Decoding Emotional Valence from Electroencephalographic Rhythmic Activity

Hande Çelikkanat[1,2], Hiroki Moriya[2], Takeshi Ogawa[2],
Jukka-Pekka Kauppi[1,3], Motoaki Kawanabe[2], Aapo Hyvärinen[1,4]

***Abstract*— We attempt to decode emotional valence from electroencephalographic rhythmic activity in a naturalistic setting. We employ a data-driven method developed in a previous study, Spectral Linear Discriminant Analysis, to discover the relationships between the classification task and independent neuronal sources, optimally utilizing multiple frequency bands. A detailed investigation of the classifier provides insight into the neuronal sources related with emotional valence, and the individual differences of the subjects in processing emotions. Our findings show: (1) sources whose locations are similar across subjects are consistently involved in emotional responses, with the involvement of parietal sources being especially significant, and (2) even though the locations of the involved neuronal sources are consistent, subjects can display highly varying degrees of valence-related EEG activity in the sources.**

I. Introduction

An important question in neuroscience is how emotions are represented in the brain. Related questions include a) how stable and discrete concepts such as 'pleasant' and 'unpleasant' can form in complex neuronal mechanisms, and b) why and how such representations are varying across individuals. While it is rather clear that high degrees of subjectivity are involved in the emotional responses, for example in their elicitation and the cognitive contents, it would seem equally clear that there exists a common enough platform that gives rise to these discrete categories—which is the core interest in affective neuroscience.

From a more technological perspective, another major challenge in emotion research is their online decoding. This is even more challenging in a naturalistic setting, i.e. outside of a laboratory. We would like to monitor and detect emotions in real-time and in real-life scenarios. In aging societies, for instance, distress situations detected by such monitoring could be important for improving caregiving.

The authors would like to thank Ken Yano for performing part of data acquisition and Takayuki Suyama, Jun-ichiro Hirayama, Atsunori Kanemura, Taiki Miyanishi, and Hiroshi Morioka for discussions. This work was funded by Japanese-Finnish Research Cooperative Program: Joint Research Activities in Information Systems for Accessibility and Support of Older People, through project 'Next generation affective life log: Machine learning with multi-modal sensor networks'.

[1]H. Çelikkanat, J.-P. Kauppi, and A. Hyvärinen are with the Department of Computer Science and HIIT, University of Helsinki, Finland.

[2]H. Çelikkanat, H. Moriya, T. Ogawa, and M. Kawanabe are with Advanced Telecommunication Research Institute International (ATR), Japan.

[3]J.-P. Kauppi is with the Faculty of Information Technology, University of Jyväskylä, Finland.

[4]A. Hyvärinen is with Gatsby Unit, University College London, UK.

E-mail: hande.celikkanat@helsinki.fi, moriyah@atr.jp, t.ogawa@atr.jp, jukka-pekka.kauppi@jyu.fi, kawanabe@atr.jp, a.hyvarinen@ucl.ac.uk

A combination of these two aims presents a unique challenge. For instance, the neuronal bases of emotions have to some degree been identified through functional magnetic resonance imaging (fMRI) studies (see *e.g.* [1]). However, the "laboratory settings" required for fMRI do not apply easily to real-life. A recent trend has been to exploit the more mobile electroencephalography (EEG) for designing brain-machine interfaces. This trend is further encouraged by the finding that distinct brain oscillations measurable by EEG are associated with distinct brain states [2]. With its high temporal resolution, EEG is also a strong candidate for monitoring the temporal evolution of emotions.

Major problems in this scenario are the relatively low signal-to-noise ratio and the low spatial resolution of EEG. Together they make EEG-based decoding a non-trivial machine learning problem. Typically, the goal in previous research has been to obtain data which is as clean as possible, by minimizing artifacts and interference from all sensory modalities in a laboratory setting (*e.g.*, [3]). By contrast, our aim is decoding emotions in naturalistic settings. Although similar previous attempts exist (*e.g.* [4], [5]), the difficult nature of the naturalistic task combined with the limitations of the EEG resulted in relatively modest decoding accuracy. Moreover, conventional approaches provide only limited insight into underlying neuroscientific processes. For instance, the coefficients of a linear classifier cannot directly provide valuable information on the neuronal representations [6]. These limitations demand more sophisticated machine learning algorithms for emotion decoding from EEG signals.

