INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0909

CLASSIFICATION OF MULTIVARIATE NON-STATIONARY SIGNALS: THE SLEX-SHRINKAGE APPROACH

BOEHM, H., OMBAO, H., von SACHS, R. and J. SANES

Classification of Multivariate Non-Stationary Signals: The SLEX-Shrinkage Approach

Hilmar Böhm[1], Hernando Ombao[2], Rainer von Sachs[1] and Jerome Sanes[3]

April 14, 2009

Abstract

We develop a statistical method for discriminating and classifying multivariate non-stationary signals. It is assumed that the processes that generate the signals are characterized by their time-evolving spectral matrix - a description of the dynamic connectivity between the time series components. Here, we address two major challenges: first, data massiveness and second, the poor conditioning that leads to numerically unstable estimates of the spectral matrix. We use the SLEX library (collection bases functions consisting of localized Fourier waveforms) to extract the best set of time-frequency features that best separate classes of time series. The SLEX approach yield readily interpretable results since it is a time-dependent analogue of Fourier approach to stationary time series. Moreover, it uses computationally efficient algorithms to enable handling of large data sets. We estimate the SLEX spectral matrix by shrinking the initial SLEX periodogram matrix estimator towards the identity matrix. The resulting shrinkage estimator has lower mean-squared error than the classical smoothed periodogram matrix. A leave-one out analysis for predicting motor intent (left vs. right movement) using electroencephalograms indicates that the proposed SLEX-Shrinkage method gives robust estimates of the evolutionary spectral matrix and good classification results.

Keywords: Classification, Discrimination, Multivariate Time Series, Shrinkage, SLEX library, SLEX spectrum.

[1]Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium. We gratefully acknowledge support from the "Projet d'Actions de Recherche Concertées" no 07/12-002 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain" as well as from the IAP research network grant P 5/24 of the Belgian government (Belgian Science Policy).

[2]Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA.

[3]Department of Neuroscience, Brown University, Providence, Rhode Island, USA.

1 Introduction

Our goal is to develop a statistical method for discriminating and classifying multivariate non-stationary signals. This project is motivated by a neuroscience experiment (conducted at the laboratory of Jerome Sanes, Brown University) to study brain network that mediate voluntary movement. In this experiment, participants performed a simple voluntary movement that required quick displacements of a hand-held joystick from a central position either to the right or to the left. Visual cues appearing on a computer monitor provided timing instructions to the participants Here, we develop a method to *discriminate* between presumed brain connectivity occurring during leftward and rightward movements, aiming to predict intentions to move by assessing the information evident in an electroencephalogram (EEG) time-series recorded contemporary with the voluntary movements. From a montage of 64 scalp electrodes, we identified a set of 11 EEG sensors that appeared to have the most relevance to the voluntary movements (Figure 1). Figure 3 illustrates time-amplitude plots of the EEG obtained from a representative participant during leftward (Figure 3, left) and rightward (Figure 3, right) joystick movements.

Discrimination and classification of time series has a long history. Shumway and Unger (1974) and Shumway (1982) developed the framework for discrimination in time series that has been adopted in most subsequent work. Shumway and colleagues applied their work to discriminate between different seismic activities (e.g., earthquake vs explosion). Kakizawa et al. (1998) concatenated the P-arrival and S-phases of the seismic signals into a bivariate time series and developed classification and discrimination methods for stationary multivariate time series.

For non-stationary time series, Shumway (2003) developed an information-theoretic classification method that treats the time series as realizations of the Dahlhaus model of locally stationary processes (see Dahlhaus 1997; 2001). Sakiyama and Taniguchi (2004) showed consistency of the classification procedure using the Kullback-Leibler criterion. Fryzlewicz

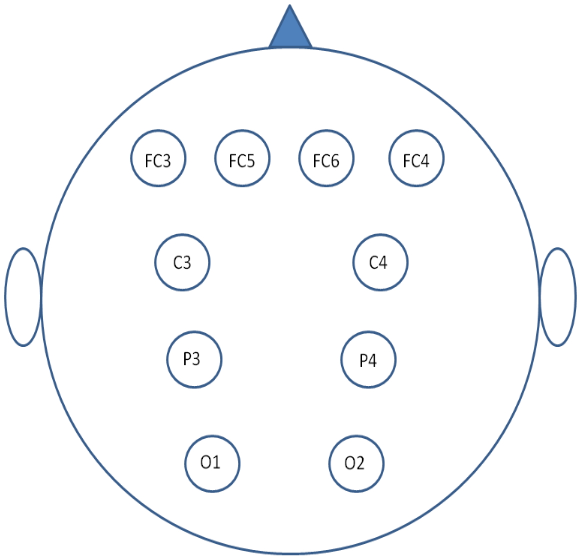

Figure 1: EEG scalp topography. The 11 channels are as follows: left frontal central ($FC3, FC5$), right frontal central ($FC4, FC6$), left central ($C3$), right central ($C4$), left parietal ($P3$), right parietal ($P4$), left occipital ($O1$), right occipital ($O2$) and central occipital (Oz, not in picture, between $O1$ and $O2$).

and Ombao (2009) developed a consistent classification method using stochastic wavelet representations. Saito (1994) developed another approach that selects a basis (from a collection of bases in a library) that gives maximal separation between classes of time series. There are a number of localized libraries that could be used for discriminating non-stationary time series. For example, one can use the localized trigonometric library, wavelet packets or the SLEX library (smoothed localized complex exponentials) which was developed in Ombao, Raz, von Sachs and Malow (2001) for analyzing non-stationary time series. Recently, Huang, Ombao and Stoffer (2004), inspired by the ideas in Saito (1994) and Shumway (1982), developed a procedure using the SLEX library to select the best time-frequency spectral features for discriminating between classes of univariate non-stationary time series.

