Dynamic Classification using Multivariate Locally Stationary Wavelet Processes

Timothy Park, Idris A. Eckley, Hernando C. Ombao

PII: S0165-1684(18)30006-9
DOI: 10.1016/j.sigpro.2018.01.005
Reference: SIGPRO 6698

To appear in: *Signal Processing*

Received date: 5 March 2017
Revised date: 22 December 2017
Accepted date: 2 January 2018

Please cite this article as: Timothy Park, Idris A. Eckley, Hernando C. Ombao, Dynamic Classification using Multivariate Locally Stationary Wavelet Processes, *Signal Processing* (2018), doi: 10.1016/j.sigpro.2018.01.005

Highlights

- An approach for dynamically classifying a multivariate, locally stationary signal is proposed.
- Our aim is to classify the signal at each time point to one of a fixed number of known classes.
- To account for uncertainty in class membership at each time point we calculate the probability of a signal belonging to a particular class.
- We prove some asymptotic consistency results for this framework.
- We validate the effectiveness of the approach using simulated data and accelerometer data.

Dynamic Classification using Multivariate Locally Stationary Wavelet Processes

Timothy Park, Idris A. Eckley[†] and Hernando C. Ombao[‡]

March 9, 2018

Abstract

Methods for the supervised classification of signals generally aim to assign a signal to one class for its entire time span. In this paper we present an alternative formulation for multivariate signals where the class membership is permitted to change over time. Our aim therefore changes from classifying the signal as a whole to classifying the signal at each time point to one of a fixed number of known classes. We assume that each class is characterised by a different stationary generating process, the signal as a whole will however be nonstationary due to class switching. To capture this nonstationarity we use the recently proposed Multivariate Locally Stationary Wavelet model. To account for uncertainty in class membership at each time point our goal is not to assign a definite class membership but rather to calculate the probability of a signal belonging to a particular class. Under this framework we prove some asymptotic consistency results. This method is also shown to perform well when applied to both simulated and accelerometer data. In both cases our method is able to place a high probability on the correct class for the majority of time points.

Keywords: Wavelets; Local stationarity; Multivariate signals; Coherence; Partial coherence.

1 Introduction

This paper focuses on a supervised signal classification problem for multivariate signals. Whilst a rich literature exists for work in the one-dimensional nonstationary setting, see for example

Statistics and Data Science, Shell Global Solutions, Amsterdam, NL
[†]Department of Mathematics & Statistics, Lancaster University, Lancaster, UK
[‡]King Abdullah University of Science, Saudi Arabia and Technology and University of California, Irvine

[12, 17, 35], in recent years there has been a concerted focus on the development of multivariate nonstationary methods [29, 31, 34]. The literature on supervised signal classification has also developed in parallel with this. For example, the canonical supervised (nonstationary) signal classification problem considered by the following: [2, 4, 11, 16, 19, 20, 23, 33, 37]. The generic approach in these articles can be summarised as follows: Assume that we are given a nonstationary signal of unknown class label, then we seek to assign the *entire* signal to one of N_c different classes, using training data. The implicit assumption within the above, of course, is that the underlying process does not switch between classes.

In practice one can conceive of several situations where such a 'mono-class' assumption might not be appropriate. For example, the nonstationary (multivariate) signal in question might be piecewise (second-order) stationary, with each stationary block representing a particular class structure. To illustrate this we introduce a motivating example using accelerometer data recorded from a movement experiment, one run of which is shown in Figure 1. The experiment involves a participant performing a series of activities, namely: walking down a corridor, up a set of stairs and down a set of stairs. The accelerometer is carefully placed on the participant throughout the experiment, with the sensor orientation known and consistent across the experiment. The interest in this setting is not to classify the whole signal, but rather to associate a class with each particular activity. As such the inference challenge we address in this article is that of dynamically classifying a nonstationary signal at a given time point into a particular pre-determined class structure.

The problem of classification of signals has a long history dating back to early work on the classification of (second-order) stationary (univariate) signals. For overviews of this area we refer the reader to [36]. In the nonstationary signal setting one could use various frameworks including nonstationary adaptations of the stationary Fourier basis, see for example [33] which adopts the locally stationary Fourier model in [7]. An alternative Fourier based approach is considered by [16] and [2] who adopt the smooth localised exponentials (SLEX) framework. Of course one need not be restricted to the Fourier basis. For example, [11] and [20] use the locally stationary wavelet approach of [27] for univariate signal classification and [22] discuss a wavelet packet based classifier. In each of these settings the focus is on classifying a signal into one class, i.e. they do not tackle the problems of nonstationarity due to class switching. Thus these approaches are inadequate for classifying many real systems.

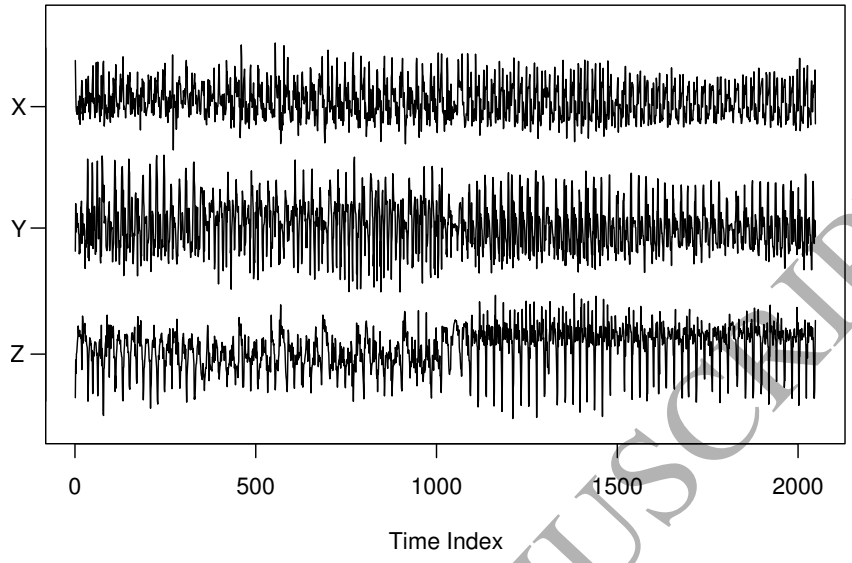


Figure 1: The X, Y and Z components of a tri-axial accelerometer signal.

A real time classification scheme is presented by [28] however they consider time points individually and so cannot account for frequency specific class differences. One possible approach that could take this into account is to segment the signal *a priori* and then assign each segment to a particular class. Such an approach is discussed in [21]. However such pre-processing can lead to some potential pitfalls. For example, in the case of a high dimensional signal, the differences between classes may be driven by only a small proportion of the channels. This can make segmentation challenging and the overall quality of classification will rely heavily on the segmentation method used. Another possible approach would be to employ a hidden Markov model (HMM). For a review of HMMs we refer the reader to [24] or [6]. Such an approach is used for classification by [1] and [5], the latter being restricted to count data. A HMM framework is also used by [26] in the related field of changepoint detection. Fitting a HMM has the drawback of being computationally intensive. It also requires the assumption that class transitions are Markovian. In other words the probability of transitioning from one class to another cannot depend on time or previous class memberships. In the absence of prior information to support these assumptions such an approach would be difficult to justify. With this in mind we introduce a novel and computationally efficient wavelet based

method for classifying a multivariate, locally stationary signal based on its local coherence structure. Our approach estimates the probability of the signal belonging to a particular class at each time point. Importantly our approach, which requires an assumption of local stationarity, requires little in the way of pre-processing save for the removal of the (time-varying) mean.

The method which we introduce is based on the Multivariate Locally Stationary Wavelet model introduced by [31]. The Multivariate Locally Stationary Wavelet model is able to account for changes in both the second order properties of the individual channels of a multivariate signal as well as the linear relationships between channels. For our classification model the nonstationarity in the signal is due to class switching causing the underlying process to change. Whilst in some applications one may have domain-specific knowledge which permits the specification of equi-duration epochs typically this is not the case. In this article we focus on the dependence *between* channels by using wavelet coherence. Wavelet coherence has the useful property of being normalised with respect to the local spectral structure. In addition it benefits from providing a scale-specific strength of the linear dependence between components, is time-dependent and can be readily computed using an efficient transform. Other methods, such as [16] or [11], normalise the spectral estimates using the global variance of the signal. In our setting, where class membership is a local rather than global characteristic, we must us a local normalisation. Our ultimate goal for classification is to identify the probability of the test signal belonging to each of the classes at a particular time given the observed data. Calculating these probabilities, as opposed to assigning whichever class is closest according some distance measure, will demonstrate the uncertainty in classification.

