Covariance Representation Analysis (CRA)
An Automatic Tool for Quality Assessment of Large-Scale EEG/MEG Data

Min-Jiun Tsai[1], Hsin-Yuan Chang[2], Ya-Lin Huang[3], Hsi-Yang Hung[4], Chun-Han Lin[4], Intan Low[5,6], Chun-Chih Huang[7], Chuan-Yu Yu[7], Tung-Ping Su[8], Jen-Chuen Hsieh[5,6,9], Li-Fen Chen[5,6,9], Chun-Shu Wei[10]

[1] *Institute of Mathematical Modeling and Scientific Computing, National Chiao Tung University, Hsinchu, Taiwan*
email: *dollars9256741.imm08g@nctu.edu.tw*

Introduction

- This work aims to develop a novel tool, covariance representation analysis (CRA) for automatic assessment of the quality of large-scale EEG/MEG data in an intuitive and instantaneous fashion.
- The clustering of EEG/MEG covariance representations provides an in-depth assessment of data quality based on the inter-/intra-subject variability.
- First, a sliding time window segments EEG/MEG data into short sections, and the covariance matrix of each section evaluates the characteristics of EEG/MEG data in a recording session. Then, the covariance matrices of EEG/MEG data across time and across subjects are mapped and visualized in a low-dimensional domain using the t-distributed stochastic neighbor embedding (t-SNE).
- Our demonstration suggests the usefulness of CRA in facilitating automatic quality assessment of large-scale EEG/MEG data.

Methodology

Band-pass filtering
In the pre-processing stage, we only apply band-pass filter with bandwidth from 1 to 40 Hz in order to preserve more original characteristics.

Time window design
We apply time window in 5-second width; each time window has an overlap of 1 second with the next time window.

Let $X_t^j \in \mathbf{R}^{ch \times k}$ be the t-th time window of the jth subject, $Cov_t^j = X_t^j X_{tT}^j \in \mathbf{R}^{ch \times ch}$ (real symmetric) be the t-th covariance matrix of the jth subject, 1_{ch} be a $ch \times ch$ matrix whose elements are all be 1.

Normalization
We provide 4 normalization methods:

- **Norm normalization**

$$Cov_t^j \leftarrow \frac{Cov_t^j}{\|Cov_t^j\|_2}$$

$$\|Cov_t^j\|_2 = \sup_{v \neq 0}\{\frac{\|Cov_t^j v\|_2}{\|v\|_2}\}$$

- **Trace normalization**
  Assume that total power of X_t^j ($trace(Cov_t^j)$) are the same for all t individually. We want to eliminate average power trend of X_t^j through this approach.

$$Cov_t^j - \frac{trace(Cov_t^j)}{ch}1_{ch}$$

- Logrithm-Trace (Log-Trace) normalization
  Log-scale version of Trace normalization, always remember that we need to take absolute value overall before taking logrithm.

$$Log(Cov_t^j) - \frac{Log(trace(Cov_t^j))}{ch}1_{ch}$$

- GeoMean-Trace normalization
  GeoMean-Trace normalization is similar to Trace normalization; Geometric mean represents average power trend of X_t^j

$$Cov_t^j - \frac{\prod_{i=1}^{ch}(Cov_t^j)_{ii}}{ch}1_{ch}$$

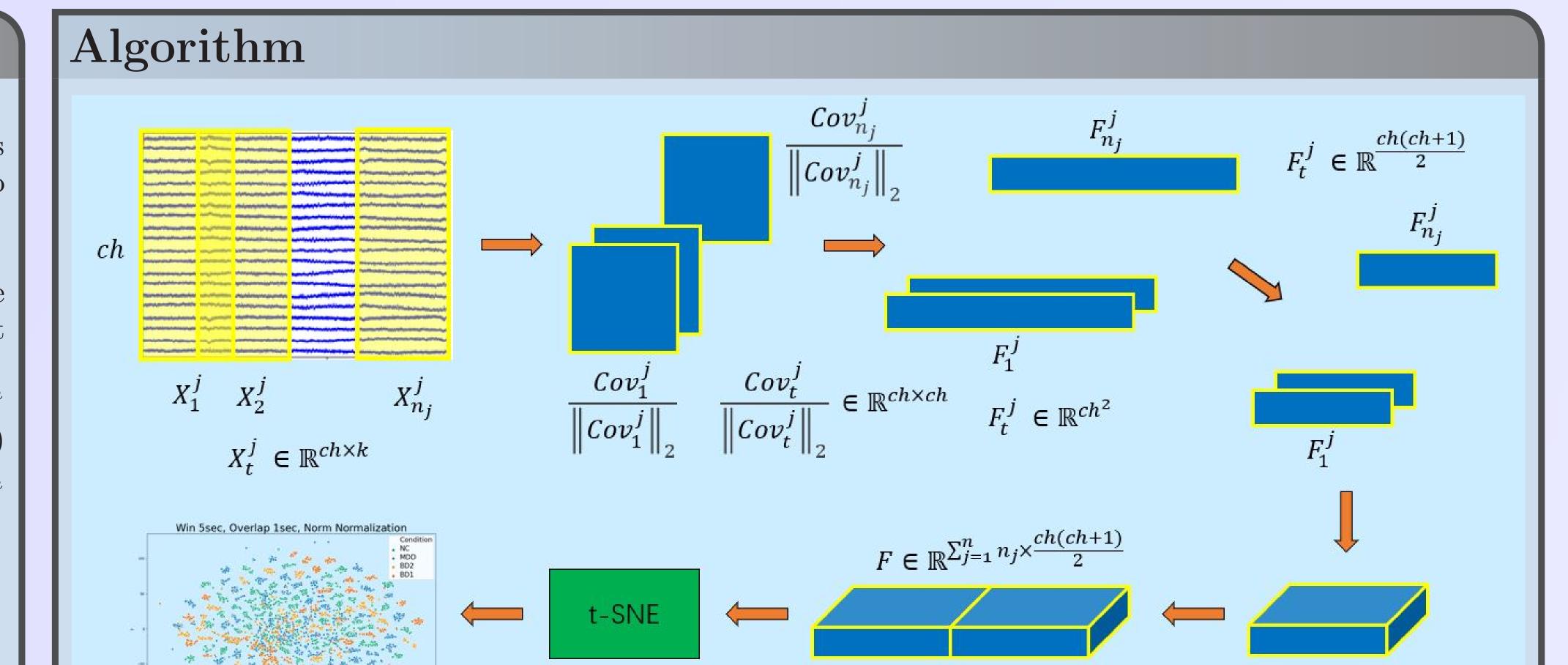### t-distributed stochastic neighbor embedding (t-SNE)
t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.

