Learning to Normalize on the SPD Manifold under Bures-Wasserstein Geometry

Supplementary Material

A. Implementation details

HDM05 dataset [39] comprises 2,337 action sequences, spanning 130 distinct classes. To ensure a fair comparison, we conduct experiments using pre-processed covariance features provided by [7], which results in a reduced dataset comprising 2,086 sequences spread over 117 classes. The size of the input SPD matrices is 93×93, reflecting the 3D coordinates of 31 body joints provided in each skeleton frame. We use half of the samples for training and the rest for testing.

NTU RGB+D [50] is another action recognition dataset which comprises 25 body joints. This large dataset concludes with 56,880 videos and 60 tasks. Following the approach of [31], we use the flattened versions of 3D joint coordinates as our feature vectors. Therefore, the size of the input SPD matrices is 75×75. We also use half of the samples for training and the rest for testing.

MAMEM-SSVEP-II dataset [45] consists of time-synchronized EEG recordings from 11 participants, collected via an EGI 300 geodesic EEG system (256 channels at a 250 Hz sampling rate). For the SSVEP-based task, participants were asked to concentrate on any one of five distinct visual stimuli flickering at different frequencies (6.66, 7.50, 8.57, 10.00, and 12.00 Hz). This was done for five seconds across a series of sessions. Each session covered five cue-triggered trials for every stimulation frequency. Each trial, in turn, was broken down into four 1-second segments from the cue's onset (1s-5s). This protocol generated a total of 100 trials per session. To ensure an equitable comparison, we adhere to the data preparation and performance evaluation procedures described in [34].

Following [31], we evaluate our GBWBN under the network structure of {93, 30} and {75, 30} for the HDM05 and NTU RGB+D datasets respectively, with the GBWBN module embedded after the BiMap layer. The learning rate, training epoch, and batch size are set to $2.5e^{-3}$, 200, 30 on the HDM05 dataset, respectively. For the NTU RGB+D dataset, we set the learning rate to $1e^{-2}$, train for 100 epochs, and use a batch size of 256.

For the MAMEM-SSVEP-II dataset, the criterion used in [58] is applied for data processing. Firstly, two convolutional layers (ConvLs) are adopted at the front of SPDNet models to extract more effective spatiotemporal representations of the original EEG signals. Then, grouping in the channel dimension results in an output feature matrix of size 15×126, from which a 15×15 SPD matrix can be derived. Moreover, the designed SPDNet-GBWBN is evaluated using an architecture of {15, 12} along with the training epoch of 200, the learning rate of $2.5e^{-3}$ and the batch size of 30 on this dataset.

B. Training the GBWBN module

Noting that the bias $\boldsymbol{\mathcal{G}}$ belongs to the SPD manifold, the traditional Stochastic Gradient Descent (SGD) algorithm fails to respect its Riemannian geometry during optimization. From a geometric viewpoint, choosing the Riemannian Stochastic Gradient Descent (RSGD) algorithm [6, 13] proves to be an effective and rational strategy for optimizing manifold parameters. Let L be the loss function of the GBWBN layer, we can have the following:

$$\boldsymbol{\mathcal{G}}_{t+1} = \mathrm{Exp}_{\boldsymbol{\mathcal{G}}_t}\left(-\mu \Pi_{\boldsymbol{\mathcal{G}}_t}\left(\nabla_{\boldsymbol{\mathcal{G}}} L|_{\boldsymbol{\mathcal{G}}_t}\right)\right), \tag{26}$$

where μ denotes the learning rate, Π signifies the projection operator used to transform the Euclidean gradient $(\nabla_{\boldsymbol{\mathcal{G}}} L|_{\boldsymbol{\mathcal{G}}_t})$, and Exp is the exponential mapping shown in Tab. 3.