The remainder of the paper is organised as follows. Section 2 provides an overview of the Multivariate Locally Stationary Wavelet model as well as the parameter estimation method which will be used. The main contribution of this paper is contained in Section 3 which gives details of our classification method and how it can be applied in practice. Section 4 contains two different examples of our method applied to simulated data while Section 5 contains an example of our method applied to accelerometer data.

2 The Multivariate Locally Stationary Wavelet Model

We now introduce the key elements of the modelling framework which will be used as the foundation of our classification model, the Multivariate Locally Stationary Wavelet model of [31]. This is a multivariate generalisation of the univariate LSW model of [27]. Following [31] let $\mathbf{X}_t = [X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(P)}]$ be a P-dimensional Multivariate Locally Stationary Wavelet process of length T where $T = 2^J$ for some $J \in \mathsf{N}$. Also let $\mathbf{V}_j(k/T)$ be a lower triangular matrix of functions known as the transfer function matrix and $\{\mathbf{z}_{jk}\}$ be a set of independent random vectors with the properties $E[\mathbf{z}_{jk}] = \mathbf{0}$ and $\mathrm{Var}\{\mathbf{z}_{jk}\} = \mathbf{1}$. Finally let $\{\psi_{j,k}\}$ be the set of discrete wavelet coefficients. $\mathbf{X}_t$ can then be represented as follows,

$$\mathbf{X}_t = \sum_{j=1}^{\infty} \sum_{k} \mathbf{V}_j(k/T)\psi_{j,t-k}\mathbf{z}_{j,k}. \tag{1}$$

A number of smoothness assumptions are also required on the elements of the transfer function matrix, $\mathbf{V}_j(k/T)$, see [31] for further details.

The transfer function matrix dictates both the auto- and cross-covariance properties of the signal. These properties can be uniquely represented by the Local Wavelet Spectral (LWS) matrix which is defined as follows: $\mathbf{S}_j(u) = \mathbf{V}_j(u)\mathbf{V}_j(u)$ for a given scale, j, and rescaled time point, $u = t/T$. As [31] describe, the diagonal elements of the LWS determine the auto-covariance structure of the individual channels of the signal, whilst the off diagonal terms determine the cross-covariance structure between pairs of channels.

Following [31], we define the wavelet coherence at scale j to be the matrix, $\boldsymbol{\rho}_j(u)$, which has the form,

$$\boldsymbol{\rho}_j(u) = \mathbf{D}_j(u)\mathbf{S}_j(u)\mathbf{D}_j(u), \tag{2}$$

where $\mathbf{D}_j(u)$ is a diagonal matrix whose elements are $S_j^{(p,p)}(u)^{(-1/2)}$. The (p,q)-th element of the coherence matrix, $\rho_j^{(p,q)}(u)$, quantifies the strength of any linear relationship between channels p and q at scale j and rescaled time point u and takes a value on the interval [-1,1]. A value close to 1 indicates a strong linear relationship whereas a value close to -1 indicates a strong negative relationship.

To estimate the LWS and coherence matrices of a process [31] introduce the empirical wavelet coefficient vector at scale j and location k, $\mathbf{d}_{jk} = \sum_t \mathbf{X}_t \psi_{jk}$. This vector can be used to define the raw wavelet periodogram matrix, $\mathbf{I}_{jk} = \mathbf{d}_{jk}\mathbf{d}_{jk}$. This is a biased estimator of the LWS matrix, $\mathbf{S}_{jk}$. Fortunately, an asymptotically unbiased and consistent estimate can be achieved by smoothing the estimate over time using a rectangular kernel smoother with window size $(2M+1)$, a commonly used approach in the time series literature see e.g. [3, 30, 32], and applying the inverse of the autocorrelation wavelet inner product matrix, $\mathbf{A}$, with elements $A_{jl} = \sum_\tau \Psi_j(\tau)\Psi_l(\tau)$ where $\Psi_j(\tau) = \sum_k \psi_{jk}(0)\psi_{jk}(\tau)$ (see [27] or [9] for further details). Hence our (asymptotically) unbiased estimate of the LWS matrix is given by $\widehat{\mathbf{S}}_{jk} = (2M+1)^{-1} \sum_{m=k-M}^{k+M} \sum_l A^{-1}_{jl} \mathbf{I}_{lm}$. The coherence matrix can then be estimated by substituting $\widehat{\mathbf{S}}_{jk}$ into equation (2). In Section 3 we will make use of wavelet coherence in order to classify a signal.

3 Dynamic Classification

We now consider the classification problem for a Multivariate Locally Stationary Wavelet signal, $\mathbf{X}_t$. The setting which we consider is the following: Assume that at any time, t, $\mathbf{X}_t$ will belong to one of $N_c \geq 2$ different classes where N_c is known. The class membership of $\mathbf{X}_t$ at time t is denoted by $C_X(t) \in \{1, 2, \ldots, N_c\}$. We do not assume that the class membership of $\mathbf{X}_t$ is constant for all time points, nor do we assume that the time spent in a particular class is fixed. Instead we assume that whilst a signal is in a given class it is second order stationary. In other words if $C_X(t) = c$, $\forall t \in \{\tau_1, \ldots, \tau_2\}$, the transfer function matrix, $\mathbf{V}_j(t)$ is a constant, i.e. $\mathbf{V}_j(t) = \mathbf{V}_j^{(c)}$, $\forall t \in \{\tau_1, \ldots, \tau_2\}$. The matrix $\mathbf{V}_j^{(c)}$ is the class specific transfer function which has the same lower triangular form as the transfer function matrix described in Section 2, however $\mathbf{V}_j^{(c)}$ is constrained to be constant over time. In effect this particular assumed representation means that we can re-express the representation in equation (1) as follows. Let $\mathsf{I}_{\{c\}}[C_X(t)]$ be an indicator function which is equal to 1 if $C_X(t) = c$ and 0 otherwise. Then $\mathbf{X}_t$ can be expressed as,

$$\mathbf{X}_t = \sum_k \sum_j \sum_{c=1}^{N_c} \mathsf{I}_{\{c\}}[C_X(k)] \mathbf{V}_j^{(c)} \psi_{jk}(t) \mathbf{z}_{jk}.$$

In effect what we have done here is to re-write the time varying transfer function matrix in terms of constant segments, $\mathbf{V}_j(k/T) = \sum_{c=1}^{N_c} \mathsf{I}_{\{c\}}[C_X(k)]\mathbf{V}_j^{(c)}$.

With this formulation in place it is readily seen that we can also write the LWS of $\mathbf{X}_t$ at rescaled time $u = t/T$ as,

$$\mathbf{S}_j(u) = \mathbf{V}_j(u)\mathbf{V}_j(u) = \sum_{c=1}^{N_c} \mathsf{I}_{\{c\}}[C_X(k)]\mathbf{S}_j^{(c)},$$

where $\mathbf{S}_j^{(c)}$ is the class specific LWS defined as $\mathbf{S}_j^{(c)} = \mathbf{V}_j^{(c)}\mathbf{V}_j^{(c)}$. Equivalently, we can express the time varying coherence matrix at rescaled time u as $\boldsymbol{\rho}_j(u) = \sum_{c=1}^{N_c} \mathsf{I}_{\{c\}}[C_X(k)]\boldsymbol{\rho}_j^{(c)}$.

In the next section we will use the coherence matrix to determine which class the signal belongs to at a particular time. In order to do this we assume that each class has a different coherence matrix, or more precisely for each pair $c_1, c_2 \in \{1, 2, \ldots, N_c\}, c_1 \neq c_2$ there exists some j such that $\boldsymbol{\rho}_j^{(c_1)} - \boldsymbol{\rho}_j^{(c_2)} \neq \mathbf{0}$.

In the following sections we will describe our dynamic classification approach. In essence this consists of five key steps which we first sketch here: (1) Estimate the coherence of a set of (labelled) training signals; (2) Transform each of the coherence estimates using the Fisher-z transform; (3) Using the known class membership of the training signals, we estimate the transformed coherence of each class; (4) We then select a set of highly discriminative coherence coefficients that will be used to classify future observed signals; (5) Using the set of discriminative coefficients and the estimated transformed coherence, we calculate the probability of an unknown signal belonging to each class at a given time point.

3.1 Training Data

To estimate the probability of signal $\mathbf{X}_t$ being in a particular class at a particular time we make use of a set of N_i labelled training signals, the i-th element of which is denoted, $\{\mathbf{Y}_t^{(i)}\}_{i \ \{1,2,\ldots,N_i\}}$. Each of the labelled signals are assumed to have a representation of the form described in Section 3. Each training signal will have an associated class function $C_{Y^{(i)}}(t)$ which is known. We estimate the LWS matrix, $\widehat{\mathbf{S}}_{jk;Y^{(i)}}$, for each training signal followed by the coherence matrix, $\widehat{\boldsymbol{\rho}}_{jk;Y^{(i)}}$. This is done using the method described in Section 2.

