NeuroImage 83 (2013) 921–936

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Decoding magnetoencephalographic rhythmic activity using spectrospatial information

Jukka-Pekka Kauppi [a,b,*], Lauri Parkkonen [b,c], Riitta Hari [b], Aapo Hyvärinen [a,d]

[a] Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland
[b] Brain Research Unit, O.V. Lounasmaa Laboratory, School of Science, Aalto University, Espoo, Finland
[c] Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, Espoo, Finland
[d] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

ARTICLE INFO

Article history:
Accepted 5 July 2013
Available online 18 July 2013

Keywords:
Decoding
Magnetoencephalography
Rhythmic activity
Time–frequency analysis
Linear discriminant analysis
Independent component analysis

ABSTRACT

We propose a new data-driven decoding method called Spectral Linear Discriminant Analysis (Spectral LDA) for the analysis of magnetoencephalography (MEG). The method allows investigation of changes in rhythmic neural activity as a result of different stimuli and tasks. The introduced classification model only assumes that each "brain state" can be characterized as a combination of neural sources, each of which shows rhythmic activity at one or several frequency bands. Furthermore, the model allows the oscillation frequencies to be different for each such state. We present decoding results from 9 subjects in a four-category classification problem defined by an experiment involving randomly alternating epochs of auditory, visual and tactile stimuli interspersed with rest periods. The performance of Spectral LDA was very competitive compared with four alternative classifiers based on different assumptions concerning the organization of rhythmic brain activity. In addition, the spectral and spatial patterns extracted automatically on the basis of trained classifiers showed that Spectral LDA offers a novel and interesting way of analyzing spectrospatial oscillatory neural activity across the brain. All the presented classification methods and visualization tools are freely available as a Matlab toolbox.

Introduction

Unveiling neuronal information processing in the human brain during real-world experiences is a central challenge in cognitive neuroscience (Spiers and Maguire, 2007). Conventionally, functional neuroimaging studies have been applied using relatively simple patterns of sensory stimuli, and little is known about how the human brain operates with real-world sensory input. More recently, the neuroimaging community has started to introduce more naturalistic experimental conditions (Hasson et al., 2004, 2008; Hejnar et al., 2007; Kauppi et al., 2010; Lahnakoski et al., 2012; Wolf et al., 2010), and even "two-person neuroscience" has been advocated to record brain activity simultaneously from two interacting subjects (for a review, see Hari and Kujala (2009)).

Due to the diversity of the stimuli and/or the complexity of the experimental settings mimicking real-world conditions, it may be necessary to use data-driven analysis methods that allow investigation of brain function without stringent assumptions about the underlying brain mechanisms (Spiers and Maguire, 2007). One of the most promising data-driven approaches to analyze complex brain-imaging signals is "decoding", which gathers information from multiple brain imaging signals to deduce the task, stimuli or brain state during the measurement. Most commonly, multivariate classifiers are used to discriminate between categories (Blankertz et al., 2011; Cox and Savoy, 2003; Kamitani and Tong, 2005; Mitchell et al., 2004; Murphy et al., 2011) but decoding can also be performed using regression in more complex experimental settings (Carroll et al., 2009; Kauppi et al., 2011).[1]

Brain-function decoding can advance our knowledge in different ways. For instance, above-chance classification performance for an independent test data set implies the presence of mutual information between the measured signals and the categories of interest (Kriegeskorte, 2011). Thus, decoding can be used to test for the presence of specific stimulus information in the region of interest or across the whole brain. Additionally, investigating how the trained models are fitted to the brain-imaging signals tells where and how information is processed and represented in the brain. For instance, the coefficients of the linear classifier may provide hints of brain regions involved in the processing and discrimination of the stimuli (see e.g. Rasmussen et al. (2012)). It is also possible to construct several decoders based on different neuroscientific hypotheses and compare their performances. *A priori* knowledge can be incorporated to the decoder design for instance in the form of neuroscientifically inspired feature transformations (see e.g. Richiardi

* Corresponding author. Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland.
E-mail addresses: jukka-pekka.kauppi@helsinki.fi (J.-P. Kauppi), lauri.parkkonen@aalto.fi (L. Parkkonen), riitta.hari@aalto.fi (R. Hari), aapo.hyvarinen@helsinki.fi (A. Hyvärinen).

