Classification of High-dimensional Time Series in Spectral Domain Using Explainable Features with Applications to Neuroimaging Data

Sarbojit Roy[1], **Malik S. Sultan**[1], **Tania R. Vallejo**[2], **Leena A. Ibrahim**[2], **and Hernando Ombao**[1]
[1]Computer, Electrical and Mathematical Science and Engineering Division
[2]Biological and Environmental Science and Engineering Division
King Abdullah University of Science and Technology, Saudi Arabia – 23955

Abstract

Interpretable classification of time series poses significant challenges in high dimensions. Traditional feature selection methods in the frequency domain often assume sparsity in spectral matrices (or their inverses) which can be restrictive for real-world applications. We propose a model-based approach for classifying high-dimensional stationary time series by assuming sparsity in the difference between spectra. The estimators for the model parameters are proven to be consistent under general conditions. We also introduce a method to select the most discriminatory frequencies, and it possesses the *sure screening property*. The novelty of our method lies in the interpretability of the parameters hence suitable for neuroscience where understanding differences in brain network connectivity across various states is crucial. The proposed approach is tested using several simulated examples and applied to EEG and calcium imaging datasets to demonstrate its practical relevance.

1 INTRODUCTION

Let $\{X_t = (\mathrm{X}_{1t}, \ldots, \mathrm{X}_{pt})^\top, 1 \leq t \leq T\}$ be a p-dimensional stationary time series of length T and $Y \in \{1, 2\}$ be the class label of associated with $\mathbf{X} = [X_1, \ldots, X_T] \in \mathbb{R}^{p \times T}$. We assume that the series is centered, i.e., $\mathrm{E}[X_t] = 0_p$. The class densities are denoted by f_1 and f_2, i.e., $\mathbf{X}|Y = l \sim f_l$ for $l = 1, 2$. Let $(\mathbf{X}_j, Y_j)$ be independent copies of $(\mathbf{X}, Y)$ such that $\mathbf{X}_j|Y_j = l \overset{iid}{\sim} f_l$ for $1 \leq j \leq n$, $l = 1, 2$. The class priors are given by $0 < \mathrm{P}[Y = l] = \pi_l < 1$ for $l = 1, 2$ with $\pi_1 + \pi_2 = 1$. We aim to predict the class label of a time series $\mathbf{z} = [z_1, \ldots, z_T]$ with $z_t = (\mathrm{z}_{1t}, \ldots, \mathrm{z}_{pt})^\top$, $1 \leq t \leq T$. The posterior probability of class l is defined as $\mathrm{P}[Y_{\mathbf{z}} = l|\mathbf{z}] = \pi_l f_l(\mathbf{z})/(\pi_1 f_1(\mathbf{z}) + \pi_2 f_2(\mathbf{z}))$ for $l = 1, 2$. The optimal classifier, namely, the Bayes classifier assigns $\mathbf{z}$ to class 1 if $\mathrm{P}[Y_{\mathbf{z}} = 1|\mathbf{z}] > \mathrm{P}[Y_{\mathbf{z}} = 2|\mathbf{z}]$, i.e., $\ln \pi_1 f_1(\mathbf{z}) - \ln \pi_2 f_2(\mathbf{z}) > 0$, and to class 2, otherwise.

A stationary time series can be represented as a superposition of Fourier waveforms with random amplitudes.

$$\mathbf{X}_t = \int_{-1/2}^{1/2} \exp(\mathrm{i}2\pi\omega t) d\mathbf{X}(\omega).$$

This representation due to Cramér (1939) is crucial, especially in neuroscience, since changes in the brain network connectivity can be characterized by the oscillatory properties of the random amplitudes, i.e., $d\mathbf{X}(\omega)$. These properties are known to differ across frequencies $\omega \in (-1/2, 1/2)$ (see, e.g., Bastos and Schoffelen, 2016). Thus, classifying in the spectral domain offers key advantages: (a) identifying relevant frequencies aids in understanding brain connectivity, and (b) frequency-specific connectivity provides deeper insights into cross-brain interactions. Furthermore, (c) excluding frequencies that have no discriminatory information leads to better classification accuracy. This approach is applicable beyond neuroscience, including fields like economics and systems biology.

In the spectral domain, the log densities can be approximated by the Whittle log-likelihood (see, e.g., Dahlhaus, 1988; Subba Rao and Yang, 2021):

$$W_l(\mathbf{z}) = \sum_{\omega_k \in \Omega_T} \left[\ln |\mathbf{\Theta}_{lk}| - z_k^* \mathbf{\Theta}_{lk} z_k\right], \tag{1}$$

where $\Omega_T = \{\omega_k = k/T,\ k \in [T']\}$ with $T' = [T/2] - 1$ is the set of fundamental Fourier frequencies, $z_k \equiv z(\omega_k) \in \mathbb{C}^p$ is Discrete Fourier Transform (DFT) of a series $z(t)$ at frequency ω_k. We denote the spectral density matrix (SDM) of class l by $\mathbf{S}_{lk} \equiv \mathbf{S}_l(\omega_k) \in$

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

$\mathbb{C}^{p\times p}$ (see, e.g., Shumway et al., 2000) and the inverse SDM as $\mathbf{\Theta}_{lk} = \mathbf{S}_{lk}^{-1}$, for $l = 1, 2$. Here z^* denotes the conjugate transpose of z. Let $\mathcal{L}(\mathbf{z}) = \ln(\pi_1/\pi_2) + W_1(\mathbf{z}) - W_2(\mathbf{z})$. A classifier based on the Whittle approximation is defined as

$$\delta_{\mathrm{W}}(\mathbf{z}) = \begin{cases} 1 & \text{if } \mathcal{L}(\mathbf{z}) > 0, \\ 2 & \text{otherwise.} \end{cases} \tag{2}$$

It is clear from equations (1)-(2) that if $\mathbf{\Theta}_{1k} = \mathbf{\Theta}_{2k}$ for some $\omega_k \in \Omega_T$, then the frequency ω_k is non-informative in terms of discriminating between the classes. Such frequencies, collectively denoted by

$$\Omega_T^0 = \{\omega_k : \mathbf{\Theta}_{1k} = \mathbf{\Theta}_{2k}, \omega_k \in \Omega_T\}, \tag{3}$$

are referred to as 'noise'. On the other hand, the set $\Omega_T^{\mathbf{D}} = \Omega_T \setminus \Omega_T^0$ contains all necessary information relevant for classification, and is referred to as 'signal'. Estimating Ω_T^0 and $\Omega_T^{\mathbf{D}}$ helps us gain valuable insight into how connectivity pattern varies across different frequencies. Subsequently, removing the irrelevant frequencies will further improve the accuracy of δ_W.

A plug-in-based estimate of the discriminant $\mathcal{L}(\mathbf{z})$ is given by $\hat{\mathcal{L}}(\mathbf{z}) = \ln(n_1/n_2) + \hat{W}_1(\mathbf{z}) - \hat{W}_2(\mathbf{z})$ where $\hat{W}_l(\mathbf{z})$ is obtained by simply plugging the estimates $\hat{\mathbf{\Theta}}_{lk}$ in (1) for $\omega_k \in \Omega_T$. However, in high dimensions, estimation of the inverse SDMs is challenging, since the sample SDMs become ill-conditioned when p is large. One needs additional assumptions on the structure of $\mathbf{\Theta}_{lk}$ for its consistent estimation. Fiecas and von Sachs (2014) proposed a data-driven shrinkage method to obtain consistent estimators of SDMs in high dimensions. Barigozzi and Farnè (2024) assumed a low-rank structure of the SDM. Also see (Ledoit and Wolf, 2020; Sun et al., 2018; Pourahmadi, 2011) for other constraint estimation methods. Another common approach is to assume sparsity on the matrix $\mathbf{\Theta}_{lk}$ for $l = 1, 2$ (see, e.g., Tony Cai and Luo, 2011; Fiecas et al., 2019). Such an assumption can be restrictive and inappropriate in many real-world situations, such as neuroimaging data where each of the regions in the brain are known to have a dense network. Moreover, the existing methods are primarily interested in obtaining a consistent estimator of the inverse SDM in high dimensions. On the other hand, we aim to obtain estimates that minimize the classification error. It is well known that consistent estimation does not always lead to improved classification accuracy (see, e.g., Cai and Liu, 2011; Mai et al., 2012). This motivates us to develop a method of estimating the model parameters from a different point of view. Instead of sparsity in $\mathbf{\Theta}_{1k}$ and $\mathbf{\Theta}_{2k}$ for each $\omega_k \in \Omega_T$, we propose a model that only requires the difference between inverse SDMs, i.e., $\{\mathbf{\Theta}_{2k} - \mathbf{\Theta}_{1k}\}$ to be sparse. In the rest of the article, we let $\pi_1 = \pi_2$. But, the results derived in this paper hold for all $\pi_1 \in (0, 1)$. Observe that the discriminant $\mathcal{L}(\mathbf{z})$ in (2) can be written as

$$\mathcal{L}(\mathbf{z}) = \sum_{\omega_k \in \Omega_T^{\mathbf{D}}} \left[z_k^* \mathbf{D}_k z_k - \ln |\mathbf{D}_k \mathbf{S}_{1k} + \mathbf{I}_p| \right], \tag{4}$$

where $\mathbf{D}_k = \mathbf{\Theta}_{2k} - \mathbf{\Theta}_{1k}$ at frequency $\omega_k \in \Omega_T^{\mathbf{D}}$. It is clear from equation (4) that the $\mathcal{L}(\mathbf{z})$ depends on the unknown parameters only through the matrices $\mathbf{D}_k$ and $\mathbf{S}_{1k}$ for $\omega_k \in \Omega_T^{\mathbf{D}}$. Expressing the discriminant $\mathcal{L}(\mathbf{z})$ this way has two advantages:

(a) *Interpretability*: In neuroimaging data, brain states like 'alert' vs. 'drowsy' differ in specific frequency bands. Our framework identifies these frequencies in $\Omega_T^{\mathbf{D}}$. The (j, l)-th element of $\mathbf{D}_k$ captures the difference in dependence between j-th and l-th covariates at ω_k.

(b) *Efficiency*: We avoid inverting high-dimensional sample SDMs by directly estimating $\mathbf{D}_k$ matrices. Furthermore, this approach allows the sparsity pattern to vary across frequencies, highlighting which covariates contribute to differences at each frequency ω_k.

Feature-based methods for multivariate time series classification (MTSC) often prioritize interpretability but struggle with scalability. Shapelet-based methods, for example, are interpretable but computationally expensive for high-dimensional time series. Karlsson et al. (2016) proposed a scalable MTSC approach using random shapelets within decision tree forests, while other shapelet-based models—such as the Shapelet Transform Classifier (STC, Hills et al., 2014) and the Random Interval Spectral Ensemble (Lines et al., 2016)—primarily target univariate time series. These methods have been extended to multivariate settings via ensemble models like the Hierarchical Vote Collective of Transformation-based Ensemble (HIVE-COTE, Middlehurst et al., 2021), but such adaptations operate independently on each dimension rather than capturing cross-dimensional dependencies. Other ensemble-based methods, such as the Canonical Interval Forest (CIF, Middlehurst et al., 2020), leverage time-series trees but lack interpretability.

Beyond ensemble methods, convolutional kernel-based classifiers like ROCKET (Dempster et al., 2020) and its more efficient variant MiniROCKET (Dempster et al., 2021) achieve state-of-the-art classification accuracy but sacrifice interpretability. Similarly, deep learning models, including InceptionTime (Ismail Fawaz et al., 2020) and Time Series Attentional Prototype Network (TapNet, Zhang et al., 2020), provide strong performance but remain largely opaque in their decision-making process. To address this, interpretable deep learning approaches have been explored. For instance, Cross Spectral Factor Analysis (CSFA) (Gallagher et al., 2017) constructs a latent space using factor mod-

els and employs binary cross-entropy loss to identify discriminative factors in EEG/LFP data. SyncNet (Li et al., 2017) similarly extracts relevant frequency bands for classification through Gaussian Process adaptors and parameterized convolutional filters.

Our Contribution: Despite the advancements, existing methods either lack interpretability, fail to model cross-dimensional dependencies explicitly, or rely on complex neural architectures. In contrast, we propose a statistically principled approach based on Bayes' rule that identifies discriminative features in the original space while directly screening relevant frequencies. This provides an interpretable alternative to high-dimensional spectral analysis without requiring extensive regularization or deep neural architectures. The main advantages of the proposed method are:

- *Interpretable model parameters*: The proposed method leads to interpretable results in the high-dimensional time series classification regime.
- *Consistent estimation*: Proposed estimators are shown to have theoretical consistency in ultrahigh-dimensional settings when the dimension may increase non-polynomially with the sample size.
- *Sparse difference assumption*: We assume sparsity in the difference between inverse spectral density matrices (SDMs) rather than in the SDMs themselves, achieving theoretical consistency under less restrictive conditions.
- *Effective frequency screening*: Our frequency screening method enjoys the *sure screening property* (Fan and Lv, 2008), and ranks frequencies by their importance in classification.
- *Flexible framework*: The method allows the importance of covariates to vary across frequencies, providing deeper insights into real-world problems such as brain network connectivity.

In Section 2, we describe the estimation procedure. The theoretical properties of the proposed estimators are presented in Section 3. We demonstrate the classifier's performance on various simulated and two real data sets in Sections 4 and 5, respectively. The article ends with a discussion on the limitations of the proposed method and potential directions for future research in Section 6. All proofs are detailed in the Supplementary.