Our ultimate goal for classification is to calculate the probability of the signal belonging to

a particular class at a particular time point. To do this we must calculate the likelihood and therefore make distributional assumptions about the estimated coherence. We find in practice that the coherence does not tend to readily fit any standard distribution. We therefore take a Fisher's-z transform of the coherence, the estimates of which are well approximated by a Gaussian distribution, see [10]. The transformed coherence for class c, $\boldsymbol{\zeta}_j^{(c)}$ is,

$$\boldsymbol{\zeta}_j^{(c)} = \tanh^{-1} \boldsymbol{\rho}_j^{(c)}. \tag{3}$$

The mean of the transformed coherence estimate for class c is thus estimated by averaging the elements of the transformed coherence estimate, $\widehat{\boldsymbol{\zeta}}_{jk;Y_i} = \tanh^{-1} \widehat{\boldsymbol{\rho}}_{jk;Y_i}$, for which $C_{Y^{(i)}}(k) = c$,

$$\widehat{\boldsymbol{\zeta}}_j^{(c)} = \frac{1}{\sum_{i=1}^{N_i} \sum_k \mathsf{I}_{\{c\}}[C_{Y^{(i)}}(k)]} \sum_{i=1}^{N_i} \sum_k \mathsf{I}_{\{c\}}[C_{Y^{(i)}}(k)] \widehat{\boldsymbol{\zeta}}_{kj;Y^{(i)}}. \tag{4}$$

In a similar way the variance can also be estimated from the training data.

3.2 Selection of Highly Discriminative Coefficients

Following [11,20] we will not use the whole set of transformed coherence coefficients for classification. Instead we use a subset of coefficients which show the greatest discrepancy between classes. Using a subset of highly discriminative coefficients will reduce the error in the class probability estimate and also reduce the computational complexity of calculating the log-likelihood. We denote such a subset, which contains the scale and channel indices (j, p, q) for $p < q$, as $\mathcal{M}$. In order to select the appropriate coefficients we rank them according to the discrepancy measure, $\Delta_{jk}^{(p,q)}$, defined as,

$$\Delta_j^{(p,q)} = \sum_{c=1}^{N_c} \sum_{g=c+1}^{N_c} \left| \frac{\widehat{\zeta}_j^{(p,q)(c)} - \widehat{\zeta}_j^{(p,q)(g)}}{\sqrt{\operatorname{var}\left(\widehat{\zeta}_j^{(p,q)(c)}\right) + \operatorname{var}\left(\widehat{\zeta}_j^{(p,q)(g)}\right)}} \right|. \tag{5}$$

This discrepancy measure is adapted from the discrepancy measure in [20] and incorporates the variance of the transformed coherence estimates which can be found empirically using the training data, though of course other approaches could be used, e.g. [13]. We select those coefficients which are found to have the largest distance measure.

3.3 Classification

Our ultimate goal is to estimate the time varying class membership of the signal, $\mathbf{X}_t$. We do this by estimating the probability of the signal belonging to a particular class at a particular time point. We first estimate the transformed coherence for $\mathbf{X}_t$ denoted as $\widehat{\boldsymbol{\zeta}}_{jk;X}$. Given this estimate we can use Bayes' theorem to obtain,

$$\begin{aligned}\Pr\left[C(k)=c|\,\widehat{\boldsymbol{\zeta}}_{jk;X}\right] \propto \\ \Pr\left[C(k)=c\right]\mathcal{L}\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\{\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\}\right),\end{aligned} \tag{6}$$

where $\mathcal{L}(\theta|x)$ is the likelihood and $\Pr\left[C(k)=c\right]$ is a prior probability.

Note In the absence of prior knowledge we assign an equal prior probability of $1/N_c$ to each class.

Due to the use of the Fisher-z transform we can assume that the distribution of the transformed coherence estimator can be approximated by a Gaussian distribution and so $\mathcal{L}(x|\theta)$ is the Gaussian likelihood function with mean vector, $\mu^{(c)}$ and variance covariance matrix, $\Sigma^{(c)}$. The elements of $\mu^{(c)}$ are the elements of $\zeta_j^{(p,q)(c)}\ \forall\ p,q,j\in\mathcal{M}$. We also define $\widehat{\mu}_k$ which contains the elements of $\widehat{\zeta}_{jk;X}^{(p,q)}\ \forall\ p,q,j\in\mathcal{M}$. The density function, up to a constant factor, can then be expressed as follows:

$$\begin{aligned}\mathcal{L}\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\right)\propto \\ \left|\Sigma^{(c)}\right|^{-\frac{1}{2}}\exp-\frac{1}{2}\left\{(\widehat{\mu}_k-\mu^{(c)})\ \left(\Sigma^{(c)}\right)^{-1}(\widehat{\mu}_k-\mu^{(c)})\right\}.\end{aligned} \tag{7}$$

Since the true mean vectors and variance covariance matrices of $\widehat{\boldsymbol{\zeta}}_{jk;X}$ are not known we substitute estimates taken from the training data described in Section 3.1. Computational considerations mean that it is easier to calculate the log-likelihood function, $\ell\left(x\,|\theta\right)=\log\left\{\mathcal{L}\left(x\,|\theta\right)\right\}$. These can be easily related to the probabilities using the following

$$\begin{aligned}\Pr\left[C(k)=c|\,\widehat{\boldsymbol{\zeta}}_{jk;X}\right]= \\ \frac{\exp\left\{\ell\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\{\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\}\right)\right\}}{\sum_{c=1}^{N_c}\exp\left\{\ell\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\{\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\}\right)\right\}}.\end{aligned} \tag{8}$$

With the above in place we can consider the probability of misclassification. To this end we define a misclassification at a particular rescaled time k/T as the highest class membership probability being placed on a class other than the true class. In the following propositions we establish the asymptotic probability of misclassifying a signal of length T.

Proposition 1 *Let $\Delta(\widehat{\boldsymbol{\mu}}_k)$ be a divergence criterion for a signal with length T. Also let M_T be the smoothing parameter used for spectral estimation. To ensure an asymptotically consistent and unbiased spectral estimate [31] make the assumptions that $M_T \to \infty$ and $M_T/T \to 0$ as $T \to \infty$. We use the divergence criterion to estimate the class membership at rescaled time k/T. In practice we place probabilities on the class memberships however in order to establish the asymptotic properties of the method we use the decision rule, $D(\widehat{\boldsymbol{\mu}}_k)$. For the case of two classes the decision rule is defined as,*

$$D(\widehat{\boldsymbol{\mu}}_k) = \begin{cases} 1 & \text{(estimate } C(k) = 1\text{) if } \Delta(\widehat{\boldsymbol{\mu}}_k) > 0 \\ 2 & \text{(estimate } C(k) = 2\text{) if } \Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0 \end{cases}.$$

We show that if the true class membership at rescaled time k/T is class 1 then the probability that $D(\widehat{\boldsymbol{\mu}}_k) = 2$ will tend to zero asymptotically, in other words,

$$\lim_{T \to \infty} Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2 | C(k) = 1) = 0$$

Proof: See Appendix 6.

This result can be generalised to the case of $N_c > 2$ by replacing class 2 with whichever class, other than class 1, has the highest likelihood at location k.

We also consider the asymptotic effect of increasing the Euclidean distance between classes on the misclassification probability.

Proposition 2 *Again using the divergence criterion, $\Delta(\widehat{\boldsymbol{\mu}}_k)$, and decision rule, $D(\widehat{\boldsymbol{\mu}}_k)$, defined in proposition 1 we consider the two class problem and the distance between classes $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$. We show that for fixed T as the distance between classes increases the probability of assigning the*

incorrect class, at rescaled time k/T, tends to zero. In other words,

$$\lim_{|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty} Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2 | C(k) = 1) = 0$$

Proof: See Appendix 6.

Again this result can be generalised to $N_c > 2$ is the same way as proposition 1.

With the above results established, we are now in a position to formally state our dynamic classification algorithm. This is described by the following pseudocode:

Input: A set of training signals, $\mathbf{Y}_i$ with known class membership and a signal $\mathbf{X}$ with unknown class membership.

1. **for** $i \in \{1, 2, \ldots, N_i\}$ **do**
   calculate $\widehat{\boldsymbol{\zeta}}_{jk;Y_i}$.

end

2. **for** $c \in \{1, 2, \ldots, N_c\}$ **do**
   from the set, $\{\widehat{\boldsymbol{\zeta}}_{jk;Y_i}\}_i$, calculate $\widehat{\boldsymbol{\zeta}}_j^{(c)}$.

end

3. Select highly discriminative coefficients based on $\Delta_j^{(p,q)}$.

4. **for** $c \in \{1, 2, \ldots, N_c\}$ **do**
   calculate $\Pr\left[C(k) = c | \widehat{\boldsymbol{\zeta}}_{jk;X}\right]$.

end

Output: The probability of signal $\mathbf{X}$ at time k, $\Pr\left[C(k) = c | \widehat{\boldsymbol{\zeta}}_{jk;X}\right]$.

