Learning Cortico-Muscular Dependence through Orthonormal Decomposition of Density Ratios

Shihan Ma[1†] **Bo Hu**[2†] **Tianyu Jia**[1] **Alexander Kenneth Clarke**[1]
Blanka Zicher[1] **Arnault H. Caillet**[1] **Dario Farina**[1] **José C. Príncipe**[2]
[†] *Equal Contribution*
[1] Department of Bioengineering, Imperial College London
[2] Department of Electrical and Computer Engineering, University of Florida

Abstract

The cortico-spinal neural pathway is fundamental for motor control and movement execution, and in humans it is typically studied using concurrent electroencephalography (EEG) and electromyography (EMG) recordings. However, current approaches for capturing high-level and contextual connectivity between these recordings have important limitations. Here, we present a novel application of statistical dependence estimators based on orthonormal decomposition of density ratios to model the relationship between cortical and muscle oscillations. Our method extends from traditional scalar-valued measures by learning eigenvalues, eigenfunctions, and projection spaces of density ratios from realizations of the signal, addressing the interpretability, scalability, and local temporal dependence of cortico-muscular connectivity. We experimentally demonstrate that eigenfunctions learned from cortico-muscular connectivity can accurately classify movements and subjects. Moreover, they reveal channel and temporal dependencies that confirm the activation of specific EEG channels during movement. Our code is available at https://github.com/bohu615/corticomuscular-eigen-encoder.

1 Introduction

The brain communicates with muscles by sending information to the spinal cord. Part of this information is directly transmitted from the cortex to spinal motor neurons via the cortico-spinal neural pathway, which is vital for motor control and movement execution. Because motor neurons are directly connected to muscles, cortical oscillations traveling through the cortico-spinal pathway are coherent with oscillations in muscle electrical activities, both in humans and non-human primates [1–3]. This relationship, known as functional cortico-muscular connectivity, is critical in neuroscience and is typically studied using concurrent recordings of neural signals such as electroencephalography (EEG) and electromyography (EMG) [4, 5] (Fig. 1). Practical applications include diagnosing and monitoring of neuromuscular disorders, such as amyotrophic lateral sclerosis [6], stroke [7], and Parkinson's disease [8], as well as developing brain-computer interfaces (BCIs) for individuals with motor impairments [9].

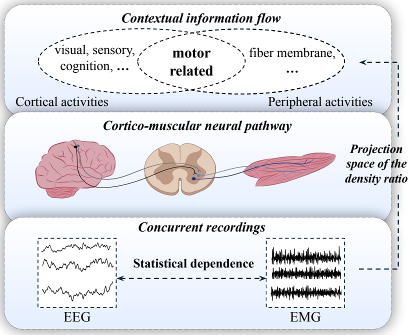

Contextual information flow
visual, sensory, cognition, ...
motor related
fiber membrane, ...
Cortical activities
Peripheral activities
Cortico-muscular neural pathway
Projection space of the density ratio
Concurrent recordings
Statistical dependence
EEG
EMG

Figure 1: The cortico-muscular pathway allows brain-muscle communication with coherent cortical and peripheral oscillations. This paper models this connectivity through the statistical dependence between their concurrent recordings of EEG and EMG.

38th Conference on Neural Information Processing Systems (NeurIPS 2024). Correspondence to Bo Hu (bohu@ufl.edu) and Shihan Ma (mmasss1205@gmail.com).

https://doi.org/10.52202/079017-4108

Despite several applications, there is still a lack of proper statistical tools to model the relationship between EEG and EMG. The predominant method, Cortico-Muscular Coherence (CMC), measures temporal and spectral coherence by computing the normalized cross-spectrum in time intervals [10]. From this analysis, it is generally accepted that the EEG's beta band (13-30 Hz) is linked to steady motor control, while the gamma band (>30 Hz) is associated to motor planning and execution [5, 11–13].

While CMC provides some relevant information on cortico-muscular connectivity, there remains a lack of generalized and higher-order statistical measures that quantify nonlinear and high-level connectivity. Can high-level contextual information, such as muscle movements and participant identifiers, be directly learned from modeling cortico-muscular connectivity? Our paper explores the potential of using statistical dependence estimators to address this problem.

Statistical dependence estimators typically follow a procedure of defining a measure preferably by probabilistic distributions, deriving a variational bound, and optimizing a variational cost of this bound using a function approximator, such as mutual information estimators [14–16] and Kernel Independent Component Analysis (KICA) [17, 18]. These measures are defined for realizations.

However, statistical dependence estimators above have rarely been successfully applied to cortico-muscular analysis, mainly due to three reasons: instability and poor scalability, lack of spatio-temporal resolution, and lack of practical contextual connections. As these estimators typically quantify dependence at the trial level, they overlook the importance of channel and temporal dependence in cortico-muscular analysis. More importantly, these measures only produce a scalar-valued score, but how this score should be used and its connection to the desired contextual factors are unclear.

This paper successfully applies statistical dependence estimators to EEG-EMG pairs using the concept of ***orthonormal decomposition of the density ratio***. Recently, there has been a shift from scalar-valued measures to decomposing density ratio as a positive definite function, and learning its eigenvalues, eigenfunctions, and associated projection spaces through neural network optimization with matrix cost functions such as $\log\det$ and nuclear norm [19–21]. This decomposition, known as the Functional Maximal Correlation Algorithm (FMCA), addresses the fundamental issue of relating dependence to contextual information: Eigenvalues define a multivariate dependence measure, and eigenfunctions span a feature projection space that captures the contextual factors affecting dependence.

This paper expands on this idea, addressing interpretability, scalability, and local-level dependence that are missing in existing dependence analyses. Sec. 2.1 explains why eigenfunctions, learned from cortico-muscular connectivity, can capture contextual factors for motor control and participant identification. Sec. 2.2 introduces FMCA-T, optimizing a new matrix trace cost for the theory, which demonstrates greater efficiency and stability than the $\log\det$ cost. Sec. 2.3 shows that while the objective estimates global dependence from trial realizations, localized channel-level and temporal-level dependencies can also be formed in a top-down manner, which are important in EEG as they indicate channel activations and synchronization of activities. Our framework is illustrated in Fig. 2.

Our main experiment demonstrates that the learned eigenfunctions, without labels, effectively capture factors such as movements and subjects that contribute to high-level cortico-muscular connectivity. After training, using EEG's eigenfunction as a feature projector noticeably improves classification accuracy over various baselines. Additionally, channel-level and temporal-level dependencies indicate that specific EEG channels are selectively activated during movements, corroborating neuroscientific findings. Simulated data further confirm that our proposed measure is invariant to nonstationary noise, including pink noise and random delays.

2 Methods

2.1 Density ratio decomposition for EEG-EMG signal pairs

Problem formulation. Consider EEG signals $\boldsymbol{X} := \boldsymbol{X}_{1:T}$ and EMG signals $\boldsymbol{Y} := \boldsymbol{Y}_{1:T}$. Denote $\boldsymbol{s}$ as the subject, $\boldsymbol{c}$ as the type of movement, and $\boldsymbol{u}$ as other auxiliary contextual factors. These are factors that could potentially affect the statistical dependence between EEG and EMG signals. Each signal is conditioned on these parameters. Denote these factors as $\boldsymbol{z} := \{\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{u}\}$ with distribution $\mathbb{P}(\boldsymbol{z})$. Distributions for EEG and EMG given these conditions are $p(\boldsymbol{X} = X|z)$ and $p(\boldsymbol{Y} = Y|z)$, respectively. Their joint distribution is given by $p(X, Y) = \int p(X|z)p(Y|z)p(z)dz$. Similarly, the marginal distributions are given by $p(X) = \int p(X|z)p(z)dz$, and likewise $p(Y) = \int p(Y|z)p(z)dz$.

https://doi.org/10.52202/079017-4108

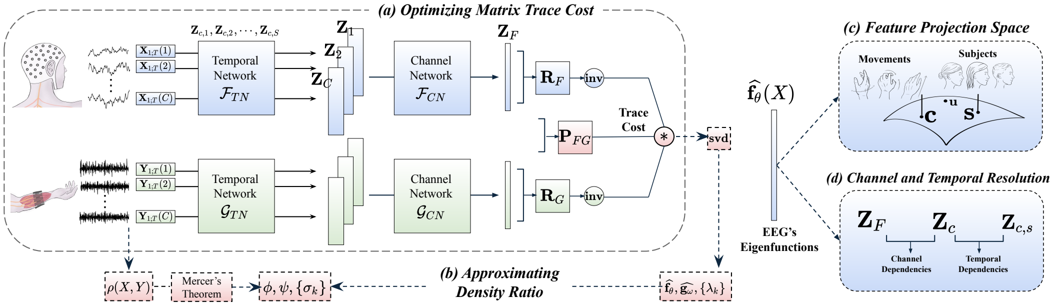

(a) Optimizing Matrix Trace Cost
$\mathbf{Z}_{c,1}, \mathbf{Z}_{c,2}, \cdots, \mathbf{Z}_{c,S}$
$\mathbf{X}_{1:T}(1)$
$\mathbf{X}_{1:T}(2)$
$\mathbf{X}_{1:T}(C)$
Temporal Network $\mathcal{F}_{TN}$
$\mathbf{Z}_1$
$\mathbf{Z}_2$
$\mathbf{Z}_C$
Channel Network $\mathcal{F}_{CN}$
$\mathbf{Z}_F$
$\mathbf{R}_F$
inv
$\mathbf{P}_{FG}$
Trace Cost
$*$
svd
$\mathbf{Y}_{1:T}(1)$
$\mathbf{Y}_{1:T}(2)$
$\mathbf{Y}_{1:T}(C)$
Temporal Network $\mathcal{G}_{TN}$
Channel Network $\mathcal{G}_{CN}$
$\mathbf{R}_G$
inv
$\rho(X, Y)$
Mercer's Theorem
$\phi, \psi, \{\sigma_k\}$
(b) Approximating Density Ratio
$\widehat{\mathbf{f}_\theta}, \widehat{\mathbf{g}_\omega}, \{\lambda_k\}$
$\widehat{\mathbf{f}}_\theta(X)$
EEG's Eigenfunctions
(c) Feature Projection Space
Movements
Subjects
c
u
s
(d) Channel and Temporal Resolution
$\mathbf{Z}_F$
$\mathbf{Z}_c$
$\mathbf{Z}_{c,s}$
Channel Dependencies
Temporal Dependencies

Figure 2: Diagram for learning cortico-muscular dependence by decomposing density ratios: (a) Network $\boldsymbol{f}_\theta$ is applied to EEG $\boldsymbol{X}_{1:T}$ and $\boldsymbol{g}_\omega$ to EMG $\boldsymbol{Y}_{1:T}$ to minimize a matrix trace cost. (b) EEG-EMG pairs are sampled from a joint distribution, from which a density ratio $\rho(X, Y)$ is defined and considered a positive definite function. Its linear operator has a spectral decomposition of eigenfunctions $\{\boldsymbol{\phi}, \boldsymbol{\psi}\}$ and eigenvalues $\{\sigma_k\}$. The networks provably approximate the dominant eigenvalues and eigenfunctions of this decomposition with network outputs $\{\widehat{\boldsymbol{f}_\theta}, \widehat{\boldsymbol{g}_\omega}\}$, and SVD results $\{\lambda_k\}$. Eigenvalues here measure multivariate statistical dependence; eigenfunctions are optimal feature projectors. (c) After training, the eigenfunctions, specifically those from EEG, form a projection space containing contextual information for motor control and participant identification. (d) To provide channel activation and activity synchronization for cortico-muscular analysis, we compute density ratios between channel-level $\boldsymbol{Z}_c$ and temporal-level features $\boldsymbol{Z}_{c,s}$ against global features $\boldsymbol{Z}_F$ to quantify channel-level and temporal-level dependencies.

Our goal is to extract factors $\boldsymbol{s}$, $\boldsymbol{c}$, $\boldsymbol{u}$ that affect the dependence between two modalities, from available sample pairs of $\boldsymbol{X}$ and $\boldsymbol{Y}$, even when $\boldsymbol{s}$, $\boldsymbol{c}$, and $\boldsymbol{u}$ are not given. We propose that this can be achieved by decomposing the density ratio of this probabilistic system.

Decomposition of EEG and EMG density ratios. Following the work on FMCA, we propose an orthonormal decomposition of the density ratio to measure the dependence between EEG and EMG:

$$\rho := \frac{p(X, Y)}{p(X)p(Y)} = \sum_{k=1}^{\infty} \sqrt{\sigma_k}\, \phi_k(X)\psi_k(Y),$$
$$\int_{\mathcal{X}} \phi_i(X)\phi_j(X)\, d\mathbb{P}(X) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}, \quad \int_{\mathcal{Y}} \psi_i(Y)\psi_j(Y)\, d\mathbb{P}(Y) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}, \tag{1}$$

for any $i, j = 1, 2, \cdots$. The density ratio $\rho(X, Y)$ is treated as a positive definite function associated with a linear operator $\boldsymbol{L}f := \int \rho(X, \cdot) f(X)\, dX$ for any measurable scalar function f. According to Mercer's theorem, this operator has a spectral decomposition with eigenvalues $\sigma_1, \sigma_2, \cdots$, and orthonormal basis functions $\phi_1, \phi_2, \cdots$ and $\psi_1, \psi_2, \cdots$. In scenarios where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent, all eigenvalues are zero. Conversely, larger eigenvalues suggest stronger dependence.

Such eigenfunctions form a linear span. Our hypothesis is that this span captures shared contextual factors such as $\boldsymbol{c}$ and $\boldsymbol{s}$ across two modalities, stated as follows.