2 METHODOLOGY

Notations and definitions: The symbol i denotes the square root of (-1). For an integer $p \geq 1$, $[p]$ denotes the set $\{1, \ldots, p\}$. For a $p \times p$ matrix $\mathbf{M}$, we write $\|\mathbf{M}\|_{\mathrm{F}} = (\sum_{i,j} \mathbf{M}^2(i,j))^{1/2}$ for its Frobenius norm, $\|\mathbf{M}\|_1 = \sum_{i,j} |\mathbf{M}(i,j)|$ and $\|\mathbf{M}\|_\infty = \max_{i,j} |\mathbf{M}(i,j)|$. In addition, we define $\|\mathbf{M}\|_{1,\infty} = \max_i \sum_j |\mathbf{M}_{i,j}|$ to be the $\ell_{1,\infty}$ norm of the matrix $\mathbf{M}$. $|\mathbf{M}|$ denotes elements wise absolute value of matrix $\mathbf{M}$. We use the symbol $\mathbf{0}_p$ to denote both the $p \times 1$ zero vector and the $p \times p$ zero matrix, depending on the context. The $p \times 1$ vector with all entries as 1 is denoted by $\mathbf{1}_p$. The $p \times p$ identity matrix is denoted by $\mathbf{I}_p$. $\mathbb{I}[A]$ denotes the indicator of A. The subscript p is sometimes dropped for brevity when the dimension is clear. For $Z = Z^{\mathcal{R}} + \mathrm{i}Z^{\mathcal{I}} \in \mathbb{C}^p$, and $\mathbf{M} = \mathbf{M}^{\mathcal{R}} + \mathrm{i}\mathbf{M}^{\mathcal{I}} \in \mathbb{C}^{p \times p}$, we define

$$\tilde{Z} = \begin{bmatrix} Z^{\mathcal{R}} \\ Z^{\mathcal{I}} \end{bmatrix} \in \mathbb{R}^{2p}, \text{ and } \widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{M}^{\mathcal{R}} & \mathbf{M}^{\mathcal{I}} \\ -\mathbf{M}^{\mathcal{I}} & \mathbf{M}^{\mathcal{R}} \end{bmatrix} \in \mathbb{R}^{2p \times 2p}.$$

We use $c, c_1, c_2, \ldots, C, C_1, C_2, \ldots,$ to denote the constants that do not depend on n, p, T, and their values may vary from place to place throughout this article. To reduce the notational complexity, we write $z(\omega_k)$, $\mathbf{S}_l(\omega_k)$, $\mathbf{\Theta}_l(\omega_k)$, and $\mathbf{D}(\omega_k)$ as z_k, $\mathbf{S}_{lk}$, $\mathbf{\Theta}_{lk}$, and $\mathbf{D}_k$, respectively, from here onwards. With the notations introduced above, the discriminant $\mathcal{L}(\mathbf{z})$ in equation (4) can be written as

$$\mathcal{L}(\mathbf{z}) = \frac{1}{2} \sum_{\omega_k \in \Omega_T^{\mathrm{D}}} \left[\tilde{z}_k^\top \widetilde{\mathbf{D}}_k \tilde{z}_k - \ln |\widetilde{\mathbf{D}}_k \widetilde{\mathbf{S}}_{1k} + \mathbf{I}| \right]. \quad (5)$$

Now, we discuss the method for estimating the model parameters. In Section 2.1, we formulate the estimation of $\widetilde{\mathbf{D}}_k$ as a problem of minimizing a convex loss function Then, in Section 2.2, we utilize the estimates of $\widetilde{\mathbf{D}}_k$ to develop a screening procedure and obtain $\hat{\Omega}_T^{\mathrm{D}}$.

2.1 Estimation of $\widetilde{\mathbf{D}}_k$

Fix $\omega_k \in \Omega_T$, and recall that $\widetilde{\mathbf{D}}_k = \widetilde{\mathbf{\Theta}}_{2k} - \widetilde{\mathbf{\Theta}}_{1k}$. A common approach to estimate $\widetilde{\mathbf{D}}_k$ is to estimate the matrix $\widetilde{\mathbf{\Theta}}_{lk}$ for $l = 1, 2$, and consider $\hat{\widetilde{\mathbf{D}}}_k$ to be the difference of the estimated matrices. Cai et al. (2011) developed a method to estimate high-dimensional precision matrices based on constrained l_1-minimization (CLIME). Fiecas et al. (2019) and Krampe and Paparoditis (2022) proposed CLIME-type estimators of a precision matrix in the spectral domain and studied its finite sample behavior. However, the consistency of these estimators is achieved under sparsity which may not be a realistic assumption in neurological data. Moreover, good estimates of inverse SDMs do not necessarily translate to better classification (Cai and Liu, 2011; Mai et al., 2012). Motivated by the work of Yuan et al. (2017), we take a direct approach to estimate the matrix $\widetilde{\mathbf{D}}_k$. Instead of assuming sparsity on $\widetilde{\mathbf{\Theta}}_{lk}$ for $l = 1, 2$, we assume that their difference $\widetilde{\mathbf{D}}_k = \widetilde{\mathbf{\Theta}}_{2k} - \widetilde{\mathbf{\Theta}}_{1k}$ is sparse. Observe that $\widetilde{\mathbf{S}}_{1k}\widetilde{\mathbf{D}}_k\widetilde{\mathbf{S}}_{2k} = \widetilde{\mathbf{S}}_{1k} - \widetilde{\mathbf{S}}_{2k}$ and $\widetilde{\mathbf{S}}_{2k}\widetilde{\mathbf{D}}_k\widetilde{\mathbf{S}}_{1k} = \widetilde{\mathbf{S}}_{1k} - \widetilde{\mathbf{S}}_{2k}$,

i.e., $(\widetilde{\mathbf{S}}_1\widetilde{\mathbf{D}}_k\widetilde{\mathbf{S}}_{2k} + \widetilde{\mathbf{S}}_{2k}\widetilde{\mathbf{D}}_k\widetilde{\mathbf{S}}_{1k})/2 = \widetilde{\mathbf{S}}_{1k} - \widetilde{\mathbf{S}}_{2k}$. We consider the *D-trace loss* function $L(\widetilde{\mathbf{D}}_k)$ defined below:

$$tr\left(\frac{1}{4}\sum_{l_1\neq l_2}\widetilde{\mathbf{S}}_{l_1k}\widetilde{\mathbf{D}}_k\widetilde{\mathbf{S}}_{l_2k}\widetilde{\mathbf{D}}_k - \widetilde{\mathbf{D}}_k(\widetilde{\mathbf{S}}_{1k} - \widetilde{\mathbf{S}}_{2k})\right). \quad (6)$$

The above loss is convex with respect to $\widetilde{\mathbf{D}}_k$ and it attains the minimum value zero at $\widetilde{\mathbf{D}}_k = \widetilde{\mathbf{\Theta}}_{2k} - \widetilde{\mathbf{\Theta}}_{1k}$. A sparse estimate of $\widetilde{\mathbf{D}}_k$ is obtained by minimizing the lasso penalized $L(\widetilde{\mathbf{D}}_k)$. First, we estimate the matrices $\widetilde{\mathbf{S}}_{1k}$ and $\widetilde{\mathbf{S}}_{2k}$. For the training sample $\chi_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, we define $\hat{\widetilde{\mathbf{S}}}_{lk}$ to be the sample covariance matrix based on the DFTs at ω_k corresponding to class l, i.e., $\hat{\widetilde{\mathbf{S}}}_{lk} = \sum_j \mathbb{I}[Y_j = l]\mathrm{E}[\tilde{X}_j(\omega_k)\tilde{X}_j^\top(\omega_k)]/n_l$ for $l = 1, 2$ and $k \in [T']$. The matrices $\widetilde{\mathbf{D}}_1, \ldots, \widetilde{\mathbf{D}}_{T'}$ are simultaneously estimated by minimizing the lasso penalized *D-trace loss*.

$$\min_{\widetilde{\mathbf{D}}_1,\ldots,\widetilde{\mathbf{D}}_{T'}}\left\{\sum_{\omega_k\in\Omega_T} L(\widetilde{\mathbf{D}}_k;\chi_n) + \lambda\sum_{\omega_k\in\Omega_T}\|\widetilde{\mathbf{D}}_k\|_1\right\}. \quad (7)$$

Here $L(\widetilde{\mathbf{D}}_k;\chi_n)$ is the loss function for given estimates of $\widetilde{\mathbf{S}}_{1k}$ and $\widetilde{\mathbf{S}}_{2k}$, and $\lambda > 0$ is a tuning parameter. This method is motivated by the sparse quadratic discriminant method by Yuan et al. (2017) where the authors estimated the parameters in the context of differential graph estimation. The minimization problem can be solved using a suitable off-the-shelf convex optimization solver. We used Adaptive Moment Estimation (Kingma and Ba, 2014) popularly known as ADAM as the optimizer to estimate the $\widetilde{\mathbf{D}}_k$-s. The details of implementation are provided in the supplementary material. The tuning parameter λ in (7) is selected by minimizing the Generalized Information Criterion (GIC) proposed by Kim et al. (2012). If Λ_n is a sequence of λ values, then we define

$$\{\hat{\widetilde{\mathbf{D}}}_1, \ldots, \hat{\widetilde{\mathbf{D}}}_{T'}\} = \underset{\lambda\in\Lambda_n}{\arg\min}\frac{1}{n}\sum_{\omega_k\in\Omega_T^{\mathbf{D}}}\Big\{L(\hat{\widetilde{\mathbf{D}}}_k(\lambda);\chi_n) + \log(\log n)\log p^2\|\hat{\widetilde{\mathbf{D}}}_k(\lambda)\|_0\Big\}, \quad (8)$$

where $\hat{\widetilde{\mathbf{D}}}_k(\lambda)$ is the estimate of $\widetilde{\mathbf{D}}_k$ for a given λ. Now we use these estimates to identify the Fourier frequencies having the most discriminatory information.

2.2 Screening of Discriminative Frequencies

Recall the definition of the sets Ω_T^0 and $\Omega_T^{\mathbf{D}}$ given in (3). We propose a data-adaptive method to estimate these sets. Define $d_k = \|\widetilde{\mathbf{D}}_k\|_{\mathrm{F}}$ for $k \in [T']$. It follows from the definition that d_k is exactly zero for all $\omega_k \in \Omega_T^0$ and is strictly positive for all $\omega_k \in \Omega_T^{\mathbf{D}}$. Consequently, we write $\Omega_T^0 = \{\omega_k : d_k = 0,\ \omega_k \in \Omega_T\}$, and $\Omega_T^{\mathbf{D}} = \{\omega_k : d_k > 0,\ \omega_k \in \Omega_T\}$. Thus the frequency screening problem is now reduced to a problem identifying strictly positive d_k-s. We now present a data-driven procedure to screen the d_ks that are significantly large.

Let $d_{(k)}$ denote the k-th minimum among $d_1, \ldots, d_{T'}$. Observe that if we arrange the values $\{d_k : k \in [T']\}$ in increasing order of magnitude, then the smallest T_0 values correspond to the set Ω_T^0 and are all equal to zero. In other words, we have $0 = d_{(1)} = \cdots = d_{(T_0)} < d_{(T_0+1)} \leq \cdots \leq d_{(T')}$. This leads to another equivalent representation of $\Omega_T^{\mathbf{D}}$ given by $\Omega_T^{\mathbf{D}} = \{d_k : d_k \geq d_{(T_0+1)} \text{ for } 1 \leq k \leq (T'-1)\}$. This definition suggests that to estimate $\Omega_T^{\mathbf{D}}$, we only need to estimate T_0, the number of 'noise' frequencies. Another key observation is that the ratio $r_k = d_{(k+1)}/d_{(k)} < \infty$ for all $(T_0+1) \leq k \leq (T'-1)$ whereas $r_{T_0} = \infty$. Let $\hat{d}_k = \|\hat{\widetilde{\mathbf{D}}}_k\|_{\mathrm{F}}$ and consider $\hat{d}_{(1)} < \ldots < \hat{d}_{(T')}$. Since the underlying distributions are absolutely continuous with respect to the Lebesgue measure, the ratios $\hat{r}_k = \hat{d}_{(k+1)}/\hat{d}_{(k)}$ for $k = 1, \ldots, (T'-1)$ are well-defined. Since $r_{T_0} = \infty$, we expect the ratio $\hat{r}_{T_0}$ to take a significantly large value when compared to the entire sequence $\{\hat{r}_k : 1 \leq k \leq (T'-1)\}$. We define

$$\hat{T}_0 = \underset{1\leq k\leq(T'-1)}{\arg\max}\ \hat{r}_k, \quad \hat{T}_{\mathbf{D}} = T' - \hat{T}_0, \text{ and}$$
$$\hat{\Omega}_T^{\mathbf{D}} = \{\omega_k : \hat{d}_k > \hat{d}_{(\hat{T}_0)},\ k = 1, \cdots, T'\}. \quad (9)$$

In practice, one may work with $\hat{T}_0 = \arg\max_{T_{min}\leq k\leq T_{max}}\hat{r}_k$, where T_{min} and T_{max} are user defined constants. In variable screening literature, similar criteria based on ratios of ordered values have been discussed in Ni and Fang (2016) and Roy et al. (2023). Note that the ordering of $\hat{d}_k$-s immediately gives the relative importance of fundamental frequencies in classification.

2.3 Classification Method

Recall the classifier δ_W and the discriminant $\mathcal{L}(\mathbf{z})$ defined in (2) and (5), respectively. Using the estimates of $\widetilde{\mathbf{D}}_k$-s and $\hat{\Omega}_T^{\mathbf{D}}$ in defined in (8) and (9), respectively, we propose the following classification rule:

$$\delta(\mathbf{z}) = \begin{cases} 1 & \text{if } \hat{\mathcal{L}}(\mathbf{z}) > 0, \\ 2 & \text{otherwise, where} \end{cases}$$
$$\hat{\mathcal{L}}(\mathbf{z}) = \frac{1}{2}\sum_{\omega_k\in\hat{\Omega}_T^{\mathbf{D}}}\left[\tilde{z}_k^\top\hat{\widetilde{\mathbf{D}}}_k\tilde{z}_k - \ln|\hat{\widetilde{\mathbf{D}}}_k\hat{\widetilde{\mathbf{S}}}_{1k} + \mathbf{I}|\right]. \quad (10)$$

Note that there does not exist a consistent estimator for the log-determinant term in (10), even when the SDMs are diagonal matrices (see Cai et al., 2015). To deal with this problem in practice, we assign $\mathbf{z}$ to class

1 if $\hat{\mathcal{L}}(\mathbf{z}) > c$ where c is tuned by cross-validation (CV). In the next section, we present the theoretical properties of the estimators $\hat{\widetilde{\mathbf{D}}}_k$, $k \in [T']$, and $\hat{\Omega}_T^{\mathbf{D}}$. Unknown prior probabilities π_1 and π_2 are estimated by $\hat{\pi}_1 = n_1/(n_1+n_2)$, $\hat{\pi}_2 = n_2/(n_1+n_2)$, respectively.