Algorithm 1: Pseudocode of the mvLSW dynamic classification approach.

4 Simulated Examples

In order to demonstrate how our method works in practice we now present a series of simulated data examples.

4.1 Example with Class Specific Autocovariance

The example presented in this section includes signals where both the auto- and cross-covariance structures are dependent upon the time varying class membership. We use a piecewise stationary

trivariate autoregressive processes of the form,

$$\mathbf{X}_t = \begin{cases} \phi_1^{(1)}\mathbf{X}_{t-1} + \phi_2^{(1)}\mathbf{X}_{t-2} + \xi_t & \text{if} \quad C_X(t) = 1 \\ \phi_1^{(2)}\mathbf{X}_{t-1} + \phi_2^{(2)}\mathbf{X}_{t-2} + \xi_t & \text{if} \quad C_X(t) = 2 \end{cases}.$$

Here $\{\phi_1^{(1)}, \phi_2^{(1)}\} = \{0.8, -0.5\}$ and $\{\phi_1^{(2)}, \phi_2^{(2)}\} = \{0.9, 0\}$ are the class specific AR coefficients. The set of random elements, $\{\xi_t\}$, are taken from a multivariate normal distribution with zero mean and class specific covariances such that,

$$\xi_t \sim \begin{cases} N(\mathbf{0}, \mathbf{\Sigma}^{(1)}) & \text{if} \quad C_X(t) = 1 \\ N(\mathbf{0}, \mathbf{\Sigma}^{(2)}) & \text{if} \quad C_X(t) = 2 \end{cases},$$

where,

$$\mathbf{\Sigma}^{(1)} = \begin{bmatrix} 1 & 0.4 & 0.6 \\ 0.4 & 1 & 0 \\ 0.6 & 0 & 1 \end{bmatrix}, \; \mathbf{\Sigma}^{(2)} = \begin{bmatrix} 1 & -0.4 & -0.6 \\ -0.4 & 1 & 0 \\ -0.6 & 0 & 1 \end{bmatrix}. \tag{9}$$

We simulate a set of 10 training signals using this model. The training signals each have the same class function which is initially in class 1 and then switches to class 2 half way through the time span. In order to test our method we simulate a group of 100 validation signals. The validation signals all have the same class function which is very different to the one used in the training set. This class function is initially in class 1 but switches 7 times at irregularly spaced intervals. We estimate the class membership probabilities for the validation signals using the method outlined in Section 3 and then take the mean.

The results of this are shown in Figure 2. We can see that the mean class probability is consistently high for the true class which shows that our method has performed well in terms of identifying the most likely class for a given time point. We also note that there is a small region of uncertainty around the class transitions which demonstrates that it is more difficult to classify in these regions. Looking at the lower plot in Figure 2 we can see that it is possible to identify the class membership visually as the signals autocovariance structure changes noticeably with class due to the changing AR coefficients. In the following sections we will explore examples where this is not the case.

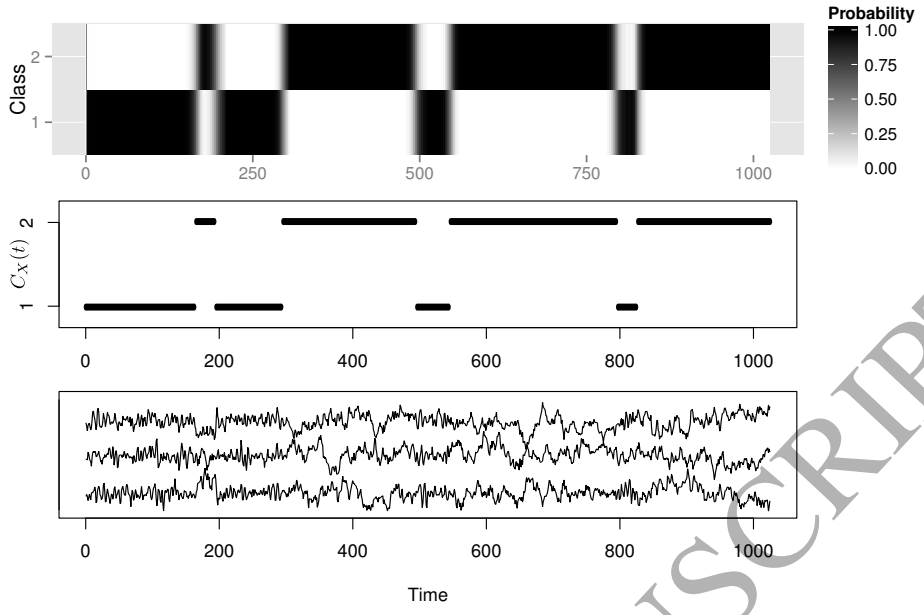


Figure 2: The upper plot shows the mean class membership probabilities for the 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time.

4.2 Example with Constant Auto-covariance

We now consider an example with a class specific cross-covariance structure and a constant auto-covariance structure. We again use an autoregressive process where, unlike the previous example, the AR coefficients are not class specific. The general form of the signals is therefore,

$$\mathbf{X}_t = 0.8\mathbf{X}_{t-1} - 0.5\mathbf{X}_{t-2} + \xi_t, \quad \forall t \in \{0, T-1\}. \tag{10}$$

The set of random elements, $\{\xi_t\}$ again follow a normal distribution with zero mean and covariances defined in equation (9).

Our example is based on a set of 10 training signals and 100 validation signals. The training signals all have the same simple class function as in the previous section, the validation signals all have the same class function which starts in class 2 and switches seven times at irregular intervals. We calculate the class membership probabilities for the validation signals and take the mean, the results are shown in Figure 3. Looking at the lower plot in Figure 3 we see that for this example it is very challenging to discern the class visually as the auto-covariance structure is constant. The upper plot indicates that despite this our method is still performing to a similar level of accuracy

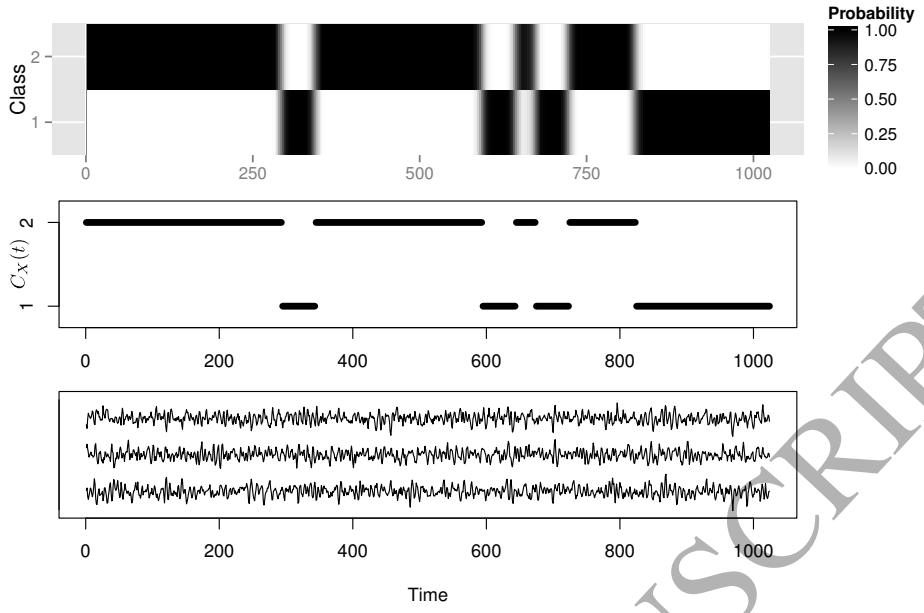


Figure 3: The upper plot shows the mean class membership probabilities for the group of 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time

as for the example in Section 4.1.