Here, we address two major challenges with classifying and predicting multivariate non-stationary time series (such as EEGs). First, we note that most EEG datasets are massive and require computationally efficient transforms that can capture localized features of the data. Second, estimates of the multivariate spectra can be poorly conditioned (i.e., the ratio of the maximum to the minimum eigenvalue can be extremely large) which could be due to strong cross-correlation between channels. Consequently, inverting the spectral matrix estimates may give imprecise results thereby adversely impacting predictive ability especially when using information-based classification criteria such as the Chernoff criterion in Equation (2.4).

A standard approach to handling highly multi-collinear data entails reducing data dimensionality via, for example, principal components analysis (PCA). Though appealing and popular in many applications, PCA may not be ideal in discrimination and classification applications since the eigenvalue-eigenvector decomposition of the spectral matrix is invariant to (spatial) permutations of the time series. Consider, for example, a simple situation where a pair of channels R1 and R2 (located on the right of the scalp topography) have a cross-dependence structure during the right-movement condition that is identical to that between a pair L1 and L2 on the left during the left movement condition. PCA cannot distinguish the location of the sources thereby rendering it ineffective to discriminating between the functional connectivity occurring during the leftward and rightward-movements conditions.

Another approach to regularize the estimators is to smooth the periodogram matrices across frequency using a larger bandwidth. The smoothing approach is discussed in Parzen (1961) for univariate time series and Brillinger (1981) for multivariate time series. However, this approach does not guarantee that the resulting spectral matrix estimates will have good condition numbers. Moreover, the spectral estimates will have poor frequency resolution that can dull the predictive ability especially when the differences between conditions are present in very narrow frequency bands.

Here, we develop a novel classification and discrimination method for multivariate time series using the SLEX library to extract the localized cross-dependence structure (brain connectivity) and the shrinkage method to estimate the spectral density matrix. The spectral shrinkage estimator is a linear combination of a mildly-smoothed periodogram matrix and the identity matrix. The spectral shrinkage procedure is developed in Böhm (2008) and Böhm and von Sachs (2009) for the stationary case. In this paper, we extend this procedure to the non-stationary setting. The shrunken spectral estimator retains excellent frequency resolution, has good condition numbers and is shown to be superior to the standard periodogram smoother in terms of the squared-error risk. Finally, as demonstrated in the simulation studies in this paper, the shrinkage approach gives excellent classification rates.

The specific features of our approach are as follows. First, we use the SLEX (**S**mooth **L**ocalized Complex **EX**ponentials) library as a tool for extracting the time localized features of the non-stationary signals. Second, we estimate the time-varying spectrum via the the shrinkage procedure (i.e., the slightly smoothed periodogram matrix is shrunk towards the identity matrix). In this paper, we employ the Chernoff criterion (see Equation 2.4) that measures the divergence between the observed time series and the classes via the spectral density matrix. Furthermore, this criterion requires the computation of the inverse and the determinant of the spectral matrices. Naturally, poorly-conditioned estimates result in unreliable Chernoff divergence values and, as demonstrated in this paper, can lead to unacceptably high misclassification rates.

2 The SLEX-Shrinkage Method

We consider the following set up. Training data for each of conditions 1 and 2 consists of N number of P-channel time series each of length T. The number of trials N need not be identical for the two conditions but we make them to be so only for ease in presenting ideas. These time series from the two conditions are denoted, respectively, by

- $\mathbf{X}_n(t) = [X_{n1}(t), \ldots, X_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$;
- $\mathbf{Y}_n(t) = [Y_{n1}(t), \ldots, Y_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$.

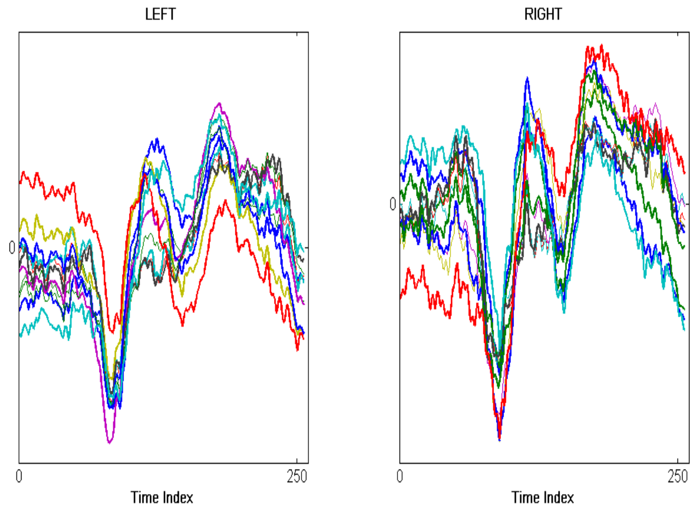

Figure 2: Left: averaged time series (across $N = 100$ trials) for the *left* condition at each of the 11 channel locations. Right: averaged time series (across $N = 100$ trials) for the *right* condition