3 THEORETICAL PROPERTIES

Let $\tilde{S}_k = \{(i,j) : \widetilde{\mathbf{D}}_k(i,j) \neq 0\}$ be the support of $\widetilde{\mathbf{D}}_k$, and $\tilde{s}_k$ is the cardinality of $\tilde{S}_k$. Suppose that $\Gamma(\mathbf{A},\mathbf{B}) = (\mathbf{A}\bigotimes\mathbf{B} + \mathbf{B}\bigotimes\mathbf{A})/2$ where $\bigotimes$ denotes the Kronecker product between two $2p\times 2p$ matrices $\mathbf{A}$ and $\mathbf{B}$. For any two subsets P_1 and P_2 of $[2p]\times[2p]$, we denote by $\Gamma_{P_1P_2}(\mathbf{A},\mathbf{B})$ the submatrix of $\Gamma(\mathbf{A},\mathbf{B})$ with rows and columns indexed by P_1 and P_2, i.e., $\Gamma_{P_1P_2}(\mathbf{A},\mathbf{B}) = \frac{1}{2}(A_{j,l}B_{k,m} + A_{k,m}B_{j,l})_{(j,k)\in P_1,(l,m)\in P_2}$. For notational simplicity, we write $\Gamma_k = \Gamma(\widetilde{\mathbf{S}}_{1k},\widetilde{\mathbf{S}}_{2k}) = (\Gamma_k(i,j))$. We always assume $\max_k \max(\|\widetilde{\mathbf{S}}_{1k}\|_\infty, \|\widetilde{\mathbf{S}}_{2k}\|_\infty) \leq M$ for a constant $M > 0$ independent of p and T, and $\max_k \tilde{s}_k < 2p$. We define the following quantities:

$$\alpha_k = 1 - \max_{e\in\tilde{S}_k^c} \|\Gamma_{k,e\tilde{S}_k}(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_1,$$
$$\kappa_{\Gamma_k} = \|(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_{1,\infty}, \text{ for } k\in[T'],$$
$$\tilde{s}_{\max} = \max_k \tilde{s}_k,\ \theta_\eta(n,p) = (\eta\ln 2p + \ln 4)/n,$$
$$\rho_\eta(n,p) = 1/(1 + C_1\theta_\eta^{-\frac{1}{2}}(n,p)), \text{ and}$$
$$\psi_\eta(n,p) = C_2M^2\left(\theta_\eta^{\frac{1}{2}}(n,p) + C_3\theta_\eta(n,p)\right)\times \max_k \tilde{s}_k\kappa_{\Gamma_k}(M^2\max_k \tilde{s}_k\kappa_{\Gamma_k} + 1). \quad (11)$$

Observe that $\tilde{S}_k = \emptyset$ for $\omega_k \in \Omega_T^0$. Therefore, $\Gamma_{k,\tilde{S}_k\tilde{S}_k}$ is essentially an empty matrix with 0 rows and 0 columns for all $\omega_k \in \Omega_T^0$. Similarly, the matrix $\Gamma_{k,e\tilde{S}_k}$ is also empty for all $e \in \tilde{S}_k^c$. Thus, $\|\Gamma_{k,e\tilde{S}_k}(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_1 = \|(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_{1,\infty} = 0$ for all k with $\omega_k \in \Omega_T^0$. Consequently, $\alpha_k = 1$ and $\kappa_{\Gamma_k} = 0$ for all k with $\omega_k \in \Omega_T^0$. Consider the following assumptions:

A1. There exists a constant $\eta_1 > 2$ such that
(a) $\min_{l,k}\max_j \mathbf{S}_{lk}(j,j) > \sqrt{2}M\theta_{\eta_1}(n,p)$,
(b) $\max_k \tilde{s}_k\kappa_{\Gamma_k} < o\left(\theta_{\eta_1}^{-\frac{1}{2}}(n,p)\right)$, and
(c) $\min_k \alpha_k > 4\max\{\rho_{\eta_1}(n,p), \psi_{\eta_1}(n,p)\}$.

We consider p to be increasing with n and T. It follows from the definition of $\theta_\eta(n,p)$ that if there exists a $0 < a < 1$, such that $\log p = n^a$, then $\theta_\eta(n,p) \to 0$ as $n\to\infty$. Therefore, assumption A1.(a) allows the variance of $X(\omega_k)$ to decrease for $l = 1,2$ and $\omega_k \in \Omega_T$, but not at a faster rate than $\theta_{\eta_1}(n,p)$. Similarly, A1.(b) implies that the quantity $\max_k \tilde{s}_k\kappa_{\Gamma_k}$ cannot grow faster than $\theta_{\eta_1}^{-1/2}(n,p)$. Both ρ_η and ψ_η are decreasing sequences in n. Thus, A1.(c) implies that for a large n, $\max_{e\in\tilde{S}_k^c}\|\Gamma_{k,e\tilde{S}_k}(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_1 < 1$ for all k which is the *irrepresentability condition* assumed by Yuan et al. (2017). Roughly speaking, A1.(b) implies that the dependence among elements of $\tilde{S}_k$ cannot grow arbitrarily with p, and A1.(c) implies that the elements of $\tilde{S}_k$ and $\tilde{S}_k^c$ are weakly correlated for all $\omega_k \in \Omega_T^{\mathbf{D}}$.

3.1 Consistent Estimation of $\widetilde{\mathbf{D}}_k$

We now present the theorem on the convergence of the proposed estimator $\hat{\widetilde{\mathbf{D}}}_k$ for $k\in[T']$. Consider the tuning parameter

$$\lambda_{\eta k} = \max[2MG_{1k}/A_k, \{128(\eta\ln p + \ln 4)\}^{\frac{1}{2}}\widetilde{M}_kG_{2k} + MG_{1k}\widetilde{M}_k]\{128(\eta\ln p + \ln 4)\}^{\frac{1}{2}} \quad (12)$$

for some $\eta > 2$ while minimizing the objective function in (7). Here G_{1k}, G_{2k}, A_k and $\widetilde{M}_k$ are constants depending on the quantities in (15). Two other constants $\bar{\sigma}_k$ and m_k that depend on $\alpha_k, \kappa_{\Gamma_k}$ and $\tilde{s}_k$ will be used in the next result. Due to space constraints, we omit the definitions of these quantities from the main text and refer the reader to page 1 of the Supplementary. Now we present the first main result of the article.

Theorem 1. *If the assumption A1 is satisfied, then* $\max_{\omega_k}\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty$ *is bounded by* $\max_{\omega_k} m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p)$ *with probability at least* $1 - \frac{T}{(2p)^{\eta_1-2}}$.

Theorem 1 shows the rates of convergence of the estimated $\widetilde{\mathbf{D}}_k$ for all ω_k. It also characterizes the relation between the length of the series T and dimension p. Clearly, under the stated conditions, the consistency holds if $T = o(p^{\eta_1-2})$.

3.2 Consistent Screening of Frequencies

In this section, we present the second main result of this article. Consider the following assumption:

A2. There exists $\eta_2 \geq \eta$ such that

$$\min_{\omega_k\in\Omega_T^{\mathbf{D}}}\left\{\min_{j,l\in\tilde{S}_k}|\widetilde{\mathbf{D}}_k(j,l)| - 2m_k\theta_{\eta_2}^{\frac{1}{2}}(n,p)\right\} > 0.$$

For every relevant Fourier frequency, assumption A2 dictates the required minimum difference between elements of the inverse SDMs that is sufficient to consistently retain that frequency. It readily follows from assumption A2 that $d_k > 2m_k\sqrt{\tilde{s}_k\theta_{\eta_2}(n,p)}$ for all $\omega_k \in \Omega_T^{\mathbf{D}}$, i.e., $d_{(T_0+1)} > 2m_k\sqrt{\tilde{s}_k\theta_{\eta_2}(n,p)}$. Define $q_\eta(T,n,p) = \max_k m_k\sqrt{\tilde{s}_k\theta_\eta(n,p)}$. Next, we assume

A3. $\max_{(T_0+1)\leq k\leq T'-1}\frac{d_{(k+1)}}{d_{(k)}} < O(d_{(T_0+1)}/q_{\eta_2}(T,n,p))$.

Assumption A3 implies that the differences between inverse SDMs cannot increase arbitrarily.

Theorem 2. *If the assumptions A1-A3 are satisfied with $\eta_1, \eta_2 > 2$, then*

$$\mathrm{P}\left[\Omega_T^{\mathbf{D}} \subseteq \hat{\Omega}_T^{\mathbf{D}}\right] > 1 - \frac{T}{p^{\eta_3 - 2}} \text{ for all } \eta_3 > \max\{\eta_1, \eta_2\}.$$

Theorem 2 shows that the estimated set of frequencies contains the true set with high probability. Therefore, if $T = o(p^{\eta_3 - 2})$, then the proposed screening method possesses *sure screening property* (Fan and Lv, 2008).

4 SIMULATION STUDY

We compare the proposed classifier δ with the linear discriminant classifier (LDA), the quadratic discriminant classifier (QDA), the 1-nearest neighbor with dynamic time warping (1NN-DTW, Leodolter et al., 2021), MiniROCKET, and SyncNet. LDA and QDA (using the SVD solver) are implemented via `scikit-learn` (Pedregosa et al., 2011). Among non-spectral MTSC methods, 1NN-DTW is known for strong classification accuracy (Ruiz et al., 2021). Based on these findings, we use 1NN-DTW as a representative baseline and omit other non-spectral methods for conciseness. It is implemented via `R`-package `IncDTW` (Leodolter et al., 2021) with Euclidean distance. MiniROCKET, identified as the new MTSC benchmark (Ruiz et al., 2021), is implemented with 10,000 kernels via `sktime` (Löning et al., 2019), followed by ridge classification. SyncNet is implemented with default arguments using the publicly available `Python` code from Li et al. (2017). Further implementation details are available in the Supplementary. All experiments were run on an Intel Core i5 CPU. Code is available at `https://github.com/MALIKSHAHIDSULTAN7/High_Dimensional_Spectral_Classification.git`.

For all simulated examples we keep the sample sizes $n_1 = n_2 = 200$ and the length of the series $T = 250$ (unless specified otherwise). Let $\tilde{X}(\omega_k) \mid Y = l \sim N_{2p}(\mathbf{0}, \widetilde{\mathbf{S}}_{lk})$ for $l = 1, 2$ and $\omega_k = 0, 1/T, \ldots, T'/T$, and X_t be the inverse Fourier transformation based on the DFTs $X(\omega_k) = \tilde{X}_{1:p}(\omega_k) + \mathrm{i}\tilde{X}_{p+(1:p)}(\omega_k)$. The relevant frequencies are $\Omega_T^{\mathbf{D}} = \{1/T, \ldots, 24/T\}$. Half of the observations are considered to constitute the training set, and the remaining half are considered as the test set. We run the experiment for 25 iterations. In this study, we consider $\mathbf{D}_k$ to be sparse on $\Omega_T^{\mathbf{D}}$.

Example 1. *We consider $p = 32$. The real and imaginary parts of the inverse SDM in class 1 are*

$$\boldsymbol{\Theta}_{1k}^{\mathcal{R}}(i,j) = -0.41^{|i-j|}, \text{ and}$$
$$\boldsymbol{\Theta}_{1k}^{\mathcal{I}}(i,j) = -0.41^{|i-j|}\{-1\}^{\mathbb{I}[j>i]} - \mathbb{I}[j=i],$$

respectively, for all $\omega_k \in \Omega_T$. Consider, $\mathbf{A}_k(i,j) = \boldsymbol{\Theta}_{1k}(i,j)\{-1\}^{\mathbb{I}[|j-i|=1]}$, and let $\lambda_{(s)}(\mathbf{A}_k)$ be the s-th smallest eigen value of the matrix $\mathbf{A}_k$. We define $\rho_k = |\lambda_{(1)}(\mathbf{A}_k)| + \left\{\frac{1}{p}\sum_{s=1}^{p}\left(\lambda_{(s)}(\mathbf{A}_k) - \frac{1}{p}\sum_{s=1}^{p}\lambda_{(s)}(\mathbf{A}_k)\right)^2\right\}^{\frac{1}{2}}$, and $\boldsymbol{\Theta}_{2k}(i,j) = \begin{cases} \boldsymbol{\Theta}_{1k}(i,j) & \text{if } \omega_k \in \Omega_T^0, \\ \mathbf{A}_k(i,j) + \rho_k \mathbb{I}[j=i] & \text{if } \omega_k \in \Omega_T^{\mathbf{D}}. \end{cases}$

Example 2. *We increase the dimension p to 128. The parameters remain the same as **Example** 1.*

Example 3. *We set $p = 200$ and $T = 100$. The parameters remain the same as **Example** 1.*

In the next examples, we add the white noise W_t to X_t (as defined in **Example** 1) and consider the classification of $U_t = X_t + W_t$. Define $W_t = (\mathrm{W}_{1t}, \ldots, \mathrm{W}_{pt})^\top$, and consider $\mathrm{W}_{jt} \overset{iid}{\sim} f_{\mathrm{W}}$ for $j = 1, \ldots, p$.

Example 4. $f_{\mathrm{W}} = t_5$.

Example 5. $f_{\mathrm{W}} = C(0,1)$.

In both **Examples** 4 and 5, the distribution of U_t is leptokurtic (heavier tails). Moreover, in Example 5, the moments of U_t do not exist. We add a linear trend with jitter to X_t in the next example.