[1] Hence, the terms "decoder" or "decoding model" may refer either to a classifier or a regression model.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.07.026

et al. (2011)) or it can be embedded more directly to the model design (see e.g. Tomioka and Müller (2010)).

So far, most decoding studies in neuroscience have used functional magnetic resonance imaging (fMRI) signals to demonstrate spatial patterns related to different tasks or stimulus categories (Haynes and Rees, 2006; Tong and Pratte, 2012). However, the poor temporal resolution of fMRI makes it inherently unsuitable for investigating the fine spectral and temporal signatures of information encoding. Instead, the brain's oscillatory electrical activity has been suggested to have a central role in information processing, and distinct oscillation frequencies and amplitudes even in the same neuronal structure reflect different brain states (Singer, 1993). Large neuronal populations can generate synchronized oscillatory electrical activity that can be enhanced or suppressed by tasks and stimuli, and the dynamics of brain oscillations associated with distinct brain states forms complex spatiotemporal patterns (Buzsáki and Draguhn, 2004). Thus, to understand brain function during real-world experiences, it seems necessary to interpret at the same time the spatial, temporal and spectral signatures of brain activity.

Magnetoencephalography (MEG) has a millisecond-range temporal resolution and has therefore potential to reveal detailed spectral and temporal characteristics of distinct brain states induced by specific tasks or stimuli. Nevertheless, decoding on the basis of MEG signals cannot be expected to be an easy task. Several factors, including the low signal-to-noise ratio (SNR) of single-epoch measurements and the high dimensionality of whole-scalp recordings, make the decoding based on MEG signals very challenging. In addition, MEG signals do not vary only between different individuals under the same experimental condition, but to some extent also within the same subject between repeated identical sessions, which makes it complicated to construct a highly generalizable classifier across sessions and/or individuals. However, training of the multivariate model based on single-epochs provides inevitable advantages over a univariate analysis based on averaged epochs. For instance, a decoding approach allows finding combinations of the most discriminative features (or sensors) among a high number of initial features, and provides a principled way of assessing the goodness of the discrimination in terms of the estimated generalization accuracy.

Previously, Besserve et al. (2007) used band-limited power and phase synchrony features to classify between MEG data recorded during a visuomotor task and rest condition. Rieger et al. (2008) used temporal features and wavelet coefficients to predict the recognition of natural scenes from single-trial MEG recordings. Ramkumar et al. (2013) used both time-resolved and time-insensitive classifiers to decode from single-epoch MEG low-level visual features in the early visual cortex. Zhdanov et al. (2007) used temporal features together with the regularized linear discriminant analysis (regularized LDA) to classify between two different visual categories (faces and houses) on the basis of MEG signals. In the "Mind Reading from MEG" challenge organized in conjunction with the International Conference on Artificial Neural Networks (ICANN 2011), the task was to design a classifier to distinguish between different movie categories on the basis of 204-channel gradiometer MEG data (Klami et al., 2011). The data were recorded from a single subject who was shown five different movie clips. The winners of the competition extracted statistical features from time-domain signals and applied sparse logistic regression for classification (Huttunen et al., 2012).

Decoding has also been applied to electroencephalographic (EEG) signals. For instance, Murphy et al. (2011) and Simanova et al. (2010) successfully decoded abstract semantic categories from EEG data. Moreover, Chan et al. (2011) used temporal features in the classification of MEG and EEG data recorded simultaneously while the subjects performed visual and auditory language tasks.