Lemma 1. *Assuming conditional independence given $\boldsymbol{z} := \{\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{u}\}$, we have $p(X, Y|z) = p(X|z)p(Y|z)$. Hence, the ratio $\rho(X, Y) := \frac{p(X,Y)}{p(X)p(Y)}$ decomposes as $\rho(X, Y) = \int \frac{p(X|z)p(z)}{p(X)p(z)} \cdot \frac{p(Y|z)p(z)}{p(Y)p(z)} \cdot p(z)dz$. Assuming $\boldsymbol{z}$ is discrete (e.g., movement patterns $\mathbf{c}$ and participant identities $\mathbf{s}$), the information of $\boldsymbol{z}$ is contained in the span of the basis functions for the density ratio $\rho(X, Y)$.*

Proof. Define the ratios $\rho_X(X, z) = \frac{p(X,z)}{p(X)p(z)}$ and $\rho_Y(Y, z) = \frac{p(Y,z)}{p(Y)p(z)}$. Considering the sample space $\mathcal{Z}$, the sets $\rho_X(X, z)$ and $\rho_Y(Y, z)$ for $z \in \mathcal{Z}$ are discrete. Under the conditional independence assumption, these ratios satisfy $\rho(X, Y) = \sum_{z \in \mathcal{Z}} p(z)\rho_X(X, z)\rho_Y(Y, z)$.

This indicates that the sets $\rho_X(X, z)$ and $\rho_Y(Y, z)$ decompose $\rho(X, Y)$, similar to the eigenfunctions ϕ_k and ψ_k. Since both decompositions represent $\rho(X, Y)$, the functions $\rho_X(X, z)$ and $\rho_Y(Y, z)$ must lie within the span of these basis functions. Hence, there exist coefficients $\alpha_{z,k}$ and $\beta_{z,k}$ such that $\rho_X(X, z) = \sum_k \alpha_{z,k}\phi_k(X)$ and $\rho_Y(Y, z) = \sum_k \beta_{z,k}\psi_k(Y)$ for each z. Thus, learning the

https://doi.org/10.52202/079017-4108

dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is implicitly learning the dependence between each of them relative to the factors $\boldsymbol{z}$, even when $\boldsymbol{z}$ is not observed. □

2.2 FMCA-T: Learning decomposition for the matrix trace

When probability densities are unavailable, we approximate eigenvalues and eigenfunctions using a learning system with two neural networks and a cost function, typically a matrix cost like $\log\det$ or nuclear norm [19–21]. These costs optimize an aggregation of eigenvalues. The networks learn the dominant eigenvalues and eigenfunctions when optimized.

Aggregation of eigenvalues. To measure the total power of the eigenspectrum, define a scalar-valued measure using a convex function $\boldsymbol{\xi} : \mathbb{R} \to \mathbb{R}$ with $\boldsymbol{\xi}(0) = 0$. Assume the eigenvalues are ranked $\sigma_1 \geq \sigma_2 \geq \cdots$. The truncated total statistical dependence measure of the top K eigenvalues is defined by $T_{\boldsymbol{\xi}} := \sum_{k=1}^{K} \boldsymbol{\xi}(\sigma_k)$. Function $\boldsymbol{\xi}(x) = -\log(1-x)$ corresponds to the $\log\det$ cost.

Prior work: log-determinant cost. Consider two networks, $\boldsymbol{f}_\theta : \mathcal{X} \to \mathbb{R}^K$ and $\boldsymbol{g}_\omega : \mathcal{Y} \to \mathbb{R}^K$, mapping realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to K-dimensional outputs, respectively. Assume $\boldsymbol{f}_\theta$ is for EEG and $\boldsymbol{g}_\omega$ for EMG. The autocorrelation (ACFs) and cross-correlation functions (CCFs) are defined as:

$$\boldsymbol{R}_F = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{f}_\theta(\boldsymbol{X})\boldsymbol{f}_\theta^\intercal(\boldsymbol{X})],\ \boldsymbol{R}_G = \mathbb{E}_{\boldsymbol{Y}}[\boldsymbol{g}_\omega(\boldsymbol{Y})\boldsymbol{g}_\omega^\intercal(\boldsymbol{Y})], \boldsymbol{P}_{FG} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[\boldsymbol{f}_\theta(\boldsymbol{X})\boldsymbol{g}_\omega^\intercal(\boldsymbol{Y})],\ \boldsymbol{R}_{FG} = \begin{bmatrix} \boldsymbol{R}_F & \boldsymbol{P}_{FG} \\ \boldsymbol{P}_{FG}^\intercal & \boldsymbol{R}_G \end{bmatrix}. \tag{2}$$

FMCA minimizes a $\log\det$ cost, which reaches the negative value of the total measure $T_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}(x) = -\log(1-x)$ when minimized. The cost is defined by:

$$\min_{\theta,\omega}\ r_L(\theta,\omega) = \log\det \boldsymbol{R}_{FG} - \log\det \boldsymbol{R}_F - \log\det \boldsymbol{R}_G,\ \ r_L^* = \sum_{k=1}^{K} \log(1-\sigma_k). \tag{3}$$

Normalization trick. After training, normalizations are needed to obtain eigenfunctions. The first step is to ensure orthonormality: $\overline{\boldsymbol{f}_\theta} = \boldsymbol{R}_F^{-\frac{1}{2}}\boldsymbol{f}_\theta, \overline{\boldsymbol{g}_\omega} = \boldsymbol{R}_G^{-\frac{1}{2}}\boldsymbol{g}_\omega$. The second step is a singular value decomposition: $\overline{\boldsymbol{P}_{FG}} = \mathbb{E}[\overline{\boldsymbol{f}_\theta}(\boldsymbol{X})\overline{\boldsymbol{g}_\omega}^\intercal(\boldsymbol{X})] = \boldsymbol{U}\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{V}$, where $\boldsymbol{S} = \text{diag}(\lambda_1, \cdots, \lambda_K)$. The third step is to normalize functions such that they are invariant to the linear operator: $\widehat{\boldsymbol{f}_\theta} = \boldsymbol{U}^\intercal\overline{\boldsymbol{f}_\theta},\ \ \widehat{\boldsymbol{g}_\omega} = \boldsymbol{V}^\intercal\overline{\boldsymbol{g}_\omega}$. Functions $\widehat{\boldsymbol{f}_\theta}, \widehat{\boldsymbol{g}_\omega}$ are the top eigenfunctions of the density ratio, and $\lambda_1, \lambda_2, \cdots$ are the top eigenvalues. An approximation of the density ratio is given by $\widehat{\rho} = \widehat{\boldsymbol{f}_\theta}^\intercal\boldsymbol{S}^{\frac{1}{2}}\widehat{\boldsymbol{g}_\omega} \approx \rho$.

Newly proposed: matrix trace cost. This paper explores alternative convex functions, specifically the simplest case $\boldsymbol{\xi}(x) = x$, The cost, in the form of a matrix trace, is described below.

Lemma 2. *Denote $\boldsymbol{P} := \boldsymbol{P}_{FG}$. Given neural nets $\boldsymbol{f}_\theta$ and $\boldsymbol{g}_\omega$, minimizing the matrix trace*

$$\min_{\theta,\omega}\ r_T(\theta,\omega) = -\mathit{Trace}(\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal), \tag{4}$$

yields $r_T^*(\theta,\omega) = -\sum_{k=1}^{K}\sigma_k$, where $r_T^*(\theta,\omega)$ is the optimal cost, reaching the sum of the top K eigenvalues of the density ratio when minimized. We name this algorithm FMCA-T.

Proof. Applying the Schur complement to r_L, we obtain $r_L = \log\det(\boldsymbol{I} - \boldsymbol{R}_F^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal\boldsymbol{R}_F^{-\frac{1}{2}})$. Denoting eigenvalues of a matrix as $\lambda_1(\cdot), \cdots, \lambda_K(\cdot)$, the cost becomes $r_L = \sum_k \log(1 - \lambda_k(\boldsymbol{M}))$, where $\boldsymbol{M} = \boldsymbol{R}_F^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal\boldsymbol{R}_F^{-\frac{1}{2}}$. Optimizing the sum of eigenvalues instead, we use $Trace(\boldsymbol{M})$ and, based on the trace property $Trace(\boldsymbol{A}\boldsymbol{B}) = Trace(\boldsymbol{B}\boldsymbol{A})$, derive the trace cost for learning multivariate statistical dependence as $r_T = -Trace(\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal)$. □

FMCA-T is more computationally efficient as it uses only matrix operations of dimension K. Directly optimizing the sum of the eigenvalues is also more stable than optimizing their logarithm.

2.3 Channel-level and temporal-level dependencies

Motivations. For EEG $\boldsymbol{X}_{1:T}$ and EMG $\boldsymbol{Y}_{1:T}$, FMCA-T applies two networks, $\boldsymbol{f}_\theta$ and $\boldsymbol{g}_\omega$, to minimize the matrix trace cost. Dependence is measured at two levels: ***random-process level***, measured by eigenvalues for the overall dataset dependence, and ***trial level***, measured by the density ratio—the higher the ratio, the greater the contribution of this pair of realizations to overall dependence. In cortico-muscular analysis, it is vital to understand how individual channels and time steps contribute to connectivity, especially in EEG signals, as they represent the temporal and spatial dynamics of brain. Hence, we propose localized density ratios to measure ***temporal-level dependence*** and ***channel-level dependence***. The core idea is computing density ratios between channel-level and temporal-level features against the global trial-level features.

Channel-level features. We design a specialized network topology to generate features for individual channels and time intervals, ensuring that the internal layers of this network quantify channel-level and temporal-level features, similar to [22–24].

Given $\boldsymbol{X}_{1:T} = [\boldsymbol{X}_{1:T}(1), \cdots, \boldsymbol{X}_{1:T}(C)]^\intercal$ for channels $c = 1, \cdots, C$, we define a temporal network $\mathcal{F}_{TN} : \mathbb{R}^T \to \mathbb{R}^K$ that maps single-channel signals to a K-dimensional feature space, and a channel network $\mathcal{F}_{CN} : \mathbb{R}^{L \times K} \to \mathbb{R}^K$ that maps concatenated channel features to global features:

$$\boldsymbol{Z}_c = \mathcal{F}_{TN}\left(\boldsymbol{X}_{1:T}(c)\right), \quad c = 1, \cdots, C; \quad \boldsymbol{Z}_F = \mathcal{F}_{CN}\left([\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_C]^\intercal\right), \tag{5}$$

where $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_C$ are channel-level features, and $\boldsymbol{Z}_F$ is global trial-level features.

Channel-level dependence $\widehat{\rho_{C,F}}(c)$. The density ratio of $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_C$ relative to $\boldsymbol{Z}_F$ measures channel-level dependence. Post-training and with fixed parameters, we compute the ACF of the channel features $\boldsymbol{R}_C = \frac{1}{C}\mathbb{E}[\sum_{c=1}^{C} \boldsymbol{Z}_c \boldsymbol{Z}_c{}^\intercal]$, the ACF of the global features $\boldsymbol{R}_F = \mathbb{E}[\boldsymbol{Z}_F \boldsymbol{Z}_F^\intercal]$, and the CCF between them $\boldsymbol{P}_{C,F} = \frac{1}{C}\mathbb{E}[\sum_{c=1}^{C} \boldsymbol{Z}_c \boldsymbol{Z}_F^\intercal]$.

Next, the features are normalized as in the Sec. 2.2: $\boldsymbol{Z}_c$ and $\boldsymbol{Z}_F$ are normalized to $\overline{\boldsymbol{Z}_c} = \boldsymbol{R}_C^{-\frac{1}{2}} \boldsymbol{Z}_c$ and $\overline{\boldsymbol{Z}_F} = \boldsymbol{R}_F^{-\frac{1}{2}} \boldsymbol{Z}_F$ for orthonormality. The SVD of $\boldsymbol{R}_C^{-\frac{1}{2}} \boldsymbol{P}_{C,F} \boldsymbol{R}_F^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{V}$ is computed. The outputs are further normalized to $\widehat{\boldsymbol{Z}_c} = \boldsymbol{U}^\intercal \overline{\boldsymbol{Z}_c}$ and $\widehat{\boldsymbol{Z}_F} = \boldsymbol{V}^\intercal \overline{\boldsymbol{Z}_F}$ to guarantee invariance in the linear operator. Finally, the density ratio can be constructed as $\widehat{\rho_{C,F}}(c) = \widehat{\boldsymbol{Z}_c}{}^\intercal \boldsymbol{S} \widehat{\boldsymbol{Z}_F}$.

This ratio $\widehat{\rho_{C,F}}(c)$ is a function of channel c and trial $\boldsymbol{X}$, implying the dependence between channel and global features. The greater the value, the stronger the activation of the channels, showing which channels contribute the most to the cortico-muscular connectivity.

Temporal-level features and dependence. To measure time-domain dependence, we compute density ratios between the internal features of the temporal network $\mathcal{F}_{TN}$ and the global features, in two steps: first, computing density ratios between ***adjacent*** network layers; second, aggregating these ratios to consider all layers.

