# Supplementary Material

## A Proof of Theorem 1

First, consider any real vector $\mathbf{s} \in \mathbb{R}^d$ and denote its radius and corresponding unit vector by $r := \|\mathbf{s}\|$ and $\mathbf{u} := \mathbf{s}/\|\mathbf{s}\|$. Then it is well-known that an integral of a density function $f(\mathbf{s})$ over $\mathbb{R}^d$ satisfies

$$\int_{\mathbb{R}^d} f(\mathbf{s})\,\mathrm{d}\mathbf{s} = \int_0^\infty \int_{U^{d-1}} f(r\mathbf{u}) r^{d-1}\,\mathrm{d}r\mathrm{d}\mathbf{u}, \tag{1}$$

where the integral about $\mathbf{u}$ is over the unit hypersphere $U^{d-1} := \{\mathbf{u} \in \mathbb{R}^d \mid \|\mathbf{u}\| = 1\}$. Notice that $r^{d-1}\mathrm{d}r\mathrm{d}\mathbf{u}$ is the infinitesimal volume element at $\mathbf{s} = r\mathbf{u}$, which depends on the radius $r$. The factor $r^{d-1}$ came from the fact that the surface area of hypersphere of radius $r$ equals $2\pi^{d/2}r^{d-1}/\Gamma(d/2)$.

By generalizing this fact to our setting which involves multiple vectors $\mathbf{s}_{[j]}$, we can readily obtain

$$p_{\mathbf{s}}(\{\mathbf{s}_{[j]}\}) \prod_j \mathrm{d}\mathbf{s}_{[j]} = p(\{r_j\}, \{\mathbf{u}_{[j]}\}) \prod_j \mathrm{d}r_j \mathrm{d}\mathbf{u}_j, \tag{2}$$

where the pdf in the right-hand side is given by

$$p(\{r_j\}, \{\mathbf{u}_{[j]}\}) = p_{\mathbf{s}}(\{r_j\mathbf{u}_{[j]}\}) \prod_j r_j^{d_j-1}. \tag{3}$$

Now denote by $\alpha_d := 2\pi^{d/2}/\Gamma(d/2)$ the surface area of unit hypersphere $U^{d-1}$, where $\Gamma$ denote the Gamma function, and let $p_{\mathbf{r}}(\cdot)$ be the joint pdf of the $m$ radius variables $r_j$. Then, by the assumption of independence and uniformity, the left-hand side of (3) reads $p_{\mathbf{r}}(\{r_j\}) \prod_j \alpha_{d_j}^{-1}$. Combining this with (3) and substituting $r_j = \|\mathbf{s}_{[j]}\|$ and $\mathbf{u}_j = \mathbf{s}_{[j]}/\|\mathbf{s}_{[j]}\|$, we obtain

$$p_{\mathbf{s}}(\{\mathbf{s}_{[j]}\}) = p_{\mathbf{r}}(\{\|\mathbf{s}_{[j]}\|\}) \prod_j \alpha_{d_j}^{-1} \|\mathbf{s}_{[j]}\|^{1-d_j}. \tag{4}$$

Finally, using the relation $p_{\mathbf{r}}(\{r_j\}) = p_{\mathbf{q}}(\{r_j^2\}) \prod_j (2r_j)$, we obtain the formula for the real-valued case:

$$p_{\mathbf{s}}(\{\mathbf{s}_{[j]}\}) = p_{\mathbf{q}}(\{\|\mathbf{s}_{[j]}\|^2\}) \prod_j \|\mathbf{s}_{[j]}\|^{2-d_j} \Gamma(d_j/2) \pi^{-d_j/2}. \tag{5}$$

For complex vector $\mathbf{s}$, the isomorphism between $\mathbb{C}^d$ and $\mathbb{R}^{2d}$ implies that we only need to replace every $d_j$ in the real-valued case with $2d_j$, which straightforwardly gives the desirable result.