Classification on the basis of MEG signals has also been studied in the context of brain–computer interfaces (BCIs), communication pathways between brains and external devices (Bahramisharif et al., 2010; Mellinger et al., 2007; Santana et al., 2012; van Gerven and Jensen, 2009). A successful BCI has to distinguish between brain signatures of the users intentions, and both temporal and spectral features have been applied, often selected based on specific *a priori* knowledge of the brain function. For instance, preparation to move a hand is associated with a brief suppression of the Rolandic mu rhythm that comprises 7–13 Hz and 15–25 Hz frequency bands. The power estimates characterizing these specific oscillations originating from the sensorimotor cortex have been successfully applied to decode motor-imagery tasks, where an individual mentally simulates different motor actions, such as hand movements (Pfurtscheller and Neuper, 2001). Even though most of the BCI literature has concentrated on classification on the basis of EEG, many technical advances in this field may also benefit MEG-based decoding; see for instance Lemm et al. (2011), Tomioka and Müller (2010), Dyrholm et al. (2007a), Liu et al. (2010), Blankertz et al. (2011), Mellinger et al. (2007), Suk and Lee (2013). On the other hand, the existing best BCI methods are not directly applicable to our setting because the goals of the analyses and experimental conditions are different. In BCI, the only goal is maximum classification accuracy, while in brain-function decoding, it is important to obtain a decoder with a meaningful interpretation to advance understanding of brain function. Consequently, many recent neuroimaging studies have concentrated on the interpretation of the decoding models (Carroll et al., 2009; de Brecht and Yamagishi, 2012; De Martino et al., 2008; Grosenick et al., 2013; Rasmussen et al., 2012; Ryali et al., 2012; van Gerven et al., 2009; van Gerven et al., 2010; Yamashita et al., 2008).

Here, we constructed a brain decoding system for MEG with the explicit goal of providing an easily interpretable decoder, as well as a general-purpose decoding toolbox for neuroscientific research. As an example of this approach, we analyzed MEG data from an experiment where the subjects were exposed to blocks of auditory, visual and tactile stimuli interspersed with rest blocks (Malinen et al., 2007; Ramkumar et al., 2012). We aimed to decode four distinct brain states, that is, "auditory", "visual", "tactile", and "rest".

The stimuli were complex, comprising video clips of people and urban scenes, speech sounds and tone beeps, as well as tactile stimuli to finger tips, all presented in brief blocks of varying duration within the same session. Because sensory stimuli are known to activate discrete projection areas, we considered this experiment well-suited for the validation of our method. However, the applied complex stimuli (speech and videos) may also activate higher-order processing. For instance, although it is plausible that variations in oscillatory activity in the visual cortex are mainly responsible for discriminating the visual category from the other categories, higher-order brain processes may involve additional neural activity in other brain regions, thereby complicating the decoding task. On the other hand, the diversity of the stimuli makes the decoding problem also more interesting, advocating the use of data-driven approaches based on relatively weak *a priori* information. As our goal was to build a classifier to infer brain function in an exploratory manner, we did not impose strong assumptions on spectral contents or spatial locations of the underlying neural activity; instead, we tried to capture the most relevant spectrospatial features automatically from a large number ($L = 204$) of MEG channels across a relatively wide frequency band (5–30 Hz).

Materials and methods

Naturalistic stimulation

We analyzed MEG data (306-channel Elekta Neuromag MEG system (Elekta Oy, Helsinki, Finland), filtered to 0–200 Hz and digitized at 600 Hz) from a previous experiment (Ramkumar et al., 2012). Eleven healthy adults (6 females, 5 males; mean age 30 years, range 23–41 years) were exposed to 6–33 s blocks of auditory, visual and tactile stimuli. Similar to Ramkumar et al. (2012), data of only nine of the eleven subjects were used in the analysis; data from two subjects were discarded due to improper delivery of auditory stimuli.