Recently, Kauppi *et al.* [7] proposed a data-driven decoding method, Spectral Linear Discriminant Analysis, for decoding brain states based on rhythmic brain signal activity. They validated the method in a multi-modal sensory stimuli classification context, but the method seems equally applicable for decoding emotions. The key advantage is its ability to reveal task-relevant spectrospatial sources of brain activity. In this paper, we apply this method to decode emotional valence from EEG signals in a naturalistic setting. Thus, our results demonstrate the feasibility of valence decoding that is robust to eye movement artifacts, in contrast to the frontal alpha-asymmetry or event-related potential studies.

II. Materials and Methods

A. Experimental Procedure and Data Acquisition

All experiments were conducted in the experimental smart house [8]. The participants sat on a comfortable chair at a distance of 1.2m from a television (Regza, Toshiba Co.,

Tokyo, Japan). Stimulus presentation was controlled by Presentation (Neurobehavioral Systems, Inc., Berkeley, CA). CPz-referenced EEG signals were recorded from 32 scalp sites using eego sports amplifier and waveguard cap (ANT Neuro, Enschede, Netherlands) at a sampling rate of 512 Hz.

To induce emotions, we used an emotional induction task by using a standardized emotional movie library library [9]. Each movie had a duration of 20-30s, and was labeled by the authors of the library with one of four emotional conditions: *positive, negative, neutral,* or *mixed* valence. The experiment consisted of 7 sessions, each including 4 emotion condition blocks. In each condition block, the subject was shown a wash-out movie for 90s, followed by a sequence of 4 movies that are labeled with the same emotional condition. The order of emotion condition blocks was randomized. In total 112 videos were shown (7 sessions $\times$ 4 conditions $\times$ 4 movies).

B. Preprocessing

We used EEGLAB [10] to preprocess raw EEG data. Data was first band-pass filtered (Butterworth, order 2) between 5-50 Hz, re-referenced in mean of all electrodes and downsampled to 128Hz. To remove physiological and environmental artifacts, we applied Independent Component Analysis and ADJUST [11]. In addition, we removed components with significantly higher high-frequency power spectrum densities compared to alpha-band as electromyography (EMG) artifacts. Altogether 5-10 components were rejected from each subject. We epoched the last 20-seconds in each video into three 8s epochs with 2s overlaps. The number of epochs for each condition was 84 (7 sessions $\times$ 4 movies $\times$ 3 epochs).

Next, Fourier-transform was applied for each epoch in every channel, and this short-time Fourier transformed data was used as input in time-frequency based Independent Component Analysis (Fourier-ICA, [12]). Fourier-ICA estimates the independent neuronal sources, or 'independent components' (ICs) of each subject's brain activity, returning the spatial distribution and frequency spectrum of ICs. (Fig. 1-A and B). Since the rank decreased to approximately 20 during artifact rejection, Principal Component Analysis was applied a priori to reduce the dimensionality to 20. The final frequency band of interest was selected as 5-20 Hz. Fourier-ICA was repeated three times with random initialization, and the estimation with the highest objective was used.

C. Classification

We used the Spectral Linear Discriminant Analysis (SLDA) by Kauppi *et al.* [7] to learn a classification between the epochs labeled with positive vs. negative valence. (The neutral and mixed-labeled epochs were discarded.)

SLDA classifier design is a two-step process which first computes a task-discriminative spectral weight vector and a corresponding spectral projection for each category and IC (feature extraction), and then finds an optimal combination of task-relevant ICs using sparse logistic regression (feature selection and classification). In SLDA, each feature corresponds to a spatial pattern of an IC with its discriminative spectral characteristics. A benefit of the SLDA is that besides classification, it automatically recovers the relevant subset of features, providing useful neuroscientific insights.

