SUBSPACE-BASED ESTIMATION OF SYMBOLIC PERIODICITIES

Johan Swärd and Andreas Jakobsson

Dept. of Mathematical Statistics, Lund University, Sweden

ABSTRACT

In this work, we propose a novel subspace-based estimator of periodicities in symbolic sequences. The estimator exploits the harmonic structure naturally occurring in symbolic sequences and iteratively forms the estimate of the periodicities using a MUSIC-like formulation. The estimator allows for alphabets of different sizes, but is here illustrated using both simulated and real DNA measurements, showing a notable performance gain as compared to other common estimators.

Index Terms— Spectrum analysis, symbolic sequences, hidden periodicities, subspace techniques.

1. INTRODUCTION

Symbolic sequences appear in numerous applications wherein measurements are formed from a finite alphabet of unordered symbols, typically lacking any form of algebraic structure. Examples of such data include genomic and proteomic sequences, text indicators, and various forms of categorical time series [1]. One commonly occurring problem for such measurements lies in the forming of an estimate of periodicities in the data; for example, the latent periodicities in DNA sequences have been shown to be correlated with various forms of functional roles [2]. For purpose of illustration, we will, without loss of generality, herein examine periodicities of DNA sequences, for which the alphabet consists of the symbols A, C, G, and T, and which governs, together with the RNA, the making of protein in an organism. In order to form an estimate of the periodicities hidden in the sequence, one needs to perform some form of initial mapping of the symbols. This is most commonly done by mapping the four symbols into numerical values, then, using this mapping, form the Fourier spectra of the resulting sequence, see e.g. [3]. An example of such a mapping is to assign the symbols as $T = 0$, $C = 1$, $A = 2$, $G = 3$ [4], or to use the complex representations $A = 1+j, C = -1-j, G = -1+j$, $T = 1 - j$, as is done in [5] and [6], although both alternatives suffers from creating an undesired ordering between the symbols, such that the Euclidian distance between the symbols are not the same, thereby creating artifacts in the spectrum [7]. Other alternatives includes using minimum entropy mappings, mapping equivalences, transformations, or maximum likelihood formulations of the cyclostationary properties of the periodicities [2, 8–10]. These forms of estimators and mappings generally suffer from a relatively high computational complexity and/or difficulties to scale the formulation to an arbitrary alphabet. Furthermore, neither of the mentioned approaches considers the natural harmonic structure occurring in the data, such that a latent periodicity of P symbols will also create periodicities of $P/2$, $P/3$, etc. This harmonic structure is illustrated in Figure 1, showing the Fourier spectrum obtained using (1), given below, for a symbolic sequence containing a periodicity of 8. In this paper, we propose a harmonically related subspace-based estimation technique that not only allows for this harmonic structure, but also scales easily to various alphabet sizes, as well as for growing sizes of the symbolic sequence, in both cases, requiring only a linear increase of complexity. The estimator exploits a greedy iterative relaxation technique to extract harmonically related structures from the symbolic sequence, by iteratively estimating the most dominant periodicity in the sequence, and then removing the found periodicity for each symbol exhibiting it [11]. The dominant periodicity of the resulting sequence is then found, and so on, until the found periodicity is too weak to be deemed to be more than a random fluctuation.

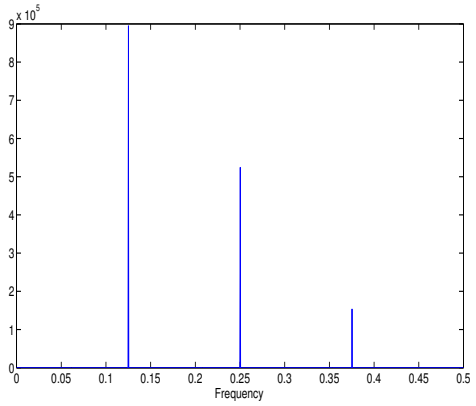

Fig. 1. The Fourier spectrum formed using (1) of a symbolic sequence with a periodicity of 8. The harmonic structure is clearly visible.

This work was supported in part by the Swedish Research Council and Carl Trygger's foundation.