The stimulus blocks were interleaved with 15-s rest periods (the entire session contained 24 rest periods). Two independent 12-min sessions were recorded from each subject. We exclusively used the first session for classifier training and the second session for the evaluation the performance of the classifier. Hence, the training and test data sets were independent from each other. Fig. 1A shows schematically the content of each 12-min stimulus sequence and the four categories: A = "auditory", V = "visual", T = "tactile", and R = "rest".

The design of visual, auditory and tactile blocks was adopted from Malinen et al. (2007). Auditory stimuli consisted of 100-ms tone bursts (at 250, 500, 1000, 2000, or 4000 Hz randomly varying in the same block; presentation rate 5 Hz) and of pre-recorded speech (a male voice narrating the history of the local university, or the same male voice providing instructions on guitar fingering). Speech and tones were presented in different blocks. Visual stimulus blocks consisted of silent home-made video clips of buildings, people with focus on hands, and people with focus on faces. Content exclusively from one of these three types was presented in each block. Tactile stimuli were delivered at 4 Hz using pneumatic diaphragms attached to the index, middle, and ring fingers of both hands. All these three fingers from both hands were stimulated within each tactile block. The order of the stimulation was random but homologous left and right fingers were always stimulated simultaneously. The alertness of the subjects was not systematically controlled because the experiment was relatively short and alerting, and it did not require notable concentration.

Preprocessing

We used only the 204 planar gradiometers because of their focal sensitivity patterns, which enables an easily interpretable visualization on the sensor helmet; for magnetometers, we would have needed a source model. We first applied the signal-space separation (SSS) method (Taulu and Kajola, 2005) to raw MEG time series to reduce artifacts and to perform head-motion correction. SSS is based solely on the physics of the measured magnetic fields and is non-adaptive. It decomposes the measured multi-channel signal to contributions from sources inside (physiological signals) and outside (environmental interference) of the MEG sensor array by modeling both spaces with specific multipole expansions. Using the generally recommended expansion orders ($L_{\text{inside}} = 8$, $L_{\text{outside}} = 3$) and SSS basis optimization, the effective dimensionality (rank) of the data was 64 after SSS processing.

We then computed time–frequency decompositions of the signals based on short-time Fourier transform (STFT). We used rectangular windowing to facilitate the estimation of the independent components (ICs) in the later stage of the analysis. The length of the time window for which Fourier coefficients were computed corresponded to the length of the epoch. Hence, the choice of the time-window length and overlap factor determined the number of epochs we obtained for training the classifier. We investigated the decoding performance for five different time-window lengths (1, 2, 3, 4, and 5 s) and for five different time-window overlaps (the fraction of overlap between two consecutive windows 0, 1/2, 2/3, 3/4, and 4/5). To optimize the estimation of classification accuracy and to keep test epochs as independent as possible, the time windows for the test data did not overlap. We obtained category labels based on the timing of stimuli. As we wanted each epoch to unambiguously represent one of the four categories, we discarded epochs on category boundaries (see Fig. 1B).

We simplified the classifier learning and interpretation of the classification results by balancing the number of epochs between the categories by discarding "redundant" epochs (see Fig. 1C). To ensure reliable estimation of the cross-validation (CV) classification accuracy (see Cross-validation section on how we computed the CV accuracy), we wanted to preserve the block structure of the epochs. Therefore, we balanced the number of epochs between the four categories by preserving the first N_{min} epochs from each category and discarding the others; N_{min} was the total number of epochs in the category containing the smallest number of epochs.

After computing the STFTs separately for each channel, we discarded low- and high-frequency components from each Fourier-transformed window to restrict our analysis to the range 5–30 Hz. A major reason for discarding low-frequency components was that we wanted to avoid decoding of the tactile category based on the 4-Hz frequency of stimulus delivery. Higher frequency components were removed because of their poor SNR; however, the limit was arbitrarily selected.

