

# Symbolic Signal Processing

Don H. Johnson and Wei Wang

**Abstract**—Symbolic signals are, in discrete-time, sequences of quantities that do not assume numeric values. In the most general case, these quantities have no mathematical structure other than that they are members of some set, but they can have a sequential structure. We show that processing such signals does not entail mapping them directly to the integers, which would impose more structure—ordering and arithmetic—than present in the data. We describe how linear estimation and prediction can be performed on symbolic sequences. We show how spectrograms can be computed from neural population responses and from DNA sequences.

## I. INTRODUCTION

Information has two basic forms: numeric and symbolic. Much of signal processing is concerned with the analysis of numeric information, such as geophysical, physical, and engineering data. Classic digital signal processing methods apply to sequences whose values have a rich mathematical structure. For example, speech signals and images can be added and scaled, which means that mathematically an algebra can be defined for such signals. *Symbolic signals* have been more difficult to analyze because the alphabet from which they are drawn does not have a rich mathematical structure. Symbolic sequences, which we define mathematically to be a sequence of members of a set drawn according to a probabilistic rule, occur in many scientific and engineering applications: Text files, when viewed as a sequence of letters, descriptive data such as daily weather (sunny, cloudy, rainy, etc.), DNA sequences (formed from four bases represented by the letters  $\{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ ), and digital communication symbols (ones and zeros for high dimensional signal sets) are but a few examples. Our applications will be drawn from neuroscience, where we are concerned with analyzing the responses of collections of neurons, and from molecular biology, where we analyze patterns in DNA sequences.

## II. DEFINITIONS

Let  $\mathbf{x}_n$  be a symbolic-valued sequence, each sample taking on values in the countable set  $\mathcal{A} = \{a_1, \dots, a_L\}$ . In this work, we consider only stochastic symbolic signals, which are generated according to a probability law. In general, this law takes the form of the conditional probability  $\Pr[\mathbf{x}_n = a_l \mid \mathbf{x}_{n-1} = a_i, \mathbf{x}_{n-2} = a_j, \dots]$ . Markov models are most prevalently used to describe symbolic sequences, with dependencies higher than first order often required. In our applications, the dependence structure of the underlying probability law is time-varying and unknown. It is this dependence structure that we want to elucidate with symbolic signal processing techniques. For DNA sequences, the symbolic signal describes which base occurred along the length of the molecule. The sequence's values are the four DNA bases, which are represented by letters of the alphabet:  $\mathcal{A} = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ .

## III. ALGORITHMS

In statistical signal processing, techniques can be broadly categorized as being detection or estimation algorithms. Interestingly, detection algorithms do *not* require the processed sequences to be numeric and deal with symbolic data on an equal footing. The reason for this fit is the fact that the optimal detection algorithm arises from the likelihood ratio, which is a ratio of probability distributions. Because probability distributions are easily defined for symbolic data, detection algorithms quickly emerge. Even in problems wherein the *a priori* probability distributions are not known (so-called empirical classification problems [1]), symbolic sequences can be handled as well. Distribution-free detectors have been developed for communications problems that have in their fundamental structure symbolic signal processing algorithms [2].

The key notion in enabling standard signal processing estimation algorithms (ones developed for numeric-valued signals) is to convert from symbols to real values in such a way that ordering and arithmetic cannot be meaningfully defined [3]. We begin by representing each symbol by an *indicator vector*  $\mathbf{e}_l$ , wherein only the  $l^{\text{th}}$  element is numerically 1 and the remainder are zero. For the neural population example,  $\mathbf{x}_n = 3 \Leftrightarrow \mathbf{e}_3 = \text{col}[0, 0, 0, 1, 0, 0, 0, 0]$ . Despite the numeric appearance of this representation, no ordering for the values can be defined, and neither scaling nor addition makes sense (indicator vectors can only have one nonzero entry and that must be unity). The symbolic sequence is equivalent to the sequence of corresponding indicator vectors, which we call  $\mathbf{y}_n$ . We then create a numeric-valued time series  $z_n$  corresponding to  $\mathbf{y}_n$  by forming the inner product between it and a fixed, but arbitrary, weight vector  $\mathbf{w}$ :  $z_n = \mathbf{w}'\mathbf{y}_n$ . Here,  $'$  denotes conjugate transpose. The weight vector is determined from signal processing considerations, which we detail subsequently. This translation from symbolic to numeric form retains *all* the information in the original symbolic sequence as well as its lack of mathematical structure (no ordering exists and no algebra can be readily defined).