Step 1: Construct density ratios $\widehat{\rho_{s-1,s,c}}(\tau_1, \tau_2)$ between *adjacent* layers. Fix a channel c and feature $\boldsymbol{Z}_c$. Consider a simple temporal network with S convolution layers with nonlinaer activation functions: $\mathcal{F}_{TN}^{(1)}, \mathcal{F}_{TN}^{(2)}, \cdots, \mathcal{F}_{TN}^{(S)}$, with kernel sizes $\Delta_1, \Delta_2, \cdots, \Delta_S$, and their outputs $\boldsymbol{Z}_{c,1}, \boldsymbol{Z}_{c,2}, \cdots, \boldsymbol{Z}_{c,S}$. Suppose the time dimensions of these layers are $T_1, T_2, \cdots, T_S$. Fix any layer s. The τ-th element of $\boldsymbol{Z}_{c,s}$, denoted as $\boldsymbol{Z}_{c,s}(\tau)$, is obtained by applying a nonlinear operation to a segment of the previous layer's output:

$$\boldsymbol{Z}_{c,s}(\tau) = \mathcal{F}_{TN}^{(s-1)}\left(\boldsymbol{Z}_{c,s-1}(\tau : \tau + \Delta_{s-1})\right). \tag{6}$$

- Define the ACF of layer $s-1$: $\boldsymbol{R}_{c,s-1} = \frac{1}{T_{s-1}}\mathbb{E}[\sum_\tau \boldsymbol{Z}_{c,s-1}(\tau)\boldsymbol{Z}_{c,s-1}^\intercal(\tau)]$
- Define the ACF of layer s: $\boldsymbol{R}_{c,s} = \frac{1}{T_s}\mathbb{E}[\sum_\tau \boldsymbol{Z}_{c,s}(\tau)\boldsymbol{Z}_{c,s}^\intercal(\tau)]$
- Define the CCF between them: $\boldsymbol{P}_{c,s-1,s} = \frac{1}{T_s}\mathbb{E}[\sum_\tau \sum_{\delta=1}^{\Delta_s} \boldsymbol{Z}_{c,s-1}(\tau+\delta)\boldsymbol{Z}_{c,s}^\intercal(\tau)]$.

https://doi.org/10.52202/079017-4108

Normalization with $\boldsymbol{R}_{c,s-1}$, $\boldsymbol{R}_{c,s}$, and $\boldsymbol{P}_{c,s-1,s}$ yields density ratios $\widehat{\rho_{s-1,s,c}}(\tau_1, \tau_2)$, which quantify the dependence between time τ_1 and τ_2 across two layers $s-1$ and s, for a given trial and channel c. A higher value indicates a stronger dependence between adjacent network layers.

Step 2: Aggregate layer-wise ratios for localized responses $\widehat{\varrho_{s,c}}(\tau)$. While $\widehat{\rho_{s-1,s,c}}(\tau_1, \tau_2)$ quantifies dependence between two layers, we aggregate these ratios to account for all network layers.

Again, fix the element $\boldsymbol{Z}_{c,s}(\tau)$ in layer s. We focus on its mapping to the next layer $s+1$. Based on Eq. (6), elements in layer $s+1$ that are mapped from $\boldsymbol{Z}_{c,s}(\tau)$ by $\mathcal{F}_{FP}^{(s)}$ are within a window of size Δ_s (the kernel size). To ensure this window stays within the feature vector's boundary, we define coordinates $\mathcal{I}_s = [\max(0, \tau - \Delta_s + 1), \min(i, T_{s+1} - 1)]$. Feature elements in layer $s+1$ that are mapped from $\boldsymbol{Z}_{c,s}(\tau)$ fall within these coordinates.

We then create a series of functions $\widehat{\varrho_{s,c}}(\tau)$ with $\tau \in [1, T_s]$ for each layer s as the aggregations of the ratios. Starting from $\widehat{\varrho_{S,c}}(\tau) = \widehat{\rho_{C,F}}(c)$ (channel-level density ratio), compute recursively

$$\widehat{\varrho_{s,c}}(\tau_1) = \sum_{\tau_2 \in I_s} \widehat{\varrho_{s+1,c}}(\tau_2) \widehat{\rho_{s,s+1,c}}(\tau_1, \tau_2), \quad \tau_1 \in [1, T_s] \tag{7}$$

That is, starting from the top layer of the network, we aggregate the density ratios within the window $\mathcal{I}_s$, layer-by-layer, until we generate a localized measurement for each element of the function $\widehat{\varrho_{s,c}}(\tau)$ at layer s, channel c, and time $\tau \in [1, T_s]$, considering all neural network layers.

The final localized responses of the density ratio, $\widehat{\varrho_{s,c}}(\tau)$, obtained in a top-down manner, are functions of the EEG trial $\boldsymbol{X}_{1:T}$, time step τ, and channel c, providing both temporal and channel-level resolution. The same analysis applies to EMG signals, differing only in the number of channels.

3 Experiments

Our experiments have three key findings: (1) Dependence measured by FMCA-T is stable against nonstationary noises and delays in simulated dataset; (2) Learning from unlabeled EEG-EMG pairs extracts movement and subject information from EEG's eigenfunctions; (3) EEG's spatio-temporal dependencies are consistent with ground truth brain activations in simulated dataset and match theoretical evidence in experimental dataset.

3.1 Datasets and baselines

Dataset 1: SinWav. We construct a simulated dataset where each data pair $\{\boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:T}\}$ has a clean sinusoidal signal $\boldsymbol{X}_t = A\sin(\omega t)$ and a noisy counterpart $\boldsymbol{Y}_t$ superimposed with various types of noise: stationary white noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, nonstationary Gaussian noise $\sigma_t^2 \propto |\boldsymbol{X}_t|$, and nonstationary pink noise $S(f) \propto 1/f^\alpha$. Random delays are added by padding the start and end of the signal with noise such that $\boldsymbol{Y}_t = \epsilon_t$ is white for $1 \leq t \leq \tau_1$ and $T - \tau_2 < t \leq T$, and $\boldsymbol{Y}_t = \boldsymbol{X}_{t-\tau_1}$ is sinusoidal for $\tau_1 < t \leq T - \tau_2$, where $\tau_1 + \tau_2 = \tau$ is fixed. Since these noises do not change the underlying sinusoid, the signal pairs are statistically independent when conditioned on the sinusoid. Thus, we expect the dependence measure to be unaffected by the noise level.

Dataset 2: EEG-EMG-Fusion. We use a public dataset [25] (approved by the Institutional Review Board at Korea University, 1040548-KU-IRB-17-181-A-2) with paired 60-channel EEG and 7-channel EMG recordings from 25 subjects. The subjects perform three ***main movements***: arm-reaching, hand-grasping, and wrist-twisting. Each main movement contains ***sub-movements***: arm-reaching along six directions, hand-grasping three objects, and wrist-twisting with two motions, and thus 11 movements in total. Subjects perform one sub-movement per trial, and 50 trials are collected per sub-movement. The same recordings are repeated for three sessions at one-week intervals. Both EEG and EMG are recorded at 2,500 Hz and downsampled to 1,000 Hz. The dataset is cleaned by removing eye-blinking artifacts and baseline wandering, and segmented into 4-second intervals, creating 41,250 paired samples of complete movement cycles.

Dataset 3: Simulated EEG-EMG Dataset. We simulate 128-channel EEG signals and 7-channel EMG signals for left/right motor and sensory activations from 20 subjects using EEGSourceSim

[26]. Motor sources are used to simulate the corresponding EMG signals. Both EEG and EMG are sampled at 1,000 Hz. White Gaussian noise at 5 dB is added to the EEG and EMG signals. FMCA-T is trained on 16 subjects and tested on 4 subjects to compare FMCA-T's spatial-level dependence representation with the ground truth activations, as shown in Fig. 6.

Classification tasks. We conduct three classification experiments: ***3-class*** (three main movements), ***11-class*** (11 sub-movements), and ***Subj*** (25 subjects). We also compare ***inter-subject*** and ***cross-subject*** classifications. Cross-subject means the test set contains only unseen subjects. Inter-subject uses an 80-20 split of trials from all subjects for training and testing, while cross-subject uses 20 subjects for training and 5 for testing with five-fold cross-validation.

Statistical dependence baselines. We compare our proposed dependence measure with established baselines: (1) KICA [17] and HSIC [18], which solve the generalized eigenvalue problem of two kernel Gram matrices, using $\sum_i \lambda_i$ for HSIC and $-\sum_i \log(1-\lambda_i)$ for KICA; (2) MINE [14], which optimizes the Donsker-Varadhan representation with a three-layer MLP; (3) CC: Pearson correlation coefficient averaged over time; (4) MIR (KNN estimator [27]), which optimizes entropies using k-nearest neighbor distances, and then computes mutual information; (5) Our method uses density ratios for trial-level dependence and eigenvalue aggregations $T_{\boldsymbol{\xi}}$ for random-process-level dependence.

For the EEG-EMG dependence study, we compare with CMC, the correlation coefficient between EEG and EMG spectra of windowed data on the alpha band of channel C4 [25]. We extend CMC by replacing linear correlation with nonlinear measures, computing CMC-KICA and CMC-MIR.

EEG feature projector baselines. After training $\boldsymbol{f}_\theta$ and $\boldsymbol{g}_\omega$ networks for dependence estimation, with parameters fixed, we train a three-layer MLP on EEG's eigenfunctions $(\widehat{\boldsymbol{f}_\theta}(X))$ for classification. This is compared with baseline EEG classifiers trained from scratch: (1) Supervised: *Vanilla Classifier*, with the same topology as ours but using a standard log-likelihood cost; *EEG-Net* [22], a specialized network for EEG-based BCI; *EEG-Conformer* [28], a compact convolutional transformer for EEG decoding and visualization; *Deep4* [29], a deep ConvNet for classification using raw EEG; and *CSP-RLDA* [30], using common spatial pattern (CSP) for feature extraction and regularized linear discriminant analysis (RLDA), adapted for multi-class classification with majority voting. (2) Self-Supervised: contrastive costs using 1-second windows from the same signal as positive pairs and from different signals as negative pairs, including *Barlow Twins* [31], *SimCLR* [32], and *VicReg* [33]. Experimental and implementation details can refer to the App. C.

3.2 Main results

Robustness of FMCA-T. Fig. 3 shows the robustness of our proposed measure in the SinWav dataset, when there are increasing levels of nonstationary noise and delays. Since EEG and EMG signals are often damaged and distorted by environmental noise and the functional coupling occurs with time delays, an effective measure should maintain its robustness against these factors.

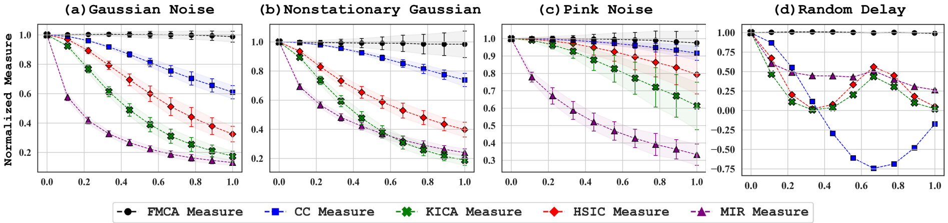


Figure 3: Density ratios from FMCA-T are robust to various noise types: (a) stationary white Gaussian noise, (b) nonstationary Gaussian noise, (c) nonstationary pink noise, and (d) random delays. FMCA-T proves the most robust estimations across all noise types and outperforms all linear and nonlinear baselines. Note that as delays increase, estimations using CC produce negative values given the opposite phase between the paired sinusoids.

In Fig. 3, FMCA-T is first trained on noisy data pairs with all four noise types and magnitudes (Sec. 3.1). Using the trained models, we measure the dependence between a clean sinusoid and its noisy counterpart. A noise level of 1.0 means the noise magnitude matches the sinusoid. The delay level determines the extent to which the clean sinusoid is shifted from its original position. A delay

https://doi.org/10.52202/079017-4108

level of 1.0 shifts the sinusoid to have no intersection with the original one. FMCA-T consistently shows invariance to noise and delays, as the dependence is determined by their frequency but not by the noise and phase shift. MINE fails to converge and produce stable results for this dataset.

Applying FMCA-T to EEG-EMG-Fusion. We confirm our primary hypothesis that the projection space defined by EEG eigenfunctions, derived from modeling the statistical dependence between EEG-EMG recordings, captures essential contextual factors like movements and subjects without requiring labels. We visualize the learned eigenfunctions and density ratios in Fig. 4.

Fig. 4(a), Fig. 4(b): eigenfunctions $\widehat{f_\theta}$. EEG's eigenfunctions effectively capture relevant contextual information. After training eigenfunctions using the entire dataset, we extract a subset of eigenfunctions that belong to a specific subject or a movement and apply t-SNE to visualize them.

Fig. 4(a) visualizes the eigenfunctions of all trials for one subject (`SUB1`). Each trial is color-coded by the type of movement (`MOV1`∼`MOV3`). Notably, the eigenfunctions form nine clusters, which are verified to correspond to the three movements recorded over three sessions. This demonstrates that the eigenfunctions contain motion-related information. The consistent clustering patterns across all 25 subjects are detailed in the App. B.
Fig. 4(b) visualizes the eigenfunctions from a single type of movement (reaching, labeled as `MOV1`) across ten different subjects (`SUB1`∼`SUB10`). Each color represents a subject. Distinct clustering patterns are observed, showing that the eigenfunctions also contain subject information which could be useful for participant identification.

Fig. 4(c), 4(d): density ratio $\widehat{\rho(X,Y)}$. Based on the t-SNE plot for `SUB1` trials, we plot the estimated density ratio values between each EEG-EMG pair in Fig. 4(c). In Fig. 4(d), we extract the mean of density ratios for trials in each cluster (`C1`∼`C9`), rank them from smallest to largest, and plot them alongside the standard deviation. These figures show that the density ratios remain consistent within each cluster (a movement during a session) while vary across different clusters. We find the highest dependence in reaching, followed by grasping, and the lowest in twisting. Our results are consistent with existing literature that links cortico-muscular connectivity with movement types [34, 35]. The results are consistent across all subjects, detailed in the App. B.

Fig. 4(e)∼(h) presents the results of MINE and CMC measures for `SUB1` trials. Only MINE produces a comparable measure that shows difference across clusters, but with higher variance and instability.

