol sets, that often map the symbolic sequence to a numerical representation, we here exploit a likelihood-based formulation in a sparse modeling framework to represent the periodic behavior of the sequence. The resulting criterion includes a restriction on the cardinality of the solution; two approximate solutions are suggested, one greedy and one using an iterative convex relaxation strategy to ease the cardinality restriction. The performance of the proposed methods are illustrated using both simulated and real DNA data, showing a notable performance gain as compared to other common estimators.**

***Index Terms*—Periodicity, symbolic sequences, spectral estimation, data analysis, DNA**

I. Introduction

SEQUENCES formed from a finite set of symbols, or *alphabet*, occur in a variety of fields, such as, for instance, in genomics, semantic analysis, and categorical time series [1], [2]. Frequently, there is an interest in determining reoccurring patterns, periodicities, in such sequences. For instance, in DNA analysis, the latent periodicities in DNA sequences, commonly assumed to be stationary in short time intervals, have been found to be correlated with various forms of functional roles of importance [3]–[11]. Traditional spectral estimation techniques are not suitable for this problem as symbolic sequences lack algebraic structures. For DNA analysis, there is no natural ordering among the four occurring symbols, A, C, G, and T. In earlier literature, several authors have addressed the problem of estimating symbolic periodicity using heuristic mappings from the symbol set to sets of complex numbers. After the transformation the periodicities are estimated through standard estimation methods like, for instance, the periodogram. However, such estimates will suffer from the well-known high variability and/or poor resolution inherent to the periodogram [12]. Other examples of methods that use a mapping to transform the symbolic data include PAM- or QPSK-based mappings, minimum entropy mapping, mapping equivalences, or other transformations [4]–[7], [9], [10], [13], [14]. Generally, these mappings are computationally intensive, and/or suffer from difficulties expanding to a larger symbol sets, and often inadvertently impose a non-existing structure on the symbols. In this work, we instead use a probabilistic approach, modeling the symbolic sequences using a categorical distribution for each observation and try to infer not only the unknown probabilities but also the unknown indices where the distribution differs, resulting in a likelihood ratio test, which, for a given index set, is equivalent with the well studied problem of testing for independence in $2 \times J$ contingency tables, where J denotes the number of categories, see, e.g., [2]. Ideally, an estimator for this problem should be able to discern not only whether the distribution differs at a certain periodicity, but also how many indices have differing distributions. If more than one statistical periodicity is considered at the same time, the number of possible combinations of index sets grows rapidly and an exact test will in many cases be computationally infeasible. By formulating the estimation of the unknown index sets, and the unknown probabilities, as a sparse logistic regression problem, we devise two approximate solutions to the combinatorial problem using sparse heuristics. Namely, one greedy approach which builds up the solution by adding the sets in a sequential manner, and one using a convex relaxation of the cardinality constraint, resulting in the well-known (reweighted) LASSO problem. The resulting methods are firmly based in statistical theory, and also easily generalized to any finite symbol set.

The remainder of the paper is organized as follows: in the next section, we introduce the considered data model and show how the problem of choosing which indices that show a periodic change in the distribution can be interpreted as a sparse estimation problem. Then, in section III, we introduce a greedy algorithm that approximately solves the sparse problem, as well as a convex relaxation of the original problem, which may be efficiently solved using convex optimization algorithms. Then, in section IV, we outline some implementation issues, including a cyclic coordinate descent algorithm for solving the resulting convex relaxation problem. In section V, we examine the performance of the discussed estimators, showing the benefits of the proposed approach as compared to previously published methods. Finally, we conclude on the work in section VI.

II. Probabilistic model for symbolic sequences

Consider a symbolic sequence, $\{s_k\}_{k=1}^N$, where each symbol, s_k, is a stochastic variable drawn from a finite set,

This work was supported in part by the Swedish Research Council, Knut and Alice Wallenberg foundation, and Carl Trygger's foundation. This work has been presented in part at the EUSIPCO 2013 conference [1].

*Dept. of Mathematical Statistics, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden, email: {sia,js,aj}@maths.lth.se.

†Dept. of Mathematical Statistics, Chalmers University, SE-412 96 Gothenburg, Sweden, email: jonwal@chalmers.se.

$\mathcal{A} = \{\alpha_1, \ldots, \alpha_B\}$, where B denotes the size of the alphabet. Assume that the symbols in the sequence are independent and identically distributed, such that

$$p_j \triangleq \text{Prob}(s_k = \alpha_j) \tag{1}$$

Then, if gathering a sequence of observations, $x_1, \ldots, x_N$, into the vector $\mathbf{x}$, the probability mass function (PMF) of $\mathbf{x}$ is given as

$$\begin{aligned} q_0(\mathbf{x}|\mathbf{p}) &\triangleq \text{Prob}(\mathbf{s} = \mathbf{x}) && (2)\\ &= \prod_{j=1}^{N}\prod_{\ell=1}^{B} p_\ell^{[x_j=\alpha_\ell]} = \prod_{\ell=1}^{B} p_\ell^{G_\ell} && (3)\end{aligned}$$

where $[\cdot]$ denotes the Iverson's bracket, which equals one if the statement inside the brackets is true, and zero otherwise, with each of the symbols appearing G_k times, and where $\mathbf{p}$ and $\mathbf{s}$ denote the vector of probabilities and the sequence of random variables, respectively, i.e.,

$$\begin{aligned} \mathbf{p} &= \begin{bmatrix} p_1 & \ldots & p_B \end{bmatrix}^T && (4)\\ \mathbf{s} &= \begin{bmatrix} s_1 & \ldots & s_N \end{bmatrix}^T && (5)\end{aligned}$$

with $(\cdot)^T$ denoting the transpose. As a result, the PMF is a function depending only on the number of times each symbol appears, and on the probability given to each symbol. In general, the probabilities, p_k, are unknown and need to be estimated from the observed sequence. This can be done using the maximum likelihood (ML) estimate, formed as

$$\hat{p}_j = \frac{G_j}{N} \tag{6}$$

for $j = 1, \ldots, B$, which is an unbiased and asymptotically efficient estimate (see, e.g., [15, p. 475]). Furthermore, note that a symbol $\alpha \in \mathcal{A}$, occurring with periodicity m, i.e., with the symbol appearing at every mth index in the sequence, implies that all elements of the sequence should be equal to the symbol α in one of the m possible (disjoint) index sets

$$I(m, \ell) = \left\{ \ell, \ell + m, \ldots, \ell + \left\lfloor \frac{N-\ell}{m} \right\rfloor m \right\} \tag{7}$$