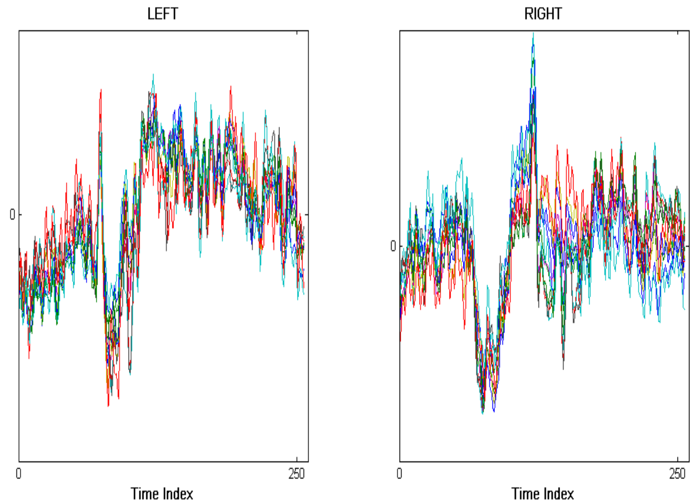

Figure 3: Left: representative 11-channel EEG recorded from one trial for the *left* condition. Right: representative 11-channel EEG recorded from one trial for the *right* condition.

Suppose that the data generated under these two conditions are modeled as zero mean multivariate non-stationary processes which are characterized by their spectral matrix denoted, respectively, as $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$. The first goal is to identify the time-frequency features (auto-spectra, cross-spectra, coherence) that can best separate the two conditions. We shall use the SLEX library to extract the localized cross-dependence features and then find the set of time blocks and frequencies that give the largest separation between $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$. The second goal is to use these features to classify a future signal whose group membership (or condition under which the signal was generated) is not known. We first provide a short discussion on the two elements of our proposed method, namely the SLEX library and the spectral shrinkage estimation method.

2.1 Brief Overview of the SLEX library

2.1.1 The SLEX waveforms.

The Fourier waveforms $\{\exp(i2\pi\omega u), \omega \in (-1/2, 1/2)\}$ may not adequately represent processes whose spectral properties evolve with time. A popular remedy uses the windowed Fourier function (Daubechies, 1992) which is of the form $\phi_F(u) = \Psi(u)\exp(i2\pi\omega u)$ where $\Psi(.)$ is a taper with compact support. Due to the Balian-Low theorem, these windowed Fourier functions are localized in time but they cannot be simultaneously smooth and orthogonal (Wickerhauser, 1994). To overcome the Balian-Low obstruction, a projection operator (rather than a taper) is applied to the Fourier waveforms resulting in localized orthonormal Fourier waveforms that we label as SLEX. The localized generalization of the Fourier functions is the SLEX functions. A SLEX waveform on time block b, denoted $\phi_{b,\omega}(u)$, has the form

$$\phi_{b,\omega}(u) \quad = \quad \Psi_+(b)\exp(i2\pi\omega u) + \Psi_-(b)\exp(-i2\pi\omega u) \tag{2.1}$$

where $\omega \in (-1/2, 1/2]$ and $u \in [-\eta, 1+\eta]$ where $0 < \eta < 0.5$. The windows Ψ_+ and Ψ_- are constructed by using rising cut-off functions. These windows come in pairs, i.e., once Ψ_+ is specified, Ψ_{-} is determined. A plot of the SLEX waveform is given in Figure 4. The SLEX waveforms (within each basis) are orthogonal. This property is important for analyzing massive time series because it is mathematically elegant, aids in the theoretical development of a model, preserves the energy of the time series and allows the use of computationally efficient tools like the best basis algorithm (BBA) of Coifman and Wickerhauser (1992). Thus, it facilitates the analysis of high dimensional massive data sets.

2.1.2 The SLEX library of bases.

The SLEX library is a collection of bases; each basis consists of the SLEX waveforms which are localized, thus they are able to capture the local spectral features of the time series. The SLEX library allows a flexible and rich representation the observed time series. To illustrate these ideas, we construct a SLEX library in Figure 4 with level $J = 2$. There are 7 dyadic blocks in this library. These are: $S(0,0)$ which covers the entire time series; $S(1,0)$ and $S(1,1)$ which are the two half blocks and $S(2,b), b = 0,1,2,3$ which are the four quarter blocks. Note that in general, for each resolution level $j = 0,1,\ldots,J$, there 2^j blocks each having length $T/2^j$. We will adopt the notation $S(j,b)$ to denote the block b on level j where $b = 0,1,\ldots,2^j - 1$. There are five possible basis from this particular SLEX library and one of them is composed of blocks $S(1,0), S(2,2), S(2,3)$ which correspond to the shaded blocks in Figure 4. We note that each basis is allowed to have multi-resolution scales, i.e., a basis can have time blocks with different lengths. This is ideal for processes whose regimes of stationarity have lengths that also vary with time. In choosing the finest time scale (or deepest level) of the transform J, the statistician will need some advice from the scientific expert. For example, neurologists can give some guidance regarding an appropriate time resolution of EEGs. In general, the blocks should be small enough so that we can be confident that the time series is stationary in these blocks. At the same time, the blocks should not be smaller than what is necessary in order to control the variance of the spectral estimator. Note that the SLEX allows for more general segmentation of the time series. For