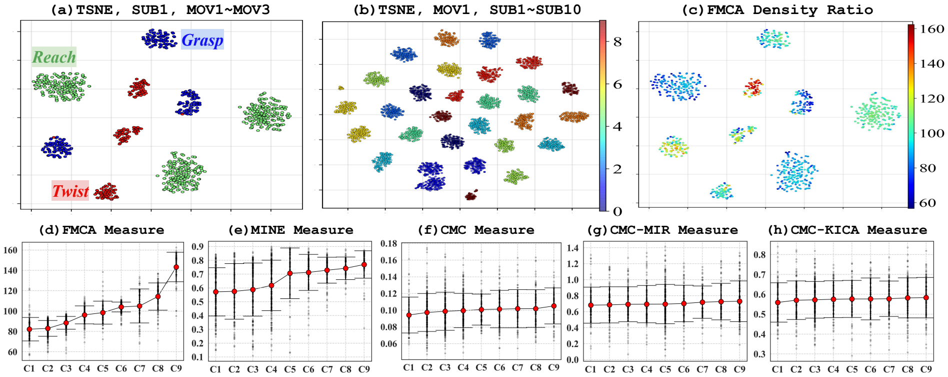


Figure 4: Visualizing eigenfunctions and density ratios in EEG-EMG fusion with FMCA-T: (a) t-SNE of EEG's eigenfunctions for a single subject (`SUB1`) show nine clusters specific to three movements (`MOV1`∼`MOV3`) across three sessions. (b) t-SNE of EEG's eigenfunctions for reaching movement (`MOV1`) across 10 subjects (`SUB1`∼`SUB10`) shows clusters specific to subjects, where each color is a subject. (c) Density ratios and (d) their mean and std of each cluster (`C1`∼ `C9`) demonstrate intra-cluster consistency and inter-cluster separability. (e-h) Comparison of baseline measures, where only MINE is comparable but with higher variance and instability.

EEG's eigenfunctions as optimal feature projector. Table 1 validates the claims in Fig. 4 with classification accuracy comparisons. We extract EEG eigenfunctions from the training set after the networks are trained on EEG-EMG pairs. The eigenfunctions are used to train a three-layer MLP for classification. This classifier predicts the class of any EEG test samples using its eigenfunctions (output of $\widehat{\boldsymbol{f}_\theta}$), without requiring EMG samples. As detailed in the Sec. 3.1, scores for ***3-class***, ***11-class***, and ***subj*** are presented in both ***inter-subject*** and ***cross-subject*** settings.

In inter-subject 3-class classification (80-20 split across trials from 25 subjects), FMCA-T exceeds the supervised baseline (EEGNet) by 7.2%, the classical EEG decoding method (CSP-RLDA) by 1.0%, and self-supervised methods by over 9.5%. Notably, CSP-RLDA is trained and tested on each individual subject, with the accuracy averaged across 25 subjects, thereby representing an upper bound for classical methods. All other methods are trained and tested on the combined subject data. For the 11-class classification task, FMCA-T surpasses all baselines with a classification accuracy of 0.32, significantly higher than the chance level of 0.09. Since CSP-RLDA uses binary classification with majority voting, it is computationally infeasible for 11 sub-movements classification.

In the more challenging cross-subject classification (trained on 20 subjects, tested on 5), FMCA-T with trace loss outperforms all baselines by over 10%, achieving an accuracy of 0.54 in the 3-class task. The highest scores from 10,000 iterations are recorded, and experiments are repeated with five-fold validation. The superior performance of FMCA-T in the cross-subject setting suggests that learning EEG-EMG dependence is robust against distribution shifts and nonstationary noise, which is consistent with the observation that self-supervised methods outperform supervised ones.

Comparing FMCA-LD ($\log\det$) and FMCA-T (matrix trace), we find that trace cost has greater stability and reduced variance. The sum of eigenvalues, especially during prolonged training, is more stable. While both costs show similar performance at the initial training stages, FMCA-T has notably reduced variance during the convergence stage of training.

Methods	**(a) Inter-Subject Acc.**			**(b) Cross-Subject Acc.**	
	3-Class	***11-Class***	***Subj***	***3-Class***	***11-Class***
Supervised					
Vanilla	0.907±0.020	0.220±0.015	0.980±0.010	0.427±0.021	0.110±0.005
EEGNet	0.904±0.015	0.246±0.028	0.988±0.007	0.405±0.019	0.095±0.021
EEG-Conformer	0.949±0.001	0.268±0.001	0.976±0.002	0.415±0.002	0.105±0.001
Deep4	0.901±0.001	0.274±0.001	0.941±0.001	0.429±0.001	**0.140±0.000**
CSP-RLDA	0.985±0.019	/	/	0.408±0.018	/
Self-Supervised					
Barlow Twins	0.893±0.018	0.269±0.012	0.987±0.008	0.437±0.018	0.115±0.004
SimCLR	0.890±0.019	0.257±0.013	0.979±0.011	0.441±0.020	0.117±0.006
VicReg	0.899±0.016	0.274±0.014	0.980±0.009	0.449±0.016	0.115±0.005
EEG-EMG Dependence					
FMCA-LD	0.985±0.003	0.257±0.011	0.989±0.007	0.509±0.014	0.115±0.003
FMCA-T	**0.994±0.002**	**0.320±0.009**	**0.998±0.004**	**0.540±0.012**	**0.121±0.002**

Table 1: Comparison of classification accuracies: supervised, self-supervised, and our EEG-EMG dependence learning. FMCA-T's eigenfunctions, trained with trace cost without labels, are optimal feature projectors for EEG. EMG is not required for testing, but only used for training.

Spatio-temporal dependencies - real data. We visualize the local density ratio responses of cluster `SUB3-C1` (reaching movement) in both spatial ($\widehat{\rho_{C,F}}(c)$) and temporal domains ($\widehat{\varrho_{s,c}}(\tau)$) in Fig. 5. The channel-level dependence is averaged across all trials and displayed in Fig. 5(a). We also randomly select nine trials from the same cluster and visualize them in Fig. 5(b). The temporal-level dependence for the first trial `T1` in `SUB3-C1` is shown in Fig. 5(c). Consistent activations are observed in other subjects, details in the App. B.

As illustrated in Fig. 5, the localized density ratio remains stable throughout the 4-second movement, corroborating the consistency of brain-muscle connectivity during stable states [36]. We also find that in channel-level dependence, the density ratio activates the fronto-central (FC) areas. The sensorimotor area is crucial for movement control, with EEG data from these regions often used to decode motor intentions. However, motor control also relies on cognitive processes [37], especially during movement planning, complex tasks, and collaborations [38]. Thus, the region of Brodmann area 6, well acknowledged to playing a role in movement planning may contribute differently during various movement tasks [39]. Our findings show that the features extracted from these fronto-central areas play an important role in classification.

Spatio-temporal dependencies - simulated data. We implement FMCA-T on paired EEG-EMG samples from the simulated dataset, using 16 subjects for training and 4 subjects for testing. We compare FMCA-T's spatial-level dependence maps for these 4 testing subjects with their ground truth brain activations computed by the motor ROI and forward matrices, shown in Fig. 6. FMCA-T's spatial-level dependencies are highly similar to the ground truth activations, indicating that the learned density ratios effectively captured the real brain activations.

https://doi.org/10.52202/079017-4108

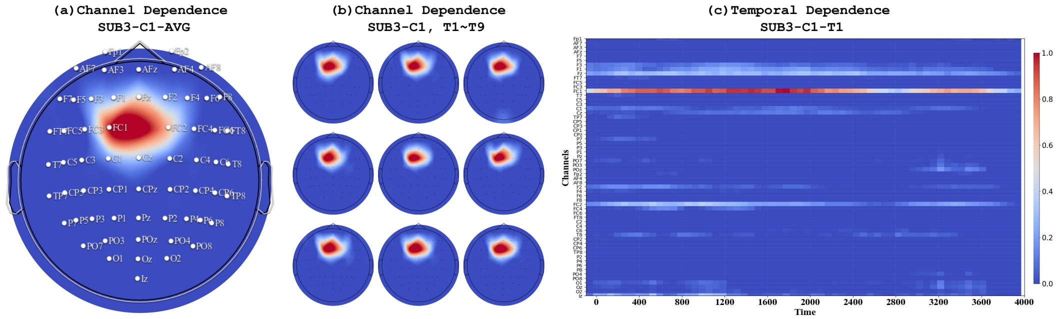


Figure 5: Localized density ratio for real data. Topographies for SUB3, C1 (reaching session) are normalized to the range $[0, 1]$. (a) displays the averaged spatial distribution of C1 across all 50 trials; (b) presents nine random trials T1-T9 from this session. Dark red indicates prominent activations around channel FC1. (c) shows stable temporal-level dependence over the 4-second movement.

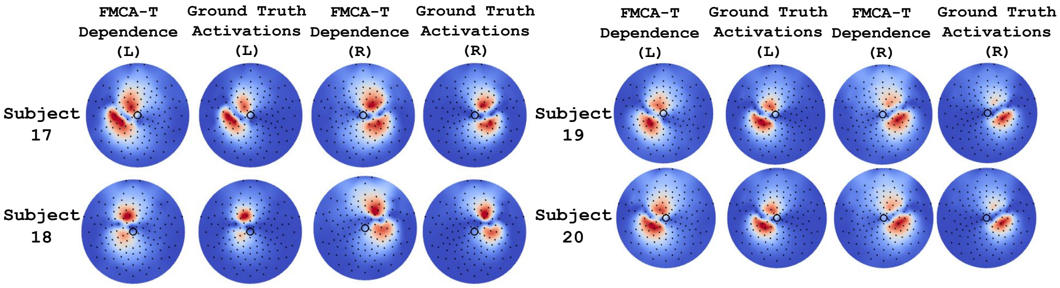


Figure 6: Localized density ratio for simulated data, showing that FMCA-T's distribution is consistent with ground truth brain activations when evaluated on four test subjects. Topographies are normalized to the range $[0, 1]$. L indicates left brain activities and R indicates right brain activities.

4 Conclusion

This paper introduces a novel approach for estimating cortico-muscular dependence through the orthonormal decomposition of the density ratio. By treating the density ratio as a positive definite function and learning its projection space from EEG and EMG, we unveil the relationship between statistical dependence, contextual factors impacting connectivity, and the spatio-temporal information shared between the brain and muscles. While our method shows promising results, challenges remain. For example, performance drops in cross-subject classification, likely due to the limited dataset of 25 participants. Future work will focus on applying our framework to larger datasets and incorporating additional bio-signal modalities to model a broader common space in neural data.

References

[1] S. N. Baker, J. M. Kilner, E. M. Pinches, and R. N. Lemon. The role of synchrony and oscillations in the motor output. *Experimental Brain Research*, 128:109–117, 1999.

[2] Stuart N. Baker. Oscillatory interactions between sensorimotor cortex and the periphery. *Current Opinion in Neurobiology*, 17:649–655, 2007.

[3] Jan Mathijs Schoffelen, Robert Oostenveld, and Pascal Fries. Neuronal coherence as a mechanism of effective corticospinal interaction. *Science*, 308:111–113, 2005.

[4] Cristina Brambilla, Ileana Pirovano, Robert Mihai Mira, Giovanna Rizzo, Alessandro Scano, and Alfonso Mastropietro. Combined use of EMG and EEG techniques for neuromotor assessment in rehabilitative applications: A systematic review. *Sensors*, 21(21), 2021.

[5] P. Grosse, M.J. Cassidy, and P. Brown. EEG−EMG, MEG−EMG and EMG−EMG frequency analysis: physiological principles and clinical applications. *Clinical Neurophysiology*, 113(10):1523–1531, 2002.

[6] Roisin McMackin, Peter Bede, Caroline Ingre, Andrea Malaspina, and Orla Hardiman. Biomarkers in amyotrophic lateral sclerosis: current status and future prospects. *Nature Reviews Neurology*, 19(12):754–768, 2023.

[7] Zhixian Gao, Shiyang Lv, Xiangying Ran, Yuxi Wang, Mengsheng Xia, Junming Wang, Mengyue Qiu, Yinping Wei, Zhenpeng Shao, Zongya Zhao, et al. Influencing factors of corticomuscular coherence in stroke patients. *Frontiers in Human Neuroscience*, 18:1354332, 2024.

[8] Nahid Zokaei, Andrew J Quinn, Michele T Hu, Masud Husain, Freek van Ede, and Anna Christina Nobre. Reduced cortico-muscular beta coupling in Parkinson's disease predicts motor impairment. *Brain Communications*, 3, 2021.

[9] Anirban Chowdhury, Haider Raza, Yogesh Kumar Meena, Ashish Dutta, and Girijesh Prasad. An EEG-EMG correlation-based brain-computer interface for hand orthosis supported neuro-rehabilitation. *Journal of Neuroscience Methods*, 312:1–11, 2019.

[10] Tatsuya Mima and Mark Hallett. Corticomuscular coherence: a review. *Journal of Clinical Neurophysiology*, 16(6):501, 1999.

[11] Gert Pfurtscheller and FH Lopes Da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999.

[12] Ole Jensen and Laura L Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, 11(7):267–269, 2007.

[13] Yuhang Xu, Verity M McClelland, Zoran Cvetković, and Kerry R Mills. Corticomuscular coherence with time lag with application to delay estimation. *IEEE Transactions on Biomedical Engineering*, 64(3):588–600, 2016.

[14] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. MINE: Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80, pages 531–540, 2018.

[15] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[16] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009.

[17] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

[18] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[19] Bo Hu and Jose C. Principe. The normalized cross density functional: A framework to quantify statistical dependence for random processes. *arXiv preprint arXiv:2212.04631*, 2024.