4.3 Example with Three Classes

Our final simulated example considers a scenario where $N_c > 2$. A third class is added to the example in Section 4.2. The AR coefficients will remain constant as in equation (10) however for time points where $C_X(t) = 3$ the random elements $\{\xi_t\}$ will be taken from a normal distribution with covariance matrix given by,

$$\mathbf{\Sigma}^{(3)} = \begin{bmatrix} 1 & 0.4 & -0.6 \\ 0.4 & 1 & 0 \\ -0.6 & 0 & 1 \end{bmatrix}$$

For this example we again use a set of 10 training signals. Each of the training signals has a class function which cycles through the three classes from 1 to 3 twice. We simulate one group of 100 validation signals which have a class function which also cycles through the class but in reverse order. Figure 4 shows the mean class probabilities for the validation signals. Note that our method is able to place a high probability on the correct class for the majority of time points. It is however

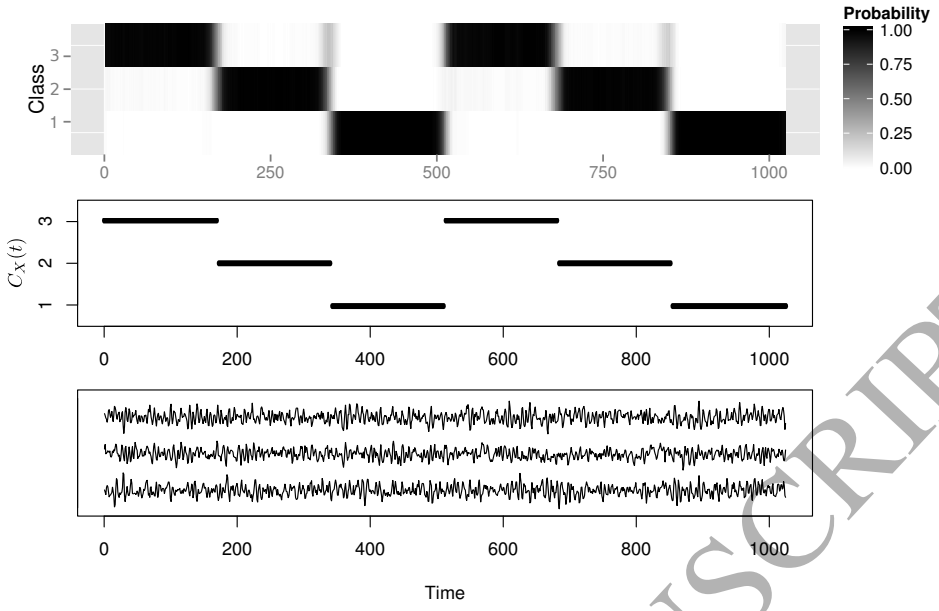


Figure 4: The upper plot shows the mean class membership probabilities for the group of 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time.

possible to see that there are slightly larger regions of uncertainty around the class transitions. This demonstrates that by adding a third class we have made the classification problem more challenging leading to greater uncertainty.

5 Accelerometer Data Example

Finally we turn to an example based on tri-axial accelerometer data. A participant is asked to walk normally following a route including a corridor and several flights of stairs whilst wearing a tri-axial accelerometer which has a recording frequency of 20Hz. The experiment is repeated 13 times in total, following three different routes. The accelerometer records continuously during each repetition. For 6 of the repetitions the participant walks along the corridor up the stairs and down the stairs before walking along the corridor again, we will refer to this as Route A. For another 6 repetitions the participant walks down the stairs, along the corridor twice and then up the stairs, we will refer to this as Route B. For the 13th repetitions the participant walks up the stairs, down the stairs and then along the corridor, we refer this as Route C. Each repetition lasts just over 100 seconds, with the accelerometer placed in the participant's pocket. For each repetition, the

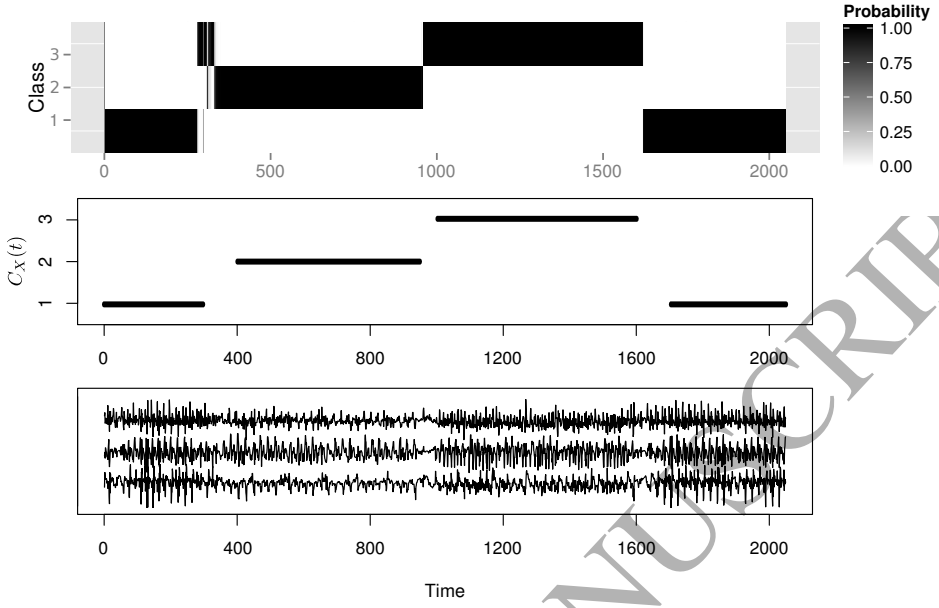
accelerometer was oriented the same way with respect to gravity and remained in place for the duration of the experiment. This ensured that the three channels could be directly compared between each repetition. Each recording is trimmed to be of length $T = 2048$. Since we do not expect the accelerometer to experience any changes in orientation, but rater oscillate around a fixed orientation, we expect the recorded signals to have a constant mean with any information relating to activity-type (walk, climb up stairs, climb down stairs) captured in the second order structure.

To illustrate our method we randomly select one repetition from each of Routes A and B, as well as the single repetition of Route C as our test set. The remaining 10 repetitions will be used as a training set. We adopt a three class model with class 1 being walking along the corridor, class 2 being walking up the stairs and class 3 being walking down the stairs. Figure 5 shows the classification results for the signals of Routes A and B. In both cases the true class is given a high probability for nearly all time points, the only exception to this being around the first transition in the Route A signal where the highest probability is placed on class 3 when the true class is either 1 or 2. The middle plots show the true class memberships, it is noticeable that there are very clear shifts in the probabilities which follow the true class memberships.

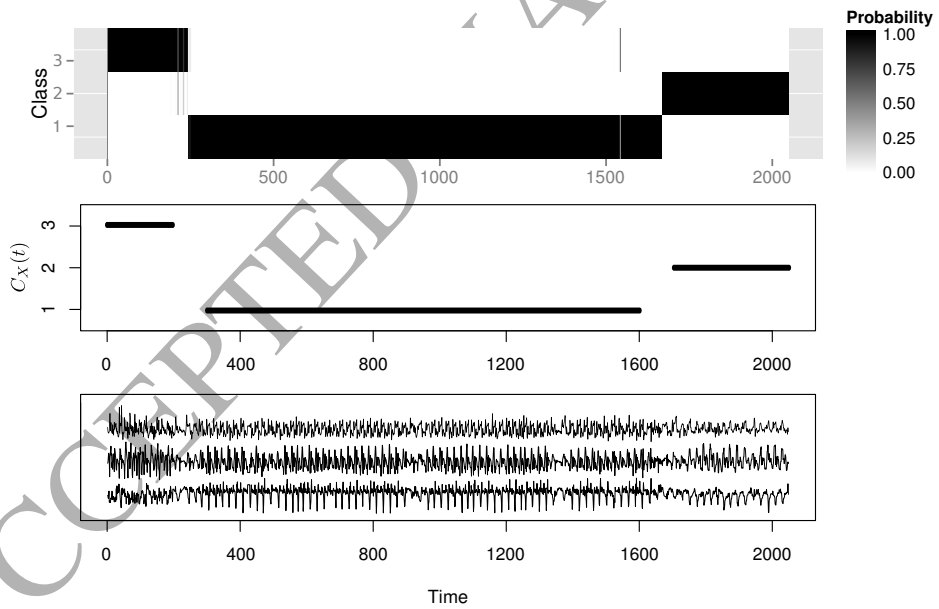
Figure 6 shows the results of our classification method performed on the Route C repetition. Since this repetition follows Route C the resulting signal is unlike any in the training data which all follow Route A or Route B. Looking at Figure 6 we see that our method is able to place a high probability on the true class for the majority of time points meaning that we can say what activities the participant is performing during Route C.

6 Discussion and Future Work

In this article we proposed a classification method for signals where the class membership is permitted to change over time. Such a model is distinct from the majority of classification methods which seek to assign a signal to one class for all time points. Our method makes use of a set of labelled training signals to estimate the true spectral properties of each class. Likelihood methods are then used to calculate the probability of the signal being in each class at a particular time point. We also demonstrate this method using both simulated data examples and a real accelerometer data set.



(a) Route A



(b) Route B

Figure 5: Class probabilities for Routes A and B. The upper plots show the estimated class probabilities. The lower plots show the accelerometer recordings. The middle plot shows the true class memberships.