In searching for direct estimation techniques, statisticians have for years dealt with symbolic sequences under the label of categorical time series or longitudinal categorical data [4]. We focus here on spectral estimation techniques. The discrete Fourier transform (DFT), among many other transforms, is a linear transformation applied to  $z_n$ . Letting  $\mathbf{f}_k$  denote the vector of DFT transform coefficients corresponding to the discrete-time frequency index  $k$ , the length- $N$  discrete Fourier transform

is  $\mathbf{f}'_k \text{col}[z_0, z_1, \dots, z_{N-1}]$ . Note that

$$\begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{N-1} \end{bmatrix} = \begin{bmatrix} \mathbf{y}'_0 \\ \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_{N-1} \end{bmatrix} \cdot \mathbf{w} . \quad (1)$$

In equation (1), the matrix formed from the  $\mathbf{y}$ s has dimension  $N \times L$ . Consequently, the Fourier transform calculation equals  $\mathbf{f}'_k(\text{col}[\mathbf{y}'_0, \mathbf{y}'_1, \dots, \mathbf{y}'_{N-1}]\mathbf{w})$ . Regrouping, the product of the first two terms corresponds to evaluating separately the DFT of each component of the sequence  $\mathbf{y}_n$ ,  $n = 0, \dots, N - 1$ , with the result forming a row vector that we represent as  $\mathbf{Y}'_k$ . Thus, the magnitude-squared of the DFT at frequency index  $k$  equals  $\mathbf{w}'\mathbf{Y}_k\mathbf{Y}'_k\mathbf{w}$ . The term  $\mathbf{Y}_k\mathbf{Y}'_k$  amounts to the cross-spectral matrix of each possible symbol's occurrence in the sequence. We can determine the weight vector according to the criterion that it maximizes the power in the Fourier transform (subject to the usual norm constraint on the weight vector). In this way, we determine the weight vector to maximize signal processing efficacy, passing from a symbolic to a numeric sequence *after* making the signal processing calculation. Formally, we want to maximize  $\mathbf{w}'\mathbf{Y}_k\mathbf{Y}'_k\mathbf{w}$  subject to the norm constraint  $\|\mathbf{w}\| = 1$ . This problem's solution corresponds to the eigenvector of the cross spectral matrix  $\mathbf{Y}_k\mathbf{Y}'_k$  having the largest eigenvalue, and the corresponding power equals this eigenvalue. This procedure is repeated for each frequency index  $k$ , meaning that we must compute the largest eigenvalues of cross-spectral matrices for each frequency of concern. Since the weight vector depends on frequency, no one weight vector relates the symbolic sequence to a numeric one, thus fulfilling the goal of not imposing mathematical structure on the data.

The cross-spectral matrix must be estimated from the data. We have suggested that it is an outer product, consisting of the data's Fourier transform computed over its entire length. The cross-spectral matrix formed this way has only one non-zero eigenvalue, and it *always* equals one. Consequently, the matrix is ill-suited for spectral estimation. Borrowing results from time-frequency analysis, a better approach is to compute shorter transforms (applying a window and overlapping the sections), and average these transforms across the data [3]. With this approach, we produce a single Fourier transform for the entire symbolic sequence; the sequence must be stationary for this procedure to be valid. Because we are concerned with non-stationary signals, this approach does not suffice. We have two alternatives. As in short-time Fourier analysis, we average cross-spectral matrices over short time segments, and produce spectrograms in the usual way. In neuroscience, we have the luxury of repeating the experiment, providing an ensemble of datasets having a common time origin with respect to the stimulus. We can therefore average across the ensemble to produce cross-spectral matrices without assuming stationarity. Either way, we can compute what amounts to a spectrogram for the symbolic sequence.

#### IV. APPLICATIONS

Spectrograms have long been used to analyze speech signals. These could be stored in Matlab format. Figure 1 shows spectrograms computed for portions of the DNA sequence of the bacterium *E. coli*. Spectral analysis reveals a weak, but persistently present spectral line having a frequency of 1/3 (upper left panel of figure 1). Such spectral lines indicate coding segments. The second spectrogram also contains a weak spectral line at the same frequency, but also a "spectral burst" near base 35,000. Examination of the sequence (bottom right) reveals a complex periodic structure that may indicate a specific coding region.

#### V. CONCLUSIONS

Symbolic signals constitute an interesting and new application area for developing signal processing algorithms. The procedure described here enable the application of more traditional, numerically oriented, signal processing methods to the analysis of symbolic information. With this approach, the spectral techniques described here and others can reveal how symbolic data represent information. The linearity of the mapping between symbolic and numeric signals allows use of other linear transforms, wavelet expansions, for example. In this linear approach, a key question is how to determine the weight vector that best reveals patterns. This choice is governed by the type of signal processing employed and by the optimization criterion.