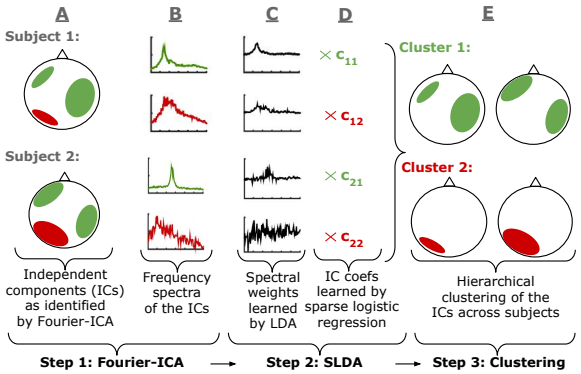

Fig. 1: The pipeline of feature extraction (Fourier-ICA), classification (SLDA), and interpretation of results (Hierarchical clustering). See text for details.

Feature Extraction: Let an 'observation' at epoch n be:

$$\mathbf{Z}(n) = [\mathbf{z}_1(n), \mathbf{z}_2(n), ...\mathbf{z}_C(n)]^T \in \mathbb{R}^{C\times F}, \quad (1)$$

where $\mathbf{z}_i(n)$ are the absolute values of the Fourier coefficients of IC i, C is the number of ICs, and F is the total number frequency bins. For each category k and IC i, SLDA estimates the maximally discriminative vector $\mathbf{f}_{ki}$ between category k and the other categories. (For the 2-class case, a single discriminative vector per IC is enough and k can be dropped.) This estimation is done by assuming a special case of regularized linear discriminant analysis (LDA), where the regularization term of the within-class scatter matrix is infinitely large. (Hence the name, "Spectral-LDA".) The short-time spectra of the observations are projected onto $\mathbf{f}_i$:

$$\mathbf{x}(n) = \left[\mathbf{f}_1^T\mathbf{z}_1(n), \mathbf{f}_2^T\mathbf{z}_2(n), ...\mathbf{f}_C^T\mathbf{z}_C(n)\right]^T \in \mathbb{R}^{C\times 1}, \quad (2)$$

where $\mathbf{x}(n)$ denotes the new representation of the data point of epoch n, and $\mathbf{f}_i$ is the spectral weight vector estimated for IC i by LDA (Fig. 1-C). In this representation, a frequency bin that is discriminative for the IC gets a large feature value.

Sparse Logistic Regression: The final classification decision is made by sparse logistic regression using a linear kernel: $h(\mathbf{c}; \mathbf{x}(n)) = \mathbf{c}^T\mathbf{x}(n) = \sum_{i=1}^{C}\sum_{j=1}^{F} c_i f_{ij} z_{ij}(n)$, in which each Fourier coefficient z_{ij} of each IC i both has its own spectral weight f_{ij} (Fig. 1-C), and is also weighted through the IC-specific coefficient c_i (Fig. 1-D). This is in accordance with SLDA's assumption that separate sources are likely to have separate spectra for different categories.

Interpretation of SLDA via Clustering: Since the ICs of each brain are slightly different, in order to interpret SLDA results, we group together the subjects whose spatial distributions of ICs are similar. This is done by an agglomerative hierarchical clustering over the spatial maps (Fig. 1-E). A cluster that is composed of a 'sufficient' number of ICs over subjects is likely to represent a common brain source. Moreover, if the ICs in this cluster have high coefficients, this cluster is likely relevant in the classification. Conversely, a cluster with close-to-zero coefficients would be less relevant.

Evaluating the Discriminability of Components: To quantify the 'discriminability' of the spectra of an IC in response