In previous literature, distinct formulas for the gradient of eigenvalue functions on SPD matrices have been independently established [30]. Considering an eigenvalue function: $\mathbf{Y} = f(\mathbf{X}) = \mathbf{U} f(\mathbf{\Sigma}) \mathbf{U}^{\mathrm{T}}$, where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$, the gradient of L w.r.t $\mathbf{X}$ can be computed as follows:

$$\frac{\partial L}{\partial \mathbf{X}} = \mathbf{U}\left(f'(\mathbf{X}) \odot \left(\mathbf{U}^T (\frac{\partial L}{\partial \mathbf{Y}}) \mathbf{U}\right)\right)\mathbf{U}^T, \tag{27}$$

with

$$f'(\mathbf{X})_{i,j} = \begin{cases} \frac{f(\sigma_i) - f(\sigma_j)}{\sigma_i - \sigma_j}, & \text{if } \sigma_i \neq \sigma_j. \\ f'(\sigma_i), & \text{otherwise}. \end{cases} \tag{28}$$

Eq. (27) is called the Dalecǩii-Kreǐn formula and $f'(\lambda_i, \lambda_j)$ is the first divided difference of f at (λ_i, λ_j). We acknowledge the work presented in [7] which illustrates the equivalence between the Dalecǩii-Kreǐn formula and the formula introduced in [30]. Given the superior numerical stability of the Dalecǩii-Krečformula, we utilize this formula throughout the backward pass.

Given the following Lyapunov equation:

$$\mathbf{PX} + \mathbf{XP} = \mathbf{S}, \text{ with } \mathbf{P} = \mathcal{L}_{\mathbf{X}}(\mathbf{S}), \tag{29}$$

according to [15], the gradients w.r.t. the Lyapunov operator in the backpropagation process can be computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}} = \mathcal{L}_{\mathbf{X}}(\frac{\partial \mathcal{L}}{\partial \mathbf{P}}), \tag{30}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = -\mathbf{P}\mathcal{L}_{\mathbf{X}}(\frac{\partial \mathcal{L}}{\partial \mathbf{P}}) - \mathcal{L}_{\mathbf{X}}(\frac{\partial \mathcal{L}}{\partial \mathbf{P}})\mathbf{P}. \tag{31}$$

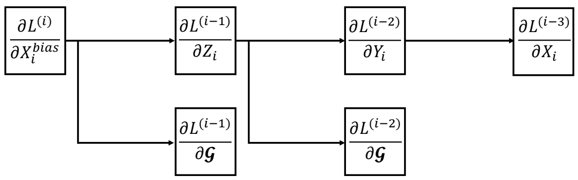

Figure 6. The backward pass of bias operation

C. Learning the bias

The bias operation $\mathbf{X}_i^{bias} = \mathrm{Exp}_{\boldsymbol{\mathcal{G}}}\left(\Gamma_{\mathbf{I}_d \to \boldsymbol{\mathcal{G}}}\left(\mathrm{Log}_{\mathbf{I}_d}\left(\mathbf{X}_i\right)\right)\right)$ can be implemented by three auxiliary layers: 1) $\mathbf{Y}_i = f^{(1)}(\mathbf{X}_i) = \mathrm{Log}_{\mathbf{I}_d}(\mathbf{X}_i)$; 2) $\mathbf{Z}_i = f^{(2)}(\mathbf{Y}_i, \boldsymbol{\mathcal{G}}) = \Gamma_{\mathbf{I}_d \to \boldsymbol{\mathcal{G}}}(\mathbf{Y}_i)$; 3) $\mathbf{X}_i^{bias} = f^{(3)}(\mathbf{Z}_i, \boldsymbol{\mathcal{G}}) = \mathrm{Exp}_{\boldsymbol{\mathcal{G}}}(\mathbf{Z}_i)$. Therefore, the bias operation can be expressed as:

$$\mathbf{X}_i \xrightarrow{f^{(1)}} \mathbf{Y}_i \xrightarrow{f^{(2)}} \mathbf{Z}_i \xrightarrow{f^{(3)}} \mathbf{X}_i^{bias}. \tag{32}$$