[20] Shao-Lun Huang, Gregory W Wornell, and Lizhong Zheng. Gaussian universal features, canonical correlations, and common information. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

[21] Shao-Lun Huang, Anuran Makur, Gregory W Wornell, and Lizhong Zheng. On universal features for high-dimensional learning and inference. *arXiv preprint arXiv:1911.09105*, 2019.

[22] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

[23] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Zhizhang Yuan, Daoze Zhang, Junru Chen, Geifei Gu, and Yang Yang. Brant-2: Foundation model for brain signals. *arXiv preprint arXiv:2402.10251*, 2024.

[25] Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeun Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions. *GigaScience*, 9(10):giaa098, 2020.

https://doi.org/10.52202/079017-4108

[26] Elham Barzegaran, Sebastian Bosse, Peter J Kohler, and Anthony M Norcia. EEGSourceSim: A framework for realistic simulation of EEG scalp data using mri-based forward models and biologically plausible signals and noise. *Journal of Neuroscience Methods*, 328:108377, 2019.

[27] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[28] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

[29] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[30] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, 6:21002, 2012.

[31] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[33] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[34] Cristian D Guerrero-Mendez and Andres F Ruiz-Olaya. Coherence-based connectivity analysis of EEG and EMG signals during reach-to-grasp movement involving two weights. *Brain-Computer Interfaces*, 9(3):140–154, 2022.

[35] Fei Ye, JinSuo Ding, Kai Chen, and Xugang Xi. Investigation of corticomuscular functional coupling during hand movements using vine copula. *Brain Sciences*, 12(6):754, 2022.

[36] Jinbiao Liu, Yixuan Sheng, and Honghai Liu. Corticomuscular coherence and its applications: a review. *Frontiers in Human Neuroscience*, 13:100, 2019.

[37] Jason P Gallivan, Craig S Chapman, Daniel M Wolpert, and J Randall Flanagan. Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, 19(9):519–534, 2018.

[38] Tianyu Jia, Jingyao Sun, Ciarán McGeady, Linhong Ji, and Chong Li. Enhancing brain–computer interface performance by incorporating brain-to-brain coupling. *Cyborg and Bionic Systems*, 5:0116, 2024.

[39] Richard P Dum and Peter L Strick. Motor areas in the frontal lobe: the anatomical substrate for the central control of movement. In *Motor Cortex in Voluntary Movements*, pages 3–48. CRC Press, 2004.

[40] Jingyao Sun, Tianyu Jia, Zhibin Li, Chong Li, and Linhong Ji. Enhancement of EEG–EMG coupling detection using corticomuscular coherence with spatial–temporal optimization. *Journal of Neural Engineering*, 20(3):036001, 2023.

[41] Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, and Malte Ebner et al. Lightly. *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020.

https://doi.org/10.52202/079017-4108

A Additional Implementation Details of FMCA-T

Adaptive estimators. Since autocorrelation functions (ACF) and crosscorrelation functions (CCF) are expectations of inner products, we prefer to compute these expectations and the cost gradient using an adaptive filter rather than a batch of data. For any parameter θ, the partial derivative of the cost $r_T(\boldsymbol{f}_\theta, \boldsymbol{g}_\omega) := Trace(\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal)$ has the following form:

$$\begin{aligned}\frac{\partial r_T}{\partial \theta} = &-Trace(\boldsymbol{R}_F^{-1}\frac{\partial \boldsymbol{R}_F}{\partial \theta}\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal) + Trace(\boldsymbol{R}_F^{-1}\frac{\partial \boldsymbol{P}}{\partial \theta}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal) \\ &-Trace(\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\frac{\partial \boldsymbol{R}_G}{\partial \theta}\boldsymbol{R}_G^{-1}\boldsymbol{P}^\intercal) + Trace(\boldsymbol{R}_F^{-1}\boldsymbol{P}\boldsymbol{R}_G^{-1}\frac{\partial \boldsymbol{P}^\intercal}{\partial \theta}).\end{aligned} \tag{8}$$

Observe that terms needed for the gradient have two classes: terms $\boldsymbol{R}_F, \boldsymbol{R}_G, \boldsymbol{P}$, as well as their inverse, and terms of the derivatives $\frac{\partial \boldsymbol{R}_F}{\partial \theta}, \frac{\partial \boldsymbol{R}_G}{\partial \theta}, \frac{\partial \boldsymbol{P}}{\partial \theta}$. We use a smoothing window over iterations to estimate the $\boldsymbol{R}_F, \boldsymbol{R}_G, \boldsymbol{P}$, and then use the estimated values of $\widetilde{\boldsymbol{R}_F}, \widetilde{\boldsymbol{R}_G}, \widetilde{\boldsymbol{P}}$ as the matrices and their inverse in the cost. For derivative terms $\frac{\partial \boldsymbol{R}_F}{\partial \theta}, \frac{\partial \boldsymbol{R}_G}{\partial \theta}, \frac{\partial \boldsymbol{P}}{\partial \theta}$, we use the derivatives of the batch. We choose smoothing coefficient $\beta = 0.9$ across all experiments.

Important parameters. When computing the matrix inverse of ACFs ($\widetilde{\boldsymbol{R}_F}^{-1}$ and $\widetilde{\boldsymbol{R}_G}^{-1}$), a small diagonal matrix scaled by a regularization parameter, $\epsilon\boldsymbol{I}$, is added. This constant is important for training stability. We choose $\epsilon = 10^{-5}$ across all experiments. The output dimension of the network is chosen as $K = 128$, which is also the number of eigenfunctions.

Pseudo-code of the algorithm. We also explain our algorithm with pseudo-code. Algorithm 1 shows the pseudo-code for FMCA-T with adaptive estimators and computing eigenfunctions. Algorithm 2 shows how to generate channel-level and temporal-level dependencies.

Algorithm 1: FMCA-T w/ Adaptive Estimators.

```
Alg. Adaptive matrix estimators
    Input: Batch estimation $\boldsymbol{A}$; Tracking matrix $\acute{\boldsymbol{A}}$; Iteration $i$;
    Step 1. $\acute{\boldsymbol{A}} \leftarrow \beta \cdot \acute{\boldsymbol{A}}$ + (1-$\beta$) $\cdot$ $\boldsymbol{A}$
    Step 2. $\widetilde{\boldsymbol{A}} = \acute{\boldsymbol{A}}/(1-\beta^i)$
    return Tracking matrix $\acute{\boldsymbol{A}}$; Smoothed estimation $\widetilde{\boldsymbol{A}}$
Alg. FMCA-T: Optimize matrix trace cost
    Initialize: Neural networks $\{\boldsymbol{f}_\theta, \boldsymbol{g}_\omega\}$
    Initialize: Tracking matrices $\acute{\boldsymbol{R}}_{\boldsymbol{F}}$, $\acute{\boldsymbol{R}}_{\boldsymbol{G}}$, $\acute{\boldsymbol{P}}$
    for $i = 1, 2, \cdots$ do
        Sample a batch of signals $\{X_n, Y_n\}_{n=1}^{bs}$;
        Compute the ACF and CCF $\{\boldsymbol{R}_{\boldsymbol{F}}, \boldsymbol{R}_G, \boldsymbol{P}\}$ with this batch;
        Apply adaptive estimators, obtain $\{\widetilde{\boldsymbol{R}_{\boldsymbol{F}}}, \widetilde{\boldsymbol{R}_{\boldsymbol{G}}}, \widetilde{\boldsymbol{P}}\}$, and update $\{\acute{\boldsymbol{R}}_{\boldsymbol{F}}, \acute{\boldsymbol{R}}_{\boldsymbol{G}}, \acute{\boldsymbol{P}}\}$.
        Estimate gradients with smoothed estimator, update networks.
    end
    return $\theta, \omega$
Alg. Retrieve eigenfunctions and density ratios
    Input: Trained networks $\boldsymbol{f}_\theta$ and $\boldsymbol{g}_\omega$. Dataset $\{X_n, Y_n\}_{n=1}^N$.
    Step 1. Obtain outputs of networks. $\{\boldsymbol{f}_{\boldsymbol{\theta}}(X_n), \boldsymbol{g}_{\boldsymbol{\theta}}(Y_n)\}$.
    Step 2. Re-estimate $\boldsymbol{R}_F$, $\boldsymbol{R}_G$, $\boldsymbol{P}$ using these outputs.
    Step 3. Normalization for orthonormality: $\overline{\boldsymbol{f}_\theta} = \boldsymbol{R}_F^{-\frac{1}{2}}\boldsymbol{f}_\theta, \overline{\boldsymbol{g}_\omega} = \boldsymbol{R}_G^{-\frac{1}{2}}\boldsymbol{g}_\omega$
    Step 4. Compute SVD: $\overline{\boldsymbol{P}} = \mathbb{E}[\overline{\boldsymbol{f}_\theta}(\boldsymbol{X})\overline{\boldsymbol{g}_\omega}^\intercal(\boldsymbol{X})] = \boldsymbol{U}\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{V}$
    Step 5. Normalization for invariance: $\widehat{\boldsymbol{f}_\theta} = \boldsymbol{U}^\intercal\overline{\boldsymbol{f}_\theta}, \ \ \widehat{\boldsymbol{g}_\omega} = \boldsymbol{V}^\intercal\overline{\boldsymbol{g}_\omega}$
    Step 6. Construct the density ratio: $\widehat{\rho} = \widehat{\boldsymbol{f}_\theta}^\intercal\boldsymbol{S}^{\frac{1}{2}}\widehat{\boldsymbol{g}_\omega} \approx \rho$.
    return Eigenfunctions $\widehat{\boldsymbol{f}_\theta}$, $\widehat{\boldsymbol{g}_\omega}$, eigenvalues $\boldsymbol{S} = diag(\lambda_1, \cdots, \lambda_K)$, density ratio $\widehat{\rho}$
```

https://doi.org/10.52202/079017-4108

Algorithm 2: Generate channel-level and temporal-level dependencies.

```
Alg. Generate temporal and channel features:
    Input: Trained network $\boldsymbol{f}_\theta$ for EEG, consisting of a temporal network $\mathcal{F}_{TN}$ and a
           channel network $\mathcal{F}_{CN}$; An arbitrary EEG trial $X$
    Step 1. Apply $\mathcal{F}_{TN}$ to each channel: $Z_c = \mathcal{F}_{TN}(X(c)),\ c = 1, \cdots, C$
    Step 2. Apply $\mathcal{F}_{CN}$ to all channel features: $Z_F = \mathcal{F}_{CN}([Z_1, \cdots, Z_C]^\intercal)$
    Step 3. Extract internal features of $\mathcal{F}_{CN}$: $\boldsymbol{Z}_{c,1}, \cdots, \boldsymbol{Z}_{c,S}$ from $S$ convolution blocks
    return Temporal features $\boldsymbol{Z}_{c,s}$, channel features $Z_c$, global features $Z_F$

Alg. Generate channel-level dependencies
    Input: Temporal, channel, global features from the dataset
    foreach channel $c$ do
        Compute ACF and CCF for global and channel features, $Z_F$ and $Z_c$
        Use ACF and CCF to compute density ratio $\widehat{\rho_{C,F}}(c)$ with normalization
    end
Alg. Generate temporal-level dependencies:
    foreach channel $c$ and layer $s$ do
        Compute ACF and CCF for temporal features between two layers, $\boldsymbol{Z}_{c,s}$, $\boldsymbol{Z}_{c,s+1}$
        Use ACF and CCF to compute density ratio $\widehat{\rho_{s-1,s,c}}(\tau_1, \tau_2)$ between two layers.
    end
    Initialize $\widehat{\varrho_{S,c}}(\tau) = \widehat{\rho_{C,F}}(c)$; Initialize each $\widehat{\varrho_{s,c}}(\tau)$ to have length $T_s$.
    for $s = S-1, \cdots, 1$ do
        For every $\boldsymbol{Z}_s(\tau)$, find elements in layer $s+1$ that are mapped from it (set $I_s$)
        Aggregate density ratios $\widehat{\varrho_{s,c}}(\tau_1) = \sum_{\tau_2 \in I_s} \widehat{\varrho_{s+1,c}}(\tau_2)\widehat{\rho_{s,s+1,c}}(\tau_1, \tau_2)$,
    end
    return Channel dependence $\widehat{\rho_{C,F}}$ and spatio-temporal dependence $\widehat{\varrho_{s,c}}(\tau)$ that can be
      applied to any arbitrary EEG trial $X$.
```

B Additional Experiments

Visualization of EEG's eigenfunctions. Building on Fig. 4 from the main paper, we further present results for the visualization of EEG's eigenfunctions on all 25 participants. In Fig. 7, each trial is color-coded by the estimated density ratios for all participants (`SUB1` to `SUB25`). In Fig. 8, each trial is color-coded by the label of their movements.

It can be observed that for each participant, the t-SNE projections form individual clusters, each corresponding to one of the three main movements during a session. The density ratios are highly similar within each cluster (intra-cluster) while varying across clusters (inter-cluster). This indicates that using density ratio as a dependence measure effectively captures each movement individually.