for all offsets $\ell \in \{1, \ldots, m\}$, where $\lfloor \cdot \rfloor$ denotes the rounding down operation. This means that if a periodicity m is present in a sequence, the sequence is clearly also periodic on the subharmonics i.e., for every mr:th symbol, for all natural numbers r [8]. To avoid ambiguity, we here refer to the period as the lowest possible such periodicity. Considering a sequence, $\mathbf{s}$, with a periodicity m in the symbol α, with offset n, this implies that all the symbols in the sequence at index k, will equal α, for $k \in I(m, n)$. Thus, it is a deterministic and not a statistical problem to determine if such a (deterministic) periodicity is present. However, of more interest are typically the statistical periodicities that occur in many forms of symbolic sequences, such as, e.g., DNA sequences. These are characterized by certain index sets having different distributions, such that the sequence may contain the periodicity over only a limited interval, and/or with some of the periodically occurring symbols occasionally being replaced by some other symbol, which may occur, for example, due to the presence of measurement noise, coding errors, or some, perhaps unknown, functional equivalence between symbols [3]. In such cases, the PMF for a symbolic sequence might instead be formed from two distribution, one for the indices, say I_1, corresponding to some unknown periodic index set $I(m, l)$, and another distribution for the complement index set, here denoted I_0. In this case, the PMF is

$$\begin{aligned} q_1(\mathbf{x}|\mathbf{p}_0, \mathbf{p}_1) &\triangleq \prod_{j=1}^{N}\prod_{\ell=1}^{B} p_{0,\ell}^{[x_j=\mathcal{A}_\ell][j\in I_0]} p_{1,\ell}^{[x_j=\mathcal{A}_\ell][j\in I_1]} \\ &= \prod_{\ell=1}^{B} p_{0,\ell}^{G_{0,\ell}} p_{1,\ell}^{G_{1,\ell}} && (8)\end{aligned}$$

where $\mathbf{p}_0$, and similarly for $\mathbf{p}_1$, is a parameter vector containing the probabilities $p_{0,k}$, denoting the probability of a symbol, α_k, occurring in the index set I_0, and with $G_{0,k}$ and $G_{1,k}$ denoting the number of times the symbol α_k occurs in the set $I(m, n)$ and in its complement, respectively. The corresponding ML estimates are found as

$$\hat{p}_{0,j} = \frac{G_{0,j}}{|I_0|} \tag{9}$$

$$\hat{p}_{1,j} = \frac{G_{1,j}}{|I_1|} \tag{10}$$

for $j = 1, \ldots, B$, where $|S|$ denotes the cardinality of a set S, i.e., the number of elements in S. In a similar fashion, the addition of more than one periodicity can be accomplished by defining the distribution on more index sets, e.g. if one considers M disjoint index sets, $I_0, \ldots, I_{M-1}$, so that their union corresponds to the entire sequence, the PMF is

$$\begin{aligned} &q_1(\mathbf{x}|\mathbf{p}_0, \ldots, \mathbf{p}_{M-1}) \triangleq && (11)\\ &\prod_{m=0}^{M-1}\prod_{k=1}^{B} p_{m,k}^{G_{m,k}} && (12)\end{aligned}$$

where $G_{m,k}$ denotes the number of times the symbol α_k occurs in the set I_m. Comparing the likelihood above with (3), it can be seen that (11) corresponds to a likelihood for i.i.d. categorical variables, within each of the M index sets. However, note this does not assume that the sequence consists of i.i.d. variables, only that knowing the index sets we can split the sequence into sub sequences of i.i.d. variables.

A similar model was considered in [8], although there they defined a statistical periodicity, say k, to be present when all index set $I(k, \ell)$, for $\ell = 1 \ldots, k$, have different distributions, and then set out to find the periodicity, k, by maximizing the log-likelihood using an information-theoretic criterion penalty term to select the correct periodicity. If doing so, and the signal has a periodicity of k, then each index set corresponding to a different offset also has a unique distribution, implying a subdivision of the data into $\lfloor N/k \rfloor$ disjoint data sets, resulting in less data to be used to estimate these probabilities. For multiple periodicities, i.e., several index sets with different distributions, this results in a necessity to consider the overall periodicity of the sequence, i.e., if periods l and k are present, then the sequence will have a periodicity of lk, resulting in the need for substantially more data to achieve a similar performance as if only a single periodicity was present, as well as the need to perform on additional analysis to identify

the factors constituting lk. Furthermore, in the case when the sequence contains more than two periodicities, the problem quickly becomes infeasible. We instead want to find the index sets where the distributions differ as much as possible from the rest of the sequence. To that end, we recast the estimation problem in a sparse modeling framework. To do so, we note that one can interpret (12) as a multi-response logistic regression problem, which, as we will show, will be particularly useful for the case of several simultaneous periodicities. Furthermore, this mapping allows us to consider sequences one symbol at a time, which is particularly useful when the periodicity in a certain symbol is sought, or if the distribution of a particular symbol deviates especially much on a given index set. This, when applicable, decreases the variance of the estimated probabilities, thus improving the detection of periodicities only occurring in one symbol, or one subset of symbols. Rewriting (12) using logistic regression is accomplished by modeling the probability of each observation separately using a logistic function to map a linear model to the interval $[0,\ 1]$. To clarify the exposition, we first consider the case of a binary symbol set, a special case which will be shown to be particularly useful. Thus, consider a binary sequence which has a statistical periodicity on the indices I_1, and some other distribution on the indices I_0, so that the PMF may be expressed as

$$q_1(\mathbf{x}|\boldsymbol{\gamma}(\mathbf{c})) \triangleq \prod_{k=1}^{N} \gamma_k(\mathbf{c})^{x_k}(1-\gamma_k(\mathbf{c}))^{1-x_k} \tag{13}$$

where $\boldsymbol{\gamma}(\mathbf{c}) \in \mathbf{R}^N$ is a vector of probabilities, such that

$$Pr(s_k = 1) = \gamma_k(\mathbf{c}) \tag{14}$$

and the vector $\mathbf{c} \in \mathbf{R}^2$ models the probabilities for the index sets I_1 and its complement, I_0, such that

$$\boldsymbol{\gamma}(\mathbf{c}) = \begin{bmatrix} \gamma_1(\mathbf{c}) & \dots & \gamma_N(\mathbf{c}) \end{bmatrix}^T \tag{15}$$

$$\gamma_k(\mathbf{c}) = \frac{e^{\mathbf{h}_k^T\mathbf{c}}}{1+e^{\mathbf{h}_k^T\mathbf{c}}} \tag{16}$$

where

$$\mathbf{h}_k = \begin{cases} \begin{bmatrix} 1 & 1 \end{bmatrix}^T & \text{if } k \in I_1 \\ \begin{bmatrix} 1 & 0 \end{bmatrix}^T & \text{if } k \notin I_1 \end{cases} \tag{17}$$