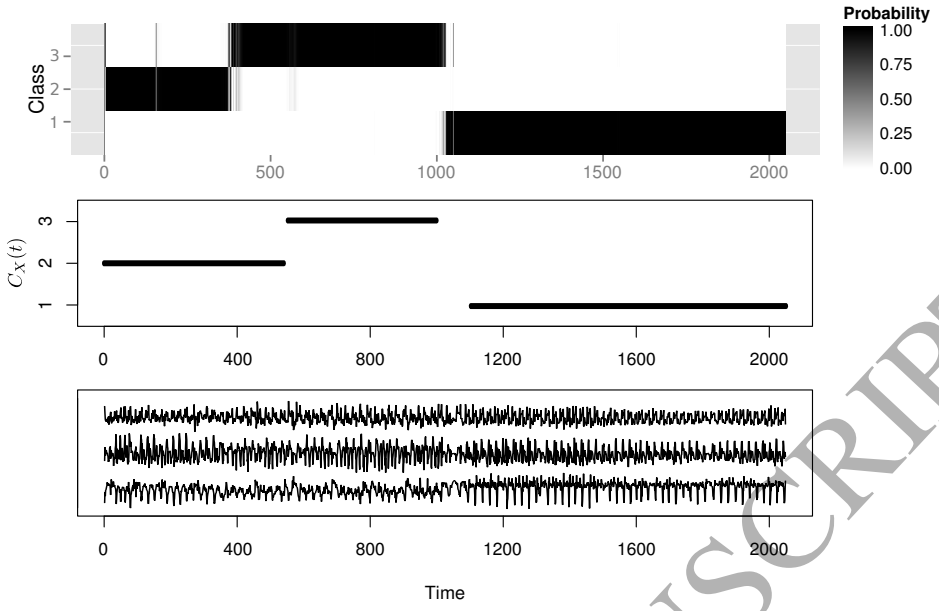


Figure 6: Class probabilities for Route C. The upper plots show the estimated class probabilities. The lower plots show the accelerometer recordings. The middle plot shows the true class memberships.

The work presented naturally gives rise to several potential avenues for future research. Primary amongst these is the important question of which wavelet family to use as an analysing wavelet in the absence of any specific knowledge about the underlying data generating process, building on the work of [14]. In practice, we have not noticed a marked difference in performance when different wavelet families are used, however a careful theoretical treatment to establish this is nevertheless an intriguing avenue for future research. It would also be interesting to contrast our locally stationary wavelet approach with other equivalent (linear) approaches such as multivariate extensions of local Fourier or time-varying AR (see [8] for further details). Finally, we are grateful to a reviewer for highlighting the very interesting prospect of extending this framework to consider series arising from multiple subjects. In such settings, being able to identify both population and subject-specific effects can be interesting. Whilst we are not aware of any pre-existing work in the multivariate LSW framework that addresses this setting, recent work by Gott, Eckley and Aston [15] on estimating the population local wavelet spectrum in a univariate setting offers some intriguing possibilities that could be extended to this multivariate setting. We leave this as an interesting prospect for future research.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive comments and suggestions that have significantly improved the quality of this manuscript. Park gratefully acknowledges funding from the EPSRC-funded STOR-i Centre for Doctoral Training and Unilever Research. Eckley's work was supported by the Engineering and Physical Sciences Research Council under grant EP/I01697X/1, whilst Ombao gratefully acknowledges funding from NSF DMS and NSF SES.

References

[1] Ainsleigh, P. L., Kehtarnavaz, N., and Streit, R. L. Hidden Gauss-Markov models for signal classification. *Signal Processing, IEEE Transactions on, 50*, 6 (2002), 1355–1367.

[2] Böhm, H., Ombao, H. C., von Sachs, R., and Sanes, J. Classification of multivariate non-stationary signals: The SLEX-shrinkage approach. *Journal of Statistical Planning and Inference 140*, 12 (Dec. 2010), 3754–3763.

[3] Brillinger, D. *Time series: data analysis and theory*, vol. 36. SIAM, 2001.

[4] Caiado, J., Crato, N., and Peña, D. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis 50*, 10 (June 2006), 2668–2684.

[5] Cappé, O. A Bayesian approach for simultaneous segmentation and classification of count data. *Signal Processing, IEEE Transactions on 50*, 2 (2002), 400–410.

[6] Cappé, O., Moulines, E., and Ryden, T. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2006.

[7] Dahlhaus, R. Fitting time series models to nonstationary processes. *The Annals of Statistics 25*, 1 (1997), 1–37.

[8] Dahlhaus, R. Locally stationary processes. *The Handbook of Statistics 30* (2012), 351–413.

[9] Eckley, I., and Nason, G. Efficient computation of the discrete autocorrelation wavelet inner product matrix. *Statistics and Computing 15*, 2 (Apr. 2005), 83–92.

[10] Fisher, R. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika 10*, 4 (1915), 507–521.

[11] Fryzlewicz, P., and Ombao, H. C. Consistent classification of nonstationary time series using stochastic wavelet representations. *Journal of the American Statistical Association 104*, 485 (Mar. 2009), 299–312.

[12] Gianfelici, F. Rbf-based technique for statistical demodulation of pathological tremor. *IEEE Transactions on Neural Networks and Learning Systems 24*, 10 (Oct 2013), 1565–1574.

[13] Gianfelici, F., and Farina, D. An effective classification framework for brain-computer interfacing based on a combinatoric setting. *IEEE Transactions on Signal Processing 60*, 3 (March 2012), 1446–1459.

[14] Gott, A. N., and Eckley, I. A. A note on the effect of wavelet choice on the estimation of the evolutionary wavelet spectrum. *Communications in statistics-simulation and computation 42* (2013), 393–406.

[15] Gott, A. N., Eckley, I. A., and Aston, J. A. D. Estimating the population local wavelet spectrum with application to non-stationary functional magnetic resonance imaging time series. *Statistics in Medicine 34* (2015), 3901–3915.

[16] Huang, H.-Y., Ombao, H. C., and Stoffer, D. S. Discrimination and classification of nonstationary time series using the SLEX model. *Journal of the American Statistical Association 99*, 467 (Sept. 2004), 763–774.

[17] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 454*, 1971 (1998), 903–995.

[18] Isserlis, L. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika 12*, 1/2 (1918), 134–139.

[19] Kakizawa, Y., Shumway, R. H., and Taniguchi, M. Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association 93*, 441 (1998), 328–340.

[20] Krzemieniewska, K., Eckley, I., and Fearnhead, P. Classification of non-stationary time series. *Stat 3*, 1 (2014), 144–157.

[21] Krzemieniewska, K. I. Classification of non-stationary time series, 2013.

[22] Learned, R., Karl, W., and Willsky, A. Wavelet packet based transient signal classification. *Time-Frequency and Time-Scale Analysis, 1992., Proceedings of the IEEE-SP International Symposium* (1992), 109–112.

[23] Liu, S., and Maharaj, E. A. A hypothesis test using bias-adjusted AR estimators for classifying time series in small samples. *Computational Statistics & Data Analysis 60* (Apr. 2013), 32–49.

[24] MacDonald, I., and Zucchini, W. *Hidden Markov and Other Models for Discrete- valued Time Series*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1997.

[25] Meenakshi, A., and Rajian, C. On a product of positive semidefinite matrices. *Linear algebra and its applications 295* (1999), 9–12.

[26] Nam, C. F. H., Aston, J. A. D., Eckley, I. A., and Killick, R. The Uncertainty of Storm Season Changes: Quantifying the Uncertainty of Autocovariance Changepoints. *Technometrics 52* (2015), 194–206.

[27] Nason, G., von Sachs, R., and Kroisandt, G. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society. Series B 62*, 2 (May 2000), 271–292.

[28] Olsen, G., and Brilliant, S. Signal processing and machine learning for real-time classification of ergonomic posture with unobtrusive on-body sensors; application in dental practice. *Complex Medical Engineering, 2009. CME. ICME International Conference on* (2009), 1–11.

[29] Ombao, H., von Sachs, R., and Guo, W. SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc. 100*, 470 (2005), 519–531.

[30] Ombao, H. C., Raz, J. A., Strawderman, R. L., and Von Sachs, R. A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika 88* (2001), 1186–1192.

[31] Park, T., Eckley, I., and Ombao, H. Estimating Time-Evolving Partial Coherence Between Signals via Multivariate Locally Stationary Wavelet Processes. *IEEE Transactions on Signal Processing 62*, 20 (2014), 5240–5250.

[32] Priestley, M. B. *Spectral Analysis and Time Series*. Academic Press, London, 1983.

[33] Sakiyama, K., and Taniguchi, M. Discriminant analysis for locally stationary processes. *Journal of Multivariate Analysis 90*, 2 (Aug. 2004), 282–300.

[34] Sanderson, J., Fryzlewicz, P., and Jones, M. Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika 97*, 2 (2010), 435–446.