Example 6. $U_t = X_t + 0.2t * \mathbf{1} + W_t$, $f_{\mathrm{W}} = N(0, 0.1)$.

In the simulation study and the real data analysis, we also consider the smoothed periodogram estimator of SDM based on $\mathbf{X}_j$, $j = 1, \ldots, n$ (see, e.g., Shumway et al., 2000), and estimate $\widetilde{\mathbf{D}}_k$. Let $\hat{\mathbf{S}}_k^j$ be the smoothed periodogram estimator with Fejer kernel (Bump et al., 2002) as the weight function. Throughout the simulation study, the kernel parameters are fixed at $m = 100$ and $r = 5$. We define $\underline{\hat{\mathbf{S}}}_{lk} = \sum_j \mathbb{I}[Y_j = l]\hat{\mathbf{S}}_k^j / n_l$ and $\underline{\widetilde{\mathbf{S}}}_{lk} = \widetilde{\underline{\hat{\mathbf{S}}}}_{lk}$ for $l = 1, 2$ and $k \in [T']$. We observe that in practice, the classifier δ performs better if $\underline{\hat{\mathbf{S}}}_{lk}$ is used instead of $\hat{\mathbf{S}}_{lk}$ (defined in 2.1) to estimate $\widetilde{\mathbf{D}}_k$, and we report the results for $\underline{\hat{\mathbf{S}}}_{lk}$ in this section.

4.1 Comparison of Classification Error

Table 1 shows that the linear classifier, LDA, fails across all seven examples due to the absence of difference between population locations. Although the underlying distributions differ in covariances, QDA struggles due to the curse of dimensionality and performs poorly. 1NN-DTW is effective in low dimensions but suffers as a consequence to the distance concentration in high dimensions (see, e.g., Aggarwal et al., 2001). In this study, both Euclidean (l_2) and Manhattan (l_1) distances are used for DTW, but neither performs well. We report the error rates of 1NNDTW-l_2 in Table 1.

MiniROCKET's performance steadily improves in Examples 1, 2, and 3 as p increases, likely because a larger

p introduces more discriminatory covariates. With a vast number of convolutional kernels at its disposal, MiniROCKET effectively exploits these features to learn the separating hyperplane. In contrast, SyncNet's performance across these examples is inconsistent, suggesting its vulnerability in high-dimensional settings and reliance on careful hyperparameter tuning. In Example 4 and 5, where the distributions are leptokurtic, both methods experience a decline in accuracy as the tails of the distributions get heavier. However, SyncNet still achieves the best performance in these cases. In Example 6, both methods perform similarly, with SyncNet holding a slight advantage over MiniROCKET.

Unlike SyncNet, the proposed classifier δ requires tuning only a single parameter, making it more straightforward to train. It outperforms all competing methods in Examples 1 and 6 and ranks second in moderate dimensional cases. This demonstrates a promising and consistent performance of δ. While the accuracy gains over state-of-the-art methods in Examples 2-5 are incremental, the true strength of δ lies in its interpretability—an advantage absent in baseline approaches. This balance between competitive accuracy and interpretability addresses a key gap in the existing literature. Additionally, its robust performance in Examples 4-5 (where U_t is non-Gaussian) and in Example 6 (where the data is non-stationary) further underscores its generalizability.

Table 1: Comparison of estimated misclassification probability of the proposed classifiers with traditional and state-of-the-art methods (standard errors in italics)

Example (p, T)	LDA	QDA	1NN DTW	Mini RCKT	Sync Net	δ
1	0.491	0.495	0.500	0.266	0.129	**0.037**
$(32, 250)$	*0.011*	*0.008*	*0.000*	*0.011*	*0.030*	*0.004*
2	0.486	0.505	0.500	**0.006**	0.256	0.068
$(128, 250)$	*0.011*	*0.007*	*0.000*	*0.001*	*0.042*	*0.006*
3	0.503	0.501	0.500	**0.000**	0.161	0.230
$(200, 100)$	*0.007*	*0.009*	*0.000*	*0.000*	*0.026*	*0.018*
4	0.494	0.501	0.500	0.265	**0.019**	0.126
$(32, 250)$	*0.008*	*0.006*	*0.000*	*0.009*	*0.004*	*0.008*
5	0.508	0.504	0.491	0.489	**0.417**	0.480
$(32, 250)$	*0.011*	*0.005*	*0.005*	*0.009*	*0.014*	*0.011*
6	0.509	0.494	0.500	0.044	0.017	**0.009**
$(32, 250)$	*0.010*	*0.008*	*0.000*	*0.004*	*0.005*	*0.003*

4.2 Performance of Frequency Screening

In this section, we study the performance of the proposed method in recovering the support of $\mathbf{D}_k$ and identifying discriminative fundamental frequencies. The true positive, true negative, and true discovery rates in the recovery of the set $\Omega_T^{\mathbf{D}}$ and the support $\cup_{\omega_k \in \Omega_T^{\mathbf{D}}} \tilde{S}_k$ are given by $(\mathrm{TP}_\omega, \mathrm{TN}_\omega, \mathrm{TD}_\omega)$ and $(\mathrm{TP}_{\mathbf{D}}, \mathrm{TN}_{\mathbf{D}}, \mathrm{TD}_{\mathbf{D}})$ in Equation (4.2), respectively. Table 2 shows that the screening method has retained the relevant Fundamental Fourier frequencies with at least 91% accuracy in Examples 1-4 and 6. The performance deteriorated in Example 5 since the distributions are heavy-tailed and the moments do not exist. We observe that less than 9% of the discriminatory covariates have been retained in the recovered support – the $\mathrm{TD}_{\mathbf{D}}$ index in all six examples needs further improvement. A reason for this could be a low signal-to-noise ratio in the examples. In contrast, $\mathrm{TP}_{\mathbf{D}}$ shows that except in Example 5, a high percent among the recovered covariates is indeed truly discriminatory. As a result, the proposed classifier shows desired performance in Examples 1-4 and 6.

$$\mathrm{TP}_{\mathbf{D}} = \sum_{k,i,j} |\tilde{S}_k| \mathbb{I}[\hat{\widetilde{\mathbf{D}}}_k(i,j) \neq 0, \widetilde{\mathbf{D}}_k(i,j) \neq 0] / \sum_k |\tilde{S}_k|,$$

$$\mathrm{TN}_{\mathbf{D}} = \sum_{k,i,j} |\tilde{S}_k^c| \mathbb{I}[\hat{\widetilde{\mathbf{D}}}_k(i,j) = 0, \widetilde{\mathbf{D}}_k(i,j) = 0] / \sum_k |\tilde{S}_k^c|,$$

$$\mathrm{TD}_{\mathbf{D}} = \sum_{k,i,j} |\hat{\tilde{S}}_k| \mathbb{I}[\hat{\widetilde{\mathbf{D}}}_k(i,j) \neq 0, \widetilde{\mathbf{D}}_k(i,j) \neq 0] / \sum_k |\hat{\tilde{S}}_k|,$$

$$\mathrm{TP}_\omega = |\hat{\Omega}_T^{\mathbf{D}} \cap \Omega_T^{\mathbf{D}}| / |\Omega_T^{\mathbf{D}}|, \quad \mathrm{TN}_\omega = |\hat{\Omega}_T^0 \cap \Omega_T^0| / |\Omega_T^0|,$$
$$\mathrm{TD}_\omega = |\hat{\Omega}_T^{\mathbf{D}} \cap \Omega_T^{\mathbf{D}}| / |\hat{\Omega}_T^{\mathbf{D}}|. \tag{13}$$

Table 2: Performance of the frequency screening method and recovery of support of $\mathbf{D}_k$-s in simulated examples (standard errors in italics).

Ex	Frequency screening			Support recovery		
	TP_ω	TN_ω	TD_ω	$\mathrm{TP}_{\mathbf{D}}$	$\mathrm{TN}_{\mathbf{D}}$	$\mathrm{TD}_{\mathbf{D}}$
1	0.917	1.000	1.000	0.911	0.402	0.089
	0.000	*0.000*	*0.000*	*0.002*	*0.001*	*0.000*
2	0.950	0.988	0.905	0.911	0.407	0.046
	0.000	*0.000*	*0.000*	*0.002*	*0.000*	*0.000*
3	1.000	0.844	0.731	0.901	0.452	0.049
	0.000	*0.156*	*0.269*	*0.001*	*0.121*	*0.047*
4	1.000	0.903	0.522	0.889	0.506	0.052
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.001*
5	0.687	0.646	0.235	0.713	0.338	0.026
	0.038	*0.086*	*0.069*	*0.079*	*0.025*	*0.010*
6	1.000	0.903	0.522	0.801	0.410	0.076
	0.000	*0.000*	*0.000*	*0.008*	*0.010*	*0.046*

4.3 Comparison of Computational Time

We compare the computational cost of our proposed classifier δ with other methods from the simulation study. Training and prediction times (in seconds) are recorded for 30 replicates using a batch size of 64 on a Google Colab Notebook (CPU). As expected, LDA and QDA are the fastest due to their simple parametric models. Both are implemented using the default

settings in scikit-learn. In contrast, 1NN-DTW is the slowest, since it computes an $n \times n$ distance matrix, making it increasingly expensive as p and T grow. Table 3 shows that MiniROCKET is the fastest among the state-of-the-art classifiers, benefiting from almost deterministic convolutional kernels to efficiently learn the separating hyperplane. Our proposed classifier δ performs comparably to benchmark methods with its efficiency closely matching SyncNet. Notably, δ slightly outperforms SyncNet when $p \times T$ is moderately large, but the trend reverses as $p \times T$ further increases.

Table 3: Average running time (based on 30 iterations) of competing methods are reported in seconds.

Method	$T = 250$		$T = 500$	
	$p = 60$	$p = 120$	$p = 60$	$p = 120$
LDA	0.44	0.79	0.56	1.44
QDA	0.24	0.32	0.27	0.72
1NN-DTW	21.62	46.12	79.97	177.86
MiniROCKET	0.67	1.46	0.86	2.22
SyncNet	0.85	1.56	3.06	5.87
δ	0.75	3.04	1.50	6.13

5 REAL DATA ANALYSIS

Alert-Drowsy (A-D) Data : Cao et al. (2019) acquired the data using a virtual reality driving simulator. This dataset includes EEG signals from 27 subjects recorded at 32 channels at 500Hz with a total of 1872 epochs of 3.2 seconds. Baseline alertness was defined as the 5th percentile of local reaction time (RT) after sudden events. Trials with RT lower than 1.5 times the baseline were labeled 'alert/normal', and those higher than 2.5 times were labeled 'drowsy.' Moderate RT trials were excluded. Our study considers 10 subjects for whom the pre-processed data is available. During pre-processing, by two reference channels are removed, and the signals are down-sampled to 128Hz.

Mice Calcium Recording (MCR) Data : We perform an experiment on adult and developing mice to investigate the role of neurogliaform cells in shaping circuits in the auditory cortex. In vivo-2, photon calcium imaging is used to monitor the activity of neurons in Layer 1 and Layer 2–3 respectively (see `http://hdl.handle.net/10754/695990` for the experiment details). The sampling rate is fixed at 15 Hz while collecting the data. We conduct 9 trials of in-sequence repetitive auditory stimulus followed by a trial of oddball unpredictable tone. Together the 10 trials constitute 1 block of experiment and each block is repeated 10 times. We analyze the data of Layer 1 neurons from adult mice using the proposed method.

For both AD and MCR data, we randomly select 50% of the trials for the training set, maintaining equal class proportions. The remaining data is split equally into validation and test sets. We run the experiment with 25 replicates. In the A-D data, we observe no significant difference in the mean of the two classes. This explains the poor performance of the linear classifier LDA (see Table 4). We observe a sparse structure in the difference between 30×30 auto-covariance matrices of class 1 and 2 at lag 0 (see Figure 1(a) in Supplementary). Due to this difference in auto-covariances, QDA performs better than LDA (see Table 4). 1NNDTW performs poorly in this data set with moderate dimension, where the scale difference dominates the location difference. This is justified since distance-based classifiers fail in high dimensions if the location difference is not larger than the difference in scales (Hall et al., 2005). Though MiniROCKET leads to the best performance in terms of classification accuracy, it has no interpretable model. SyncNet too performs poorly on this dataset. A reason for this could be the requirement of rigorous tuning of the hyper-parameters, e.g., number of filters, learning rate, pooling size, etc. In contrast to SyncNet, the proposed method has one tunable parameter to control the sparsity and it is easily tuned by CV.

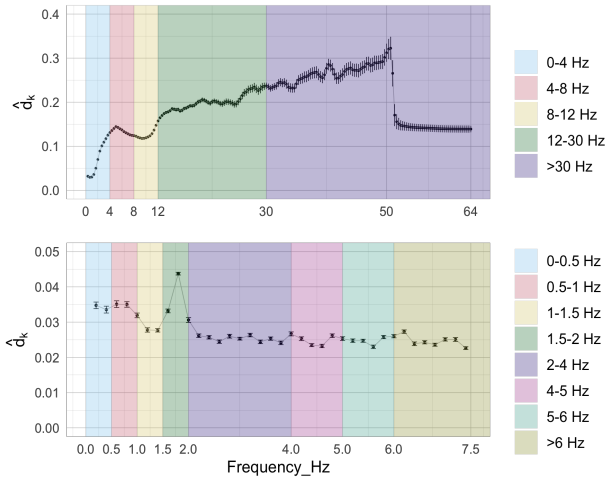


Figure 1: Average $\hat{d}_k$ with standard error based on 25 replicates are plotted against Fourier frequencies ω_k for the A-D (top) and MCR data (bottom).