Thus, there is a simple relationship between the parameters $p_{0,1}$ and $p_{1,1}$ in the original model in (8), i.e.,

$$P(s_k = 1) = p_{0,1} \quad \text{for } k \in I_0 \tag{18}$$

$$P(s_k = 1) = p_{1,1} \quad \text{for } k \in I_0 \tag{19}$$

and the parameter vector, $\mathbf{c}$, introduced in (13), i.e.,

$$\log\left(\frac{p_{0,1}}{1-p_{0,1}}\right) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T \mathbf{c} \tag{20}$$

$$\log\left(\frac{p_{1,1}}{1-p_{1,1}}\right) = \begin{bmatrix} 1 & 1 \end{bmatrix}^T \mathbf{c} \tag{21}$$

It should be noted that (20) implies that the probability of a symbol appearing in the set I_0 is given by the first element of the vector $\mathbf{c}$, and, similarly, one may by substituting (20) into (21) and simplifying, note that

$$\log\left(\frac{p_{1,1}}{1-p_{1,1}}\right) - \log\left(\frac{p_{0,1}}{1-p_{0,1}}\right) = \begin{bmatrix} 0 & 1 \end{bmatrix}^T \mathbf{c} \tag{22}$$

Thus, the second element in $\mathbf{h}_k$ control the change in probability on the index set, I_1, as compared to the indices in the set, I_0, e.g., if the second element is zero, then the probabilities are the same for both sets, whereas a positive or negative second element implies higher or lower probabilities on the set I_1, respectively. Extending the model to allow for the possibility of several periodicities using the logistic regression parameterization can be achieved by adding elements to the $\mathbf{c}$ vector such that each new element adjusts the probability for an additional index set. To that end, consider the case with M index sets, I_j, for $j = 1, \dots, M$, corresponding to some specific periodicities with their different offsets, then $\mathbf{c} \in \mathbf{R}^M$ and every element of $\mathbf{h}_k^T \in \mathbf{R}^M$ is zero except the elements where k is in the corresponding index set, i.e.,

$$h_{k,j} = \begin{cases} 1 & k \in I_j \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

for $j = 1, \dots, M$, and $d_{k,j}$ denotes element j of the vector $\mathbf{d_k}$. The resulting model can then be seen as the solution of the following optimization criterion

$$\begin{aligned} \underset{\mathbf{c}}{\text{maximize}} \quad & \prod_{k=1}^{N} \gamma_k(\mathbf{c})^{x_k}(1-\gamma_k(\mathbf{c}))^{1-x_k} \\ \text{subject to} \quad & \begin{cases} ||\mathbf{c}||_0 \leq L \\ \gamma_k(\mathbf{c}) = \frac{e^{\mathbf{h}_k^T\mathbf{c}}}{1+e^{\mathbf{h}^T\mathbf{c}}} \end{cases} \end{aligned} \tag{24}$$

where $||\cdot||_0$ denotes the ℓ_0 (pseudo) norm, which counts the number of nonzero elements of a vector, and L is the maximum number of periodicities that will be included in the model. It is worth noting that the expression for $\gamma_k(\mathbf{c})$ does not pose a restriction to the minimization, but has been included to emphasize that the probabilities for each observation are being modeled explicitly. Solving (24) for a given L, i.e., finding the maximum allowed number of simultaneous periodic sets, can be accomplished using an exhaustive search, since for each fixed k there are $(M)!/\left((M-j)!j!\right)$ index sets. For each such set, the ML estimates may then be found using (6). However, the dimension of the parameter vector will grow quadratically with the maximum periodicity considered, since

$$M = \sum_{k=1}^{m_{max}} k = \frac{m_{max}(m_{max}+1)}{2} \tag{25}$$

where m_{max} is the maximum allowed periodicity, since each period k has k corresponding index sets, one for each possible offset. Thus, to evaluate the likelihood for all combinations of index sets will soon lead to a computationally infeasible problem. Generalization to larger symbol sets may be carried out in a similar manner, leading to the multi-response logistic regression model (see, e.g., [2] for a further discussion on multi-response logistic regression). The corresponding optimization problem is therefore given as the maximum of the

log-likelihood with a cardinality constraint [16]

$$\begin{aligned} \underset{\mathbf{c}_1,\ldots,\mathbf{c}_B}{\text{maximize}} \quad & \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\ell=1}^{B} x_{i\ell}(\mathbf{h}_i^T\mathbf{c}_\ell) - \log\left(\sum_{\ell=1}^{B} e^{\mathbf{h}_i^T\mathbf{c}_\ell}\right)\right] \\ \text{subject to} \quad & ||\mathbf{C}_{k\cdot}||_0 \leq L, \quad \text{for } k = 1,\ldots,R \end{aligned} \tag{26}$$

where $\mathbf{C}$ is a matrix constructed such that its k:th column is formed by the vector $\mathbf{c}_k$, and R is the number of considered index sets, with $\mathbf{C}_{k\cdot}$ denoting the restriction that $||\mathbf{C}_{k\cdot}||_0$ forces the solution to adjust the B parameters corresponding to every index set simultaneously. Thus, the distributions can be changed on at most L index sets. As a result, the framework allows for flexibility in what is deemed a periodicity, e.g., one might test for a high probability of a certain symbol appearing, or even for if some symbols appear with low probability. Both of these ideas will be explored further in the following, where we outline a couple of possible algorithms for estimating periodicities for some commonly occurring situations, namely, estimation of an unknown periodicity, detection of an unknown periodicity, and, finally, estimation of multiple periodicities.

III. Relaxation of the cardinality constraint

For cardinality constrained, or sparse, least squares problems, there are a wide range of tools for forming approximate solutions, with many methods falling into two broad categories, namely greedy methods that build up a solution one variable at a time until either fitting criterion is satisfied, or the number of variables reaches the constraint, or methods that replace the cardinality constraint with a penalty function that promotes solutions that have few non-zero variables [17]. This implies that the optimization can be carried out without the combinatorial computation complexity inherent in cardinality constrained optimization problems. Typically, the penalty function is selected as the ℓ_1 norm, leading to a simple convex optimization problem. In the following two subsections, we propose both kinds of algorithms, first a greedy approach and then an iterative convex relaxation.