[35] Santhanam, B., and Maragos, P. Multicomponent am-fm demodulation via periodicity-based algebraic separation and energy-based demodulation. *IEEE Transactions on Communications 48*, 3 (Mar 2000), 473–490.

[36] Shumway, R. Discriminant analysis for time series. In *Classification Pattern Recognition and Reduction of Dimensionality*, P. Krishnaiah and L. Kanal, Eds., vol. 2 of *Handbook of Statistics*. Elsevier, 1982, pp. 1 – 46.

[37] Shumway, R. H. Time-frequency clustering and discriminant analysis. *Statistics & Probability Letters 63*, 3 (July 2003), 307–314.

Appendix

Proof of Proposition 1

We begin by reminding the reader of a result established by [31] which is relevant to this proof, namely that the variance of the LWS estimate, $\widehat{S}_{jk}^{(p,q)}$, can be expressed as,

$$\mathrm{Var}\left\{\widehat{S}_{jk}^{(p,q)}\right\} = \mathcal{O}(M_T^{-1}) + \mathcal{O}(T^{-1}).$$

Here M_T is the smoothing bandwidth used to calculate $\widehat{S}_{jk}^{(p,q)}$. For this estimate to be both asymptotically unbiased and consistent [31] make the assumptions that $M_T \to \infty$ and $M_T/T \to 0$ in the limit as $T \to \infty$. Given this we can express M_T in the form $M_T = \mathcal{O}(T^{\alpha})$ for some $\alpha \in (0,1)$. The variance of $\widehat{S}_{jk}^{(p,q)}$ can then be expressed as a single order term, $\mathrm{var}\left(\widehat{S}_{jk}^{(p,q)}\right) = \mathcal{O}(T^{-\alpha})$.

We now consider the asymptotics of our classification procedure. Let $\widehat{\boldsymbol{\mu}}_k$ be a vector of length N which contains the elements of $\widehat{\boldsymbol{\zeta}}_{j,k;X}$ which will be used to distinguish the different classes. For simplicity we will consider the two class problem however the results are easily generalised to the more general case. We define the divergence criterion to be,

$$\begin{aligned}\Delta(\widehat{\boldsymbol{\mu}}_k) = \frac{1}{2}\Big\{&(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2) \\ &-(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\,\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\Big\}. \end{aligned} \tag{11}$$

This divergence criterion is simply the difference between log-likelihoods under the two classes. We also define the classification decision rule,

$$D(\widehat{\boldsymbol{\mu}}_k) = \begin{cases} 1 & (\text{estimate } C(k)=1) \text{ if } \quad \Delta(\widehat{\boldsymbol{\mu}}_k) > 0 \\ 2 & (\text{estimate } C(k)=2) \text{ if } \quad \Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0 \end{cases}.$$

Suppose that the true class membership, $C(k)$, is equal to 1. Here we want to show that the probability of misclassification goes to 0 as $T \to \infty$. That is we want to show, $\Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k)=1) \to 0$, or equivalently, $\Pr(\Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k)=1) \to 0$.

What we will actually show is that for the scaled divergence, $\delta_T(\widehat{\boldsymbol{\mu}}_k) = \Delta(\widehat{\boldsymbol{\mu}}_k)/T^{\alpha}$ for some $\alpha \in (0,1)$, that $\Pr(\delta_T(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k)=1) \to 0$ as $T \to \infty$, and consequently that $\Pr(D(\widehat{\boldsymbol{\mu}}_k) =$

$2|C(k)=1) \to 0$ in the same limit. This results immediately follows if we can establish that as $T \to \infty$ then $\delta_T(\widehat{\boldsymbol{\mu}}_k) \xrightarrow{P} K \geq 0$, which is satisfied by the following two conditions in the limit a $T \to \infty$: **A1:** $E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] \to K$ where $K > 0$ and **A2:** $\text{var}\left(\delta_T(\widehat{\boldsymbol{\mu}}_k)\right) \to 0$,

Expectation of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$

We first consider the expectation of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$,

$$\begin{aligned} E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = &-\frac{1}{2T^\alpha} E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_1^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\right] \\ &+\frac{1}{2T^\alpha} E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right] \\ &+\frac{1}{2T^\alpha} E\left[\log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right]. \end{aligned}$$

We note that the first term follows a chi-squared distribution with N degrees of freedom, the expectation of which is equal to N. We now focus on the second term,

$$\begin{aligned} &E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right] \\ =&E\left[\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}\, \boldsymbol{\Sigma}_2^{-1} \{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}\right], \\ =&E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\right] + 2E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] \\ &\qquad + E\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]. \\ =&E\left[\text{tr}\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\}\right] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ =&E\left[\text{tr}\{\boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\,\}\right] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ =&\text{tr}\{\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \end{aligned} \tag{12}$$

Therefore,

$$\begin{aligned} E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2T^\alpha} \Big\{ &-N + \text{tr}\{\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1\} \\ &+(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \Big\}. \end{aligned}$$

Using the results of Proposition 5 from [31], the variance covariance matrices can be expressed as,

$$\begin{aligned} \boldsymbol{\Sigma}_c =& \frac{\boldsymbol{A}_c}{T^\alpha}, \\ \boldsymbol{\Sigma}_c^{-1} =& \boldsymbol{B}_c T^\alpha. \end{aligned}$$

Here $\boldsymbol{A}_c$ and $\boldsymbol{B}_c$ are constant symmetric positive definite matrices. The expectation can then be written as,

$$E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2T^{\alpha}}\Bigg\{ -N + \mathrm{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\} + T^{\alpha}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\,\boldsymbol{B}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \log\frac{|\boldsymbol{A}_2|}{|\boldsymbol{A}_1|}\Bigg\}.$$

Since $\boldsymbol{B}_2$ is positive definite then we can say $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\,\boldsymbol{B}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = C$ for some constant $C>0$. Also, since $\boldsymbol{A}_1$ and $\boldsymbol{B}_2$ are symmetric and positive definite, [25] showed that $\boldsymbol{A}_1\boldsymbol{B}_2$ will be positive definite and so we can say that $\mathrm{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\} = D$ for some $D>0$. Finally the term $|\boldsymbol{A}_2|/|\boldsymbol{A}_1|$ must be positive as both $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are positive definite. Without loss of generality we assume that this term is equal to $G\in(0,\infty)$. We can therefore express the expectation as,

$$\begin{aligned} E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] =& \frac{1}{2T^{\alpha}}\left\{-N + CT^{\alpha} + D + \log G\right\}, \\ =& \frac{C}{2} + \frac{D-N+\log G}{2T^{\alpha}}. \end{aligned}$$

Clearly as $T\to\infty$ then $E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] \to C/2$ where $C/2$ is a positive constant therefore condition **A1** is satisfied.

Variance of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$

We now consider the variance of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$,

$$\mathrm{var}\left(\delta_T(\widehat{\boldsymbol{\mu}}_k)\right) = \frac{1}{4T^{2\alpha}}\,\mathrm{var}\left((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\,\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\right)$$

$$\begin{aligned} &+ \frac{1}{4T^{2\alpha}}\,\mathrm{var}\left((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right) \\ &+ \frac{1}{2T^{2\alpha}}\mathrm{cov}\big((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\,\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1), \\ &\qquad (\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\big). \end{aligned} \tag{13}$$

The first term is simply the variance of a chi-squared random variable with N degrees of freedom so is equal to $2N$. We therefore focus on the second and third terms. Looking at the second term,

$$\begin{aligned}\text{var}\left((\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right) \\ = E\left[\left\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right\}^2\right] \\ - \left\{E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right]\right\}^2.\end{aligned}$$

The second term in the above equation, $\left\{E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right]\right\}^2$, is simply the square of the term found in equation (12). We therefore focus on the first term, $E\left[\left\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right\}^2\right]$. For simplicity we make the substitution $\boldsymbol{\mu} = (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)$,

$$\begin{aligned}E\left[\left\{\boldsymbol{\mu}\ \ \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}\ \right\}^2\right] &= E\left[\boldsymbol{\mu}\ \ \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}\ \boldsymbol{\mu}\ \ \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}\ \right], \\ =E&\left[\sum_i\sum_j \boldsymbol{\mu}_i\ \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\boldsymbol{\mu}_j \sum_i\sum_j \boldsymbol{\mu}_i\ \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i\,j}\ \boldsymbol{\mu}_j\right], \\ =&\sum_i\sum_j\sum_i\sum_j \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{i\,j}\ E\left[\boldsymbol{\mu}_i\boldsymbol{\mu}_j\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right], \\ =&\sum_i\sum_j\sum_i\sum_j \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{i\,j}\ \left\{E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right]E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right]\right. \\ &\left.+E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_i\right]E\left[\boldsymbol{\mu}_j\boldsymbol{\mu}_j\right] + E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right]E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right]\right\}.\end{aligned}$$

In the final step above we have used Isserlis' theorem, [18], to split the expression into three terms. We label these as D_1, D_2 and D_3. We also note that $E\left[\boldsymbol{\mu}_i\,\boldsymbol{\mu}_j\right] = (\boldsymbol{\Sigma}_1)_{ij} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j$.