Subj	SLDA	Random	Subj	SLDA	Random
11*	$\mathbf{75 \pm 10\%}$	$50 \pm 11\%$	**1***	$\mathbf{56 \pm 13\%}$	$49 \pm 10\%$
2*	$\mathbf{74 \pm 12\%}$	$50 \pm 11\%$	10	$56 \pm 11\%$	$50 \pm 8\%$
3*	$\mathbf{70 \pm 14\%}$	$50 \pm 10\%$	**8***	$\mathbf{55 \pm 9\%}$	$49 \pm 8\%$
6*	$\mathbf{64 \pm 12\%}$	$50 \pm 8\%$	13	$55 \pm 13\%$	$50 \pm 10\%$
16*	$\mathbf{63 \pm 13\%}$	$50 \pm 11\%$	17	$54 \pm 13\%$	$50 \pm 10\%$
5*	$\mathbf{63 \pm 11\%}$	$50 \pm 9\%$	14	$53 \pm 8\%$	$49 \pm 10\%$
12*	$\mathbf{60 \pm 9\%}$	$50 \pm 8\%$	15	$49 \pm 12\%$	$50 \pm 8\%$
4*	$\mathbf{59 \pm 12\%}$	$50 \pm 9\%$	7	$46 \pm 11\%$	$49 \pm 10\%$
9*	$\mathbf{57 \pm 10\%}$	$50 \pm 10\%$	**Mean**	**59%**	**50%**

TABLE I: Individual decoding accuracies (mean±standard deviation). The statistically significant (permutation test, p < 0.05) results are indicated with bold text and an asterisk.

to positive vs. negative stimuli, we define a measure:

$$disc = \frac{\sum_{z=Fr_{min}}^{Fr_{max}} [\mu_p(z) - \mu_n(z)]^2}{\sum_{z=Fr_{min}}^{Fr_{max}} \sigma_p^2(z) + \sigma_n^2(z)}, \quad (3)$$

where z is a frequency in [Fr$_{min}$=5Hz, Fr$_{max}$=20Hz], $\mu_p(z)$ and $\mu_n(z)$ denote the mean Fourier coefficients in positive and negative valence epochs, and $\sigma_p^2(z)$ and $\sigma_n^2(z)$ denote the corresponding variances over epochs. This measure is high when there is a significant difference between the two spectra, and when they are consistent over trials.

Training Settings: The training of SLDA was conducted for each subject by a leave-one-block-out paradigm, using 6 of the subject blocks for training, and the remaining block for testing. This procedure was repeated 7 times by changing the test block, and the obtained 7 classification results were averaged to obtain our final estimate of the test accuracy. It is important to use the leave-one-block-out scheme to avoid bias in the test accuracy due to temporal dependencies between adjacent epochs. To see if our results were significantly above the chance level (50%), we conducted a permutation test [13] based on 100 permutations of the category labels.

III. Results

First, we tried to decode subjects' emotional valence during positive vs. negative valence videos. Table I shows the decoding accuracies for positive vs. negative valence in individual subjects, which were from 75% to 46% (highest: S11, 75±10%; lowest: S7, 46±10%, mean: 59%). SLDA was significantly above chance level (permutation test, p $<$ 0.05) for 11 subjects out of 17 (64.7%). We found a high variability of decoding accuracy across subjects.

Next, we attempt to understand individual differences of spatial-spectral features between subjects with higher and lower decoding accuracies. One of the major advantages of SLDA is offering an interpretation of the spectral characteristics of different brain sources, in terms of their contributions to different states. Specifically, the trained classifiers can provide insight on (1) the brain sources related to emotional valence, and (2) the reasons for the variance across subjects.

Fig. 2 presents the most relevant cluster that SLDA has identified. Sample ICs belonging to this cluster are also shown. When an IC has a high coefficient, it is relevant for the classification of that subject. High spectral weights for any given frequency band similarly mean this band is relevant to the classification. Finally, a big difference between the spectra of the positive and negative valence epochs hints this IC is displaying different characteristics for different states.

The ICs in Fig. 2 show a consistently parietal spatial distribution across subjects, and have in general a significant peak in the alpha band. Yet, there is also significant variability across subjects. The subjects whose emotional states can be effectively decoded (*valence-discriminable* subjects) all show a clear peak, and have high discriminability between the spectra. These ICs also have high coefficients, and their