A. Greedy approach

In order to form a greedy estimate of the minimization in (26), one may note the analogy between this formulation and that of simple hypothesis test for testing if a distribution is different on some index sets (see also [3]). Thus, one may form a test to determine the hypothesis that a given sequence has a different distribution for the indices corresponding to $I(m,\ell)$, i.e., that the PMF is formed using (8), against the null hypothesis that the entire sequence has the same categorical distribution, such that the PMF instead follows (3), i.e.,

$$\mathrm{H}_0 : \mathbf{p}_0 = \mathbf{p}_1 \tag{27}$$

$$\mathrm{H}_1 : \mathbf{p}_0 \neq \mathbf{p}_1 \tag{28}$$

Such a test may be formed as a likelihood ratio (LR) test (see, e.g., [18, p. 375])

$$\lambda_{m,\ell}(\mathbf{x}_N) = \frac{q_0(\mathbf{x}_N|\mathbf{p}_0,\mathrm{H}_0)}{q_1(\mathbf{x}|\mathbf{p}_0,\mathbf{p}_1,\mathrm{H}_1)} \tag{29}$$

where the probabilities are determined using (6) under H_0, and using (9) and (10) under H_1. Thus, if one only seek to find a single index set, a suitable choice would be the one maximizing the LR, i.e.,

$$\underset{m,\ell}{\arg\min} \quad \lambda_{m,\ell}(\mathbf{x}_N) \tag{30}$$

If the number of periodicities is unknown, i.e., the problem is one of detection and not estimation, one can allow for the possibility of no set being added by considering that if H_0 is true, it holds asymptotically that [18, p. 489]

$$-2\log(\lambda_{m,\ell}(\mathbf{x}_N)) \xrightarrow{d} \chi^2_{B-1} \tag{31}$$

where $\xrightarrow{d}$ denotes convergence in distribution and χ^2_k denotes the chi-squared distribution with k degrees of freedom. Thus, if no periodicity is present, a critical value, denoted T_α, for the likelihood ratio, below which no periodicity is deemed to be present, can be constructed for the likelihood ratio for each of the tests. Since M tests are formed in order to compute (30), and if assuming that these are independent, the critical value may be well approximated using extreme value theory as a quantile of the random variable

$$\psi = max\left(z_1,\ldots,z_M\right) \tag{32}$$

where each z_k is χ^2 distributed, implying that ψ will follow a Gumbel distribution (see, e.g., [19, p. 156]). In the case when multiple periodicities may be present, one can extend this procedure using a step-wise approach. To do so, first define I_1 as the index set containing all the indices in the sequence. Then, the initial step is performed by using the above algorithm to determine an index set $I_2 = I_{m_1,\ell_1}$, where m_1 and ℓ_1 denote the initially estimated periodicity and offset, respectively, found in the minimization of (30). In order to determine the next periodicity, the H_0 distribution is formed from (12), using one distribution for the found index set I_2 and one for all the other indices, $I_1 \setminus I_2$, where $\setminus$ denotes set subtraction operation. The second phase, m_2, and periodicity, ℓ_2, may be determined using (30). This procedure can then be repeated until the zero hypothesis can not be rejected using a suitable quantile of (32), i.e., at iteration s the corresponding likelihood ratio test may be formed as

$$\lambda^{(s)}_{m,\ell}(\mathbf{x}_N) = \frac{q_0(\mathbf{x}_N|\mathbf{p}_0,\ldots,\mathbf{p}_{s-1},\mathrm{H}_0)}{q_1(\mathbf{x}|\mathbf{p}_0,\ldots,\mathbf{p}_s,\mathrm{H}_1)} \tag{33}$$

Note that this assumes that the sets I_k being added to the zero hypothesis are disjoint, otherwise the likelihood would include some data points more than once. To ensure this we propose to only consider the indices that have not already been added to H_0 when evaluating $q_1(\mathbf{x}|\mathbf{p}_0,\mathbf{p}_1,\mathrm{H}_1)$ in (29), i.e., at iteration k the sets $I(m,\ell)$ are replaced with $I(m,l) \leftarrow I(m,l) \setminus I_{k-1}$, for all m and ℓ, where $\leftarrow$ denotes that the quantity on the left is replaced with the one on the right. The resulting greedy algorithm, here termed the greedy *P*eriodicity *E*stimation of *C*ategorical *S*equences (PECS$_G$) estimator, is outlined in Algorithm 1 below, with each iteration requiring at most $\mathcal{O}(Bm_{max}N)$ operations.

Algorithm 1 The PECS$_{\mathrm{G}}$ estimator

Given a categorical sequence, $\mathbf{x}$ of length N
$I_0 = \{1, \ldots, N\}$
for $s = 1, \ldots$ **do**
  $\{m_s, \ell_s\} = \arg\max_{m,\ell} \lambda_{m,\ell}(\mathbf{x}_N)$
  if $\lambda_{m,\ell}(\mathbf{x}_N) > C_\alpha$ **then**
    $I_s = I_{m_s,\ell_s}$
  else
    break
  end if
  $I(m,l) \leftarrow I(m,l) \setminus I_s$ for all m and l
  $I_0 \leftarrow I_0 \setminus I_s$
  H_0 distribution is replaced with (12) using $I_0, \ldots, I_s$
end for

B. Iterative Convex Relaxation

It is worth noting that the optimization criterion in (24) is not convex as it restricts the parameter space to lie in a non-convex set. A commonly used relaxation for problems of this kind is to replace the ℓ_0 restriction with the convex ℓ_1 ball, which by taking the negative logarithm and using the Lagrange duality, results in the relaxed convex optimization criterion

$$\underset{\mathbf{c}}{\text{minimize}} \quad \sum_{k=1}^{N} -x_k \mathbf{h}_k^T \mathbf{c} + \log(1 + e^{\mathbf{h}_k^T \mathbf{c}}) + \lambda ||\mathbf{c}||_1 \tag{34}$$

where we have exploited the equality constraint for $p_k(\mathbf{c})$ and where $\lambda > 0$ is a tuning parameter, which may be set using, for example, cross validation (see e.g., [20]), or by an heuristic choice using the observation following equation (44). Some adjustments may be done to this criterion; firstly, the penalty on $\mathbf{c}$ includes the first element. This is not appropriate since the first element controls the probability for all observations, and we have no reason to want to bias that probability towards $1/2$. This is easily accomplished by only penalizing the other elements of the vector, i.e., replacing $||\mathbf{c}||_1$ with $||\underline{\mathbf{c}}||_1$, where $\underline{\mathbf{c}}$ denotes the resulting vector once the first element of $\mathbf{c}$ is removed. However, the resulting expression will also have an undesirable ambiguity due to the lack of distinction being made between if the probability is higher or lower on the periodic indices. For instance, consider a case when every third index starting with 1 has the probability 0.1 of being 1, and all other indices have probability 0.9 of being 1. Should this be considered two periodicities of 3 with probability 0.9, or one periodicity of 3 with probability 0.1? Such a distinction is of course not a problem specific for this model. However, since one is commonly interested in finding periodic indices where the probability is either higher or lower, such an ambiguous result would result in a non-consistent interpretation of the estimates. Fortunately, this can be easily handled by adding a constraint on $\underline{\mathbf{c}}$, ensuring that only periodicities with greater probability of a symbol appearing are considered, i.e., $c_k > 0$, for $k = 2, \ldots, M$, where c_i is the i:th element of the vector $\mathbf{c}$. This yields

$$\begin{aligned} \underset{\mathbf{c}}{\text{minimize}} \quad & \sum_{k=1}^{N} -x_k \mathbf{h}_k^T \mathbf{c} + \log(1 + e^{\mathbf{h}_k^T \mathbf{c}}) + \lambda ||\underline{\mathbf{c}}||_1 \\ \text{subject to} \quad & c_k \geq 0 \quad \text{for } k = 2, \ldots, M \end{aligned} \tag{35}$$