Looking at these terms individually we have,

$$
\begin{aligned}
D_1 = & \sum_{ijj'i'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} E\left[\boldsymbol{\mu}_i \boldsymbol{\mu}_j\right] E\left[\boldsymbol{\mu}_{i'} \boldsymbol{\mu}_{j'}\right], \\
= & \sum_{ijj'i'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} \Big\{ (\boldsymbol{\Sigma}_1)_{ij} (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} (\boldsymbol{\Sigma}_1)_{ij} \\
& + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} \Big\}, \\
= & \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\Sigma}_1)_{ij} \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\Sigma}_1)_{ij} \\
& + \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j \\
& \times \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'}, \\
= & \operatorname{tr}\left\{\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1\right\}^2 + 2 \operatorname{tr}\left\{\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1\right\} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
& + \left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2, \\
= & E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2) \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right]^2.
\end{aligned}
$$

Following a similar procedure for D_2,

$$\begin{aligned}
D_2 &= \sum_{iji\,j} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i\,j}\; E\left[\boldsymbol{\mu}_i \boldsymbol{\mu}_i\right] E\left[\boldsymbol{\mu}_j \boldsymbol{\mu}_j\right], \\
&= \sum_{iji\,j} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i\,j} \Big\{ (\boldsymbol{\Sigma}_1)_{ii}\; (\boldsymbol{\Sigma}_1)_{jj} \\
&\quad + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i\; (\boldsymbol{\Sigma}_1)_{jj} \\
&\quad + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j\; (\boldsymbol{\Sigma}_1)_{ii} \\
&\quad + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i\, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j \Big\}, \\
&= \mathrm{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} \\
&\quad + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\quad + \left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2.
\end{aligned}$$

Similarly $D_3 = D_2$. Putting together we obtain,

$$\begin{aligned}
\mathrm{var}\left((\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right) = 2\mathrm{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} \\
+ 4(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
+ 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).
\end{aligned}$$

We now consider the covariance term in equation (13),

$$\begin{aligned}
&\mathrm{cov}\left((\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1), (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right) \\
&= E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right] \\
&\quad - E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\right] E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right].
\end{aligned}$$

Looking at the first term,

$$\begin{aligned}
&E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\, \boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\, \boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right] \\
&= \sum_{iji'j'} (\boldsymbol{\Sigma}_1^{-1})_{ij} (\boldsymbol{\Sigma}_2^{-1})_{i'j'} \\
&\quad \times E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_i (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_j (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_{i'} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_{j'}\right].
\end{aligned}$$

We again split this up into three terms, C_1, C_2 and C_3,

$$\begin{aligned}
C_1 = &\sum_{iji\,j} (\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i\,j}\, E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_i(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_j\right] \\
&\times E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_i\,(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_j\,\right], \\
= &\sum_{iji\,j} (\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i\,j}\,\left\{(\boldsymbol{\Sigma}_1)_{ij}(\boldsymbol{\Sigma}_1)_{i\,j}\right. \\
&\left.+(\boldsymbol{\Sigma}_1)_{ij}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j\,\right\}, \\
= &N\mathrm{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\} + N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).
\end{aligned}$$

Similarly for C_2,

$$\begin{aligned}
C_2 = &\sum_{iji\,j} (\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i\,j}\, E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_i(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_i\,\right] \\
&\times E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)_j(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)_j\,\right], \\
= &\sum_{iji\,j} (\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i\,j}\,(\boldsymbol{\Sigma}_1)_{ii}\,(\boldsymbol{\Sigma}_1)_{jj}\,, \\
= &\mathrm{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\} = \mathrm{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}.
\end{aligned}$$

It can also be shown that $C_2 = C_3$ therefore,

$$\begin{aligned}
&\mathrm{cov}\left((\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\,\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1), (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right) \\
&\qquad = 2\mathrm{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{var}\left(\delta_T(\widehat{\boldsymbol{\mu}}_k)\right) = &\frac{1}{2T^{2\alpha}}\left[N + \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\}\right. \\
&+ 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\left.+(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \text{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}\right], \\
=&\frac{1}{2T^{2\alpha}}\left[N + \text{tr}\left\{\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2\boldsymbol{A}_1\right\}\right. \\
&+ 2T^{\alpha}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\left.+\ T^{\alpha}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2)\,\boldsymbol{B}_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + 4\text{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\}\right].
\end{aligned}$$

Using similar arguments as for the expectation we can say that $\text{tr}\left\{\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2\boldsymbol{A}_1\right\} = F > 0$ and $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = H > 0$ and so we can say,

$$\begin{aligned}
\text{var}\left(\delta_T(\widehat{\boldsymbol{\mu}}_k)\right) =&\frac{1}{2T^{2\alpha}}[N + F + 2T^{\alpha}H + T^{\alpha}C + 4D], \\
=&\frac{2H + C}{2T^{\alpha}} + \frac{N + F + D}{2T^{2\alpha}}.
\end{aligned}$$

Clearly as $T \to \infty$ then $\text{var}\left(\delta_T(\widehat{\boldsymbol{\mu}}_k)\right) \to 0$ and so condition **A2** is satisfied. Since both conditions are now satisfied we have established that $\Pr(\delta_T(\widehat{\boldsymbol{\mu}}_k) \leq 0 | C(k) = 1) \to 0$ as $T \to \infty$.

□

Proof of Proposition 2

We now consider the case of the distance between classes diverging, i.e. $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ for a fixed T. Here we define $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| = \sqrt{\sum_{i=1}^{N}|(\boldsymbol{\mu}_1)_i - (\boldsymbol{\mu}_2)_i|}$. To this end we define a different scaling of the divergence criterion,

$$\begin{aligned}
\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k) =&\frac{\Delta(\widehat{\boldsymbol{\mu}}_k)}{|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2}, \\
=&\frac{1}{2\,|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2}\left\{+(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right. \\
&\left.-(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)\,\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right\}.
\end{aligned}$$

Following a similar logic to the proof of Proposition 1 we aim to show that in the limit $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ the probability of misclassification, $\Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1)$ will tend to 0. This is equivalent to showing that $\Pr(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k) = 1) \to 0$ in the same limit which immediately follows if we satisfy the following conditions in the limit as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$: **B1:** $E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] \to K$ where $K > 0$ and **B2:** $\text{var}\left(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right) \to 0$.

Expectation

We first consider the expected value of $\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)$. Using the results from the proof of Proposition 1 it is readily seen that,

$$E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^2}\Big\{ - N + \text{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\Big\}.$$

We assume that the terms N, $\text{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\}$ and $\log|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|$ do not depend upon the distance between classes and so we will replace these three terms by the constant Q. We now consider the third term in the bracket, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. First we rewrite $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in the form,

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|\,\boldsymbol{v},$$

where $\boldsymbol{v} = [v_1, \ldots, v_N]$ a vector of constants. The third term can then be rewritten,

$$\begin{aligned}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) &= |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2\,\boldsymbol{v}\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{v} \\ &= |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2\,R,\end{aligned}$$

where R is a positive constant due to $\boldsymbol{\Sigma}_2$, and therefore $\boldsymbol{\Sigma}_2^{-1}$ being positive definite. Putting these terms into the expectation we get,

$$E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{Q}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^2} + \frac{R}{2}.$$

Clearly as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ then $E\left[\delta(\widehat{\boldsymbol{\mu}}_k)\right] \to \frac{R}{2}$ which is a positive constant thus condition **B1** is satisfied.

Variance

We now consider the variance of $\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)$. Using the results from Section 6 we can say that,

$$\begin{aligned}\text{var}\left(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right) = \frac{1}{2\left|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right|^4}\left[N + \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\}\right.\\ + 2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\ \left.+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + 2\text{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}\right].\end{aligned}$$

We first look at the terms which do not depend on $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|$ namely N, $\text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\}$ and $\text{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}$. These terms can again be collected into one constant term, U. We have already stated that the fourth term in the brackets can be written as $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2 R$. The third term can also be re written,

$$\begin{aligned}2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)&\\ = |\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2\boldsymbol{v}\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{v}&\\ = |\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2 V.&\end{aligned}$$

Putting these terms into the variance we get,

$$\text{var}\left(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right) = \frac{U}{2\left|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right|^4} + \frac{R+V}{2\left|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right|^2}.$$

Clearly in the limit $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|\to\infty$ then $\text{var}\left(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right)\to 0$ and so the condition **B2** is satisfied. We have therefore satisfied both conditions for this proof and have established that $\Pr(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\leq 0|C(k)=1)\to 0$ as $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|\to\infty$.

□