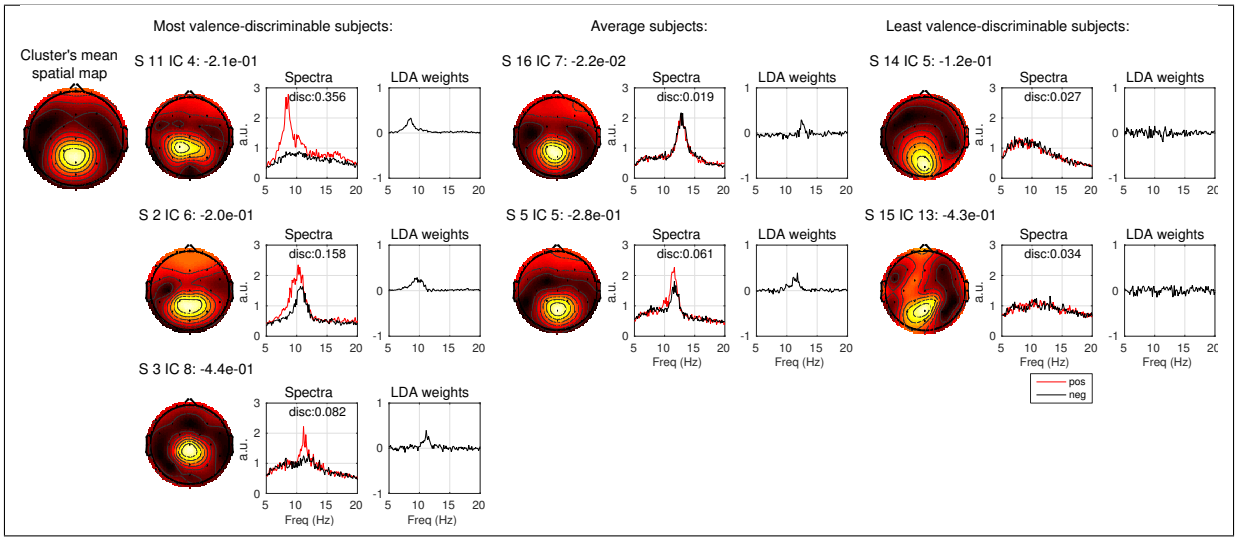

Fig. 2: A relevant cluster. *(Far-left)* The mean spatial map of the cluster. *(Middle-left)* ICs from the most valence-discriminable subjects that are associated with the cluster. *(Middle-right)* ICs from average valence-discriminable subjects. *(Far-right)* ICs from the least valence-discriminable subjects. Within each IC: *(Left)* spatial distribution and coefficient, *(Middle)* mean spectra during positive (red) vs. negative (black) valence epochs, *(Right)* spectral weights. Note that y-axis for the spectra denotes arbitrary units: The scale represents the absolute value of the ICs and has no physical meaning.

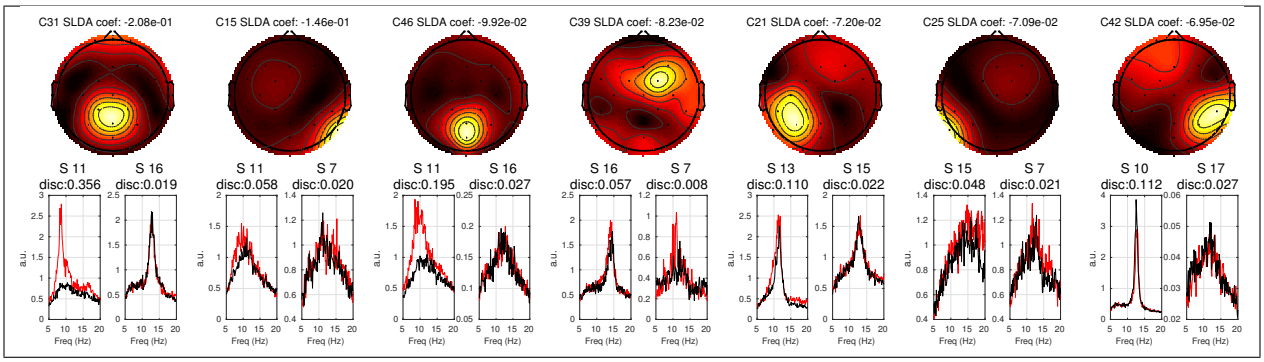

Fig. 3: The 7 most relevant clusters as identified by SLDA. For each cluster, the mean SLDA coefficient and spatial map (top), and spectra of its most (bottom left) vs. least (bottom right) discriminative ICs are shown. The red and black colors in the spectra denote the mean of positive and negative valence epochs respectively. y-axis is in arbitrary units.

spectral weights are selective to the discriminative regions of the spectrum. For average subjects, the spectral signatures show a less-discriminable peak. (Especially in S16, whose IC consequently has a lower coefficient.) For the least discriminable subjects, both the peak and discriminability of the spectra diminish, and spectral weights are not selective.