Next, we transformed the three-dimensional multichannel time–frequency data (channel × time × frequency) into a two-dimensional matrix by combining (collapsing) the time and frequency dimensions into one. After this, we had a complex-valued data matrix $\mathbf{X} \in \mathbb{C}^{L \times NF}$ (one matrix for classifier training and one for testing), where $L = 204$ was the number of channels, N the total number of epochs, and F the number of Fourier coefficients in each epoch after discarding components outside of 5–30 Hz. We applied complex-valued independent component analysis (ICA) to the data matrix of the training session to obtain $C = 64$ ICs as described by Hyvärinen et al. (2010); 64 was the effective dimensionality of the data matrix after applying the SSS preprocessing method. The ICs for the test

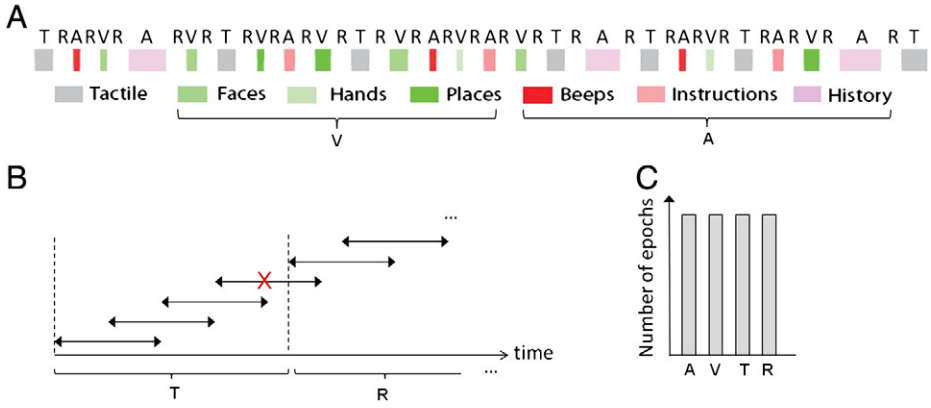

Fig. 1. Labeling and preprocessing of the data used in the decoding analysis: A) The 12-min stimulus sequence (modified from Ramkumar et al. (2012) and Malinen et al. (2007)). The four categories used in the decoding analysis were: A = "auditory", V = "visual", T = "tactile", and R = "rest". White spaces denote rest blocks and other colors correspond to different stimuli. "Visual" and "auditory" categories consisted of different subcategories, making the decoding task more challenging. B) An illustration of the extraction of short-time epochs from the training data. Epochs falling on category boundaries were discarded to avoid ambiguities in category labeling. C) After extracting the epochs, the number of epochs in each category was made equal so that the chance level of the correct classification was 0.25 for each class.

data were obtained using the transformation estimated based on the training data.

Note that we computed STFTs already before estimating ICs and not *vice versa*. The reason is that the distribution of the MEG signals is expected to become more sparse after the STFT, enhancing the estimation of ICs (the ensuing method is called "Fourier-ICA", see Hyvärinen et al. (2010)). The method is related to Fourier-domain methods for convolutive ICA (Anemüller et al., 2003; Dyrholm et al., 2007b) except that the convolutive ICA estimates a separate mixing matrix for distinct frequency bands whereas we estimated only one mixing matrix across the whole time–frequency representation. Further details of the estimation are provided here:

1. *Outlier removal*: We rejected outliers to improve robustness of the estimation of ICs from $\mathbf{X}$ by setting all Fourier coefficients in a specific epoch to zero if the logarithm of the norm of the data within it was larger than the mean plus three standard deviations.
2. *Dimension reduction and whitening*: We used PCA to whiten the data and to reduce the dimensionality from 204 to 64, which was the effective dimensionality of the data after SSS.
3. *Estimation*: We estimated ICs from the whitened data using the complex-valued FastICA algorithm based on the logarithmic measure of non-Gaussianity (Bingham and Hyvärinen, 2000). Estimation was repeated three times using random initialization, and the solution having the highest objective function value was selected as the final one.