Let $L : \mathbb{R}^d \to \mathbb{R}$ be the loss function of the designed GBWBN. The backpropagation of the bias operation is illustrated in Fig. 6, where $L^{(i)} = L \circ f^{(K)} \circ \cdots \circ f^{(i)}$ represents the loss of the i-th layer. It is crucial to note that the computation of the partial derivatives of $f^{(3)}$ w.r.t. $\mathbf{Z}_i$ and $\boldsymbol{\mathcal{G}}$ involves the backpropagation of Lyapunov operator. The partial derivative of $f^{(2)}$ w.r.t. $\boldsymbol{\mathcal{G}}$ is computed automatically by PyTorch. Based on the above computations, we can deduce the Euclidean gradient ($\nabla_{\boldsymbol{\mathcal{G}}} L|_{\boldsymbol{\mathcal{G}}_t}$) in the proposed BWBN module. Then, Eq. (26) can be used to update the bias. Since the partial derivative of $f^{(1)}$ w.r.t. $\mathbf{X}_i$ involves the computation of square root $(\cdot)^{\frac{1}{2}}$, following SPDNetBN [7], we use Eq. (28) to solve it.

D. Proofs of the propositions and theories in the main paper

D.1. Proof of Prop. 3.1

Using Riemannian distance in Tab. 3, we can deduce that

$$\begin{aligned} d_{\mathrm{BW}}(\mathbf{X}, \mathbf{I}_d) &= \left(\mathrm{tr}(\mathbf{X}) + \mathrm{tr}(\mathbf{I}_d) - 2\mathrm{tr}(\mathbf{X}^{\frac{1}{2}}\mathbf{I}_d\mathbf{X}^{\frac{1}{2}})^{\frac{1}{2}}\right)^{\frac{1}{2}} \\ &= \left(\mathrm{tr}(\mathbf{X}) + \mathrm{tr}(\mathbf{I}_d) - 2\mathrm{tr}(\mathbf{X})^{\frac{1}{2}}\right)^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left(\left(\mathbf{X}^{\frac{1}{2}} - \mathbf{I}_d\right)^2\right)\right)^{\frac{1}{2}}. \end{aligned} \tag{33}$$

Similarly, using the exponential map and logarithmic map in Tab. 3, we can deduce that

$$\mathrm{Exp}_{\mathbf{I}_d}\mathbf{S} = \mathbf{I}_d + \mathbf{S} + \mathcal{L}_{\mathbf{I}_d}(\mathbf{S})\mathbf{I}_d\mathcal{L}_{\mathbf{I}_d}(\mathbf{S}), \tag{34}$$

$$\mathrm{Log}_{\mathbf{I}_d}\mathbf{X} = 2\mathbf{X}^{\frac{1}{2}} - 2\mathbf{I}_d. \tag{35}$$

Since $\mathbf{I}_d\mathcal{L}_{\mathbf{I}_d}(\mathbf{S}) + \mathcal{L}_{\mathbf{I}_d}(\mathbf{S})\mathbf{I}_d = \mathbf{S}$, we can obtain: $\mathcal{L}_{\mathbf{I}_d}(\mathbf{S}) = \frac{1}{2}\mathbf{S}$. Then, Eq. (34) can be reformulated as:

$$\mathrm{Exp}_{\mathbf{I}_d}\mathbf{S} = \mathbf{I}_d + \mathbf{S} + \frac{1}{4}\mathbf{S}^2 = (\mathbf{I}_d + \frac{1}{2}\mathbf{S})^2. \tag{36}$$

Combining Eq. (35) and Eq. (36), we have

$$\begin{aligned} \boldsymbol{\psi}_{\mathbf{s}}(\mathbf{X}) &= \mathrm{Exp}_{\mathbf{I}_d}\left(\mathbf{s}\mathrm{Log}_{\mathbf{I}_d}(\mathbf{X})\right) \\ &= \mathrm{Exp}_{\mathbf{I}_d}\left(\mathbf{s}\left(2\mathbf{X}^{\frac{1}{2}} - 2\mathbf{I}_d\right)\right) \\ &= \left(\mathbf{s}\left(\mathbf{X}^{\frac{1}{2}} - \mathbf{I}_d\right) + \mathbf{I}_d\right)^2 \end{aligned} \tag{37}$$