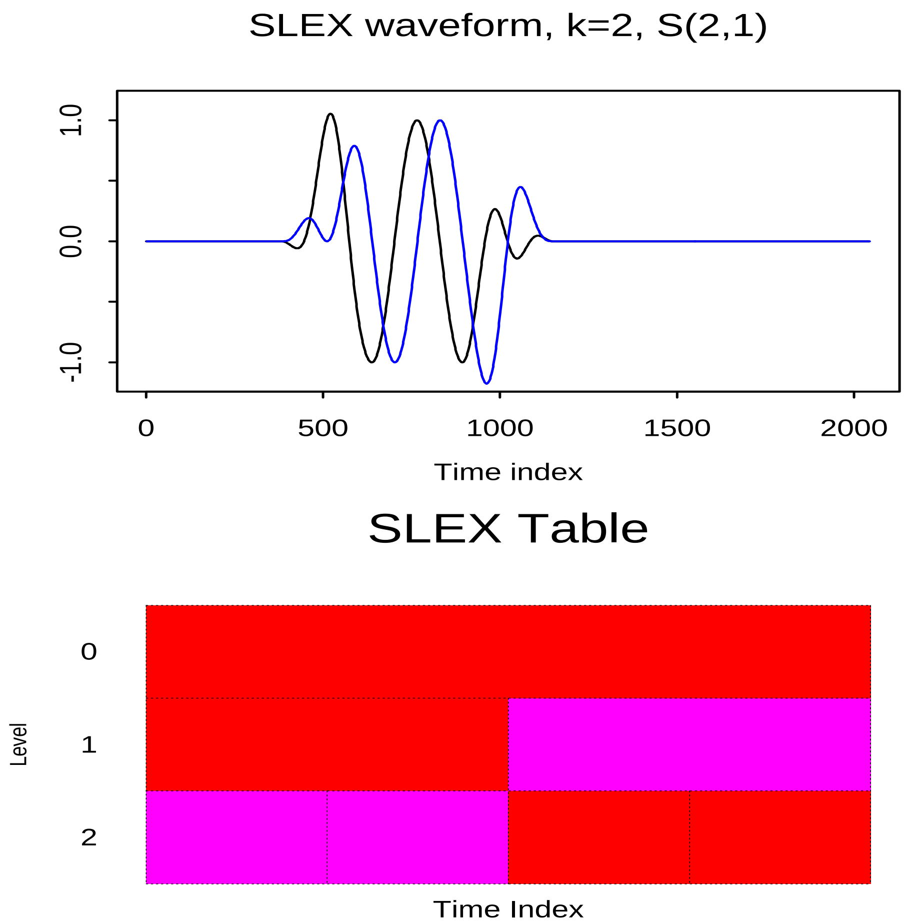

Figure 4: Top: A SLEX waveform localized at time block $[0.25, 0.50]$ and oscillating at frequency of two cycles over the time block. Bottom: A SLEX library with level $J = 2$. The shaded blocks represent one basis from the SLEX library.

simplicity and computational efficiency, we adopt the dyadic segmentation.

2.1.3 Computing the SLEX transform.

The SLEX transform is a collection of coefficients corresponding to the SLEX waveforms in the SLEX library. We will demonstrate that the SLEX coefficients can be computed using the fast Fourier transform (FFT). Let $X_\ell(t)$ be one component of a P-variate time series $\mathbf{X}(t)$ of length T. The SLEX coefficients (corresponding to $X_\ell(t)$) on block $S(j,b)$ are defined as:

$$\begin{aligned} d^{\ell}_{j,b}(\omega_k) &= (M_j)^{-1/2} \sum_t X_\ell(t) \overline{\phi_{j,b,\omega_k,t}} \\ &= (M_j)^{-1/2} \sum_t \Psi_+(\frac{t-\alpha_0}{|S|}) X_\ell(t) \; \exp[-i2\pi\omega_k(t-\alpha_0)] + \\ & \quad (M_j)^{-1/2} \sum_t \Psi_-(\frac{t-\alpha_0}{|S|}) X_\ell(t) \; \exp[i2\pi\omega_k(t-\alpha_0)] \end{aligned}$$

where $M_j = |S(j,b)| = T/2^j$ and $\phi_{j,b,\omega_k,t} = \phi_{S(j,b),\omega_k,t}$ is the SLEX basis vector on block $S(j,b)$ oscillating at frequency $\omega_k = k/M_j$ where $k = -M_j/2+1,\ldots,M_j/2$. We also note the "edge" blocks in each level j, namely $S(j,0)$ and $S(j,2^j-1)$, are padded with zeros when we compute the SLEX transform. Finally, by using the FFT, the number of operations needed to compute the SLEX transform has order of magnitude $O[T(\log_2 T)^2]$.