In Figure 2 we plot the average difference $\sum_{\omega_k \in \Omega} |\hat{\mathbf{D}}(\omega_k)|/|\Omega|$ for specific frequency bands $\Omega \subset \Omega_T$. It reveals some crucial findings. First of all, we see that the differences are sparse. Second, the pattern of sparsity varies across frequency bands, and third, Figure 1 (top row) clearly shows that the overall magnitude of the difference varies with frequencies. For example, the most prominent difference is observed in the pattern of interaction between the frontal lobe channels for the Gamma band which is associated with an elevated

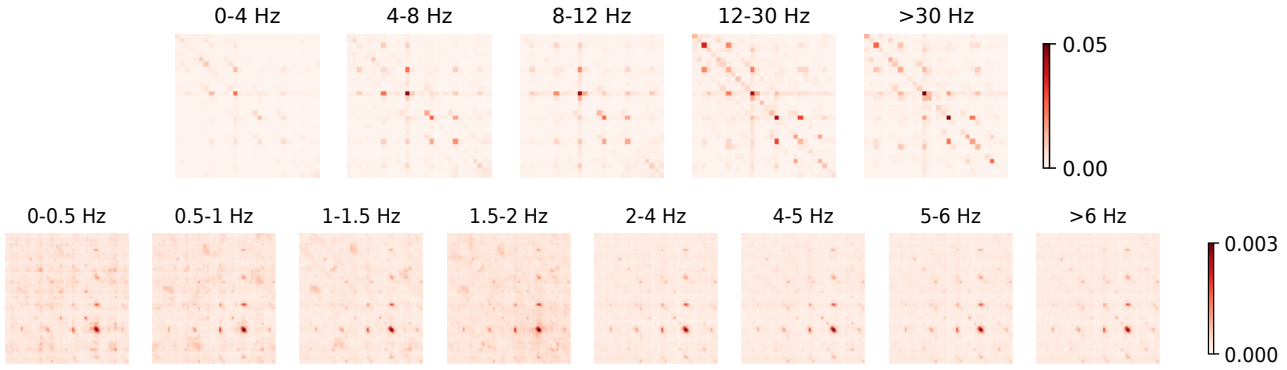


Figure 2: The heatmaps in the top row represent an average difference in pairwise interactions for frequency bands $\delta, \theta, \alpha, \beta$, and γ for the Alert-Drowsy data. The i, j-th cell of a heatmap is $\sum_{\omega_k \in \Omega} |\hat{\mathbf{D}}_k(i,j)|$ for frequency band Ω. The heatmaps in the bottom row represent the same for the MCR data with different frequency bands.

state of vigilance and cognitive activity. The proposed method, unlike its competitors, successfully extracts this information and leverages it for classification.

In the MCR data, we observe a difference in average calcium density in neurons from Layer 1 when the stimulus is in sequence vs when it is oddball. This justifies the performance of LDA. For the same reason, an improved error rate of 1NNDTW is also observed. Figure 1(b) in Supplementary shows a dense structure in the difference between 70×70 auto-covariance matrices of in-sequence and oddball at lag 0. QDA leverages this discriminatory information in the interaction of neurons and performs better than LDA and 1NN. We observe that at different frequencies, some neurons behave differently to in-seq and out-of-sequence auditory stimuli. Both MiniROCKET and SyncNet yield near-perfect classification with the former having a slight edge between the two. Our method performs well on the task of classifying in-sequence and oddball trials and identifies a sparse set of neurons that carry the most discriminatory information (see Figure 2, bottom row). Figure 1 reveals that all frequency bands except (1.5-2] Hz contribute equally to the classification. The role of Layer 1 neurons in information processing remains largely unknown, but the responses observed in this classification may be an indicator of the role these neurons play in the processing of expected and unexpected auditory signals.

6 CONCLUDING REMARKS

In this article, we present a statistical method to classify high-dimensional stationary time series in the spectral domain, with an emphasis on parameter interpretability. The classifier shows promising performance in a wide range of simulated settings, demonstrating the generalizability of the method. We establish the consistency

Table 4: Comparison of estimated misclassification probability of the proposed classifier with traditional and state-of-the-art methods (standard errors in italics)

Data	LDA	QDA	1NN DTW	Mini RCKT	Sync Net	δ
A-D	0.501	0.327	0.473	0.129	0.453	0.259
	0.024	*0.020*	*0.001*	*0.021*	*0.030*	*0.020*
MCR	0.219	0.168	0.179	0.023	0.061	0.065
	0.069	*0.052*	*0.008*	*0.026*	*0.041*	*0.059*

of the proposed estimators of the model parameters under fairly general conditions. The consistency of our proposed classifier hinges on the difference between the actual likelihood and its Whittle approximation. Investigating this difference in high dimensions remains an open problem and is beyond the scope of this article.

The proposed framework can be extended to more complex scenarios. A natural extension is to non-stationary time series, where a simple approach could involve assuming blockwise stationarity and estimating parameters within each block. A more challenging direction would be to adapt the method for multivariate locally stationary wavelet processes by screening relevant wavelets and scales—an extension whose theoretical properties remain unexplored.

If sparsity varies across frequencies, incorporating a hierarchical group lasso penalty with the *D-trace loss* function could enhance classification accuracy while preserving interpretability.

Additionally, the assumed sparsity of the difference matrix $\mathbf{D}_k$ may not always hold in practice. A more flexible approach would be to assume a low-rank structure instead. Extending our theoretical results to accommodate this relaxation is an important direction for future research but is beyond the scope of this article.

Acknowledgements

We thank the reviewers for their insightful comments, which have helped us improve the manuscript. We also gratefully acknowledge the financial support provided by King Abdullah University of Science and Technology for this research.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8, pages 420–434. Springer.

Barigozzi, M. and Farnè, M. (2024). An algebraic estimator for large spectral density matrices. Journal of the American Statistical Association, 119(545):498–510.

Bastos, A. M. and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Frontiers in Systems Neuroscience, 9:175.

Bump, D., Diaconis, P., and Keller, J. B. (2002). Unitary correlations and the Fejér kernel. Mathematical Physics, Analysis and Geometry, 5:101–123.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association, 106(496):1566–1577.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494):594–607.

Cai, T. T., Liang, T., and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. Journal of Multivariate Analysis, 137:161–172.

Cao, Z., Chuang, C.-H., King, J.-K., and Lin, C.-T. (2019). Multi-channel eeg recordings during a sustained-attention driving task. Scientific Data, 6(1):19.

Cramér, H. (1939). On the representation of a function by certain fourier integrals. Transactions of the American Mathematical Society, 46(2):191–201.

Dahlhaus, R. (1988). Small sample effects in time series analysis: A new asymptotic theory and a new estimate. The Annals of Statistics, 16(2):808–841.

Dempster, A., Petitjean, F., and Webb, G. I. (2020). Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery, 34(5):1454–1495.

Dempster, A., Schmidt, D. F., and Webb, G. I. (2021). Minirocket: A very fast (almost) deterministic transform for time series classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 248–257.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society Series B: Statistical Methodology, 70(5):849–911.

Fiecas, M., Leng, C., Liu, W., and Yu, Y. (2019). Spectral analysis of high-dimensional time series. Electronic Journal of Statistics, 13(2):4079–4101.

Fiecas, M. and von Sachs, R. (2014). Data-driven shrinkage of the spectral density matrix of a high-dimensional time series. Electronic Journal of Statistics, 8(2):2975 – 3003.

Gallagher, N., Ulrich, K. R., Talbot, A., Dzirasa, K., Carin, L., and Carlson, D. E. (2017). Cross-spectral factor analysis. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(3):427–444.

Hills, J., Lines, J., Baranauskas, E., Mapp, J., and Bagnall, A. (2014). Classification of time series by shapelet transformation. Data Mining and Knowledge Discovery, 28:851–881.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F. (2020). Inceptiontime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery, 34(6):1936–1962.

Karlsson, I., Papapetrou, P., and Boström, H. (2016). Generalized random shapelet forests. Data Mining and Knowledge Discovery, 30:1053–1085.

Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. Journal of Machine Learning Research, 13:1037–1057.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krampe, J. and Paparoditis, E. (2022). Frequency domain statistical inference for high-dimensional time series. arXiv preprint arXiv:2206.02250.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. The Annals of Statistics, 48(5):3043–3065.

Leodolter, M., Plant, C., and Brändle, N. (2021). Incdtw: An r package for incremental calculation of dynamic time warping. Journal of Statistical Software, 99(9):1–23.

Li, Y., Dzirasa, K., Carin, L., Carlson, D. E., et al. (2017). Targeting EEG/LFP synchrony with neural nets. Advances in Neural Information Processing Systems, 30.

Lines, J., Taylor, S., and Bagnall, A. (2016). Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In 2016 IEEE 16th International Conference on Data Mining (ICDM), pages 1041–1046. IEEE.

Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., and Király, F. J. (2019). sktime: A unified interface for machine learning with time series. arXiv preprint arXiv:1909.07872.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika, 99(1):29–42.

Middlehurst, M., Large, J., and Bagnall, A. (2020). The canonical interval forest (cif) classifier for time series classification. In 2020 IEEE International Conference on Big Data (Big Data), pages 188–195. IEEE.

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., and Bagnall, A. (2021). Hive-cote 2.0: a new meta ensemble for time series classification. Machine Learning, 110(11):3211–3243.

Ni, L. and Fang, F. (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. Journal of Nonparametric Statistics, 28(3):515–530.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Pourahmadi, M. (2011). Covariance estimation: the GLM and regularization perspectives. Statistical Science, 26(3):369–387.

Roy, S., Sarkar, S., Dutta, S., and Ghosh, A. K. (2023). On exact feature screening in ultrahigh-dimensional binary classification. Journal of Computational and Graphical Statistics, pages 1–15.

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., and Bagnall, A. (2021). The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 35(2):401–449.

Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). Time Series Analysis and its Applications, volume 3. Springer.

Subba Rao, S. and Yang, J. (2021). Reconciling the gaussian and whittle likelihood with an application to estimation in the frequency domain. The Annals of Statistics, 49(5):2774–2802.

Sun, Y., Li, Y., Kuceyeski, A., and Basu, S. (2018). Large spectral density matrix estimation by thresholding. arXiv preprint arXiv:1812.00532.

Tony Cai, W. L. and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494):594–607.

Wainwright, M. J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Yuan, H., Xi, R., Chen, C., and Deng, M. (2017). Differential network analysis via lasso penalized d-trace loss. Biometrika, 104(4):755–770.

Zhang, X., Gao, Y., Lin, J., and Lu, C.-T. (2020). Tapnet: Multivariate time series classification with attentional prototypical network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 6845–6852.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Material for
"Classification of High-dimensional Time Series in Spectral Domain Using Explainable Features with Applications to Neuroimaging Data"

Sarbojit Roy[1] **Malik S. Sultan**[1] **Taniya R. Vallejo**[2] **Leena A. Ibrahim**[2] **Hernando Ombao**[1]

[1]Computer, Electrical and Mathematical Science and Engineering Division
[2]Biological and Environmental Science and Engineering Division
King Abdullah University of Science and Technology, Saudi Arabia – 23955

PROOFS AND MATHEMATICAL DETAILS

For a stationary time series $X_t,\ t \in [T]$, the discrete Fourier transform coefficients at fundamental frequencies are given by

$$X_k \equiv X(\omega_k) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t e^{-\mathrm{i}2\pi\omega_k t},\ \omega_k \in \{k/T : k = 0, \ldots, \lfloor (T-1)/2 \rfloor\} \text{ (see, e.g., Shumway et al., 2000). (14)}$$

For each p and T, we assume that the zero-mean random vector $X_k|Y = l$ is sub-Gaussian, i.e., there exists constants $0 < a_{1k}, a_{2k} < \infty$ such that

$$\mathrm{E}[e^{\gamma X_{kj} \mathbf{S}_{lk}^{-\frac{1}{2}}(j,j)} | Y = l] \leq e^{\frac{1}{2}\gamma^2 a_{lk}^2} \text{ for all } \gamma \in \mathbb{R} \text{ and } j \in [2p], \text{ with } k \in [T'].$$

We believe this assumption to be fairly general catering to a wide variety of time series. For example, consider X_t to be a zero-mean $p \times 1$ generalized linear process $X_t = \sum_{s \in \mathbb{Z}} \mathbf{A}_s W_s$, where W_s is a $p \times 1$ dimensional white noise process with $p \times p$ covariance $\mathrm{E}[W_t W_t^\top] = \mathbf{\Sigma}_W$ and the $p \times p$ matrices of filter coefficients bA_s satisfy

$$\sum_{s \in \mathbb{Z}} \mathrm{trace}\{\mathbf{A}_s \mathbf{A}_s^\top\} < \infty.$$

Note that stable ARMA processes satisfy the above conditions. For each $p < \infty$, it now follows from Theorem C.7 of Shumway et al. (2000) that the distribution of $\tilde{X}_k|Y = l$ converges to a $2p$-dimensional Gaussian distribution with mean $\mathbf{0}_{2p}$ and covariance $\widetilde{\mathbf{S}}_{lk}/2$ for all $\omega_k \in \Omega_T$ as T increases. Also, X_k and $X_{k'}$ are asymptotically independent for $k \neq k'$.