The resulting optimization is thus a sum of an affine function and the logarithm of a sum of exponential functions, and is thus a convex function. (see, e.g., [21, p. 93]). Thus, since the constraints can be seen as inequalities involving inner products with the Cartesian coordinate basis vectors, they are affine, and therefore convex functions, and the criterion is as a result a convex optimization problem in the standard form, as defined in [21, p. 136]. However, the criterion in (35) will not yield sufficiently sparse estimates, as a result of the rather coarse approximation of the ℓ_1 norm to the desired ℓ_0 norm. Recently, interest in non-convex penalties that are closer, in some sense, to the ℓ_0 norm have been suggested, such as the use of the ℓ_q norm, for $0 < q < 1$ (see e.g., [22], [23]). Herein, we consider an alternative approach where the ℓ_1 penalty is replaced with the concave $\log(\cdot)$ penalty. The resulting optimization is then solved with an iteratively re-weighted ℓ_1 minimization, using a technique suggested in [24]. The resulting algorithm thus solves, at iteration $j+1$, the minimization

$$\begin{aligned} \underset{\mathbf{c}}{\text{min.}} \quad & \sum_{k=1}^{N} -x_k \mathbf{h}_k^T \mathbf{c} + \log(1 + e^{\mathbf{h}_k^T \mathbf{c}}) + \lambda \sum_{k=2}^{M} \frac{|c_k|}{|\hat{c}_k^{(j)}| + \epsilon} \\ \text{s. t.} \quad & c_k \geq 0 \quad \text{for } k = 2, \ldots, M \end{aligned} \tag{36}$$

where $\hat{c}_k^{(j)}$ is the k:th element of the $\mathbf{c}$ estimate resulting from the j:th iteration, and ϵ is set as a small number to avoid numerical problems as well as to enable zero valued elements of $\mathbf{c}$ to transition from zero to non-zero values (see also [24]). The resulting sequence of convex minimizations yields a sufficiently sparse estimate of the periodicities (although at a high a computational complexity if implemented directly using a standard interior point-based solver). The resulting estimator is in the following referred to as the *P*eriodicity *E*stimation of *C*ategorical *S*equences using Logistic regression, PECS$_L$.
Comparing the two methods, PECS$_G$ offers a faster solution, whereas PECS$_L$ yields better results in the case of multiple periodicities. This is due to the fact that the iterative greedy procedure in PECS$_G$ does not take into account the overlap between the two index sets, e.g., the index sets $I(k,1) \cap I(l,1) = I(kl,1)$, whereas, the logistic regression approach also takes the overlap into account in the estimation procedure.

IV. Efficient Implementation

In order to form an efficient solver for the minimization in (36), we proceed to develop a cyclic coordinate descent (CCD) algorithm. The CCD algorithm minimize the cost function in (36) one variable at a time, in a cyclical fashion, holding the other variables fixed at their most recent estimates. This will thus transform the $M-$dimensional optimization problem into a scheme where one instead repeatedly solves simpler one-dimensional problems.

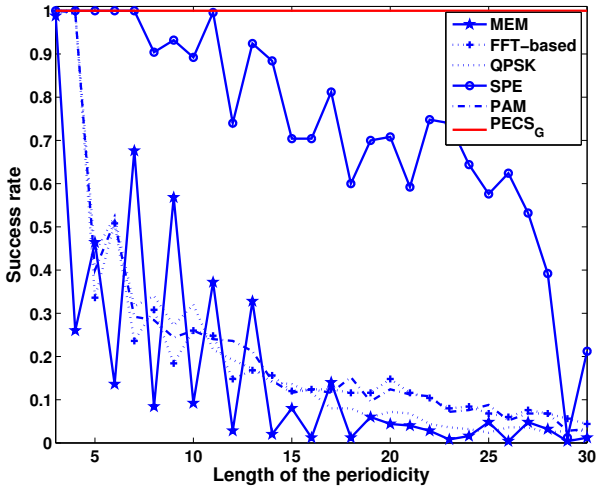


Fig. 1. Rate of success in estimating deterministic periods.

It should be noted that such an approach is, in general, converging notoriously slowly, or in some cases, not at all. However, for the optimization problems encountered in sparse modeling, this does no longer hold, as in fact, convergence proofs exist [20], [25], and in many applications, CCD implementations have emperically been shown to be the fastest algorithms available [16], [26]. Below, we outline the steps involved in a CCD algorithm for the case of $c_k \geq 0$, with the other case being handled in a similar manner. Thus, consider $c_i^{(r)}$ as the r:th estimate of element i of the vector $\boldsymbol{c}$, then, for $i > 1$,

$$\begin{aligned} c_i^{(r+1)} &= \arg\min_{c_i} \sum_{k=1}^{N} -x_k \mathbf{h}_k^T \mathbf{c} + \log(1 + e^{\mathbf{h}_k^T \mathbf{c}}) + \lambda ||\underline{\mathbf{c}}||_1 \\ &= \arg\min_{c_i} -\mathbf{x}^T \mathbf{H}_{(\cdot,i)} c_i + \lambda |c_i| + \sum_{k=1}^{N} \log(1 + a_{k,i} e^{h_{k,i} c_i}) \end{aligned} \tag{37}$$

The notation $\mathbf{H}_{(\cdot,i)}$ denotes the i:th column of the matrix $\mathbf{H}$, $h_{k,i}$ the i:th element of the vector $\mathbf{h}_k$, and

$$\boldsymbol{x} = \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}^T \tag{38}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_N \end{bmatrix}^T \tag{39}$$

$$\mathbf{c} = \begin{bmatrix} c_1^{(r+1)} & \dots & c_{(i-1)}^{(r+1)} & c_i^{(r)} & \dots & c_N^{(r)} \end{bmatrix}^T \tag{40}$$

$$a_{k,i} = \exp\left(\sum_{j,\, j \neq i} h_{k,j} c_j \right) \tag{41}$$

If the maximum value of the subdifferential set

$$\partial f_0 = -\mathbf{x}^T \mathbf{H}_{(\cdot,i)} + \lambda w + \sum_{k=1}^{N} \frac{a_{k,i} h_{k,i} e^{h_{k,i} c_i}}{1 + a_{k,i} e^{h_{k,i} c_i}} \tag{42}$$