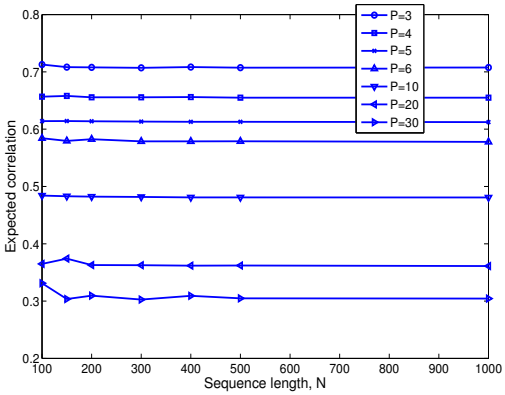

Fig. 2. The upper bound of the expected correlation for different values of P and N.

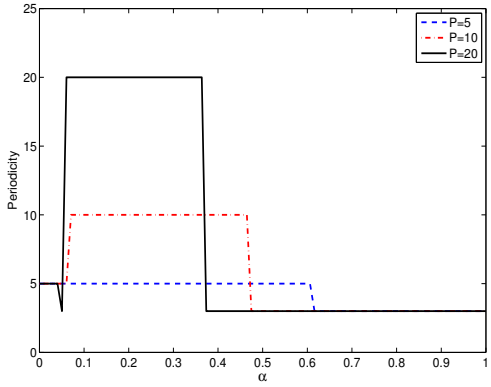

Fig. 3. Detection ability for three different periodicities, as a function of the threshold level α.

2. HARMONICALLY RELATED SUBSPACE-BASED PERIODICITY ESTIMATION

Consider an N-sample (reasonable stationary) symbolic sequence, $\mathbf{S}$, made up from a set of B symbols, $\{s_k\}_{k=1}^{B}$. Each such symbol, s_k, is here mapped onto the $B \times 1$ (column) indicator vector $\mathbf{e}_k$, having a one at position k and zeros elsewhere, forming the $B \times N$ matrix $\mathbf{Y}$. Each row of $\mathbf{Y}$ can thus be interpreted as the occurrences of the corresponding symbol in the sequence $\mathbf{S}$. Let $\mathbf{y}_k$ denote the k:th row of $\mathbf{Y}$. One option for forming an estimate of the symbolic (Fourier) spectrum would then be using the Discrete Fourier Transform (DFT), by transforming each row $\mathbf{y}_k$ separately, which then, when combined, yield the spectral estimate for the whole sequence, i.e.,

$$\mathbf{F}_{\mathbf{S}}(f) = \sum_{k=1}^{B} \left| \sum_{a=1}^{N} y_k(a) e^{-i2\pi f a} \right|^2 \tag{1}$$

A similar mapping was used in [9], wherein the symbolic spectrum was then formed using a weighted Fourier transform, where the weights were used for finding which symbols made up which periodicities. As is well known, in both cases, the resulting periodogram estimate will suffer from poor resolution as well as high variability, making it difficult to separate higher order periodicities reliably. Herein, we instead use the resulting rows to form an estimate of the corresponding covariance matrix estimate

$$\hat{\mathbf{R}}_k = \frac{1}{N-M+1} \sum_{t=1}^{N-M+1} \tilde{\mathbf{y}}_k(t)^T \tilde{\mathbf{y}}_k(t) \tag{2}$$

where $(\cdot)^T$ and M denote the transpose and the length of the considered subvectors, respectively, with

$$\tilde{\mathbf{y}}_k(t) = \left[\begin{array}{ccc} y_k(t) & \dots & y_k(t+M-1) \end{array} \right] \tag{3}$$

for $t = 1, \dots, N-M+1$. The choice of the vector length, M, should be made as a trade-off between variance and bias, with shorter vectors yielding a more reliable estimate of $\hat{\mathbf{R}}_k$, although the resulting lower spectral resolution will also make it harder to separate and detect closely spaced frequencies. Combining these covariance matrix estimates, the joint covariance matrix of the sequence may thus be estimated as

$$\hat{\mathbf{R}} = \sum_{k=1}^{B} \hat{\mathbf{R}}_k \tag{4}$$

It is worth noting that slightly better performance can be obtain by using each covariance matrix separately, although this gain comes at the cost of increasing complexity. For simplicity, we will use the summed covariance in the following. We now proceed by initially assuming the presence of just a single periodicity, $P_1 = 1/f_1$, which appears with L_1 harmonics. Let $\mathbf{G}$ denote the matrix spanning the estimated noise subspace, formed from the $(M \times M - L_1)$ eigenvectors corresponding to the $M - L_1$ least significant eigenvalues, i.e.,