Therefore,

$$\begin{aligned} d_{\mathrm{BW}}\left(\boldsymbol{\psi}_{\mathbf{s}}\left(\mathbf{X}\right), \mathbf{I}_d\right) &= \left(\mathrm{tr}\left(\left(\mathbf{s}\left(\mathbf{X}^{\frac{1}{2}} - \mathbf{I}_d\right) + \mathbf{I}_d - \mathbf{I}_d\right)^2\right)\right)^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left(\mathbf{s}^2\left(\mathbf{X}^{\frac{1}{2}} - \mathbf{I}_d\right)^2\right)\right)^{\frac{1}{2}} \\ &= \mathbf{s}\left(\mathrm{tr}\left(\left(\mathbf{X}^{\frac{1}{2}} - \mathbf{I}_d\right)^2\right)\right)^{\frac{1}{2}} \\ &= \mathbf{s}d_{\mathrm{BW}}\left(\mathbf{X}, \mathbf{I}_d\right) \end{aligned} \tag{38}$$

D.2. Proof of Prop. 3.2

As shown in [14], when $\theta \to 0$, for all $\mathbf{X} \in \mathcal{S}_{++}^d$ and all $\mathbf{S} \in T_{\mathbf{X}}\boldsymbol{\mathcal{S}}_{++}^{\boldsymbol{d}}$, we have

$$g_{\mathbf{X}}^{(\theta)-\mathrm{GBW}}(\mathbf{S}, \mathbf{S}) \longrightarrow g_{\mathbf{I}_d}^{\mathrm{GBW}}\left(\log_{*,\mathbf{X}}(\mathbf{S}), \log_{*,\mathbf{X}}(\mathbf{S})\right). \tag{39}$$

Using Eq. (4), we can deduce that

$$\begin{aligned} g_{\mathbf{I}_d}^{\mathrm{GBW}}(\mathbf{S}, \mathbf{S}) &= \frac{1}{2}\mathrm{tr}\left(\mathcal{L}_{\mathbf{I}_d, \mathbf{M}}(\mathbf{S})\mathbf{S}\right) \\ &= \frac{1}{2}\mathrm{tr}\left(\mathcal{L}_{\mathbf{M}}(\mathbf{S})\mathbf{S}\right) \\ &= \frac{1}{2}\left\langle\mathcal{L}_{\mathbf{M}}(\mathbf{S}), \mathbf{S}\right\rangle \end{aligned} \tag{40}$$

Combining Eq. (39) and Eq. (40), we have

$$\begin{aligned} g_{\mathbf{X}}^{(\theta)-\mathrm{GBW}}(\mathbf{S}, \mathbf{S}) &\xrightarrow{\theta \to 0} g_{\mathbf{I}_d}^{\mathrm{GBW}}\left(\log_{*,\mathbf{X}}(\mathbf{S}), \log_{*,\mathbf{X}}(\mathbf{S})\right) \\ &= \frac{1}{2}\left\langle\log_{*,\mathbf{X}}\left(\mathcal{L}_{\mathbf{M}}(\mathbf{S})\right), \log_{*,\mathbf{X}}(\mathbf{S})\right\rangle \end{aligned} \tag{41}$$

D.3. Proof of Prop. 3.3

For (θ)-GBW, We have

$$\begin{aligned} & g_{\mathbf{X}}^{(\theta)-\mathrm{GBW}}(\mathbf{S}_1, \mathbf{S}_2) \\ &= \frac{1}{\theta^2} g_{\mathbf{X}^\theta}^{\mathrm{GBW}}\left(\tilde{\mathbf{S}}_1, \tilde{\mathbf{S}}_2\right) \\ &= \frac{1}{4} \cdot \frac{1}{\theta^2}\mathrm{vec}(\tilde{\mathbf{S}}_1)^\top(\tilde{\mathbf{X}} \otimes \tilde{\mathbf{X}})^{-1}\mathrm{vec}(\tilde{\mathbf{S}}_2) \\ &\overset{(1)}{=} \frac{1}{4} \cdot \frac{1}{\theta^2} g_{\tilde{\mathbf{X}}}^{\mathrm{AI}}(\mathbf{S}_1, \mathbf{S}_2) \\ &= \frac{1}{4} g_{\mathbf{X}}^{(\theta)-\mathrm{AI}}(\mathbf{S}_1, \mathbf{S}_2), \end{aligned} \tag{42}$$

where $\tilde{\mathbf{S}}_1 = (\phi_\theta)_{*,\mathbf{X}}(\mathbf{S}_1), \tilde{\mathbf{S}}_2 = (\phi_\theta)_{*,\mathbf{X}}(\mathbf{S}_2), \tilde{\mathbf{X}} = \mathbf{X}^\theta$, and (1) follows from Eq. (3).