## B Adaptive Subspace Partitioning

First, observe that

$$\|\mathbf{\Omega} - \mathbf{Z}^\mathsf{T}\mathbf{\Gamma}\mathbf{Z}\|^2 = -2\mathrm{tr}[\mathbf{\Gamma}^\mathsf{T}\mathbf{Z}\mathbf{\Omega}\mathbf{Z}^\mathsf{T}] + \mathrm{tr}[\mathbf{D}\mathbf{\Gamma}^\mathsf{T}\mathbf{D}\mathbf{\Gamma}] + \text{const.}, \tag{6}$$

where $\mathbf{D} := \mathrm{diag}(d_1, d_2, \ldots, d_m) = \mathbf{Z}\mathbf{Z}^\mathsf{T}$. As readily seen, $\mathbf{\Gamma} = \mathbf{D}^{-1}\mathbf{Z}\mathbf{\Omega}\mathbf{Z}^\mathsf{T}\mathbf{D}^{-1}$ minimizes (6) for any given $\mathbf{Z}$. Substituting this, we eventually obtain $-\|\mathbf{D}^{-1/2}\mathbf{Z}\mathbf{\Omega}\mathbf{Z}^\mathsf{T}\mathbf{D}^{-1/2}\|^2 + \text{const.}$ to be minimized with respect to $\mathbf{Z}$. This is equivalent that we maximize $\|\widetilde{\mathbf{Z}}\mathbf{\Omega}\widetilde{\mathbf{Z}}^\mathsf{T}\|$ with respect to $\widetilde{\mathbf{Z}} := \mathbf{D}^{-1/2}\mathbf{Z}$, as desired.

# C Details of EEG analysis

The EEG data were measured during the subjects were performing two-class cued motor imagery task (Blankertz et al., 2007). The two classes were possibly different for each subject, selected out of left hand, right hand or foot. The data were already downsampled at 100Hz. As a common preprocessing, we first re-referenced all the sensor channels to the common average and slightly reduced the number of channels from 59 to 41 to reduce computation (see below for channel names and layout). Then we applied standard Morlet wavelet filter in each channel to convert the data into complex time-frequency domain, with the center frequencies at every 0.5Hz in the range of 8-30Hz and at the reduced sampling rate of 10Hz. Finally, we vectorized the spectra (45 discrete frequencies) and concatenated them in all the 41 channels, resulting in 1845-dimensional complex data vectors $\mathbf{x}_t$.

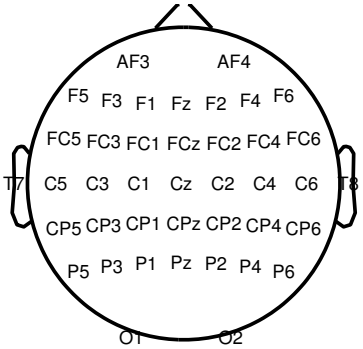


Figure S1: Layout of 41 EEG channels

The database provided two types of data, namely, the "calibration" and "evaluation" datasets, originally used for training and testing the classifiers, respectively. In our analysis, we used only the "calibration" data for unsupervised learning. The "calibration" data contained 200 trial epochs (the half for one class). To avoid transient or cue-related effects, we extracted only the last 2s (out of the duration of 4s) of each trial epoch. Thus, we eventually obtained the dataset $\{\mathbf{x}_t\}$ of sample size 4000 (i.e., 20 time points per trial), for which we trained the model by each of the three methods: EBM-NCE, SPLICE-LW and SPLICE-ML.

After learning, every second-layer components $s'_k$ was averaged within each trial, and then regarded as a discriminant score for the two classes. Note that the sign of each discriminant score was actually arbitrary due to the indeterminacy of sign in ICA. Thus, we computed the AUC score by first computing the two scores by flipping the sign and taking the greater one (Fig. 4).

To investigate the generalization ability of the model, we further evaluated the same AUC scores by transferring the model to the other "evaluation" data. In this EEG data, the duration of task trials are irregular (1.5-8s). Instead of averaging the second-layer components $s'_k$ within each trial of varying length, we thus computed the average value within each of the sliding time windows of $2s$. The window was sampled only within the task periods except for the initial $1s$ after the cue. Thus, every time point except for the initial $1s$ in each trial was given a single discriminant score per second-layer component. The AUC scores were then computed for every component.

# References

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., and Curio, G. The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2): 539–550, 2007.