Let $\tilde{S}_k = \{(i,j) : \widetilde{\mathbf{D}}_k(i,j) \neq 0\}$ be the support of $\widetilde{\mathbf{D}}_k$, and $\tilde{s}_k$ is the cardinality of $\tilde{S}_k$. Suppose that $\Gamma(\mathbf{A}, \mathbf{B}) = (\mathbf{A} \bigotimes \mathbf{B} + \mathbf{B} \bigotimes \mathbf{A})/2$ where $\bigotimes$ denotes the Kronecker product between two $2p \times 2p$ matrices $\mathbf{A}$ and $\mathbf{B}$. For any two subsets P_1 and P_2 of $[2p] \times [2p]$, we denote by $\Gamma_{P_1 P_2}(\mathbf{A}, \mathbf{B})$ the submatrix of $\Gamma(\mathbf{A}, \mathbf{B})$ with rows and columns indexed by P_1 and P_2, i.e., $\Gamma_{P_1 P_2}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}(A_{j,l} B_{k,m} + A_{k,m} B_{j,l})_{(j,k) \in P_1, (l,m) \in P_2}$. For notational simplicity, we write $\Gamma_k = \Gamma(\widetilde{\mathbf{S}}_{1k}, \widetilde{\mathbf{S}}_{2k}) = (\Gamma_k(i,j))$. We always assume $\max_k \max(\|\widetilde{\mathbf{S}}_{1k}\|_\infty, \|\widetilde{\mathbf{S}}_{2k}\|_\infty) \leq M$ for a constant $M > 0$ independent of p and T, and $\max_k \tilde{s}_k < 2p$. We define the following quantities:

$$\alpha_k = 1 - \max_{e \in \tilde{S}_k^c} \|\Gamma_{k,e\tilde{S}_k} (\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_1,\ \kappa_{\Gamma_k} = \|(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_{1,\infty},\ \text{for } k \in [T'],\ \tilde{s}_{\max} = \max_k \tilde{s}_k,$$

$$\hat{\Gamma}_k = \Gamma(\hat{\widetilde{\mathbf{S}}}_{1k}, \hat{\widetilde{\mathbf{S}}}_{2k}),\ \kappa_{\Gamma_k^\top} = \|(\Gamma_{k,\tilde{S}_k\tilde{S}_k}^\top)^{-1}\|_{1,\infty},\ \tilde{\sigma}_{lk}^2 = (1 + 4a_{lk}^2)^2 \max_j \widetilde{\mathbf{S}}_{lk}^2(j,j).$$

$$G_{1k} = \frac{\tilde{\sigma}_{1k}}{\sqrt{n_1}} + \frac{\tilde{\sigma}_{2k}}{\sqrt{n_2}},\ G_{2k} = \frac{\tilde{\sigma}_{1k}}{\sqrt{n_1}} \frac{\tilde{\sigma}_{2k}}{\sqrt{n_2}},$$

$$A_k = M \frac{\alpha_k}{4 - \alpha_k},\ C_G = 3200M^2,\ \widetilde{M}_k = \frac{24M}{\alpha_k} \tilde{s}_k \kappa_{\Gamma_k} (2M^2 \tilde{s}_k \kappa_{\Gamma_k} + 1),$$

$$\bar{\sigma}_k = \min \left\{ -M + \left[M^2 + (6\tilde{s}_k \kappa_{\Gamma_k})^{-1}\right]^{\frac{1}{2}}, -M + \left[M^2 + M\widetilde{M}_k^{-1}\right]^{\frac{1}{2}}, A_k, 16\tilde{\sigma}_{1k}, 16\tilde{\sigma}_{2k} \right\},$$

$$m_k = 240\sqrt{2}M^2 \tilde{s}_k \kappa_{\Gamma_k}^2 (2M + A_k) + 40\sqrt{2}M \left\{\kappa_{\Gamma_k} + 3\tilde{s}_k \kappa_{\Gamma_k}^2 A_k (2M + A_k)\right\} \left[2 + \max\left\{\widetilde{M}_k (2M + A_k), \frac{4M}{A_k}\right\}\right],$$

$$\theta_\eta(n,p) = \frac{1}{n}(\eta \ln 2p + \ln 4),\ q_\eta(T,n,p) = \max_k m_k \tilde{s}_k^{\frac{1}{2}} \theta_\eta^{\frac{1}{2}}(n,p),\ \rho_\eta(n,p) = \frac{1}{1 + C_1 \theta_\eta^{-\frac{1}{2}}(n,p)},$$
$$\psi_\eta(n,p) = C_2 M^2 \left(\theta_\eta^{\frac{1}{2}}(n,p) + C_3 \theta_\eta(n,p)\right) \times \max_k \tilde{s}_k \kappa_{\Gamma_k} (2M^2 \max_k \tilde{s}_k \kappa_{\Gamma_k} + 1),$$
$$\lambda_\eta = 8\sqrt{2} n^{\frac{1}{2}} \theta_\eta^{\frac{1}{2}}(n,p) \max_k \max \left\{ 2M \frac{G_{1k}}{A_k}, 8\sqrt{2} n^{\frac{1}{2}} \theta_\eta^{\frac{1}{2}}(n,p) \widetilde{M}_k G_{2k} + M G_{1k} \widetilde{M}_k \right\}. \quad (15)$$

Observe that $\tilde{S}_k = \emptyset$ for $\omega_k \in \Omega_T^0$. Therefore, $\Gamma_{k,\tilde{S}_k \tilde{S}_k}$ is essentially an empty matrix with 0 rows and 0 columns for all $\omega_k \in \Omega_T^0$. Similarly, the matrix $\Gamma_{k,e\tilde{S}_k}$ is also empty for all $e \in \tilde{S}_k^c$. Thus, $\|\Gamma_{k,e\tilde{S}_k}(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_1 = \|(\Gamma_{k,\tilde{S}_k\tilde{S}_k})^{-1}\|_{1,\infty} = 0$ for all k with $\omega_k \in \Omega_T^0$. It is clear from the above definitions that,

$$\alpha_k = 1,\ \kappa_{\Gamma_k} = \kappa_{\Gamma_k^\top} = 0,\ \widetilde{M}_k = 0,\ \bar{\sigma}_k = \min\{A_k, 16\tilde{\sigma}_{1k}, 16\tilde{\sigma}_{2k}\},\ \text{and } m_k = 0 \text{ for all } \omega_k \in \Omega_T^0 = \Omega_T \backslash \Omega_T^{\mathbf{D}}. \quad (16)$$

Recall the following assumptions that are introduced in the main text:

A1. There exists a constant $\eta_1 > 2$ such that
(a) $\min_{l,k} \max_j \mathbf{S}_{lk}(j,j) > \sqrt{2} M \theta_{\eta_1}(n,p)$,
(b) $\max_k \tilde{s}_k \kappa_{\Gamma_k} < o\left(\theta_{\eta_1}^{-\frac{1}{2}}(n,p)\right)$, and
(c) $\min_k \alpha_k > 4 \max\{\rho_{\eta_1}(n,p), \psi_{\eta_1}(n,p)\}$.

A2. There exists $\eta_2 \geq \eta_1$ such that $\min_{j,l \in \tilde{S}_k} |\widetilde{\mathbf{D}}_k(j,l)| > 2\tilde{s}_k^{-\frac{1}{2}} q_{\eta_2}(T,n,p)$ for all $\omega_k \in \Omega_T^{\mathbf{D}}$.

A3. $\max_{(T_0+1) \leq k \leq T'-1} \frac{d_{(k+1)}}{d_{(k)}} < \frac{d_{(T_0+1)}}{3q_{\eta_2}(T,n,p))}$.

We now present a technical lemma that will be later used to prove Theorem 1.

Lemma 1. *If Assumption A1 is satisfied, then there exists a constant $C > 0$ such that $\min_k \bar{\sigma}_k > MC\theta_{\eta_1}^{\frac{1}{2}}(n,p)$.*

Proof: Recall the definition of $\bar{\sigma}_k$ in (15) and observe that the condition $\min_k \bar{\sigma}_k > MC\theta_\eta^{\frac{1}{2}}(n,p)$ is satisfied if there exists constants $c_1, \ldots, c_5 > 0$ such that

$$\min_k \left(-M + \left[M^2 + (6\tilde{s}_k \kappa_{\Gamma_k})^{-1}\right]^{\frac{1}{2}}\right) > Mc_1 \theta_{\eta_1}^{\frac{1}{2}}(n,p),$$
$$\min_k \left(-M + \left[M^2 + \frac{\alpha_k}{24\tilde{s}_k(2\tilde{s}_k M^2 \kappa_{\Gamma_k}^2 + \kappa_{\Gamma_k})}\right]^{\frac{1}{2}}\right) > Mc_2 \theta_{\eta_1}^{\frac{1}{2}}(n,p),$$
$$\min_k A_k > Mc_3 \theta_\eta^{\frac{1}{2}}(n,p),\ \min_k \tilde{\sigma}_{1k} > Mc_4 \theta_{\eta_1}^{\frac{1}{2}}(n,p),\ \text{and}\ \min_k \tilde{\sigma}_{2k} > Mc_5 \theta_{\eta_1}^{\frac{1}{2}}(n,p). \quad (17)$$

It follows from assumption A1(c) that

$$\min_k \alpha_k > \frac{4}{1 + C_1 \theta_{\eta_1}^{-\frac{1}{2}}(n,p)},$$
$$\text{i.e.,}\ \ 1 - \frac{1}{4}\min_k \alpha_k < \frac{C_1 \theta_{\eta_1}^{-\frac{1}{2}}(n,p)}{1 + C_1 \theta_{\eta_1}^{-\frac{1}{2}}(n,p)} = \frac{1}{1 + C_1^{-1} \theta_{\eta_1}^{\frac{1}{2}}(n,p)},$$
$$\text{i.e.,}\ \ \max_k \frac{4 - \alpha_k}{4} < \frac{1}{1 + C_1^{-1} \theta_{\eta_1}^{\frac{1}{2}}(n,p)},$$
$$\text{i.e.,}\ \ \min_k \frac{4}{4 - \alpha_k} > 1 + C_1^{-1} \theta_{\eta_1}^{\frac{1}{2}}(n,p),$$
$$\text{i.e.,}\ \ \min_k \frac{\alpha_k}{4 - \alpha_k} > C_1^{-1} \theta_{\eta_1}^{\frac{1}{2}}(n,p),$$
$$\text{i.e.,}\ \ \min_k A_k > MC_1^{-1} \theta_{\eta_1}^{\frac{1}{2}}(n,p).$$

It again follows from A1(c) that

$$\begin{aligned}
&\min_k \alpha_k > C_2 M^2 \left(\theta_{\eta_1}^{\frac{1}{2}}(n,p) + C_3\theta_{\eta_1}(n,p)\right) \times \max_k \tilde{s}_k\kappa_{\Gamma_k}(2M^2 \max_k \tilde{s}_k\kappa_{\Gamma_k} + 1),\\
\text{i.e., } &\frac{24\max_k \tilde{s}_k\kappa_{\Gamma_k}(2M^2\max_k \tilde{s}_k\kappa_{\Gamma_k}+1)}{\min_k \alpha_k} < \frac{24}{C_2M^2\left(\theta_{\eta_1}^{\frac{1}{2}}(n,p) + C_3\theta_{\eta_1}(n,p)\right)},\\
\text{i.e., } &\max_k \frac{24\tilde{s}_k\kappa_{\Gamma_k}(2M^2\tilde{s}_k\kappa_{\Gamma_k}+1)}{\alpha_k} < \frac{1}{\frac{C_2}{24}M^2\left(\theta_{\eta_1}^{\frac{1}{2}}(n,p) + \frac{C_3}{24}\theta_{\eta_1}(n,p)\right)},\\
\text{i.e., } &\min_k \frac{\alpha_k}{24\tilde{s}_k(2\tilde{s}_kM^2\kappa_{\Gamma_k}^2 + \kappa_{\Gamma_k})} > \frac{C_2}{24}M^2\theta_{\eta_1}^{\frac{1}{2}}(n,p) + \frac{C_3}{24}M^2\theta_{\eta_1}(n,p),\\
\text{i.e., } &M^2 + \min_k \frac{\alpha_k}{24\tilde{s}_k(2\tilde{s}_kM^2\kappa_{\Gamma_k}^2 + \kappa_{\Gamma_k})} > M^2 + \frac{C_2}{24}M^2\theta_{\eta_1}^{\frac{1}{2}}(n,p) + \frac{C_3}{24}M^2\theta_{\eta_1}(n,p),\\
\text{i.e., } &\min_k \left[M^2 + \frac{\alpha_k}{24\tilde{s}_k(2\tilde{s}_kM^2\kappa_{\Gamma_k}^2 + \kappa_{\Gamma_k})}\right]^{\frac{1}{2}} > M + C_4M\theta_{\eta_1}^{\frac{1}{2}}(n,p), \text{ (for some constant } C_4>0)\\
\text{i.e., } &\min_k \left(-M + \left[M^2 + \frac{\alpha_k}{24\tilde{s}_k(2\tilde{s}_kM^2\kappa_{\Gamma_k}^2 + \kappa_{\Gamma_k})}\right]^{\frac{1}{2}}\right) > C_4M\theta_{\eta_1}^{\frac{1}{2}}(n,p). \qquad (18)
\end{aligned}$$

Observe that $24(2M^2\max_k \tilde{s}_k\kappa_{\Gamma_k}+1) > 1$ and $\min_k \alpha_k < 1$ by definition. Therefore,

$$\begin{aligned}
&\max_k \tilde{s}_k\kappa_{\Gamma_k} < \frac{24\max_k \tilde{s}_k\kappa_{\Gamma_k}(2M^2\max_k \tilde{s}_k\kappa_{\Gamma_k}+1)}{\min_k \alpha_k},\\
\text{i.e., } &\max_k \tilde{s}_k\kappa_{\Gamma_k} < \frac{1}{\frac{C_2}{24}M^2\left(\theta_{\eta_1}^{\frac{1}{2}}(n,p) + \frac{C_3}{24}\theta_{\eta_1}(n,p)\right)}, \text{ (due to assumption A1(c))}\\
\text{i.e., } &\min_k(6\tilde{s}_k\kappa_{\Gamma_k})^{-1} > C_2'M^2\left(\theta_{\eta_1}^{\frac{1}{2}}(n,p) + C_3'\theta_{\eta_1}(n,p)\right),\\
\text{i.e., } &M^2 + \min_k(6\tilde{s}_k\kappa_{\Gamma_k})^{-1} > M^2 + C_2'M^2\theta_{\eta_1}^{\frac{1}{2}}(n,p) + C_3'M^2\theta_{\eta_1}(n,p),\\
\text{i.e., } &\min_k \left[M^2 + (6\tilde{s}_k\kappa_{\Gamma_k})^{-1}\right]^{\frac{1}{2}} > M + C_5M\theta_{\eta_1}^{\frac{1}{2}}(n,p), \text{ (for some constant } C_5>0)\\
\text{i.e., } &\min_k \left(-M + \left[M^2 + (6\tilde{s}_k\kappa_{\Gamma_k})^{-1}\right]^{\frac{1}{2}}\right) > C_5M\theta_{\eta_1}^{\frac{1}{2}}(n,p). \qquad (19)
\end{aligned}$$

The inequality $\min_k \tilde{\sigma}_{lk} > Mc_l\theta_{\eta}^{\frac{1}{2}}(n,p)$ is trivially satisfied for $l=1,2$ due to A1(a). This completes the proof. □

Recall the definition of λ in (15), the regularization parameter in the lasso-penalized *D-trace loss*.