2.1.4 Some Applications.

The SLEX library has been used in a number of methods for non-stationary time series including evolutionary spectral and coherence estimation (Ombao, Raz, von Sachs and Malow (2001)); stochastic representation of non-stationary time series (Ombao, Raz, von Sachs and Guo (2002)); classification and discrimination (Huang, Ombao and Stoffer, 2004) and time-dependent principal components analysis (Ombao, von Sachs and Guo, 2005). In this paper, we will utilize SLEX, in conjunction with the shrinkage method, to classify multi-channel non-stationary time series.

2.2 Overview of the Shrinkage Procedure for Spectral Estimation

The possibility of the data being highly multi-collinear makes it necessary to regularize the estimate of the SLEX spectrum in order to permit matrix inversion without running into numerical instability. One possible solution could be to reduce the overall dimensionality of the data using PCA, as proposed in Ombao, von Sachs and Guo (2005). However, there are a number of disadvantages in using PCA. First, the choice of the number of dimensions to take into account is an arbitrary one. Second, with a non-stationary signal, a sensible choice of dimension may even vary over different time points or, worse, between different instances of the training data set. Third, as already discussed in the introduction, PCA abstracts from the spatial structure of the data, which is not desirable here because the cross dimensional stochastic dependence may be critical for discrimination.

In this paper, we follow the shrinkage procedure for spectral estimation. The philosophy behind shrinkage is to regularize a possibly badly conditioned, matrix-valued estimator by constructing a new estimator which is a linear combination of the estimator and a well conditioned matrix. In this setting, where we make no model assumptions on the data, the identity matrix is used as the latter and referred to as the *shrinkage target*.

2.2.1 Shrinkage for Stationary Time Series

We summarize the basic ideas on the shrinkage procedure for spectral estimation for stationary time series. For deeper insights and technical details, the readers are referred to Böhm (2008) and Böhm and von Sachs (2009). Let $\mathbf{X}(t) = [X_1(t), \ldots, X_P(t)]'$, $t = 1, \ldots, T$, be a stationary time series with spectral density matrix $\mathbf{f}(\omega)$. Define the vector of Fourier coefficients to be $\mathbf{d}_T(\omega) = [d_1(\omega), \ldots, d_P(\omega)]'$ where

$$d_p(\omega) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_p(t) \exp(-i2\pi\omega t), \quad p = 1, \ldots, P.$$

The Fourier $P \times P$ periodogram matrix is

$$\mathbf{I}_T(\omega) = \mathbf{d}_T(\omega)\mathbf{d}_T^*(\omega).$$

The classical estimator is the smoothed periodogram (with span m_T) which we denote as

$$\widetilde{\mathbf{f}}_T(\omega) = \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathbf{I}_T(\omega + \omega_k) \text{ where } \omega_k = k/T.$$

Call the elements of $\widetilde{\mathbf{f}}_T(\omega)$ to be $\widetilde{f}_{pq,T}(\omega)$. Define $\widehat{\mu}_T(\omega) = \frac{1}{P}\sum_{p=1}^{P} \widetilde{f}_{pp,T}(\omega)$ and $\mathbf{1}$ to be the $P \times P$ identity matrix. The shrinkage estimator for $\mathbf{f}(\omega)$ takes the form

$$\widehat{\mathbf{f}}_T(\omega) = \frac{\widehat{\beta}^2_T(\omega)}{\widehat{\delta}^2_T(\omega)} \widehat{\mu}_T(\omega)\mathbf{1} + \frac{\widehat{\alpha}^2_T(\omega)}{\widehat{\delta}^2_T(\omega)} \widetilde{\mathbf{f}}_T(\omega) \tag{2.2}$$

where the weights are as follows.

First, denote $\| \mathbf{A} \|^2$ to be the Hilbert-Schmidt norm of the matrix $\mathbf{A}$ (i.e., $\| \mathbf{A} \|^2 = \frac{1}{P}$ trace$(\mathbf{A}\mathbf{A}')$). Next, define

$$\widehat{\delta}^2_T(\omega) = \| \widetilde{\mathbf{f}}_T(\omega) - \widehat{\mu}_T(\omega)\mathbf{1} \|^2$$

which is a measure of empirical divergence (i.e., Hilbert-Schmidt norm) between the classical smoothed periodogram and the scaled identity matrix. Define $\overline{\beta}^2_T(\omega)$ to be

$$\overline{\beta}^2_T(\omega) = \frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \| \mathbf{I}_T(\omega + \omega_k) - \widetilde{\mathbf{f}}_T(\omega) \|^2 .$$

Finally, $\widehat{\beta}^2_T(\omega)$ and $\widehat{\alpha}^2_T(\omega)$ are

$$\begin{aligned} \widehat{\beta}^2_T(\omega) &= \min\{\overline{\beta}^2_T(\omega), \widehat{\delta}^2_T(\omega)\} \\ \widehat{\alpha}^2_T(\omega) &= \widehat{\delta}^2_T(\omega) - \widehat{\beta}^2_T(\omega). \end{aligned}$$