Furthermore, the coefficients of the clusters can give insight on the common sources involved, higher coefficients meaning more relevant clusters. Fig. 3 presents the seven most relevant ones. Clusters with less than 9 ICs across subjects (the ones that are not common enough) are excluded, moreover only a single IC from each subject is allowed in a cluster. Therefore, all presented clusters are consistent over a large number of subjects (each between 9 and 11), and represent the 'common neuronal sources'. Mostly parietal regions are found to be involved consistently, although many ICs also have activation in more anterior regions. As in Fig. 2, there is high variability between the spectral signatures, although the locations of the sources are highly consistent.

IV. Conclusion

We demonstrated that it is possible, at least to some degree, to decode positive vs. negative emotional valence from EEG measurements in a naturalistic setting. [1] We analyzed the EEG signals from subjects watching naturalistic move clips in a real house instead of a laboratory. The decoding method was the recently proposed Spectral Linear Discriminant Analysis, combining preprocessing by Fourier-ICA, optimum frequency band estimation by LDA, IC selection by sparse logistic regression, and hierarchical clustering of the selected ICs across subjects.

An analysis of decoding weights pointed out the important role of the parietal regions in emotional reactions. The proposed methodology enables more-detailed future investigations of the neuronal loci related to emotional processing.

We further found the decoding accuracy varies strongly across subjects. Such variability can be due to differences in the neural representations of emotions, or differences in the quality of measurements over subjects, among other factors.

Clustering subject-wise ICs allowed us to compare neuronal sources across subjects. We found that although the locations of the sources involved are rather similar across subjects, subjects are highly variable in their valence-related rhythmic activity in these sources. This is presumably a major reason for the individual differences in decoding accuracy. The methodology could be used in future to provide insight on the detailed individual differences of emotional responses.

[1] Though it should be noted that our results do not completely rule out the possibility that the decoding could have been affected by the arousal differences between the training instances.

References

[1] P. A. Kragel and K. S. LaBar, "Decoding the nature of emotion in the brain," *Trends Cogn Sci*, vol. 20, pp. 444–455, 2016.

[2] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, pp. 1926–1929, 2004.

[3] E. Harmon-Jones and J. J. Allen, "Anger and frontal brain activity: EEG asymmetry consistent with approach motivation despite negative affective valence," *J Pers Soc Psychol*, vol. 74, pp. 1310–1316, 1998.

[4] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in *CW*, 2010, pp. 262–269.

[5] G. Garcia-Molina, T. Tsoneva, and A. Nijholt, "Emotional brain-computer interfaces," *Int J Auton Adap Comm Sys*, vol. 6, 2013.

[6] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, , and F. Biessmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.

[7] J.-P. Kauppi, L. Parkkonen, R. Hari, and A. Hyvärinen, "Decoding magnetoencephalographic rhythmic activity using spectrospatial information," *NeuroImage*, vol. 83, pp. 921–936, 2013.

[8] T. Ogawa, J. Hirayama, P. Gupta, H. Moriya, S. Yamaguchi, A. Ishikawa, Y. Inoue, M. Kawanabe, and S. Ishii, "Brain-machine interfaces for assistive smart homes: A feasibility study with wearable near-infrared spectroscopy," in *EMBC*, 2015, pp. 1107–1110.

[9] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cognition and Emotion*, vol. 30, 2016.

[10] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J Neurosci Meth*, vol. 134, pp. 9–21, 2004.

[11] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, pp. 229–240, 2011.

[12] A. Hyvärinen, P. Ramkumar, L. Parkkonen, and R. Hari, "Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis," *Comp. Intell. Neurosci.*, vol. 49, 2010.

[13] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *J. Neurosci. Meth.*, 1989.