D.4. Proof of Thm. 3.4

Our proof is inspired by [14, Thm. 5.3], which characterize the calculation of LieBN [14] under the Riemannian isometry between Lie groups. However, our RBN does not involve Lie group structures but relies on Riemannian operators. In the following, we will use the properties of Riemannian operators under the Riemannian isometry.

We denote $\mathrm{Exp}, \mathrm{Log}, \Gamma, d(\cdot,\cdot)$ and WFM are Riemannian exponentiation, logarithm, parallel transportation, geodesic distance, and weighted Fréchet mean on $(\mathcal{S}^d_{++}, g^{\mathrm{BW}})$, while $\widetilde{\mathrm{Exp}}, \widetilde{\mathrm{Log}}, \widetilde{\Gamma}, \widetilde{\mathrm{d}}(\cdot,\cdot)$ and $\widetilde{\mathrm{WFM}}$ are the counterparts on $(\mathcal{S}^d_{++}, g^{\theta\text{-GBW}})$. Since $f: (\mathcal{S}^d_{++}, g^{BW}) \to (\mathcal{S}^d_{++}, g^{\theta\text{-GBW}})$ is a Riemannian isometry, for $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}^d_{++}$, $\mathbf{S}_1, \mathbf{S}_2 \in T_{\mathbf{X}_1}\mathcal{S}^d_{++}$, and $\{\mathbf{X}_i\}_{i=1}^N \in \mathcal{S}^d_{++}$ with weights $\{\omega_i\}_{i=1}^N$ satisfying $\omega_i \geq 0$ and $\sum_{i\leq N}\omega_i = 1$ .we have the following:

$$\widetilde{\mathrm{Exp}}_{\mathbf{X}_1}\mathbf{S}_1 = f\left(\mathrm{Exp}_{f^{-1}(\mathbf{X}_1)} {f^{-1}}_{*,\mathbf{X}_1}(\mathbf{S}_1)\right), \tag{43}$$

$$\widetilde{\mathrm{Log}}_{\mathbf{X}_1}\mathbf{X}_2 = (f^{-1}_{*,\mathbf{X}_1})^{-1}\left(\mathrm{Log}_{f^{-1}(\mathbf{X}_1)} f^{-1}(\mathbf{X}_2)\right), \tag{44}$$

$$\begin{aligned} &\widetilde{\Gamma}_{\mathbf{X}_1\to\mathbf{X}_2}\mathbf{S}_1 \\ &= \left(f^{-1}_{*,\mathbf{X}_2}\right)^{-1}\left(\Gamma_{f^{-1}(\mathbf{X}_1)\to f^{-1}(\mathbf{X}_2)} {f^{-1}}_{*,\mathbf{X}_1}(\mathbf{S}_1)\right), \end{aligned} \tag{45}$$

$$\widetilde{d}(\mathbf{X}_1, \mathbf{X}_2) = d\left(f^{-1}(\mathbf{X}_1), f^{-1}(\mathbf{X}_2)\right), \tag{46}$$

$$\widetilde{\mathrm{WFM}}(\{\mathbf{X}_i\}, \{\omega_i\}) = f(\mathrm{WFM}(\{f^{-1}(\mathbf{X}_i)\}, \{\omega_i\})). \tag{47}$$