#### REFERENCES

- [1] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Info. Th.*, vol. 35, pp. 401–408, 1989.
- [2] D.H. Johnson et al., "Type-based detection for unknown channels," in *Proc. ICASSP*, Atlanta, GA, 1996.
- [3] D.S. Stoffer, D.E. Tyler, and A.J. McDougall, "Spectral analysis for categorical time series: Scaling and the spectral envelope," *Biometrika*, vol. 80, pp. 611–622, 1993.
- [4] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York, 1990.

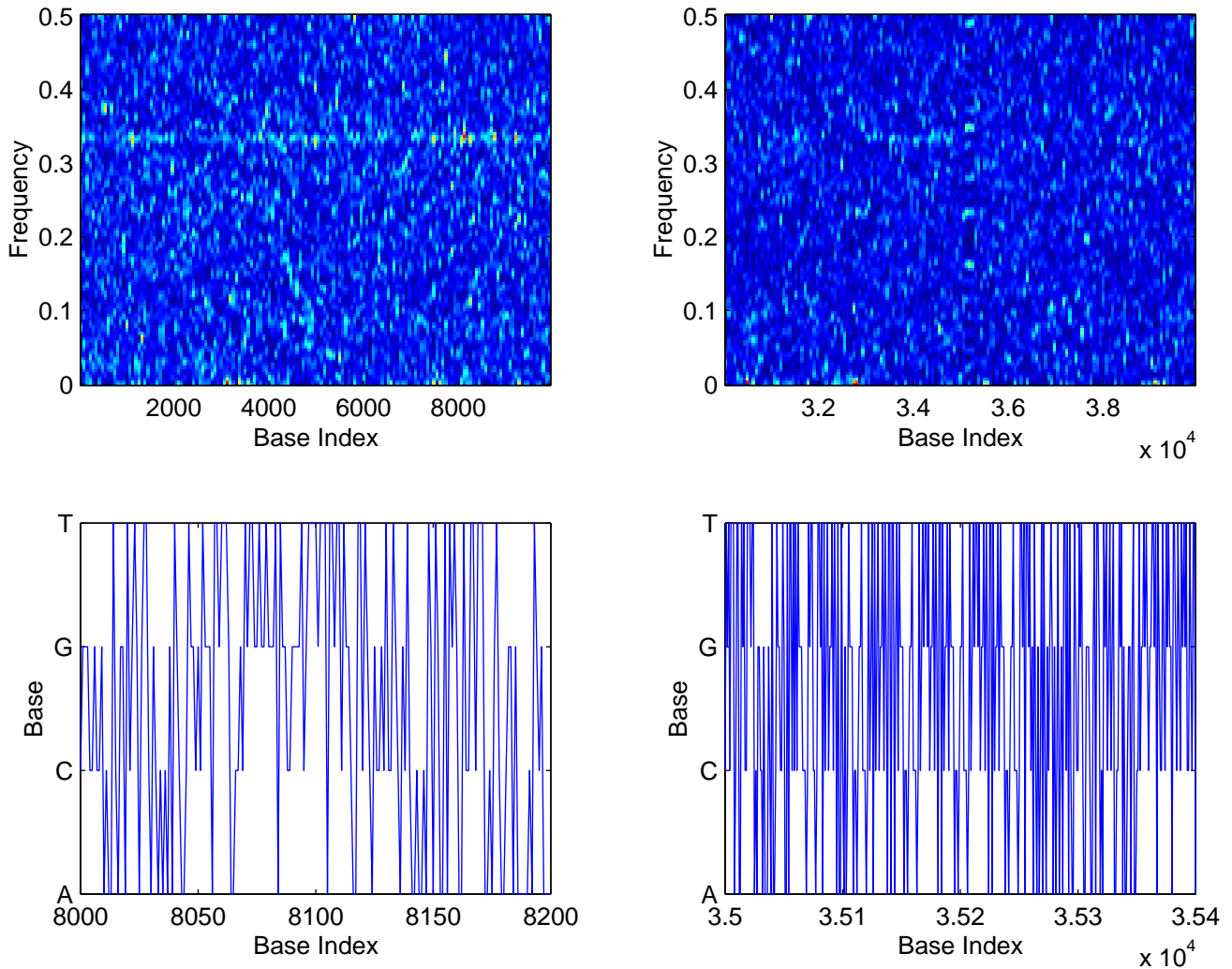


Fig. 1

The upper panels show spectrograms for two length-10000 segments of the *E. coli* DNA sequence obtained from the National Library of Medicine. Section length was 128 bases and we employed a Hanning window in calculating the spectra (256-point DFT). Sections overlapped by 64 bases (half-section). The bottom plots show “interesting portions” of the original data as expressed by the spectrograms.