Lemma 2. *If assumption A1 is satisfied for $\eta_1 > 2$, then the regularization parameter $\lambda_{\eta_1} < \infty$.*

Proof: Note that

$$\lambda_{\eta_1} \leq \max\left\{8\sqrt{2}n^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\max_k 2M\frac{G_{1k}}{A_k},\ 8\sqrt{2}n^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\max_k\left(8\sqrt{2}n^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\widetilde{M}_kG_{2k} + MG_{1k}\widetilde{M}_k\right)\right\}$$

Let us consider the first term in the right hand side of the above inequality.

$$
\begin{aligned}
& n^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\max_k 2M\frac{G_{1k}}{A_k} \\
&= 8\sqrt{2}(\eta\ln 2p+\ln 4)^{\frac{1}{2}}\max_k\left(\frac{8}{\alpha_k}-2\right)\left(\frac{\tilde{\sigma}_{1k}}{\sqrt{n_1}}+\frac{\tilde{\sigma}_{2k}}{\sqrt{n_2}}\right) \\
&\le c_1\frac{(\eta\ln 2p+\ln 4)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\left(\frac{4}{\min_k\alpha_k}-1\right)\max_k(\tilde{\sigma}_{1k}+\tilde{\sigma}_{2k}) \\
&= c_1\theta_\eta^{\frac{1}{2}}(n,p)\left(\frac{4}{\min_k\alpha_k}-1\right)\max_k(\tilde{\sigma}_{1k}+\tilde{\sigma}_{2k}) \\
&\le c_2\max_k(\tilde{\sigma}_{1k}+\tilde{\sigma}_{2k})\ \text{[follows from A1 for sufficiently large } n] \\
&\le c_3\max_k\left(\max_j\ \mathbf{S}_{1k}(j,j)+\max_j\ \mathbf{S}_{2k}(j,j)\right) \\
&\le c_3\left(\max_k\max_j\ \mathbf{S}_{1k}(j,j)+\max_k\max_j\ \mathbf{S}_{2k}(j,j)\right)\le 2c_3M. \qquad (20)
\end{aligned}
$$

Now, consider the second term in $\max_k\lambda_k$.

$$
\begin{aligned}
& \max_k\left(\{128(\eta\ln 2p+\ln 4)\}^{\frac{1}{2}}\widetilde{M}_kG_{2k}+MG_{1k}\widetilde{M}_k\right)\{128(\eta\ln 2p+\ln 4)\}^{\frac{1}{2}} \\
=\ & \max_k\left(\{128(\eta\ln 2p+\ln 4)\}\widetilde{M}_kG_{2k}+MG_{1k}\widetilde{M}_k\{128(\eta\ln 2p+\ln 4)\}^{\frac{1}{2}}\right) \\
\le\ & c_1\frac{(\eta\ln 2p+\ln 4)}{n}\max_k\left(\max_j\ \mathbf{S}_{1k}(j,j)\max_j\ \mathbf{S}_{2k}(j,j)\right)\frac{\max_k\tilde{s}_k\kappa_{\Gamma_k}}{\min_k\alpha_k}(\max_k\tilde{s}_k\kappa_{\Gamma_k}+1) \\
& + c_2\frac{(\eta\ln 2p+\ln 4)^{\frac{1}{2}}}{n^{\frac{1}{2}}}\max_k\left(\max_j\ \mathbf{S}_{1k}(j,j)+\max_j\ \mathbf{S}_{2k}(j,j)\right)\frac{\max_k\tilde{s}_k\kappa_{\Gamma_k}}{\min_k\alpha_k}(\max_k\tilde{s}_k\kappa_{\Gamma_k}+1) \\
=\ & \left\{c_1M^2\theta_\eta(n,p)+2c_2M\theta_\eta^{\frac{1}{2}}(n,p)\right\}\frac{\max_k\tilde{s}_k\kappa_{\Gamma_k}}{\min_k\alpha_k}(\max_k\tilde{s}_k\kappa_{\Gamma_k}+1) \\
\le\ & \left\{c_1M^2\theta_\eta(n,p)+2c_2M\theta_\eta^{\frac{1}{2}}(n,p)\right\}\frac{\max_k\tilde{s}_k\kappa_{\Gamma_k}(\max_k\tilde{s}_k\kappa_{\Gamma_k}+1)}{4\max\{\rho_{\eta_1}(n,p),\psi_{\eta_1}(n,p)\}}\ \text{[follows from A1(c)]} \\
\le\ & \left\{c_1M^2\theta_\eta(n,p)+2c_2M\theta_\eta^{\frac{1}{2}}(n,p)\right\}\frac{\max_k\tilde{s}_k\kappa_{\Gamma_k}(\max_k\tilde{s}_k\kappa_{\Gamma_k}+1)}{4\psi_{\eta_1}(n,p)}<\infty. \qquad (21)
\end{aligned}
$$

Combining (20) and (21) we conclude that for sufficiently large n and p, λ_{η_1} is finite. □

Recall that we estimated $\widetilde{\mathbf{S}}_{lk}$ by $\hat{\widetilde{\mathbf{S}}}_{lk}=2\sum_j\mathbb{I}[Y_j=l]\mathrm{E}[\tilde{X}_k\tilde{X}_k^\top]/n_l$ for $l=1,2$ and $k\in[T']$.

Proof of Theorem 1: Recall the definition of $\bar{\sigma}_k$ given in (15). If assumption A1 is satisfied, then it follows from Lemma 1 that

$$
\begin{aligned}
\min_k\bar{\sigma}_k &> MC\theta_{\eta_1}^{\frac{1}{2}}(n,p), \\
\text{i.e., } n &> MC\max_k\bar{\sigma}_k^{-2}(\eta_1\ln 2p+\ln 4),\ \text{for some } \eta_1>2.
\end{aligned}
$$

Following arguments similar to Theorem 1 of Yuan et al. (2017), we have

$$
\begin{aligned}
\mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k-\widetilde{\mathbf{D}}_k\|_\infty-m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p)\le 0\right] &> 1-\frac{2}{(2p)^{\eta_1-2}}\ \text{for all } k=1,\ldots,T', \\
\text{i.e., }\ \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k-\widetilde{\mathbf{D}}_k\|_\infty-m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p)>0\right] &\le \frac{2}{(2p)^{\eta_1-2}}\ \text{for all } k=1,\ldots,T'.
\end{aligned}
$$

Now, $\max_k \left\{\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty - m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right\} > 0 \Rightarrow \|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty \geq m_k\theta_{\eta}^{\frac{1}{2}}(n,p)$ for some k. Therefore,

$$\begin{aligned}
&\mathrm{P}\left[\max_k \left\{\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty - m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right\} > 0\right] \\
&\leq \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty - m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p) > 0 \text{ for some } k\right] \\
&\leq \sum_{k=1}^{T'} \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty - m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p) > 0\right] \leq \frac{T}{(2p)^{\eta_1-2}}.
\end{aligned}$$

This completes the proof. □

Corollary 1. *If assumption A1 is satisfied, then* $\mathrm{P}\left[\max_k \left\{\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}} - m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right\} > 0\right] \leq \frac{T}{(2p)^{\eta_1-2}}$.

Proof: Note that $\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}} \leq \tilde{s}_k^{\frac{1}{2}}\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty$. Therefore,

$$\begin{aligned}
&\mathrm{P}\left[\max_k \left\{\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}} - m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right\} > 0\right] \\
&\leq \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}} - m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p) > 0 \text{ for some } k\right] \\
&\leq \sum_{k=1}^{T'} \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}} - m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p) > 0\right] \\
&\leq \sum_{k=1}^{T'} \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_\infty - m_k\theta_{\eta_1}^{\frac{1}{2}}(n,p) > 0\right] \leq \frac{T}{(2p)^{\eta_1-2}}.
\end{aligned}$$

□

Now we proceed to develop the mathematical arguments for proving the sure screening property of the proposed method. Recall the definition of $q_\eta(T,n,p) = \max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta}^{\frac{1}{2}}(n,p)$ and consider the following result:

Lemma 3. *If assumption A1 is satisfied, then under the conditions and notations of Theorem 1 we have*

1. $\mathrm{P}\left[\max_k |\hat{d}_k - d_k| \leq q_{\eta_1}(T,n,p)\right] > 1 - \frac{T}{(2p)^{\eta_1-2}}$.

2. $\mathrm{P}\left[\max_k |\hat{d}_{(k)} - d_{(k)}| \leq q_{\eta_1}(T,n,p)\right] > 1 - \frac{T}{(2p)^{\eta_1-2}}$.

Proof:

1. Recall the definition of d_k and observe that

$$\begin{aligned}
|\hat{d}_k - d_k| = \left|\|\hat{\widetilde{\mathbf{D}}}_k\|_{\mathrm{F}} - \|\widetilde{\mathbf{D}}_k\|_{\mathrm{F}}\right| \leq \|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}}, \\
\text{i.e.,} \quad \max_k |\hat{d}_k - d_k| \leq \max_k \|\hat{\widetilde{\mathbf{D}}}_k - \widetilde{\mathbf{D}}_k\|_{\mathrm{F}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
1-\frac{T}{(2p)^{\eta_1-2}} &\leq \mathrm{P}\left[\max_k\left\{\|\hat{\widetilde{\mathbf{D}}}_k-\widetilde{\mathbf{D}}_k\|_{\mathrm{F}}-m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right\}\leq 0\right]\\
&= \mathrm{P}\left[\|\hat{\widetilde{\mathbf{D}}}_k-\widetilde{\mathbf{D}}_k\|_{\mathrm{F}}-m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\leq 0 \text{ for all } k\right]\\
&\leq \mathrm{P}\left[|\hat{d}_k-d_k|-m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\leq 0 \text{ for all } k\right]\\
&= \mathrm{P}\left[|\hat{d}_k-d_k|\leq m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p) \text{ for all } k\right]\\
&\leq \mathrm{P}\left[\max_k|\hat{d}_k-d_k|\leq \max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right].
\end{aligned}$$

2. Let $(u_1,\ldots,u_{T'})^\top$ and $(v_1,\ldots,v_{T'})^\top$ denote two vectors in $\mathbb{R}^{T'}$. Then, for any $1\leq k\leq T'$, we have $|u_{(k)}-v_{(k)}|\leq|u_i-v_j|$ for $1\leq i,j\leq T'$, where i and j are such that $u_i\geq u_{(k)}$ and $v_j\leq v_{(k)}$. There are $(T'-k+1)$ and k such choices for i and j, respectively. It follows from the *pigeon-hole principle* that for each $1\leq k\leq T'$ there exists at least one l satisfying $1\leq l\leq T'$ such that $|u_{(k)}-v_{(k)}|\leq|u_l-v_l|$ (see Wainwright (2019)). Therefore,

$$\begin{aligned}
&|u_{(k)}-v_{(k)}|\leq \max_{1\leq l\leq T'}|u_l-v_l| \text{ for all } 1\leq k\leq T'\\
\Rightarrow\ &\max_{1\leq k\leq T'}|u_{(k)}-v_{(k)}|\leq \max_{1\leq l\leq T'}|u_l-v_l|. \qquad (22)
\end{aligned}$$

Using this result for the vectors $(\hat{d}_1,\ldots,\hat{d}_{T'})^\top$ and $(d_1,\ldots,d_{T'})^\top$, we obtain the following:

$$\begin{aligned}
&\max_k|\hat{d}_{(k)}-d_{(k)}|>\max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\Rightarrow\max_k|\hat{d}_k-d_k|>\max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\\
\text{i.e.,}\ \ &P\left[\max_k|\hat{d}_{(k)}-d_{(k)}|>\max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right]\leq P\left[\max_k|\hat{d}_k-d_k|>\max_k m_k\tilde{s}_k^{\frac{1}{2}}\theta_{\eta_1}^{\frac{1}{2}}(n,p)\right].
\end{aligned}$$

The proof follows from part (1). □

Define $r_k=\frac{d_{(k+1)}}{d_{(k)}}$ and $\hat{r}_k=\frac{\hat{d}_{(k+1)}}{\hat{d}_{(k)}}$ for $k=1,\ldots,T'-1$.