Conclusion

CRA with Norm normalization reaches the best performance overall since it could preserve individual power characteristics without scaling effect.

CRA with Trace and GeoMean-Trace normalization methods eliminate power trend individually; however, they can't deal with scaling problem appropriately.

CRA with Log-Trace normalization copes with scaling problem but captures less individual power characteristics than CRA with Norm normalization.

Algorithm

Dataset

Magnetoencephalography (MEG) dataset is provided by Brain Mapping Laboratory at National Yang-Ming University, collected from 65 normal controls (NC), 51 Major Depressive Disorder (MDD) patients, 30 Bipolar I Disorder (BD1) patients, and 30 Bipolar II Disorder (BD2) patients. In this research, We only use data recorded by 102 magnetometers and 204 gradiometers.

Cluster Validation Index

Fisher Discriminant Ratio (FDR)
Let N be the total number of clusters, M_p be the total number of element in the p-th cluster, m_p be the mass of the p-th cluster, c_p be the center of the p-th cluster, and c be the center of whole data.

$$FDR = \sqrt{\frac{\sum_{p=1}^{N} \|m_p - c\|_2^2}{\sum_{p=1}^{N}\sum_{q=1}^{M_p} \|x_{pq} - c_p\|_2^2}}$$

Silhouette Coefficient $shilhoutte(x_i)$
The Silhouette Coefficient is a measure of how similar an object x_i is to its own cluster C^k (cohesion $a(x_i)$) compared to other clusters $C^t, t \neq k$ (separation $b(x_i)$). Let $d(x_i, x_j) = \|x_i - x_j\|_2$, For any data point $x_i \in C^k$:

- **cohesion**: mean intra-cluster distance
  $a(x_i) = \frac{1}{|C^k|-1}\sum_{x_j \in C^k, x_j \neq x_i} d(x_i, x_j)$

- **separation**: mean nearest-cluster distance
  $b(x_i) = \min_{t \neq k}\{\frac{1}{|C^t|}\sum_{x_j \in Ct} d(x_i, x_j)\}$

- $-1 \leq shilhoutte(x_i) = \frac{b(x_i)-a(x_i)}{\min\{b(x_i),a(x_i)\}} \leq 1, \forall i$

Note that we calculate mean $shilhoutte(x_i)$ for whole data which is denoted by mSH.

Performance

4 Conditions Normalization	before t-SNE FDR	before t-SNE mSH	after t-SNE FDR	after t-SNE mSH
Norm	0.0028	-0.0068	0.0019	-0.0207
Trace	0.0007	-0.2289	0.0045	-0.0274
Log-Trace	0.0037	-0.0252	0.0047	-0.0326
GeoMean	0.0007	-0.2382	0.0030	-0.0232

176 Subjects Normalization	before t-SNE FDR	before t-SNE mSH	after t-SNE FDR	after t-SNE mSH
Norm	0.1133	-0.019	0.1410	0.0744
Trace	0.0191	-0.310	0.1326	-0.2310
Log-Trace	0.1236	-0.037	0.1329	-0.1671
GeoMean	0.0191	-0.255	0.1378	-0.0024

Result

The gray-shaded area is composed of data points of several subjects in the beginning of session.

For those subjects whose intra-subject cluster has better cohesion, their intra-subject clusters seem to near outside of whole data distribution. For instance, We block 26-th (in MDD-condtion) subject's data points in a red frame.

Win 5sec, Overlap 1sec, Norm Normalization