2.2.2 Extension of Shrinkage Procedure for Non-Stationary Time Series

For a given non-stationary time series, the shrinkage estimator of the SLEX spectrum at time block b and frequency ω_k is derived by extending the result above. Let $\mathbf{I}_T(b,k)$ be the SLEX periodogram at block b and frequency index k. Denote the smoothed SLEX periodogram to be

$$\widetilde{\mathbf{f}}_T(b, \omega_k) = \frac{1}{m_T} \sum_{\ell=-(m_T-1)/2}^{(m_T-1)/2} \mathbf{I}_T(b, k+\ell)$$

and whose elements are denoted $\widetilde{f}_{pq,T}(b, \omega_k)$. Denote

$$\widehat{\mu}_T(b, \omega_k) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{f}_{pp,T}(b, \omega_k).$$

The shrinkage estimator for $\mathbf{f}(b, \omega_k)$ takes the form

$$\widehat{\mathbf{f}}_T(b, \omega_k) = \frac{\widehat{\beta}_T^2(b, \omega_k)}{\widehat{\delta}_T^2(b, \omega_k)} \widehat{\mu}_T(b, \omega_k)\mathbf{1} + \frac{\widehat{\alpha}_T^2(b, \omega_k)}{\widehat{\delta}_T^2(b, \omega_k)} \widetilde{\mathbf{f}}_T(b, \omega_k) \tag{2.3}$$

where the weights are derived analogously as follows:

$$\begin{aligned}
\widehat{\delta}_T^2(b, \omega_k) &= \parallel \widetilde{\mathbf{f}}_T(b, \omega_k) - \widehat{\mu}_T(b, \omega_k)\mathbf{1} \parallel^2 \\
\widehat{\beta}_T^2(b, \omega_k) &= \min\{\overline{\beta}_T^2(b, \omega_k), \widehat{\delta}_T^2(b, \omega_k)\} \\
\widehat{\alpha}_T^2(b, \omega_k) &= \widehat{\delta}_T^2(b, \omega_k) - \widehat{\beta}_T^2(b, \omega_k)
\end{aligned}$$

where

$$\overline{\beta}_T^2(b, \omega_k) = \frac{1}{m_T^2} \sum_{\ell=-(m_T-1)/2}^{(m_T-1)/2} \parallel \mathbf{I}_T(b, k+\ell) - \widetilde{\mathbf{f}}_T(b, \omega_k) \parallel^2 .$$

2.3 Discussion

Böhm and von Sachs (2009) use a double-asymptotic framework is used to derive the mean-squared error of the shrinkage estimator of the spectrum of a stationary time series. The benchmark is the smoothed periodogram which is a consistent estimator of the spectral density matrix under mixing conditions in Brillinger (1981). However, the fact that the smoothed periodogram matrix is consistent does not mean that it has good finite sample properties. As a matter of fact, it is shown to be suboptimal with respect to L_2 risk, compared to a novel shrinkage estimator. In addition, the averaged periodogram has a bad condition number unless the smoothing span is much larger than the dimensionality. However, the smoothing span cannot be chosen to be arbitrarily large without losing high frequency resolution which can be problematic when the spectra of two processes are different at very narrow frequency band(s).

The new shrinkage estimator is shown, in the stationary case, to have asymptotically minimal L_2 risk in a class of estimators that is chosen to compensate for the bias of the eigenvalues of the averaged periodogram. Thus, the shrinkage estimator is not only more precise, but at the same time is dramatically superior in terms of condition number. As a matter of fact, the shrinkage estimator remains well-conditioned even when the dimensionality exceeds the smoothing span. Finally, intensive Monte Carlo studies (in Böhm, 2008) indicate that it is not only asymptotically superior, but that, even for very small sample size, it works far better than the averaged periodogram. This makes the shrinkage estimator an ideal candidate for the analysis of possibly high frequency, non-stationary multivariate time series.

2.4 The Algorithm of the SLEX-Shrinkage Method

Suppose that we have two multivariate non-stationary processes which are characterized by the spectra denoted as $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$ where $\mathbf{f}^g(u,\omega)$ is the $P \times P$ time-varying spectral density matrix of condition g. As stated, we need to identify time-localized spectral (auto-spectral, cross-spectral and coherence) features that best separate the two conditions. The second is to use these discriminant features to classify a *future* multivariate time series.

Goal 1: Feature Extraction and Selection.

Step 1.1 Compute the spectral matrix estimate at time-block and frequency-index (b,k) .

Let $\mathbf{X}_n(t) = [X_{n1}(t), \ldots, X_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$; be the multivariate time series recorded from N trials for condition 1. The SLEX-shrinkage spectral estimate at time block b and frequency ω_k is

$$\widehat{\mathbf{f}}^1(b,\omega_k) = \frac{1}{N}\sum_{n=1}^{N} \widehat{\mathbf{f}}_n^1(b,\omega_k)$$

where $\widehat{\mathbf{f}}_n^1(b, \omega_k)$ is the SLEX-shrinkage spectral estimate for the n-th trial of condition 1.