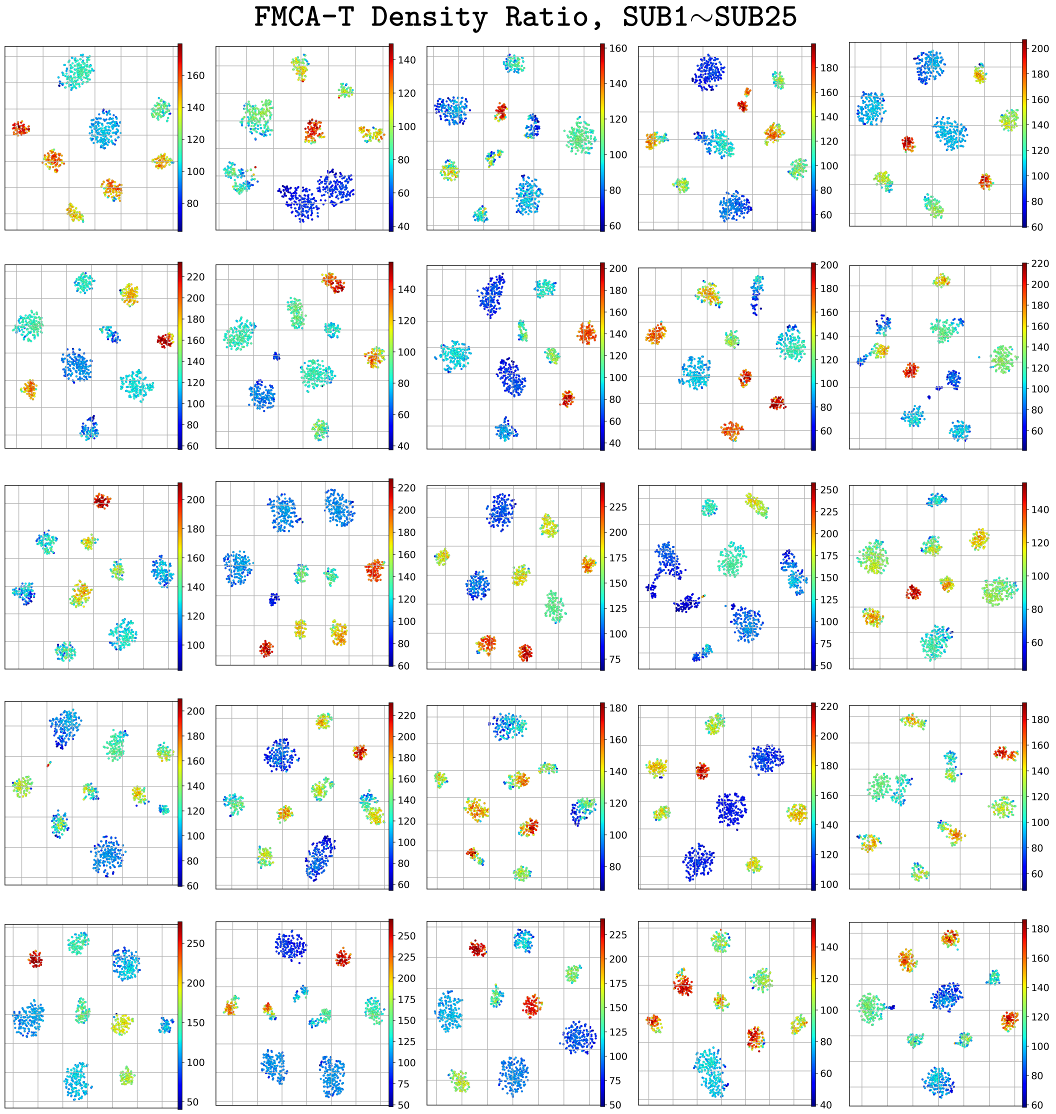


Figure 7: t-SNE visualization of eigenfunctions from all trials of 25 participants, color-coded by their corresponding density ratios. Each subplot represents one participant. The density ratios for the three movements (reaching, grasping, and twisting) show consistent values: reaching is the lowest, grasping is in the middle, and twisting is the highest. For cluster labels, refer to Fig. 8.

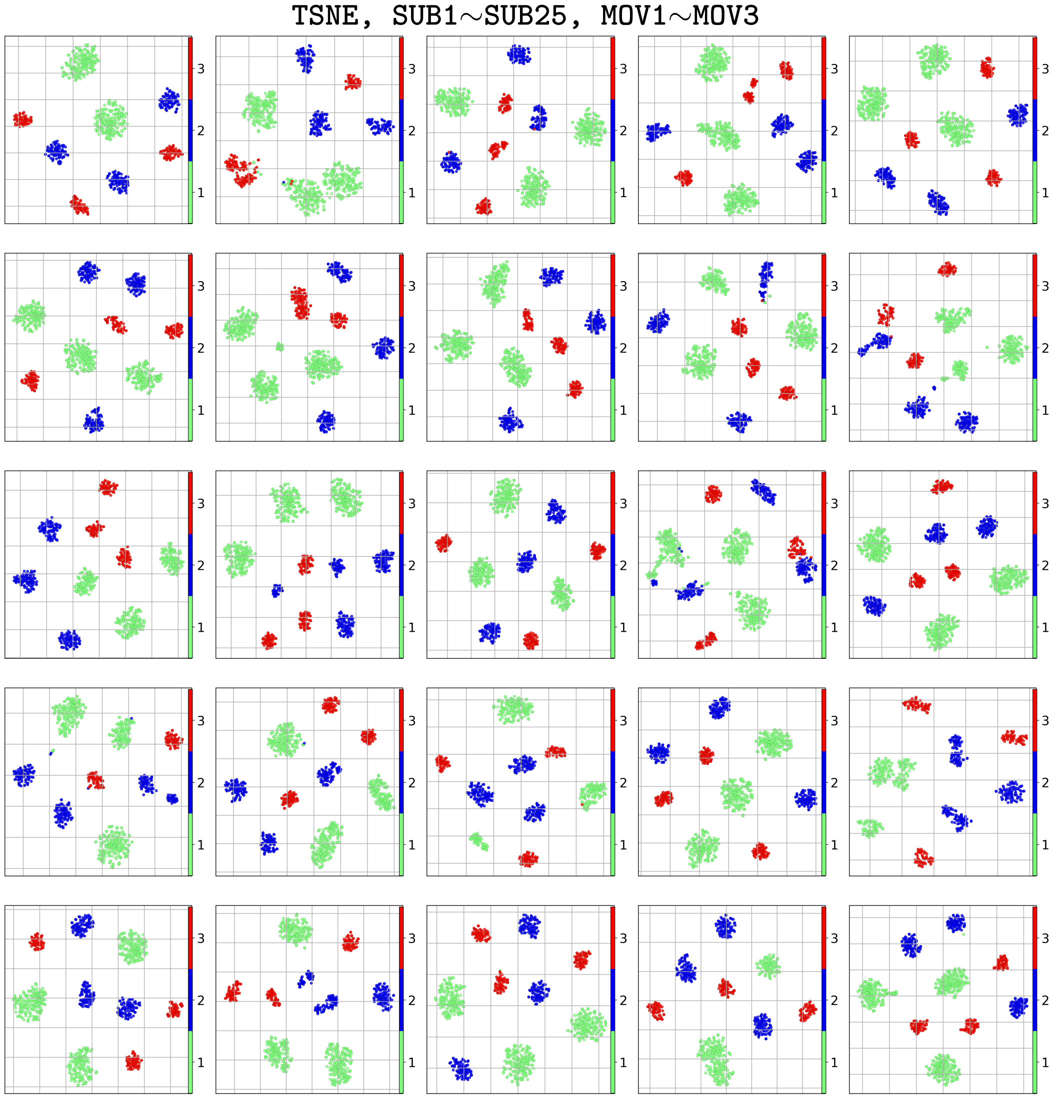


Figure 8: t-SNE visualization of all trials from 25 participants, color-coded by the type of movement: reaching (green), grasping (blue), and twisting (red). Each participant has distinct clusters, with each cluster corresponding to one of the three types of movements observed during the sessions.

From the figure, we observe that the reaching movement consistently has the lowest dependence, twisting the highest, and grasping falling in between. This pattern matches the differences and complexity of different movement tasks [34, 35]. This variability reflects the differences in connectivity between the brain and target muscles, which is the basis of movement recognition and measurement by using CMC [40].

Mean and variance of density ratios in clusters. Same as in Fig. 4 in the main paper, we extract the nine clusters shown in the t-SNE projections (Fig. 8) and compute the mean and standard deviation of the density ratios for each cluster (C1∼C9). This process is repeated for all 25 participants (SUB1∼SUB25), with results shown in Fig. 9. The results further demonstrate the effectiveness of using density ratio as a dependence measure, as the density ratios are similar within each cluster but vary across different clusters (movements and sessions).

Participant identification. Similar to Fig. 4 in the main paper, where we visualize the projected EEG eigenfunctions from 10 subjects during the reaching movement, we extend this analysis to include all three movements: reaching, grasping, and twisting. The projections shown in Fig. 9 illustrate that participant information is consistently contained in the projection space of the eigenfunctions across all movements, not just limited to reaching.

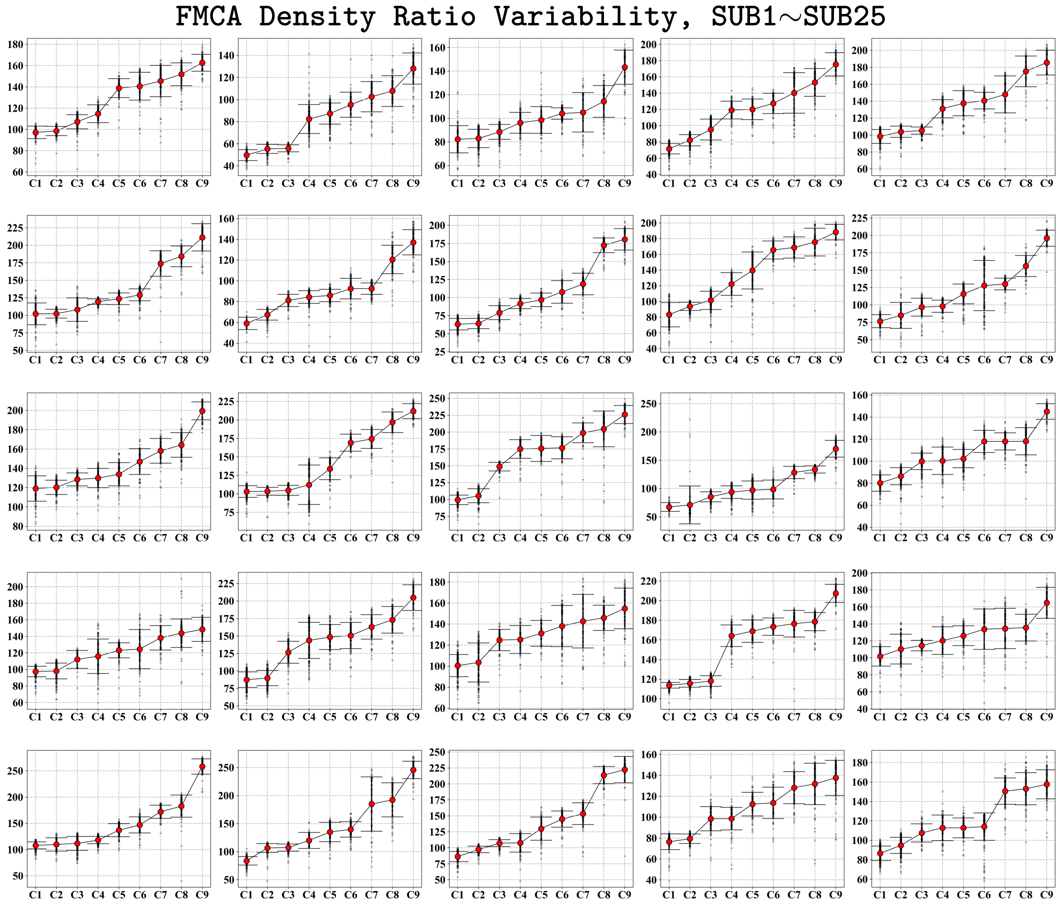


Figure 9: Similar to Fig. 4, we extract the density ratios of each cluster for the subject, corresponding to one movement and one session, rank their means, and plot them along with their standard deviations. It can be seen that for each subject, the density ratio value for each cluster is different, indicating that our measures provide a unique measurement for each movement of the subject. This figure is extended from Fig. 4 and shows that this observation applies to all participants.