Lemma 4. *If assumption A1 is satisfied for $\eta_1>2$, and A2-A3 are satisfied for $\eta_2>\eta_1$, then*

$$\mathrm{P}\left[\hat{r}_{T_0}<\max_{1\leq k\leq T_{\mathbf{D}}-1}\hat{r}_{T_0+k}\right]<\frac{T}{(2p)^{\eta_2-2}}.$$

Proof: Recall that $d_{(k)}=0$ for all $k=1,\ldots,T_0$. Since A2 is satisfied for $\eta_2>\eta_1$, we have

$$\begin{aligned}
&\min_{j,l\in\tilde{S}_k}|\widetilde{\mathbf{D}}_k(j,l)|>2\tilde{s}_k^{-\frac{1}{2}}q_{\eta_2}(T,n,p) \text{ for all } \omega_k\in\Omega_T^{\mathbf{D}}\\
\Rightarrow\ &d_k>2\tilde{s}_k^{\frac{1}{2}}\min_{j,l\in\tilde{S}_k}|\widetilde{\mathbf{D}}_k(j,l)|>2q_{\eta_2}(T,n,p) \text{ for all } \omega_k\in\Omega_T^{\mathbf{D}}\\
\Rightarrow\ &d_{(T_0+1)}>2q_{\eta_2}(T,n,p). \qquad (23)
\end{aligned}$$

Now, note that

$$\begin{aligned}
&\max_k|\hat{d}_{(k)}-d_{(k)}|\leq q_{\eta_1}(T,n,p)\leq q_{\eta_2}(T,n,p)\\
\Rightarrow\ &\hat{d}_{(T_0)}<q_{\eta_2}(T,n,p), \text{ and}\\
&\qquad d_{(T_0+k)}-q_{\eta_2}(T,n,p)<\hat{d}_{(T_0+k)}<d_{(T_0+k)}+q_{\eta_2}(T,n,p) \ \text{ for all } k=1,\ldots,T_{\mathbf{D}}-1,\\
\Rightarrow\ &\frac{\hat{d}_{(T_0+1)}}{\hat{d}_{(T_0)}}>\frac{d_{(T_0+1)}-q_{\eta_2}(T,n,p)}{q_{\eta_2}(T,n,p)}, \text{ and } \frac{\hat{d}_{(T_0+k+1)}}{\hat{d}_{(T_0+k)}}<\frac{d_{(T_0+k+1)}+q_{\eta_2}(T,n,p)}{d_{(T_0+k)}-q_{\eta_2}(T,n,p)} \text{ for all } k=1,\ldots,T_{\mathbf{D}}-1,\\
\Rightarrow\ &\hat{r}_{(T_0)}>\frac{d_{(T_0+1)}-q_{\eta_2}(T,n,p)}{q_{\eta_2}(T,n,p)}, \text{ and } \max_{1\leq k\leq T_{\mathbf{D}}-1}\frac{\hat{d}_{(T_0+k+1)}}{\hat{d}_{(T_0+k)}}<\max_{1\leq k\leq T_{\mathbf{D}}-1}\frac{d_{(T_0+k+1)}+q_{\eta_2}(T,n,p)}{d_{(T_0+k)}-q_{\eta_2}(T,n,p)} \qquad (24)
\end{aligned}$$

We have already established in (23) that $d_{(T_0+1)} - q_{\eta_2}(T,n,p) > 0$. Fix a $k \in \{1,\ldots,(T_{\mathbf{D}}-1)\}$. It follows from A3 that

$$\begin{aligned}
&\frac{d_{(T_0+k+1)}}{d_{(T_0+k)}} < \frac{d_{(T_0+1)}}{3q_{\eta_2}(T,n,p)} \\
\Rightarrow\ & 3q_{\eta_2}(T,n,p)d_{(T_0+k+1)} < d_{(T_0+1)}d_{(T_0+k)} \\
\Rightarrow\ & q_{\eta_2}(T,n,p)(d_{(T_0+k+1)} + d_{(T_0+k)} + d_{(T_0+1)}) < d_{(T_0+1)}d_{(T_0+k)} \\
\Rightarrow\ & d_{(T_0+k+1)}q_{\eta_2}(T,n,p) < d_{(T_0+1)}d_{(T_0+k)} - q_{\eta_2}(T,n,p)\left(d_{(T_0+1)} + d_{(T_0+k)}\right) \\
\Rightarrow\ & d_{(T_0+k+1)}q_{\eta_2}(T,n,p) + q^2_{\eta_2}(T,n,p) < d_{(T_0+1)}d_{(T_0+k)} - q_{\eta_2}(T,n,p)\left(d_{(T_0+1)} + d_{(T_0+k)}\right) + q^2_{\eta_2}(T,n,p) \\
\Rightarrow\ & q_{\eta_2}(T,n,p)\left(d_{(T_0+k+1)} + q_{\eta_2}(T,n,p)\right) < \left(d_{(T_0+k)} - q_{\eta_2}(T,n,p)\right)\left(d_{(T_0+1)} - q_{\eta_2}(T,n,p)\right) \\
\Rightarrow\ & \frac{d_{(T_0+k+1)} + q_{\eta_2}(T,n,p)}{d_{(T_0+k)} - q_{\eta_2}(T,n,p)} < \frac{d_{(T_0+1)} - q_{\eta_2}(T,n,p)}{q_{\eta_2}(T,n,p)}.
\end{aligned}$$

Therefore,

$$\max_{1\le k\le T_{\mathbf{D}}-1} \frac{d_{(T_0+k+1)} + q_{\eta_2}(T,n,p)}{d_{(T_0+k)} - q_{\eta_2}(T,n,p)} < \frac{d_{(T_0+1)} - q_{\eta_2}(T,n,p)}{q_{\eta_2}(T,n,p)} \tag{25}$$

Combining (24) and (25), we obtain

$$\begin{aligned}
&\max_k |\hat{d}_{(k)} - d_{(k)}| \le q_{\eta_2}(T,n,p) \Rightarrow \max_{1\le k\le T_{\mathbf{D}}-1} \frac{\hat{d}_{(T_0+k+1)}}{\hat{d}_{(T_0+k)}} < \hat{r}_{(T_0)} \\
\text{i.e., }\ & \mathrm{P}\left[\max_k |\hat{d}_{(k)} - d_{(k)}| \le q_{\eta_2}(T,n,p)\right] \le \mathrm{P}\left[\max_{1\le k\le T_{\mathbf{D}}-1} \frac{\hat{d}_{(T_0+k+1)}}{\hat{d}_{(T_0+k)}} < \hat{r}_{(T_0)}\right].
\end{aligned}$$

The proof follows from Lemma 3(2). □

Proof of Theorem 2: Let us assume that $\Omega_T^{\mathbf{D}}$ is not a subset of $\hat{\Omega}_T^{\mathbf{D}}$. Now, $\Omega_T^{\mathbf{D}} \not\subseteq \hat{\Omega}_T^{\mathbf{D}}$ means that the set $\{\hat{d}_k \le \hat{d}_{(\hat{T}_0)}$ for some k with $\omega_k \in \Omega_T^{\mathbf{D}}\}$ is non-empty. Thus,

$$\begin{aligned}
& P\left[\Omega_T^{\mathbf{D}} \not\subseteq \hat{\Omega}_T^{\mathbf{D}}\right] \\
=\ & P\left[\hat{d}_k \le \hat{d}_{(\hat{T}_0)} \text{ for some } k \text{ with } \omega_k \in \Omega_T^{\mathbf{D}}\right] \\
\le\ & P\left[\hat{d}_k \le \hat{d}_{(\hat{T}_0)} \text{ for some } k \text{ with } \omega_k \in \Omega_T^{\mathbf{D}}, \max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| \le q_{\eta_2}(T,n,p)\right] \\
& + P\left[\max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| > q_{\eta_2}(T,n,p)\right] \\
\le\ & \sum_{\omega_k\in\Omega_T^{\mathbf{D}}} P\left[\hat{d}_k \le \hat{d}_{(\hat{T}_0)}, \max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| \le q_{\eta_2}(T,n,p)\right] + P\left[\max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| > q_{\eta_2}(T,n,p)\right].
\end{aligned}$$

Using Lemma 3(2), we have $P\left[\max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| > q_{\eta_2}(T,n,p)\right] < \frac{T}{(2p)^{\eta_2-2}}$. Consequently,

$$P\left[\Omega_T^{\mathbf{D}} \not\subseteq \hat{\Omega}_T^{\mathbf{D}}\right] \le \sum_{\omega_k\in\Omega_T^{\mathbf{D}}} P\left[\hat{d}_k \le \hat{d}_{(\hat{T}_0)}, \max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| \le q_{\eta_2}(T,n,p)\right] + \frac{T}{(2p)^{\eta_2-2}}. \tag{26}$$

Under the conditions in Lemma 4, we have

$$\begin{aligned}
&\max_{1\le l\le T'} |\hat{d}_{(l)} - d_{(l)}| \le q_{\eta_2}(T,n,p) \Rightarrow \hat{r}_{T_0} > \max_{1\le l\le (T_{\mathbf{D}}-1)} \hat{r}^d_{T_0+l} \\
\Rightarrow\ & \operatorname*{arg\,max}_{1\le l\le (T'-1)} \hat{r}^d_l \le T_0 \Rightarrow \hat{T}_0 \le T_0 \Rightarrow \hat{d}_{(\hat{T}_0)} \le \hat{d}_{(T_0)}.
\end{aligned} \tag{27}$$

Therefore, it follows from (26) and (27) that

$$\begin{aligned} & P[\Omega_T^{\mathbf{D}} \not\subseteq \hat{\Omega}_T^{\mathbf{D}}] \\ & \leq \sum_{\omega_k \in \Omega_T^{\mathbf{D}}} P\left[\hat{d}_k \leq \hat{d}_{(\hat{T}_0)}, \max_{1 \leq l \leq T'} |\hat{d}_{(l)} - d_{(l)}| \leq q_{\eta_2}(T,n,p)\right] + \frac{T}{(2p)^{\eta_2 - 2}} \\ & \leq \sum_{\omega_k \in \Omega_T^{\mathbf{D}}} P\left[\hat{d}_k \leq \hat{d}_{(T_0)}, \max_{1 \leq l \leq T'} |\hat{d}_{(l)} - d_{(l)}| \leq q_{\eta_2}(T,n,p)\right] + \frac{T}{(2p)^{\eta_2 - 2}}. \end{aligned} \tag{28}$$

Fix a k with $\omega_k \in \Omega_T^{\mathbf{D}}$. Recall (22) and observe that

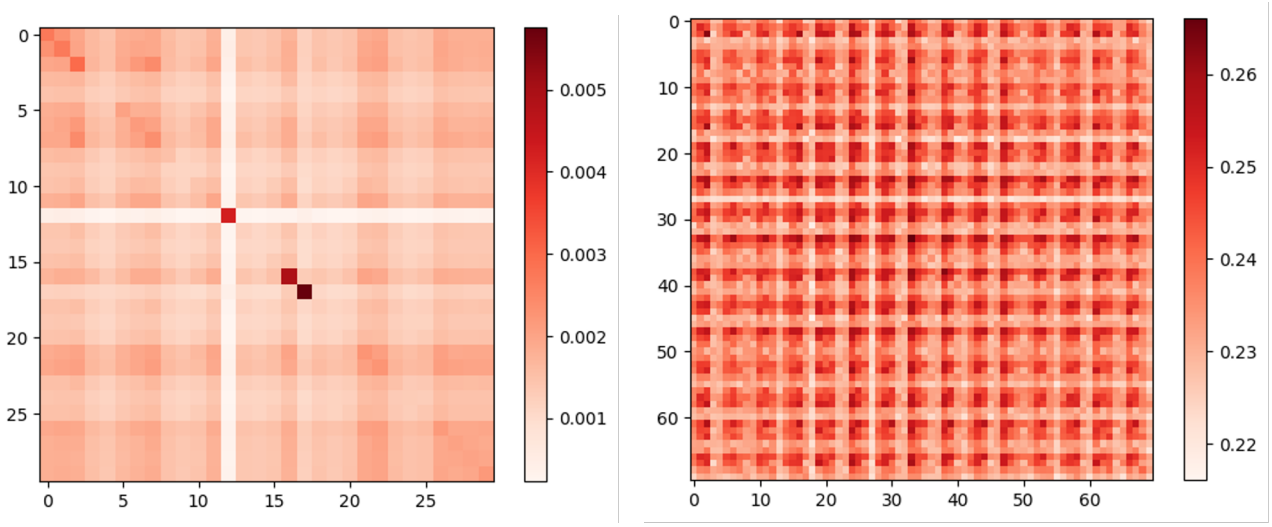
$$\begin{aligned} & P\left[\hat{d}_k \leq \hat{d}_{(T_0)}, \max_{1 \leq l \leq T'} |\hat{d}_{(l)} - d_{(l)}| \leq q_{\eta_2}(T,n,p)\right] \\ & \leq P\left[\hat{d}_k \leq \hat{d}_{(T_0)}, \max_{1 \leq l \leq T'} |\hat{d}_l - d_l| \leq q_{\eta_2}(T,n,p), \max_{1 \leq l \leq T'} |\hat{d}_{(l)} - d_{(l)}| \leq q_{\eta_2}(T,n,p)\right] \\ & \leq P\left[d_k - q_{\eta_2}(T,n,p) \leq \hat{d}_k \leq \hat{d}_{(T_0)} \leq q_{\eta_2}(T,n,p)\right] \\ & \leq \mathbb{I}[d_k - q_{\eta_2}(T,n,p) \leq q_{\eta_2}(T,n,p)] \\ & = \mathbb{I}[d_k \leq 2q_{\eta_2}(T,n,p)] \leq \mathbb{I}[d_k \leq 2q_{\eta_2}(T,n,p)] = 0 \text{ [follows from A2].} \end{aligned} \tag{29}$$

Therefore, combining (28) and (29), we obtain

$$P[\Omega_T^{\mathbf{D}} \not\subseteq \hat{\Omega}_T^{\mathbf{D}}] \leq \frac{T}{(2p)^{\eta_2 - 2}}.$$

This completes the proof. □

7 ADDITIONAL FIGURES



(a) Alert vs Drowsy

(b) Mice Calcium Recording

Figure 3: Heatmap of the difference between auto-covariance matrices at lag zero.

8 ADDITIONAL DETAILS

Model	Hyperparameters
Proposed	Optimizer = Adam Learning Rate = eta = 1e-3 Lamda = [1e-3, 1e-2, 1e-1,0.5,0.75,1] Epochs = 40 Minibatch Size = 32
QDA	reg_param = [0,0.5,1]
LDA	Solver = svd
Minirocket	Default parameters followed by the ridge classifier with CV to tune regularization parameter $\lambda \in [0.2, 1]$
SyncNet	K = 8 Minibatch Size = 32 Training Steps = 50 Validation Steps = 50

Table 5: Details of hyperparameters used in the analysis of simulated data

Model	Hyperparameters
Proposed	Optimizer = Adam Learning Rate = eta = 1e-3 Lamda = [1e-3, 1e-2, 1e-1,0.5,0.75,1] Epochs = 40 Minibatch Size = 32
QDA	reg_param = [0,0.5,1]
LDA	Solver = svd and $\lambda \in \{0, 0.5, 1\}$
Minirocket	Default parameters followed by ridge classifier with CV to tune regularization parameter $\lambda \in [0.2, 1]$
SyncNet	K = 32 Pool Size = 75 Minibatch Size = 32 Training Steps = 50 Validation Steps = 50

Table 6: Details of hyperparameters used in the analysis of real data sets