Let $\mathbf{Y}_n(t) = [Y_{n1}(t), \ldots, Y_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$; be the multivariate time series recorded from N trials for condition 2. The SLEX-shrinkage spectral estimate at time block b and frequency ω_k is

$$\widehat{\mathbf{f}}^2(b, \omega_k) = \frac{1}{N} \sum_{n=1}^{N} \widehat{\mathbf{f}}_n^2(b, \omega_k)$$

where $\widehat{\mathbf{f}}_n^2(b, \omega_k)$ is the SLEX-shrinkage spectral estimate for the n-th trial of condition 2.

Step 1.2. Compute the *Chernoff divergence* between the two conditions at time-block b and frequency ω_k:

$$\mathcal{D}(b, \omega_k) = \ln \frac{|\lambda \widehat{\mathbf{f}}^1(b, \omega_k) + (1 - \lambda) \widehat{\mathbf{f}}^2(b, \omega_k)|}{|\widehat{\mathbf{f}}^2(b, \omega_k)|} - \lambda \ln \frac{|\widehat{\mathbf{f}}^1(b, \omega_k)|}{|\widehat{\mathbf{f}}^2(b, \omega_k)|} \tag{2.4}$$

where $|\mathbf{A}|$ denotes the determinant of the matrix $\mathbf{A}$ and $\lambda \in (0, 1)$ is the regularization parameter . Thus, the total Chernoff divergence at time block b is

$$\mathcal{D}(b) = \sum_{k=1}^{M_b} \mathcal{D}(b, \omega_k)$$

where M_b is the number of coefficients in this block.

Step 1.3. Select the most discriminant basis.

The choice of the best discriminant basis is accomplished via the best basis algorithm (BBA) of Coifman and Wickerhauser (1992) which is a bottom-up algorithm. A parent block $S(b, j)$ is selected in favor of the children blocks $S(2b, j+1) \bigcup S(2b+1, j+1)$ if the Chernoff divergence at the parent block exceeds the sum of the divergence at the children blocks. Denote the best basis to be the collection of blocks which we denote by $\mathcal{B}$.

Goal 2: Classification.

Consider a new time vector-valued series to be $\mathbf{Z} = [\mathbf{Z}(1), \ldots, \mathbf{Z}(T)]$ with estimated spectral matrix $\widehat{\mathbf{f}}_{\mathbf{Z}}$. The goal is to classify $\mathbf{Z}$ to the condition (either 1 or 2) to which it is least dissimilar according to the Chernoff divergence criterion. The Chernoff divergence between $\mathbf{Z}$ and conditions 1 and 2, denoted $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, is

$$\begin{aligned}
\mathcal{D}_1 &= \sum_{b\in\mathcal{B}}\sum_{k} \ln \frac{|\lambda\widehat{\mathbf{f}}^1(b,\omega_k) + (1-\lambda)\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|} - \alpha \ln \frac{|\widehat{\mathbf{f}}^1(b,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|} \\
\mathcal{D}_2 &= \sum_{b\in\mathcal{B}}\sum_{k} \ln \frac{|\lambda\widehat{\mathbf{f}}^2(b,\omega_k) + (1-\lambda)\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|} - \lambda \ln \frac{|\widehat{\mathbf{f}}^2(b,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(b,\omega_k)|}.
\end{aligned}$$

If $\mathcal{D}_1 > \mathcal{D}_2$ then we classify $\mathbf{Z}$ into condition 2. Otherwise, it is classified to condition 1. In our analysis, we used $\lambda = 0.50$.

3 Data Analysis

Electroencephalograms (EEGs) were recorded in an experiment for which five participants moved the joystick from a central position to the right when a cursor flashed on the right side of a computer monitor (or left, accordingly). There were $N = 100$ trials for each condition (right and left) and the EEG trace for each trial is a 500 millisecond interval with time 0 as the stimulus onset. In our analysis, we focused on the $P = 11$ channels identified from the standard 10-20 EEG topography that are believed to be most highly involved in brain motor networks engaged in visual-motor actions, namely,

- On the left side of the scalp topography: $P3$, $C3$, $FC3$, $FC5$
- On the right side of the scalp topography: $P4$, $C4$, $FC4$, $FC5$
- Occipital channels: $O1$, Oz and $O2$.

Our analysis showed that difference between the right and left conditions is best captured by the partial coherence between $C3$ and $FC3$ channels at the alpha frequency band ($8-12$ Hz) which is significantly larger in magnitude for the left condition than the right condition (see Figure 5). This difference appears to be consistent across all five participants. We evaluated the predictive ability of the best discriminant features via a leave-one-out procedure,

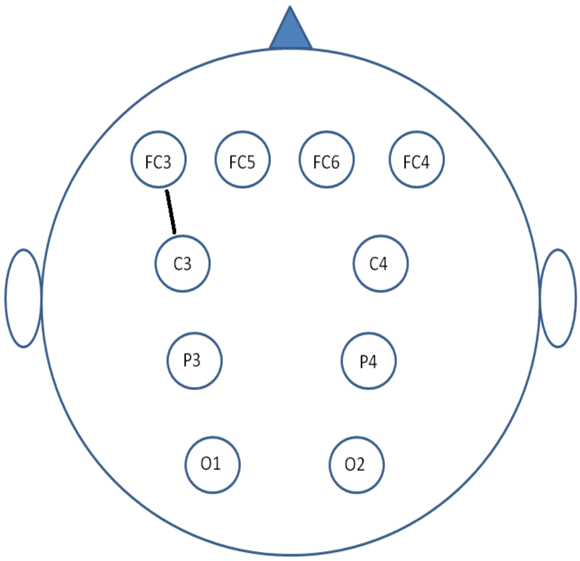

Figure 5: The most highly discriminant network feature is the alpha-band ($8 - 12$ Hertz) coherence between the $C3$ and $FC3$ channels which is significantly greater for the left condition than the right condition.

comparing the SLEX with vs. without shrinkage procedures. The obtained classification rates correctly identifying leftward or rightward movements are shown in the table below.