Classifier design

With Fourier-ICA, we decomposed the STFT data matrix as $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A} \in \mathbb{R}^{L \times C}$ contains the spatial patterns and $\mathbf{S} \in \mathbb{C}^{C \times NF}$ contains the time–frequency decompositions of the $C = 64$ ICs (N and F depend on the choices of the time-window length and of overlap parameters, as well as on the selected frequency range of interest). The length of each time window in the decomposition corresponds to the length of the epoch we use in the classification.

For classification, we transformed ICs to three-dimensional multichannel time–frequency data (IC × time × frequency). Let us denote the absolute values of the Fourier coefficients of the ith IC for the nth time-window as $\mathbf{z}_i^{(n)}$. Then, a data "point" in the training dataset can be given as follows:

$$\mathbf{Z}(n) = [\mathbf{z}_1(n), \mathbf{z}_2(n), \ldots \mathbf{z}_C(n)]^T \in \mathbb{R}^{C \times F}, \quad \text{for } n = 1, 2, \ldots, N, \tag{1}$$

where T denotes a non-conjugate transpose. We trained a multicategory classifier using the matrices $\mathbf{Z}(n)$ and their corresponding binarized category labels, given by:

$$y_{nk} = \begin{cases} 1, & \text{if } \mathbf{Z}(n) \text{ belongs to the } k\text{th category} \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

We applied different feature transformations to matrices $\mathbf{Z}(n)$ prior to classification which we denote as $\mathbf{x}(n)$. We extracted features utilizing detailed spectrospatial information of the MEG signals. Basically, we estimated discriminative directions in the spectral domain of the ICs using the LDA, and obtained the features by projecting the spectra on these directions. For decoding, we adopted a symmetric version of the penalized multinomial logistic regression classifier (Friedman et al., 2010). We consider the use of logistic regression well-motivated because it extends naturally to multicategory classification problems and allows the use of suitable priors and kernel functions. Moreover, in our pilot study with several classifiers (including non-linear and linear SVM, minimum-distance classifier, k-nearest neighbor classifier, two-layer nonlinear Perceptron and random forest classifier), logistic regression yielded the best performance. Logistic regression has previously shown very good performance in decoding problems based on MEG signals (see e.g. Huttunen et al., 2012; Santana et al., 2012).

Spectral feature extraction

Because different brain states can be characterized by complex spectrospatial neuronal patterns, it is important to extract features that can capture this information. A plausible assumption is that different neuronal sources of the MEG signals have specific spectral characteristics, which may vary together with brain states. Hence, we allowed each IC to have its own "spectral signature" that additionally depended on the category. For each category and IC, we estimated *spectral-weight vectors* as the direction of maximum discrimination between the given category and other categories:

$$\mathbf{f}_{ki} = \mu_{ki} - \mu_{\bar{k}i}, \tag{3}$$

where $\mathbf{f}_{ki}$ denotes a spectral-weight vector (or a discriminative direction) for ith IC and kth category, μ_{ki} is the mean of the short-time spectra belonging to category k, and $\mu_{\bar{k}i}$ is the mean of the short-time spectra belonging to categories other than k. This estimation method coincides with LDA under the assumption that covariance matrices of the categories are spherical (Blankertz et al., 2011). The method can also be seen as a special case of regularized LDA, where the regularization term of the within-class scatter matrix is infinitely large. Very heavy regularization is expected to work well here because overlearning is a major concern due to session-to-session variability of MEG. After estimating spectral weights, we projected the short-time spectra of each IC to one dimension:

$$\mathbf{v}_i(n) = \left[\mathbf{f}_{1i}^T \mathbf{z}_i(n), \quad \mathbf{f}_{2i}^T \mathbf{z}_i(n), \ldots, \mathbf{f}_{Ki}^T \mathbf{z}_i(n) \right]^T, \tag{4}$$

where $K = 4$ is the number of categories. We then formed the final feature vectors by combining the information across ICs as

$$\mathbf{x}(n) = \left[\mathbf{v}_1(n)^T, \quad \mathbf{v}_2(n)^T, \ldots, \mathbf{v}_C(n)^T \right]^T \in \mathbb{R}^{KC}.$$