Note that:

$$\begin{aligned} &\widetilde{\mathrm{Exp}}_{\mathbf{X}_2}\left(\widetilde{\Gamma}_{\mathbf{X}_1\to\mathbf{X}_2}\widetilde{\mathrm{Log}}_{\mathbf{X}_1}(\mathbf{S}_1)\right) \\ &\overset{(1)}{=} \widetilde{\mathrm{Exp}}_{\mathbf{X}_2}\left(\widetilde{\Gamma}_{\mathbf{X}_1\to\mathbf{X}_2}(f^{-1}_{*,\mathbf{X}_1})^{-1}\left(\mathrm{Log}_{f^{-1}(\mathbf{X}_1)} f^{-1}(\mathbf{S}_1)\right)\right) \\ &\overset{(2)}{=} \widetilde{\mathrm{Exp}}_{\mathbf{X}_2}(f^{-1}_{*,\mathbf{X}_2})^{-1}\left(\Gamma_{f^{-1}(\mathbf{X}_1)\to f^{-1}(\mathbf{X}_2)}(\boldsymbol{\mathcal{A}})\right) \\ &\overset{(3)}{=} f\left(\mathrm{Exp}_{f^{-1}(\mathbf{X}_1)}\left(\Gamma_{f^{-1}(\mathbf{X}_1)\to f^{-1}(\mathbf{X}_2)}(\boldsymbol{\mathcal{A}})\right)\right), \end{aligned} \tag{48}$$

where $\boldsymbol{\mathcal{A}} = \mathrm{Log}_{f^{-1}(\mathbf{X}_1)} f^{-1}(\mathbf{S}_1)$.

The derivation comes from the following.

(1) follows from Eq. (44).

(2) follows from Eq. (45).

(3) follows from Eq. (43).

Then, we denote Eq. (11) and Eq. (13) on $(\mathcal{S}^d_{++}, g^{BW})$ of the main paper as $\xi(\cdot|\boldsymbol{\mathcal{B}}, \nu^2, \boldsymbol{\mathcal{G}}, \mathrm{s})$, while $\widetilde{\xi}(\cdot|\boldsymbol{\mathcal{B}}, \nu^2, \boldsymbol{\mathcal{G}}, \mathrm{s})$ is the counterpart on $(\mathcal{S}^d_{++}, g^{\theta\text{-GBW}})$. We can deduce that:

$$\begin{aligned} &\widetilde{\xi}(\mathbf{X}_i|\boldsymbol{\mathcal{B}}, \nu^2, \boldsymbol{\mathcal{G}}, \mathbf{s}) \\ &= f\left(\xi\left(f^{-1}(\mathbf{X}_i)|f^{-1}(\boldsymbol{\mathcal{B}}), \nu^2, f^{-1}(\boldsymbol{\mathcal{G}}), \mathbf{s}\right)\right). \end{aligned} \tag{49}$$

Since f is a Riemannian isometry, we can directly deduce that the Fréchet variance and Fréchet mean are both the same. Therefore, we can obtain that:

$$\begin{aligned} &(\theta)\text{-GBWBN}(\mathbf{X}_i, \boldsymbol{\mathcal{G}}, \omega, \boldsymbol{\epsilon}, \mathbf{s}) \\ &= f\left(\mathrm{BWBN}\left(\boldsymbol{f^{-1}}(\mathbf{X}_i), f^{-1}(\boldsymbol{\mathcal{G}}), \omega, \boldsymbol{\epsilon}, \mathbf{s}\right)\right). \end{aligned} \tag{50}$$

E. EEG model interpretation

Fig. 7 and Fig. 8 are the visualization results generated by the SPDNet-BN [7]. Compared with Fig. 2 and Fig. 3, the gradient responses of our method and SPDNet-BN are both concentrated in the OZ channel, clearly demonstrating the effectiveness of the Riemannian method. However, our method shows a concentration between 0.25 and 0.60 seconds, dovetailed with previous studies on the relationship between SSVEP and Oz in EEG research [23, 26], whereas SPDNet-BN exhibits a less focused response. This observation further supports that the proposed power-deformed GBWM-based RBN can extract more essential geometric features than AIM-based methods.

Fig. 9 illustrates the spatial distribution across distinct epochs corresponding to each visual stimulus in the MAMEM-SSVEP-II dataset. Each epoch displays similar spatial topo maps with a consistently strong gradient response located at the Oz spanning all epochs. Moreover, despite varying frequencies of visual stimulation, the gradient response location across the scalp remained steady throughout the given period.