$$\mathbf{G} = \left[\begin{array}{ccc} \mathbf{u}_{L_1+1} & \dots & \mathbf{u}_M \end{array} \right] \tag{5}$$

where $\mathbf{u}_k$ denotes the k:th eigenvector of $\hat{\mathbf{R}}$, and let $\mathbf{Z}_1$ denote the Vandermonde matrix

$$\mathbf{Z}_1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1^1 & z_1^2 & \cdots & z_1^{L_1} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{(M-1)} & z_1^{(M-1)2} & \cdots & z_1^{(M-1)L_1} \end{bmatrix} \tag{6}$$

where $z_1 = e^{2\pi i f_1}$. It is worth noting that the number of harmonics, L_1, will only depend on the periodicity and on the largest frequency in the frequency grid, f_{max}, i.e., $L_1 = \lfloor f_{max}/f_1 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operator and $f_{max} \leq 1/2$ to avoid aliasing. As is well-known, for the

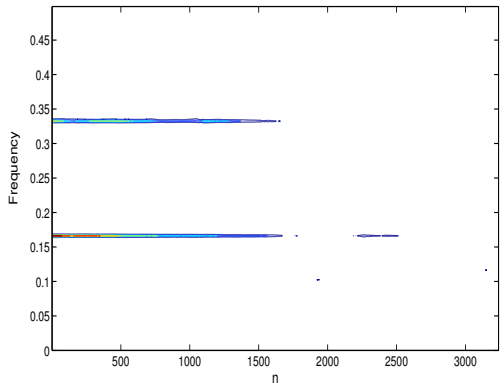

Fig. 4. The local MEM spectra for the repeated sequence CACCCG, with increasing levels of perturbation.

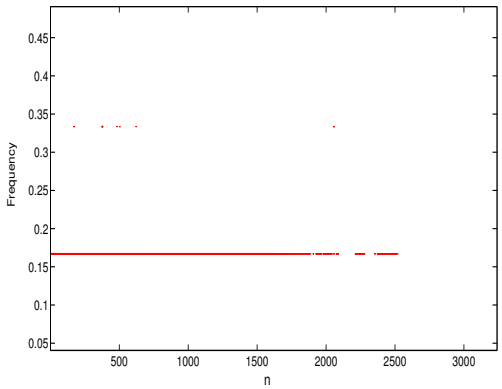

Fig. 5. The local SPE spectra for the repeated sequence CACCCG, with increasing levels of perturbation.

frequency f_1, the matrix $\mathbf{Z}_1$ will be orthogonal to the noise subspace, such that [12]

$$\mathbf{Z}_1^H \mathbf{G} = \mathbf{0} \tag{7}$$

where $(\cdot)^H$ denotes the conjugate transpose. Reminiscent to the harmonic MUSIC-like algorithm formulated for pitch estimation of speech and audio signals in [13], the presented formulation allows for the forming of an estimate of the hidden periodicity, using a simple 1-D search, as

$$\hat{f}_1 = \arg\min_{f_1} \frac{1}{MK} ||\mathbf{Z}_1^H \mathbf{G}||_F^2 \tag{8}$$

$$= \arg\min_{f_1} \frac{1}{MK} \sum_{l=1}^{L_1} ||\mathbf{z}^H(f_1)\mathbf{G}||_2^2 \tag{9}$$