with $c_i = 0$ is positive and $\{w \in [-1, 1]\}$, then the optimum is attained at $c_i = 0$ for the constrained optimization problem. On the other hand, if the maximum is negative, the stationary point may be found using a gradient approach (since the cost function is differentiable for all positive c_i). Note that this analysis gives insight into both the sparsity promoting effect of the ℓ_1 norm as well as the role of the tuning parameter λ, in fact, rewriting (42) as

$$\partial f_0 = -\mathbf{x}^T \mathbf{H}_{(\cdot,i)} + \lambda w + \mathbf{r}_i^T \mathbf{H}_{(\cdot,i)} \tag{43}$$

where $\mathbf{r}_i = \begin{bmatrix} \frac{a_{1,i}}{1+a_{1,i}} & \cdots & \frac{a_{N,i}}{1+a_{N,i}} \end{bmatrix}$ can be interpreted as probabilities for each index. Furthermore, $\mathbf{r}_i^T \mathbf{H}_{(\cdot,i)}$ is the expected number of symbols on the periodicity corresponding to i and $\mathbf{x}^T \mathbf{H}_{(\cdot,i)}$ is the observed number of symbols on that periodicity, thus if

$$|\mathbf{r}_i^T \mathbf{H}_{(\cdot,i)} - \mathbf{x}^T \mathbf{H}_{(\cdot,i)}| < \lambda \tag{44}$$

implying that, if the expectation for the model with $c_i = 0$ is closer than λ to the observed number in the data, then set $c_i^{(r+1)} = 0$. The resulting CCD algorithm is outlined in Algorithm 2.The computational cost of one iteration of the outer loop is $\mathcal{O}(m_{max}^2 N)$. Note that a significant performance increase is often possible in batch applications, where a recursive algorithm is needed, by the so called *active set strategy* [20]. The strategy simply involves not updating the parameters that are currently zero in every iteration, and perhaps only doing so once every tenth iteration or so.

Algorithm 2 The PECS$_L$ estimator

```
Initiate c = c_0
for r = 1, ... do
    for i = 1, ..., M do
        if maximum of (42) ≥ 0 then
            c_i^(r) = 0
        else
            Update c_i^(r) according to (37)
        end if
    end for
end for
```

V. Numerical Results

We proceed to examine the performance of the proposed likelihood-based estimators using simulated DNA sequences, binary sequences, and measured DNA data. For DNA sequences, only $B = 4$ different symbols are present, namely A, C, G, and T. Initially, we examine a simulated DNA sequence containing one deterministic periodicity. Figure 1 illustrates the rate of successfully determining this periodicity as a function of the length of the periodicity, comparing the proposed PECS$_G$ estimator with the MEM [10], PAM [7], QSPK [5], and SPE [27] estimators, as well as with a Fourier-based estimator detailed in [27]. As the simulated sequence is stationary, the window length used for the DFT-based methods were selected to be equal to the length of the sequence. Here, and in the following, the success rate has been determined using 250 Monte-Carlo simulations using $N = 1000$ equiprobable symbols, with the sought periodicity being inserted appropriately. As is clear from Figure 1, the proposed estimator succeeds in successfully determining all the considered periodicities, whereas all the other methods

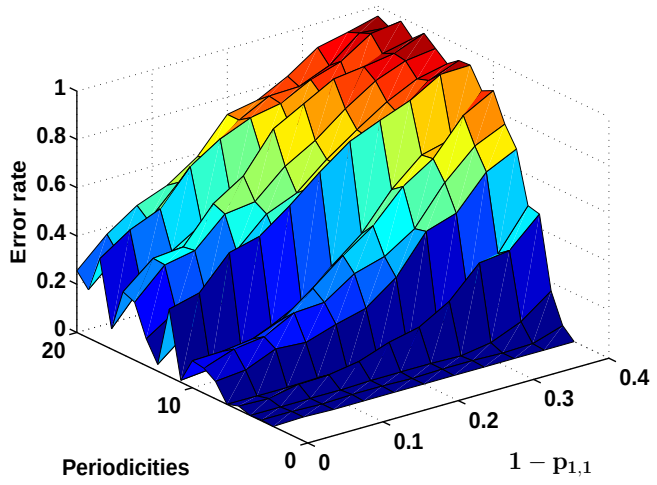


Fig. 2. The error rate of finding the periodicity as a function of the negative probability, $1 - p_{1,1}$, and the periodicity for the SPE algorithm.

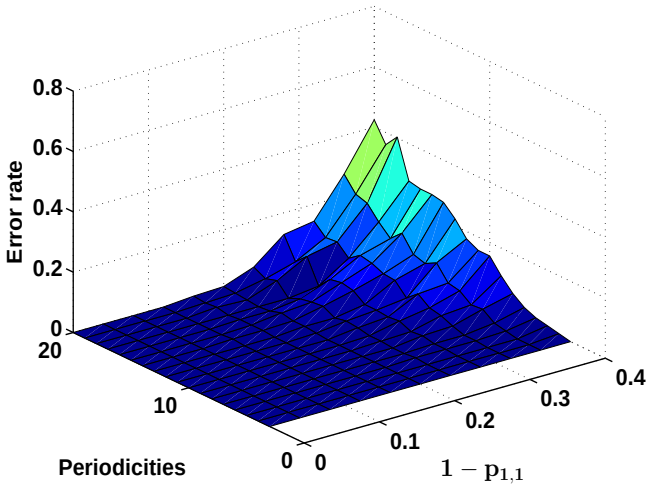


Fig. 3. The error rate of finding the periodicity as a function of $1 - p_{1,1}$, and the periodicity for the proposed PECS$_G$ method.

lose performance as the length of the periodicity grows. Of the other examined estimators, the SPE estimator seems to offer the second best performance, and we will for this reason only show the results for this estimator in the following comparisons, noting that all the other discussed estimators exhibits a notably worse performance than the SPE estimator in all the considered cases (see also [1]). Proceeding to examine also statistical periodicities, we vary $p_{1,1}$ for the index set corresponding to the generated periodicity, with $p_{0,1} = 1/4$ on the complement set. It may be noted that $p_{1,1} = 1$ corresponds to a perfect periodicity, whereas $p_{1,1} < 1$ corresponds to a statistical periodicity with a probability of each symbol being eroded, i.e., a non-perfect periodicity, being $1-p_{1,1}$. Similarly, $p_{0,1}$ is the corresponding probability for the complement set. Figures 3 and 2 show the resulting success rate for the SPE and PECS$_G$ estimators as a function of the periodicity and the probability $p_{1,1}$, again clearly illustrating how PECS$_G$ outperform SPE (and thus also all the other mentioned estimators) for all periodicities and $p_{1,1}$.