In conclusion, SPDNet-GBWBN can identify subtle discrepancies hidden within similar spatial distributions represented by the topographic map of each epoch among the five frequencies used to decode SSVEP-EEG signals. The visualization results validate the proficiency and capability of the proposed SPDNet-GBWBN in identifying and capturing the elusive non-stationarity observed in dynamic brain activity.

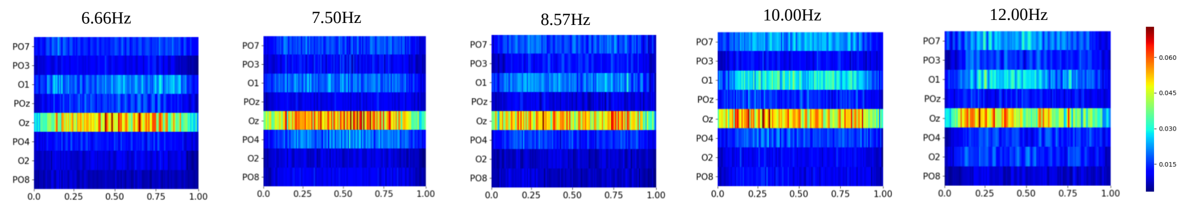

Figure 7. The heatmaps of five frequency classes of the MAMEM-SSVEP-II dataset demonstrated by SPDNet-BN.

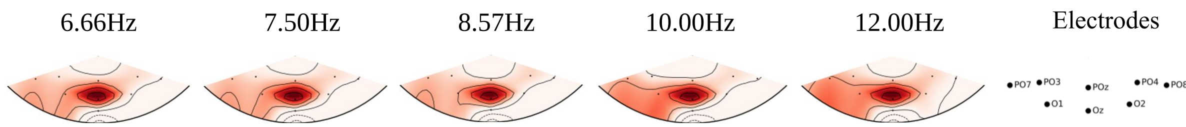

Figure 8. The brain topology maps obtained on the MAMEM-SSVEP-II dataset demonstrated by SPDNet-BN.

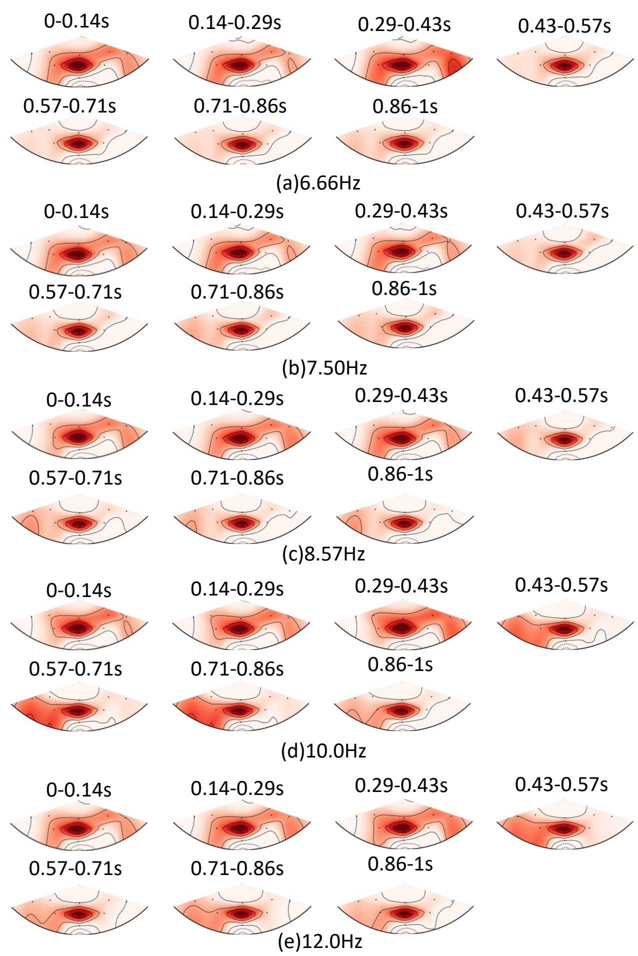

Figure 9. The spatial topomaps at different epochs and frequencies of visual stimulation for the S11 model on the MAMEM-SSVEP-II dataset (dark red signifies strong gradient response).