where

$$K = \min\{L_1, M - L_1\} \tag{10}$$

and $L_1 = \lfloor f_{max}/f_1 \rfloor$. The denominator, MK, is a scaling factor introduced to ensure a proper normalization of the norm for the different model orders and frequencies. It should be stressed that the resulting frequency estimate assumes that only a single periodicity was present in $\mathbf{S}$. In order to avoid this limitation, we now proceed to extend the above formulation to a relaxation-based iterative formulation, wherein, after finding the most dominant periodicity, this is removed from the measurement to allow for the estimation of the second most dominant cycle, and so on. This is done by forming an indicator sequence $\mathbf{t}_1$ such that every $P_1 = 1/f_1$ element is set to one, whereas all remaining elements are set to zero. The correlation between each vector $\mathbf{y}_k$ and $\mathbf{t}_1$ yields which of the B symbols that contained the periodicity, and with which phase, thereby enabling its removal by forming the updated data vector $\mathbf{y}_k^{(2)} = \mathbf{y}_k^{(1)} - \mathbf{t}_1$, where $\mathbf{y}_k^{(1)}$ denotes the k:th symbol vector, wherein the periodicity was found, with the updated symbol vector $\mathbf{y}_k^{(2)}$ being formed with this periodicity removed. This removal is done for each symbol vector having a correlation with $\mathbf{t}_1$, which is above the threshold α, discussed further below. The resulting updated symbol vectors are then used to form an updated covariance matrix reminiscent to (4), followed by a new estimation using an estimate formed reminiscent to (9), and so on. This procedure will continue until no further periodicities are found, which occurs when the found periodicity in iteration k, P_k, has a correlation with the corresponding indicator vector, $\mathbf{t}_k$, that is weaker than α. The threshold α therefore both determines if a found periodicity is deemed to be significant, and as a stopping criteria when no further periodicities occurs in the sequence. Therefor, if set too high, the algorithm will miss longer periodicities, whereas, if set too low, the robustness of the algorithm will be effected. In order to select α, an upper bound is determined by computing the correlation resulting from a random vector, containing only a periodicity of P, and the indicator sequence $\mathbf{t}_k$. The resulting correlation, as illustrated in Figure 2, indicates an upper bound on how strong correlation that should be expected for a sequence with a periodicity of P, if found throughout the entire sequence. It is worth noting that this periodicity is made up by just one symbol, which produces a smaller correlation than a periodicity containing more than one symbol. Furthermore, depending how many symbols that can be allowed to be replaced by other symbols while still being deemed to indicate a relevant periodicity, the used threshold α should be selected as the corresponding fraction of the obtained bound. In our experience, a reasonable choice of α seems to be somewhere in between 60%-80% of the upper bound. For example, if a periodicity of 20 is expected in an sequence with length 1000, a reasonable α would be $\alpha = 0.25$. The effect that α has on the performance of the algorithm is depicted in Figure 3, where the detection ability for three different periodicities are shown as a function of α. We term the resulting estimator the Symbolic Periodicity Estimation (SPE) algorithm.

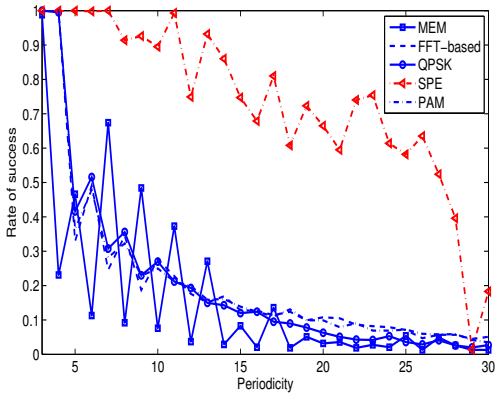

Fig. 6. Likelihood of correctly determining a hidden periodicity with period P.

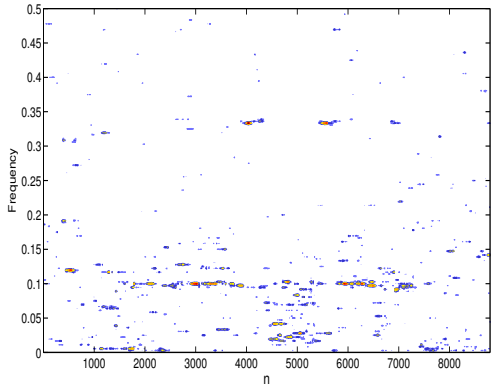

Fig. 7. The local MEM spectrogram for the gene C. elegans.