Next, we investigate how well PECS$_G$ and PECS$_L$ are able to resolve two periodicities in a binary sequence. In this case, some care needs to be taken when setting up the simulations, as when generating two periodicities, these may overlap or combine to create a new periodicity, e.g., if generating two periodicities of period six, these may be placed such that they instead form just a single periodicity with period three. Similarly, two periodicities with period four and twelve may cause the resulting sequence to have only a single periodicity of four. In order to avoid such ambiguities in the resulting performance measure, the test data has been generated such that it avoids this form of ambiguities. Figure 4 illustrates the success rate of determining both periodicities correctly, as a function of the length of the two periodicities, with $N = 500$ and again using $p_{1,1} = 3/4$ and $p_{0,1} = 1/4$. Each point on the x-axis should be interpreted as the average error for all combinations of periodicities within the brackets, i.e., for instance $(14, 14 - 17)$ denotes all combinations $(14, 14), (14, 15), (14, 16)$ and $(14, 17)$. As may be seen from the figure, even when the sequence contains two periodicities of lengths up to 12, when most of the other discussed estimators completely fail to find even a single perfect periodicity, both PECS algorithms have a very low proportion of errors. From the figure, one can also observe that, as expected, the PECS$_L$ outperforms the PECS$_G$ when there is more than one periodicity present in the sequence. For the last simulated data experiment, we recreate a simulation experiment similar to the one that was used in [8], where a deterministic periodicity of 11 and 31 are present simultaneously in a signal generated from a 4 element set being uniformly distributed on the other indices. As can be seen in Figure 5, the PECS$_G$ estimator achieves almost 100 % success rate even before the method presented in [8] can start to be used, since it requires a minimum of $11 \times 31 = 341$ data points. Finally, we examine the performance of the PECS$_G$ estimator on measured genomic data, in the form of the gene C. elegans F56F11.4 [28]. Since genomic data is generally not stationary, the estimate has been formed using a sliding window with length $N = 360$. The results obtained by PECS$_G$ are shown in Figure 6, where the periodicities with a likelihood ratio greater than the 95% quantile of the maximum of $M = 465$ χ^2 distributed random variables are shown for each symbol. In earlier work, such as [10] and [27], a period of three was found at around index 7000. This period was also found when using PECS$_G$, but when looking at the corresponding $\tilde{p}$, one may note that this periodicity is actually constituted by the lack of the symbol C, i.e., this period is detected since the symbols A, G, and T are alternating in a non-periodic fashion, and since C is always absent at these indices, this apparently causes the Fourier based methods to indicate a periodicity of three. If one is not interested in finding these sorts of periodicities, one may restrict $p_{1,1}$ to be in $[1/2, 1]$, in the same manner as mentioned above. This will ensure that PECS$_G$ only finds

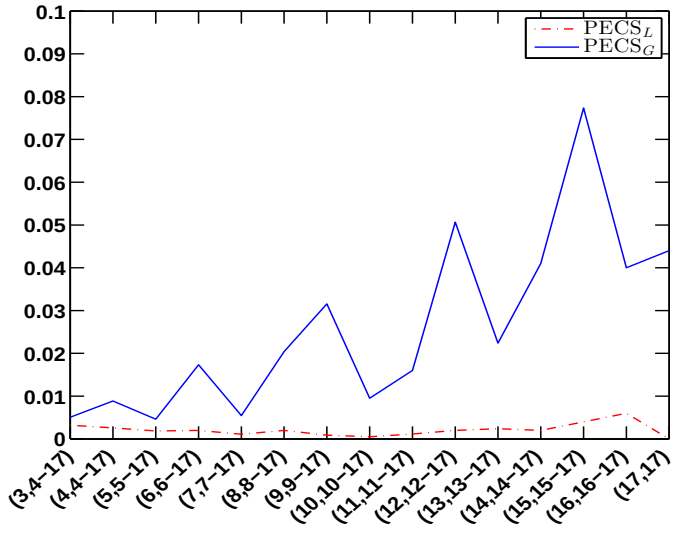


Fig. 4. The proportion of incorrect estimations of two periodicities for the PECS algorithms. Each point on the x-axis represent average error for all combination of that point and smaller (or equal) periodicities.

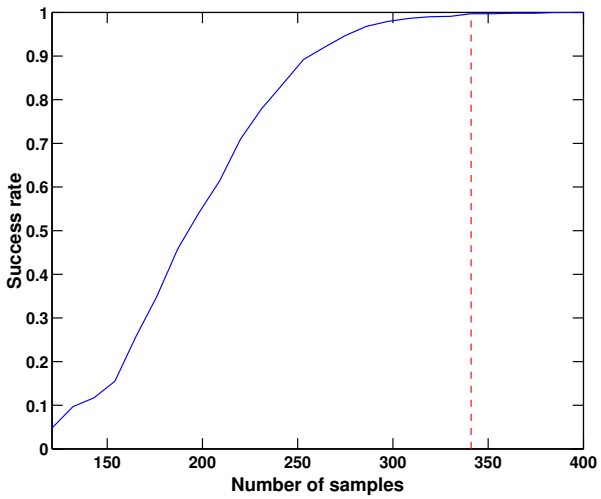


Fig. 5. Rate of success for PECS$_G$ in estimating the periodicities of a signal with periodicities at 11 and 31, as a function of signal length. The dashed line denotes the minimum data needed for using [8].

periodicities that are made up by an increased probability in the presence of a symbol.

VI. Conclusion

In this work, we have presented a likelihood-based approach for modeling periodicities in symbolic sequences. Modeling the observations using a categorical distribution with periodic indices, possibly having a different distribution, leads to a difficult combinatorial problem. Here, we have proposed two algorithms to relax the problem using sparse heuristics: namely, one fast greedy approach which builds up the solution set in an iterative fashion, and one based on convex relaxation ideas, which has the benefit of a more efficient usage of the data. Finally, we show the benefits of the proposed algorithms as compared to previously published methods using simulation experiments as well as with real DNA data examples.

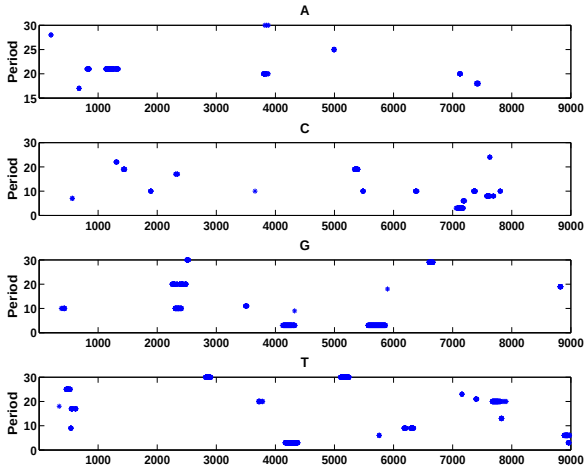


Fig. 6. The periodicities of each symbol in the gene C.elegans F56F11.4 computed using a sliding window.