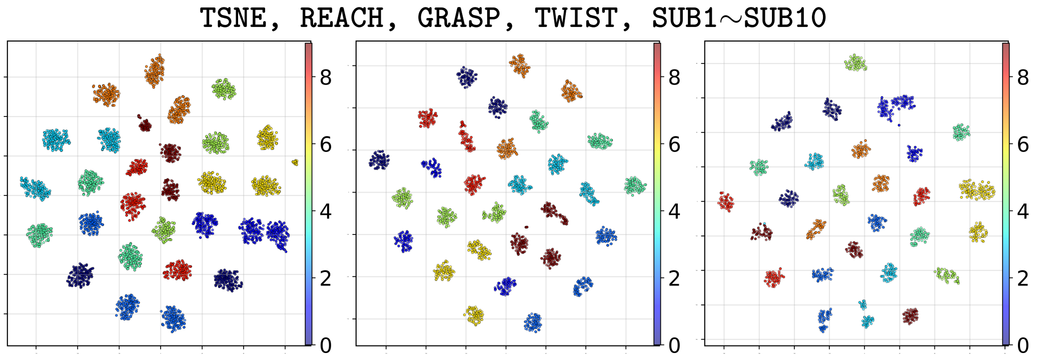
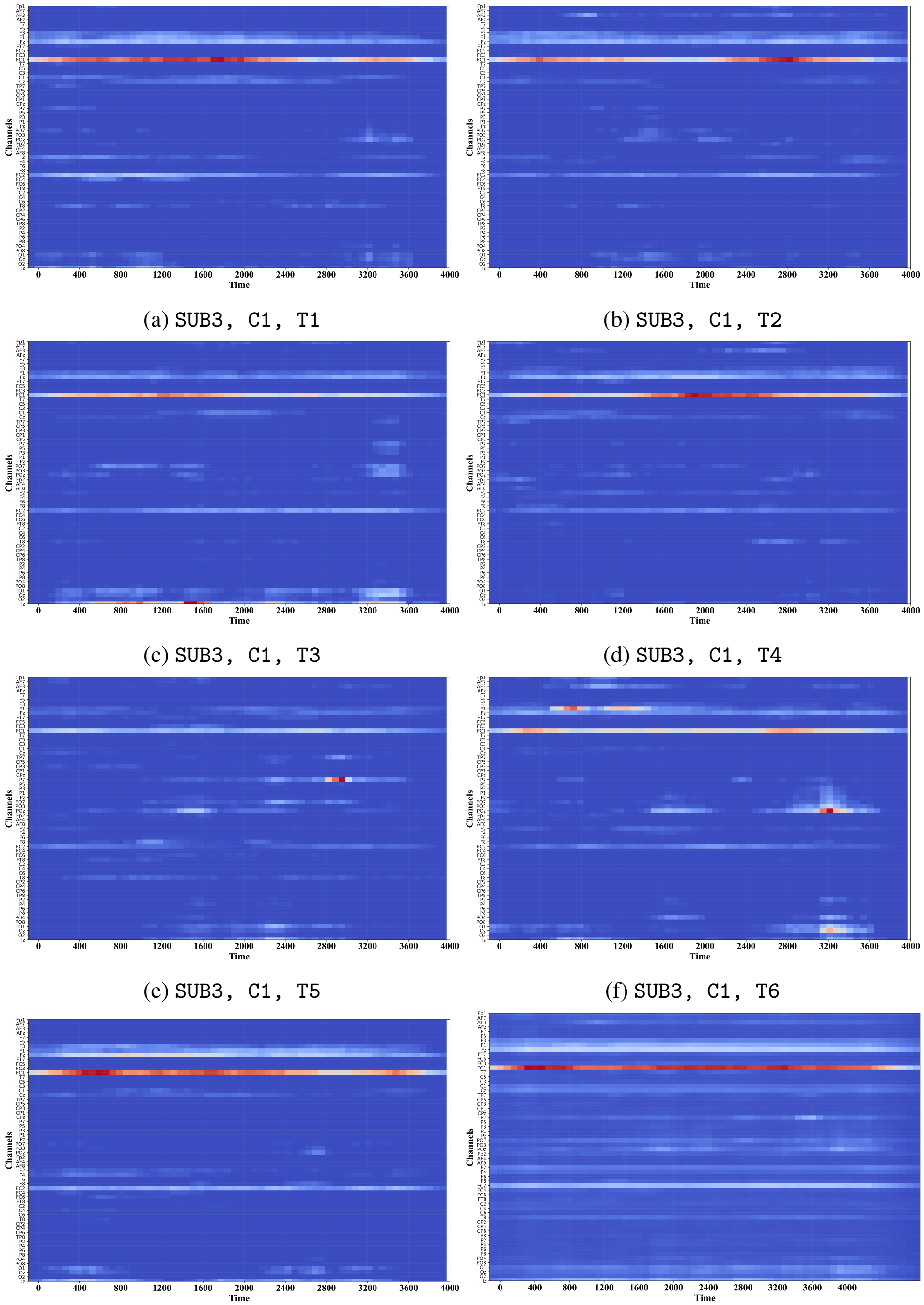


Figure 10: t-SNE projection of trials from ten participants. Subplots from left to right show the projected EEG eigenfunctions from reaching, grasping, and twisting movements, respectively. In all movements, participant-specific information is consistently captured in the clusters of the EEG's eigenfunctions.

Temporal-level dependence. Extending the analysis in Fig. 5 of the main paper, where the temporal-level dependence was shown for a single subject, we randomly select another seven trials from subject `SUB3`'s reaching movement (`C1`) and visualize the temporal-level dependence in Fig. 11. We also plot the average temporal dependence across all trials in Fig. 11h. We find consistent activations of fronto-central (FC) channel during the 4-second movement in each trial. When we compare each trial's result with the averaged one, still we observe that the activation patterns are highly similar. This indicates that our dependence measure captures information that is consistent and generalizable across subjects.



(a) `SUB3, C1, T1`

(b) `SUB3, C1, T2`

(c) `SUB3, C1, T3`

(d) `SUB3, C1, T4`

(e) `SUB3, C1, T5`

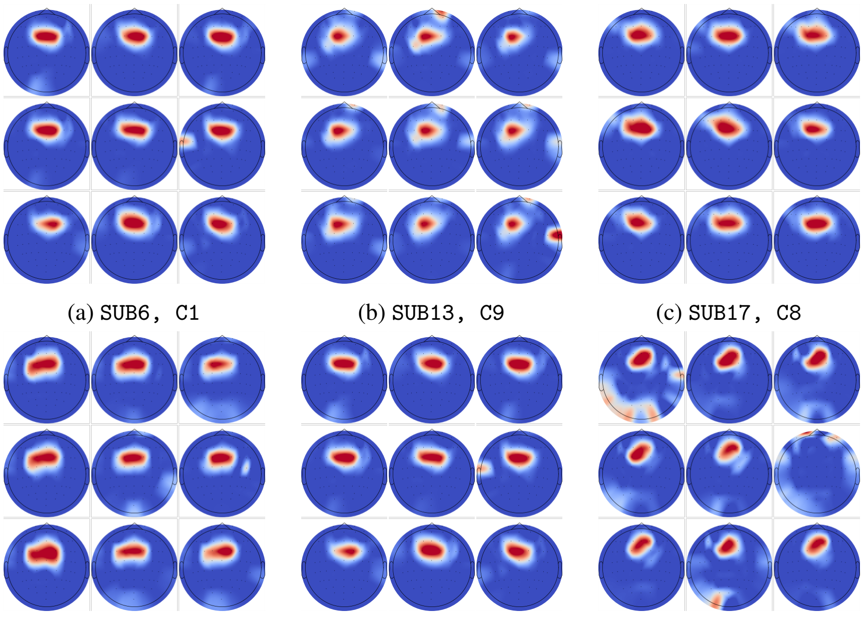
(f) `SUB3, C1, T6`

(g) `SUB3, C1, T7`

(h) `SUB3, C1, Averaged`

Figure 11: Temporal-level dependence for nine trials from `SUB3`'s reaching movement, confirming consistent activation of the frontal brain region and stable temporal dependence over the movement.

Channel-level dependence. Similar to Fig. 5 of the main paper, we also quantify the channel-level statistical dependence from other subjects, not just `SUB3`. We randomly select clusters from subjects and visualize the results in Fig. 12. We find a consistent pattern across subjects that the FC channels are activated most strongly. It can also be observed that within each cluster (each subplot), channel activations are highly similar across trials. This indicates that our dependence measure robustly captures the cortical-muscular connectivity of the same movement.

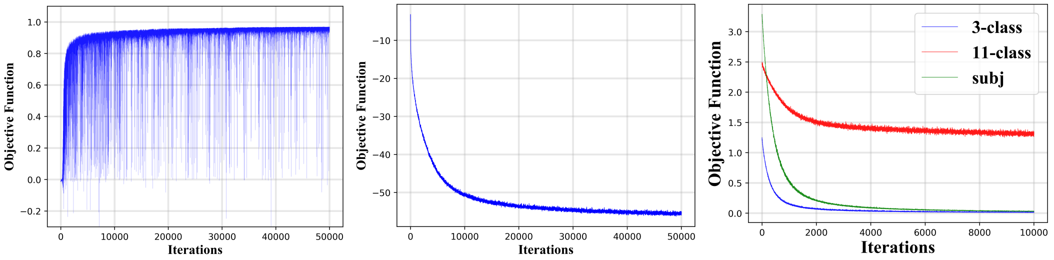


(a) `SUB6, C1` (b) `SUB13, C9` (c) `SUB17, C8`

(d) `SUB17, C4` (e) `SUB6, C1` (f) `SUB24, C2`

Figure 12: Channel-level dependence for visualizing activation patterns. We find that there are strong activations around the FC area across subjects, not just `SUB3`. We show that multiple clusters from various subjects demonstrate similar activation channels in the FC area. This suggests that these channels are overall the most important to classifying movements and contribute the most to connectivity.

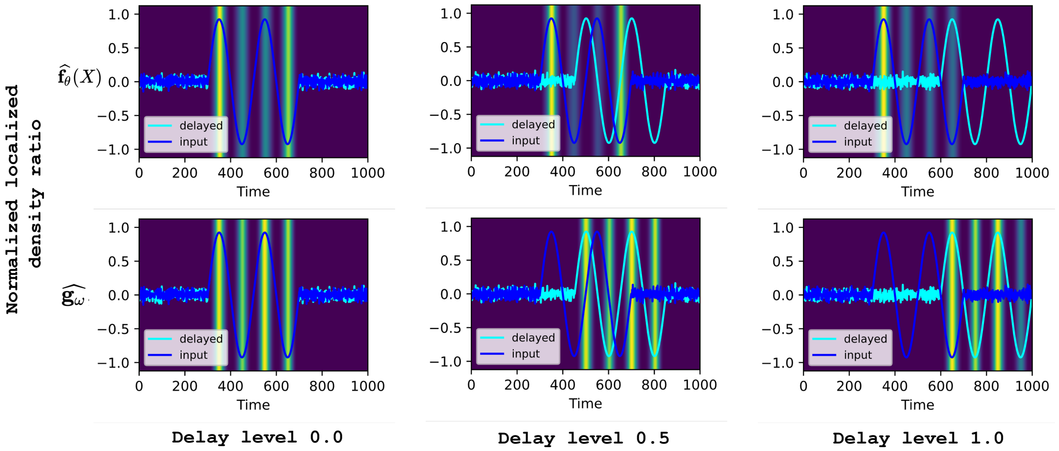
Learning curve comparison. We show the smoothness of the training stage of FMCA-T, comparing its learning curve with the learning curve of MINE when applied both on EEG-EMG-Fusion. As shown in Fig. 13, FMCA-T demonstrates superior stability, whereas MINE suffers from greater instability even when smoothing windows are applied to estimate the gradient of the variational cost in the Donsker-Varadhan representation.



(a) `MINE` (b) `FMCA-T` (c) `Eigenfunction Classifiers`

Figure 13: Comparison of learning curves on EEG-EMG-Fusion dataset. MINE: Variational cost; FMCA-T: Matrix trace cost; Eigenfunction Classifiers: Training errors. MINE suffers from high noise even when smoothing windows are applied to estimate the gradient of the variational cost. MINE is unable to produce stable and comparable results on SinWav. The classifier of eigenfunctions is trained separately from the rest of the networks.

Temporal-level dependence on SinWav. We analyze the temporal activations of the learned dependence measure on SinWav by visualizing its localized density ratios. This is performed by computing the density ratios between adjacent layers of feature projectors. The layer-wise density ratios are then aggregated for visualization. We find that the localized density ratios exhibit higher activations at the hills and valleys, and correctly capture the period and phase of the sinusoids when there is an increasing delay between the two sinusoids (Fig. 14). As shown in Fig. 15, when Gaussian noise with a standard deviation equal to 1.0 is added to the clean sinusoidal signal (signal-to-noise ratio less than 0 dB), the density ratio can still correctly identify the hills and valleys. This indicates that our proposed dependence measure is robust to random noise and delay by filtering out trivial factors like noise while focusing on the primary signals.

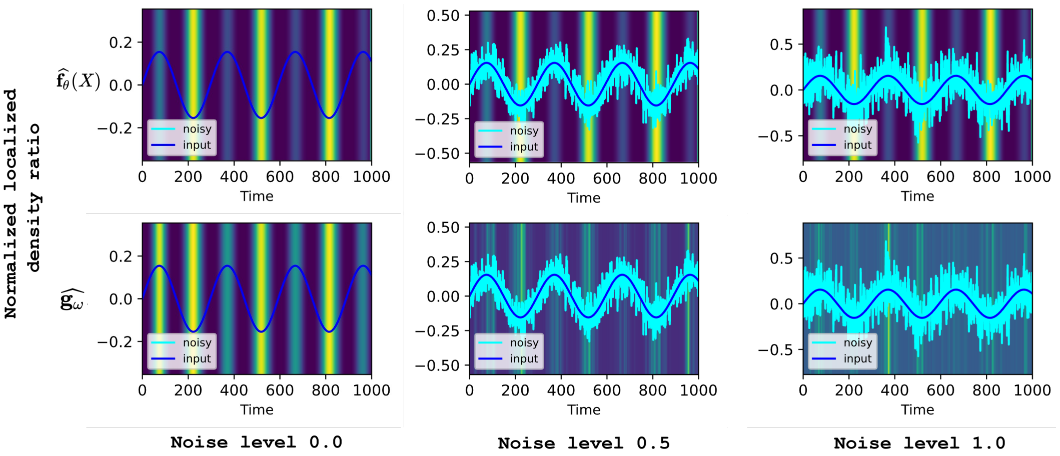


Delay level 0.0

Delay level 0.5

Delay level 1.0

Figure 14: Visualization of localized density ratios for samples from SinWav under three delay levels. The first row shows localized density ratios based on the eigenfunctions of the input (clean) signal and the second row shows the localized density ratios based on the eigenfunctions of the delayed signal. The density ratio successfully captures the period and phase of the two signals.



Noise level 0.0

Noise level 0.5

Noise level 1.0

Figure 15: Visualization of localized density ratios for samples from SinWav under three noisy levels. The first row shows localized density ratios based on the eigenfunctions of the input (clean) signal and the second row shows the localized density ratios based on the eigenfunctions of the noisy signal (Gaussian noise). The localized density ratio successfully captures the period and phase of the noisy signals even when the noise level reaches 1.0, where the signal-to-noise ratio is less than 0 dB.