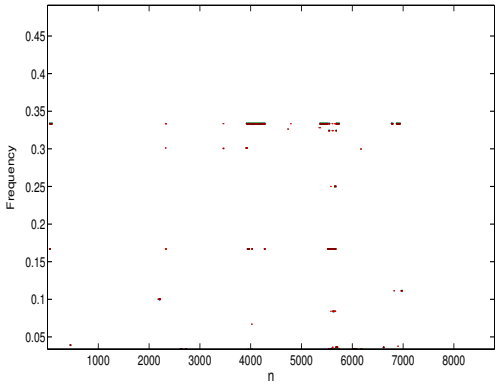

Fig. 8. The local SPE spectrogram for the gene C. elegans

3. EXPERIMENTAL RESULTS

We proceed to examine the performance of the proposed estimator in comparison with four other algorithms developed for genomic sequences, namely PAM [7], QSKP [5], MEM[1] [3], and the Fourier-based estimator given in (1). First, the algorithms are examined using a simulated random DNA sequence, with different periodicities. All the symbols have then been formed with uniform probability. After thus forming the random vector, a randomly selected symbol is inserted at every P:th index to creating a periodicity of P in the sequence. The performance of the discussed estimators is shown in Figure 6, illustrating the probability of correctly detecting the periodicity in a $N = 1000$ long sequence, using 1000 Monte-Carlo simulations. Here, we set the MEM user parameter $N_0 = 11$ (as in [3]), and for the SPE algorithm, use a subvector length of $M = 50$ and $\alpha = 0.2$ as to reflect the maximum expected periodicity. As can be seen from the figure, the reference methods all have similar success rate in finding the periodicity in the sequence, whereas the proposed algorithm is showing preferable performance, having a success rate of over 60% for periodicities below 20. Next, we examine the sensitivity of the estimator using the test introduced in [3]. In this test, a non-stationary sequence, $\mathbf{S}$, with length $N = 3600$, is made up by the sequence CACCCG, repeated over and over again. As in [3], the sequence is perturbed by a random replacement of the symbols, such that the probability of changing a symbol is n/N. The symbol is then replaced by one of the possible symbols, which is drawn uniformly (and can thus remain unchanged). Following [3], a sliding-window spectrogram is computed for both the MEM and SPE algorithms (to allow for non-stationary signals), using a window length of $N_w = 360$, with the latter using $\alpha = 0.3$. Figures 4 and 5 show the results, clearly illustrating that the MEM algorithm will capture both the periodicity and its harmonics, whereas SPE will only capture the periodicity as such. Furthermore, SPE can be seen to offer somewhat higher robustness to the perturbations as compared to MEM. Finally, we examine the performance of the proposed method for measured genomic data, examining the gene C. elegans F56F11.4 [14], which should occasionally contain a periodicity of 3, indicating the presence of exonic regions, i.e., regions in the DNA that contains protein codings, as well as regions containing a periodicity close to 10 [3]. Again, we use a window length of $N_w = 360$. Figures 7 and 8 show the resulting MEM and SPE spectrograms, with the former not showing the least dominant peaks in the estimate, and the latter using $\alpha = 0.15$ and $M = 150$. As could be expected, the MEM estimate is quite noisy, making it difficult to detect the exonic regions,whereas the SPE estimate is notably cleaner, making it easy to detect these regions. The 10-periodicity seems to be more pronounced in the MEM estimate than in the SPE estimate; we are unaware to which extent this reflect any real periodicity.

[1]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 [3].

4. REFERENCES

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

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

[3] L. Galleani and R. Garello, "The Minimum Entropy Mapping Spectrum of a DNA Sequence," *IEEE Transactions on Information Theory*, vol. 56, pp. 771–783, Feb. 2010.

[4] P. D. Cristea, "Genetic signal representation and analysis," *Proc. SPIE COnf., Int. Biomedical Optics Symp. (BIOS02)*, pp. 77–84, 2002.

[5] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 8–20, July 2001.

[6] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences," *IEEE Trans. Signal Process.*, vol. 51, pp. 2280–2287, Sep 2003.

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

[8] L. Wang and D. Schonfeld, "Mapping Equivalence for Symbolic Sequences: Theory and Applications," *IEEE Transactions on Signal Processing*, vol. 57, pp. 4895–4905, Dec. 2009.

[9] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Transactions on Signal Processing*, vol. 50, pp. 628–634, March 2002.

[10] R. Arora, W. A. Sethares, and J. A. Bucklew, "Latent Periodicities in Genome Sequences," *IEEE J. Sel. Topics in Signal Processing*, vol. 2, pp. 332–342, June 2008.

[11] J. Li and P. Stoica, "Efficient Mixed-Spectrum Estimation with Applications to Target Feature Extraction," *IEEE Transactions on Signal Processing*, vol. 44, pp. 281–295, February 1996.

[12] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, N.J.: Prentice Hall, 2005.

[13] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. Morgan & Claypool, 2009.

[14] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/nuccore/FO081497.1.