VII. Acknowledgement

The authors would like to thank Prof. Lorenzo Galleani and Dr. Roberto Garello at Politecnico di Torino, Italy, for providing us with the their implementation of MEM-algorithm detailed in [10].

References

[1] S. I. Adalbjörnsson, J. Swärd, and A. Jakobsson, "'Likelihood-based Estimation of Periodicities in Symbolic Sequences," in *Proceedings of the 21th European Signal Processing Conference*, Marrakesh, 2013.

[2] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, second edition, 2007.

[3] M. B. Chaley, E. V. Korotkov, and K. G. Skryabin, "Method Revealing Latent Periodicity of the Nucleotide Sequences Modified for a Case of Small Samples," *DNA Res.*, vol. 6, no. 3, pp. 153–163, 1999.

[4] E. Korotkov and N. Kudryaschov, "Latent periodicity of many genes," *Genome Informatics*, vol. 12, pp. 437–439, 2001.

[5] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, July 2001.

[6] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 628–634, March 2002.

[7] G. L. Rosen, *Signal Processing for Biologically-Inspired Gradient Source Localization and DNA Sequence Analysis,*, Ph.D. thesis, Georgia Institute of Technology, 2006.

[8] R. Arora, W. A. Sethares, and J. A. Bucklew, "Latent Periodicities in Genome Sequences," *IEEE J. Sel. Topics in Signal Processing*, vol. 2, no. 3, pp. 332–342, June 2008.

[9] L. Wang and D. Schonfeld, "Mapping Equivalence for Symbolic Sequences: Theory and Applications," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4895–4905, Dec. 2009.

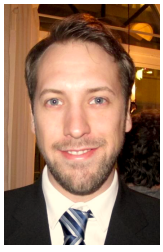
[10] L. Galleani and R. Garello, "The Minimum Entropy Mapping Spectrum of a DNA Sequence," *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 771–783, Feb. 2010.

[11] J. Epps, H. Ying, and G. Huttley, "Statistical methods for detecting periodic fragments in dna sequence data," *Biology Direct*, vol. 6, no. 21, pp. 1–16, 2011.

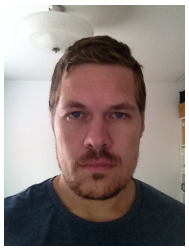
[12] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall,

Upper Saddle River, N.J., 2005.
[13] D. D. Muresan and T. W. Parks, "Orthogonal, exactly periodic subspace decomposition," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2270–2279, Sept. 2003.
[14] W. A. Sethares and T. W. Staley, "Periodicity transforms," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 2953–2964, Nov 1999.
[15] E. L. Lehmann and G. Casella, *Theory of Point Estimation (Springer Texts in Statistics)*, Springer, 2nd edition, 1998.
[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
[17] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.
[18] G. Casella and R. Berger, *Statistical Inference*, Duxbury, 2nd edition, 2002.
[19] P. Emberchts, C. Klüppelberg, and T. Mikosch, "Fluctuations of Maxima," in *Modelling Extremal Events*, vol. 33 of *Applications of Mathematics*, pp. 113–179. Springer Berlin Heidelberg, 1997.
[20] P. G. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*, Springer Series in Statistics. Springer, 2011.
[21] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[22] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.
[23] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, 2010.
[24] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[25] P. Tseng, "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.
[26] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.
[27] J. Swärd and A. Jakobsson, "Subspace-based estimation of symbolic periodicities," in *38th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 26-31 2013.
[28] National Center for Biotechnology Information, "Genome sequence of the nematode C. elegans: a platform for investigating biology," http://www.ncbi.nlm.nih.gov/nuccore/FO081497.1.

Stefan I. Adalbjörnsson (S'09) received his B.Sc. in Electrical and Computer Engineering from the University of Iceland in 2004, and his M.Sc. in Engineering Mathematics and Ph.D. in Mathematical Statistics from Lund University in 2009 and 2014, respectively. Currently, he is working as a post doctoral researcher in applied mathematics at the Centre for Mathematical Sciences at Lund University, working on an interdisciplinary project with the Lund University Humanities laboratory, as well as with an industry partner, Quanox, on recommendation systems and related large data problems. He has been a visiting researcher at the Spectral Analysis Laboratory in University of Florida, Gainesville. His research interest include big data analytics, recommendation systems, exploratory data analysis, and applications of sparse and convex modeling in statistical signal processing and spectral analysis.



Johan Swärd (S'12) received his M.Sc. from Lund University in Industrial Engineering and Management in 2012, Sweden, and is currently working towards a Ph.D. in Mathematical Statistics at Lund University. He has been a visiting researcher at the Department of Systems Innovations at Osaka University, Japan. His research interests include time-frequency analysis and applications of sparse and convex modeling in statistical signal processing and spectral analysis.



Jonas Wallin received his Ph.D. in Mathematical statistics at Lund University, 2014. He is currently a Postdoc at Chalmers, Gothenburg. Main research areas are Spatial statistics, Bayesian modelling and estimation theory, with application towards, mainly, precipitation, medical imaging and Flow cytometry classification.

Andreas Jakobsson (S'95-M'00-SM'06) received his M.Sc. from Lund Institute of Technology and his Ph.D. in Signal Processing from Uppsala University in 1993 and 2000, respectively. Since, he has held positions with Global IP Sound AB, the Swedish Royal Institute of Technology, King's College London, and Karlstad University, as well as held an Honorary Research Fellowship at Cardiff University. He has been a visiting researcher at King's College London, Brigham Young University, Stanford University, Katholieke Universiteit Leuven, and University of California, San Diego, as well as acted as an expert for the IAEA. He is currently Professor and Head of Mathematical Statistics at Lund University, Sweden. He has published his research findings in over 150 refereed journal and conference papers, and has filed five patents. He has also authored a book on time series analysis (Studentlitteratur, 2013), and co-authored (together with M. G. Christensen) a book on multi-pitch estimation (Morgan & Claypool, 2009). He is a member of The Royal Swedish Physiographic Society, a Senior Member of IEEE, and an Associate Editor for Elsevier Signal Processing. He has previously also been a member of the IEEE Sensor Array and Multichannel (SAM) Signal Processing Technical Committee (2008-2013), an Associate Editor for the IEEE Transactions on Signal Processing (2006-2010), the IEEE Signal Processing Letters (2007-2011), the Research Letters in Signal Processing (2007-2009), and the Journal of Electrical and Computer Engineering (2009-2014). His research interests include statistical and array signal processing, detection and estimation theory, and related application in remote sensing, telecommunication and biomedicine.