Participant	With Shrinkage (%)	Without Shrinkage (%)
1	71	65
2	72	67
3	74	66
4	74	66
5	68	71

The results are very promising - shrinkage in general gives a better classification rate than non-shrinkage. The SLEX-Shrinkage method discriminates between the two conditions through their spectral matrices. For zero mean Gaussian processes, all information in the time series are contained in their spectra. For non-Gaussian processes, it is necessary to investigate differences in the higher order moments.

4 Conclusion

We developed a statistical method for discrimination and classification of multivariate non-stationary signals. We addressed two major challenges namely massiveness of typical data and the poor conditioning which leads to numerically instable estimates of the spectral matrix. We used the SLEX library to extract the best set of time-frequency features that best separate classes of time series. We estimate the SLEX spectral matrix by shrinking the initial SLEX periodogram matrix estimator towards the identity matrix. The resulting shrinkage estimator is superior to (i.e., has mean-squared error than) the classical smoothed periodogram matrix and is demonstrated in this paper to produce better correct classification rates.

Acknowledgements

H. Böhm and R. von Sachs acknowledge support from the "Projet d'Actions de Recherche Concertées" no 07/12-002 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain" as well as from the IAP research network grant P 5/24 of the Belgian government (Belgian Science Policy). H. Ombao has been supported by the NSF Division of Mathematical Sciences and J. Sanes by the National Institutes of Health.

References

BÖHM, H. (2008). Shrinkage Methods for Multivariate Spectral Analysis. Ph.D. Dissertation, Université catholique de Louvain, Institut de statistique.

BÖHM, H. AND VON SACHS, R. (2009). Shrinkage Estimation in the Frequency Domain of Multivariate Time Series. *Journal of Multivariate Analysis*, **100**, 913-935.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory,* Oakland, CA: Holden-Day.

Coifman, R. and Wickerhauser, M. (1992). Entropy Based Algorithms For Best Basis Selection. *IEEE Transactions on Information Theory,* **32**, 712–718.

Dahlhaus, R. (1997). Fitting Time Series Models to Nonstationary Processes. *Annals of Statistics,* **25**, 1–37.

Dahlhaus, R. (2001). A Likelihood Approximation for Locally Stationary Processes. *Annals of Statistics,* **28**, 1762–1794.

Daubechies, I. (1992). *Ten Lectures on Wavelets,* Society for Applied and Industrial Mathematics, Philadelphia, PA.

Fryzlewicz, P. and Ombao, H. (2009). Consistent Classification of Non-Stationary Time Series Using Stochastic Wavelet Representations. *Journal of the American Statistical Association,* in press.

Huang, H-Y., Ombao, H. and Stoffer, D. (2004). Discrimination and Classification of Nonstationary Time Series Using the SLEX Model. *Journal of the American Statistical Association*, **99**, 763-774.

Kakizawa, Y., Shumway, R. and Taniguchi, M. (1998). Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association*, **93**, 328-340.

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (2001). Automatic Statistical Analysis of Bivariate Nonstationary Time Series. *Journal of the American Statistical Association,* **96**, 543–560.

Ombao, H., Raz, J., von Sachs, R. and Guo, W. (2002). The SLEX Model of a Non-Stationary Random Process. *Annals of the Institute of Statistical Mathematics*, **54**, 171-200.

Ombao, von Sachs, R. and Guo, W. (2005). The SLEX Analysis of Multivariate Nonstationary Time Series. *J. Amer. Statist. Assoc.* **100**, 519-531.

Parzen, E. (1961). Mathematical Considerations in the Estimation of Spectra. *Technometrics*, **3**, 167-190.

Saito, N. (1994). Local Feature Extraction and Its Applications. Ph.D. Dissertation, Yale University, Department of Mathematics.

Sakiyama, K. and Taniguchi, M. (2004). Discriminant Analysis for Locally Stationary Processes. *Journal of Multivariate Analysis,* **90**, 282-300.

Shumway, R. (1982). Discriminant Analysis for Time Series. *Handbook of Statistics,*, Vol. 2, eds. P. R. Krishnaiah and L. N. Kanal, New York: Holland.

Shumway, R.H. (2003). Time-Frequency Clustering and Discriminant Analysis. *Statistics and Probability Letters,* , 948-956.

Shumway, R.H. and Unger, A.N. (1974). Linear Discriminant Functions for Stationary Time Series, *Journal of the American Statistical Association,* **69**, 948-956.

Wickerhauser, M. (1994). *Adapted Wavelet Analysis from Theory to Software.* IEEE Press, Wellesley, MA.