C Implementation Details

This section includes details of data preprocessing, implementation of baselines, network structures, and training configurations. Our code is available at https://github.com/bohu615/corticomuscular-eigen-encoder.

C.1 EEG and EMG preprocessing

We performed standard preprocessing procedures for the 60-channel EEG signals, including 1) Down-sampling from 2500 Hz to 1000 Hz, 2) Band-pass filtering at 1-48 Hz, 3) Removal of signal artefacts with independent component analysis (ICA), and 4) Segmenting EEG at 0-4s of the onset of the movement cue for each trial. 7-channel EMG signals were preprocessed according to the following procedure: 1) Down-sampling from 2500 Hz to 1000 Hz, 2) High-pass filtering at 5 Hz, 3) Baseline correction, and 4) Segmenting EMG at 0-4s of the onset of the movement cue for each trial. The raw dataset (http://gigadb.org/dataset/100788) is distributed under a CC0 license.

C.2 Baseline implementations

CMC baselines. Cortico-muscular coherence (CMC) measures the linear synchronization between sensorimotor rhythms (present in EEG) and muscular activities (reflected in EMG) to analyze brain-muscle coupling. Given N trials, for each i-th EEG channel $X_{1:T}(n,i)$ and j-th EMG channel $Y_{1:T}(n,j)$, signals are filtered using a Butterworth filter and segmented into $\widehat{X_{t:t+\tau}}(n,i)$ and $\widehat{Y_{t:t+\tau}}(n,j)$. Coherence is computed as the correlation coefficient ($\boldsymbol{cc}$) between the spectral densities of the windows, $\boldsymbol{cc}(FX_{t:t+\tau}, FY_{t:t+\tau})$, averaged over t to $t+\tau$.

It can be seen that $\boldsymbol{cc}$ is the correlation coefficients between two variables. We replace this linear measure by nonlinear measures, such as measures from KICA and mutual information estimated by KNN (MIR), producing CMC-KICA and CMC-MIR.

MINE implementation [14]. We adapt MINE with the same topology as ours but operating on the joint space of EEG and EMG $\mathcal{X} \times \mathcal{Y}$, producing a one-dimensional output. We find that using a a sigmoid activation function stabilizes training. MINE computes $h_\theta(\boldsymbol{X}, \boldsymbol{Y})$ for samples from the joint distribution and $h_\theta(\boldsymbol{X}', \boldsymbol{Y}')$ for samples from respective marginal distributions, and minimizes the variational cost $\min_\theta \mathbb{E}[h_\theta(\boldsymbol{X}, \boldsymbol{Y})] - \log \mathbb{E}[e^{h_\theta(\boldsymbol{X}', \boldsymbol{Y}')} + 10^{-5}]$. The network is optimized by a Adam optimizer with a learning rate of 10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. MINE's trial-level solution follows the Donsker-Varadhan representation: $\log \rho(X,Y) + \gamma$, where γ can be any constant.

KICA and HSIC [17, 18]. KICA and HSIC are implemented in the following steps. First, individual Gram matrices $\boldsymbol{R}_X$ and $\boldsymbol{R}_Y$ are estimated for given Gaussian kernel $\mathcal{K}(X_i, X_j) = \mathcal{N}(X_i - X_j; \delta)$ with δ the standard deviation. Then, we construct the normalization matrix $\boldsymbol{N}_{i,j}$ and normalize the two Gram matrices as $\widehat{\boldsymbol{R}_X} = \boldsymbol{N}\boldsymbol{R}_X\boldsymbol{N}$ and $\widehat{\boldsymbol{R}_Y} = \boldsymbol{N}\boldsymbol{R}_Y\boldsymbol{N}$. For KICA-KGV, matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are constructed:

$$\begin{gathered}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & 0 \\ 0 & \boldsymbol{A}_2 \end{bmatrix}, \quad \boldsymbol{A}_1 = \widehat{\boldsymbol{R}_X}\,\widehat{\boldsymbol{R}_Y}, \quad \boldsymbol{A}_2 = \widehat{\boldsymbol{R}_Y}\,\widehat{\boldsymbol{R}_X}, \\ \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 & 0 \\ 0 & \boldsymbol{B}_2 \end{bmatrix}, \quad \boldsymbol{B}_1 = (\widehat{\boldsymbol{R}_X} + \epsilon\boldsymbol{I})(\widehat{\boldsymbol{R}_X} + \epsilon\boldsymbol{I}), \quad \boldsymbol{B}_2 = (\widehat{\boldsymbol{R}_Y} + \epsilon\boldsymbol{I})(\widehat{\boldsymbol{R}_Y} + \epsilon\boldsymbol{I}),\end{gathered} \tag{9}$$

Then, solve the generalized eigenvalue problem for KICA $\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{B}\boldsymbol{v}_i$, where $i = 1, \cdots, 2N$. This generalized eigenproblem generates $2N$ eigenvalues that are symmetric over the real line. Only N positive eigenvalues of them are used to compute the measure, obtaining KICA's Kernel Generalized Variance (KGV) measure. For HSIC, we construct matrix $\boldsymbol{C}$:

$$\boldsymbol{C} = \boldsymbol{B}_1^{-\frac{1}{2}} \boldsymbol{A}_1 \boldsymbol{B_2}^{-\frac{1}{2}}, \tag{10}$$

and solve the eigenvalue problem $\boldsymbol{C}\boldsymbol{v}_i = \sigma_i \boldsymbol{v}_i$, where $i = 1, \cdots, N$. Compute $T_{HSIC} = Trace(\boldsymbol{C})$. HSIC's measure is named the Normalized Cross-Covariance Operator (NOCCO). Hyperparameters are set as kernel size $\delta = 0.1$ and regularization constant $\epsilon = 0.1$.

https://doi.org/10.52202/079017-4108

Self-Supervised baselines: Self-Supervised Learning (SSL) methods are also implemented to compare the classification performance, including Barlow Twins [31], SimCLR [32], and VICReg [33]. We mainly use their cost functions. SSL experiments use a window size of $1,000$ with windows from the same trial as positive pairs, and windows from different trials as negative pairs. The cost, hyper-parameters, and implementations follow the Lightly package [41].

EEGNet [22]. EEGNet is a convolutional neural network commonly used for EEG signal classification. After random search hyperparameter optimization, we find optimal performance with settings close to the original paper's recommendations. Validation set is split from the training set to enable early stopping regularization. As in the original paper, we use the Adam optimizer with a learning rate of 0.0001 and a decay factor of 0.1 every 20 epochs. We train and test EEGNet on the same train-test splits as the proposed algorithm for all inter-subject and cross-subject experiments.

CSP-RLDA [30]. Common Spatial Pattern (CSP) has been proven to effectively discriminate two classes of EEG by constructing optimal spatial filters. We use CSP to extract features and Regularized Linear Discriminant Analysis (RLDA) as a classifier. If class sample volumes are unbalanced, the CSP-based classifier may be biased towards the larger sample volume category. Thus particularly for CSP, one participant with one wrist-twisting session with bad EEG quality was discarded from the analysis. We use two pairs of CSP filters and extract four feature dimensions. CSP is trained and tested in both inter-subject and cross-subject settings. To achieve three-class classification with CSP, three classifiers are trained for each pair of the three classes, and a voting strategy determines the results.

C.3 Network structures

The structure of the temporal network is illustrated in Table 2, which consists of four convolutional blocks and max pooling. Each of these blocks and max pooling are treated as a layer for computing localized density ratio responses. We apply the temporal network to each channel of the signal, obtaining $\boldsymbol{Z}_{1,9}, \boldsymbol{Z}_{2,9}, \cdots, \boldsymbol{Z}_{C,9}$, where the total number of channels is $C = 60$ for EEG and $C = 7$ for EMG and 9 indicates the ninth layer. The output of the temporal network for each channel is a vector with dimension $K = 128$. In the paper, we use the localized density ratios of $\boldsymbol{Z}_{c,6}$ to visualize the temporal resolution.

The channel network is a three-layer MLP that takes $[\boldsymbol{Z}_{1,9}, \boldsymbol{Z}_{2,9}, \cdots, \boldsymbol{Z}_{C,9}]^\intercal$ (dimension of $K \times C$) as input and also produces an output of dimension $K = 128$. Each layer uses BN and ReLU with $2,000$ units per layer. The classifier used for eigenfunctions is also a three-layer MLP with 500 units per layer.

Layer	In Ch.	Out Ch.	Kernel Size	Padding	Output
Conv, BN, ReLu	1	32	11	5	$\boldsymbol{Z}_{c,1}$
Maxpool	32	32	4	-	$\boldsymbol{Z}_{c,2}$
Conv, BN, ReLu	32	64	11	5	$\boldsymbol{Z}_{c,3}$
Maxpool	64	64	4	-	$\boldsymbol{Z}_{c,4}$
Conv, BN, ReLu	64	128	11	5	$\boldsymbol{Z}_{c,5}$
Maxpool	128	128	4	-	$\boldsymbol{Z}_{c,6}$
Conv, BN, ReLu	128	256	11	5	$\boldsymbol{Z}_{c,7}$
Maxpool	256	256	4	-	$\boldsymbol{Z}_{c,8}$
Linear BN, ReLu	256 × 15	1024	-	-	
Linear BN, ReLu	1024	512	-	-	-
Linear, Sigmoid	512	K	-	-	$\boldsymbol{Z}_{c,9}$

Table 2: Architecture of Temporal Network.

C.4 Training configurations

SinWav experiments were conducted on an NVIDIA GeForce RTX 3090. EEG-EMG-Fusion experiments were conducted on an NVIDIA GeForce A5000. Both SinWav and EEG-EMG-Fusion used an Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ for network optimization.

D Limitation

We demonstrate the effectiveness of using FMCA to learn the dependence between EEG and EMG signals and have shown that the learned eigenfunctions embed subject and movement information after optimization. We conducted the experiments on a public EEG and EMG dataset, which only contains 11 discrete upper extremity movements from 25 subjects. What we leave in the future is to use the meaningful eigenfunctions for regression tasks, i.e., continuously predicting the kinematics and contraction forces during the movement. Another limitation of the study is that we did not include patients' data due to a dearth of such large datasets that collect patients' multi-modal bio-signals. We hope that the promise offered by using our dependence measurement to evaluate cortico-muscular connectivity will further stimulate experimental research in this direction. Last, from a technical point of view, we only used convolutional neural networks with a concatenated MLP as the backbone of our networks in this study. Although we suppose that a CNN model is sufficient for the current scope as the temporal information is processed with CNN and spatial information can be leveraged by the final MLP projection layer, more advanced network structures, such as the attention in the transformer, could be potentially useful when aiming for more complex tasks.

E Broader Impact

This paper proposes to use the statistical dependence between the densities of neural data to evaluate cortico-muscular connectivity. Compared with the traditional method (CMC) that computes the linear correlation between EEG and EMG spectra, our measurement shows less variance within movement and subject while is more distinguishable across movements and subjects. This helps in exploring the neural information pathways from the brain to the muscle, which could further be used in the field of patient rehabilitation. Moreover, the learned eigenfunctions can be used as "common information" decoders, for example to decode movement and subject, and are more robust to distribution shift when tested on unseen subjects. This could be very useful for developing brain-machine interfaces under inter-subject conditions.

All datasets in this paper are publicly available and are not associated with any privacy or security concerns. Usages of the datasets strictly follow the corresponding licenses.

https://doi.org/10.52202/079017-4108

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we claim that 1) learning the dependence between EEG and EMG signals by using FMCA produces a robust dependence measure of cortico-muscular connectivity, 2) the learned eigenfunctions embed rich information of movement and subject, and 3) the temporal and spatial dependence can be visualised by computing the localized density ratios. The claims are well justified in the Sec. 2 with theoretical support and experimentally demonstrated in the Sec. 3.2, where we show that our dependence measure is more consistent than other dependence measurements and classifiers fed with EEG eigenfunctions outperform other baseline methods. We also show temporal and channel activations that indicate the temporal and spatial distribution of the density ratio in the Sec. 3.2.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Conclusion and App. D for discussions on limitations of this study.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the proof in the Sec. 2.1 to demonstrate the eigenfunctions decomposed from the density ratio form a linear span. in the Sec. 2.2, we provide proof to derive the matrix trace cost.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe matrix trace cost in detail in the Sec. 2.2, which is the critical component to reproduce our work. Network structures and training configurations are described in the App. C. Pseudocodes for the algorithms are provided in the App. A. We include an anonymous link (see App. C) that provides the source codes with all implementation details and implementation of baselines.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include an anonymous link (see App. C) that provides the source codes with all implementation details and implementation of baselines. Details of data preparation are provided in the App. C.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of experimental settings are provided in the App. C.

https://doi.org/10.52202/079017-4108

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We run experiments for five times and report the average value with standard deviation. See Table 1, Figure 3, Figure 4 in the main text and App. B for more details.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See App. C for information on computer resources.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See App. E for discussion on broader impacts of the work.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used one public EEG and EMG dataset in this study. We cited the original paper and provided the URL and the license of the asset in the App. C.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We used one toy dataset that generates pairs of sinusoids for a preliminary study. We do not regard this toy dataset as a new asset. However, we described the details of the dataset in the Sec. 3.1 and provided the codes via the link in the Appendix.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

https://doi.org/10.52202/079017-4108

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We used one public EEG and EMG dataset in this study. We have acknowledged the secondary use of this public human dataset and included the IRB approval of the original public dataset (approved by the Institutional Review Board at Korea University, 1040548-KU-IRB-17-181-A-2) in the manuscript.