Eq. (4) shows that each IC is transformed to K category-specific features. For instance, if the spectrum of an IC is useful for discriminating mth category from the other categories, the value of the mth feature for this IC will be positive (by definition of the LDA). This feature will contribute to the final classification performance more than the other features and therefore a linear classifier is expected to give a positive classification coefficient (for category m) for the mth feature and coefficient values with low magnitudes for the other $K - 1$ features. The investigation of the spatial and spectral characteristics of the corresponding IC as well as its associated spectral-weight vector may advance our understanding of neural processes in the brain. In practice, it is likely that the learned model combines information from a combination of informative ICs for each category.

Penalized symmetric multinomial logistic regression

We used the symmetric version of multinomial logistic regression model to perform classification (Friedman et al., 2010). The symmetric multinomial logistic model is:

$$p_k(\mathbf{x}(n)) = \frac{\exp[h(\boldsymbol{\theta}_k; \mathbf{x}(n)) + b_k]}{\sum_{j=1}^{K} \exp\left[h\left(\boldsymbol{\theta}_j; \mathbf{x}(n)\right) + b_k\right]}, \quad \text{for } k = 1, 2, \ldots, K, \tag{5}$$

where $p_k(\mathbf{x}(n))$ is the posterior probability that features $\mathbf{x}(n)$ belong to category k, b_k is a bias term, and $h(\theta_k; \mathbf{x}(n))$ is a kernel function that projects $\mathbf{x}(n)$ onto a real line $\mathbb{R}$. The kernel function depends on category-specific coefficients θ_k which need to be estimated from the training data together with b_k. We used the most widely applied kernel that projects features linearly onto $\mathbb{R}$.

Table 1
The summary of the evaluated classifiers ordered according to increasing spectrospatial complexity.

Name	Features	Key assumption(s)
Baseline	Total energies	Spectral information is irrelevant
Statistical	Standard deviation of the spectra of the ICs	Spectral information is relevant but unspecific in nature
Bilinear	Entire spectrospatial matrix	Spectral information is relevant; spectral information is specific to each category but common to each IC
Spectral PCA	Projections based on PCA	Spectral information is relevant; spectral information is common to each category but specific to each IC
Spectral LDA	Projections based on LDA	Spectral information is relevant; spectral information is both category- and IC-specific

To show explicitly how Spectral LDA depends on category- and IC-specific spectral weights, we can write the kernel function in terms of the Fourier-coefficients $z_{ij}(n)$ as:

$$h(c_{kmi}|i=1,2,\ldots,C,m=1,2,\ldots,K;\mathbf{x}(n)) = \mathbf{c}_k^T\mathbf{x}(n) = \sum_{m=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{F} c_{kmi} f_{mij} z_{ij}(n), \quad (6)$$

where $\mathbf{c}_k \in \mathbb{R}^{KC}$ are classification coefficients for category k to be estimated and f_{mij} are spectral weights estimated beforehand using the LDA. The key observation here is that each spectral weight is specific to both IC (index i) and category (index m). The index of the Fourier coefficients is denoted by j.

The classification coefficients can be estimated by maximizing a sample log-likelihood L_θ of Eq. (5); see Eq. (A.2) in the Appendix A for the exact functional form of the sample log-likelihood. We controlled overlearning of the classifier by incorporating a suitable regularization term P_θ to the objective function besides log-likelihood. The objective function then became:

$$J_\theta = L_\theta - \lambda P_\theta, \quad (7)$$

where the hyperparameter λ controls the extent of regularization and needs to be fixed beforehand (see Cross-validation section how we estimated λ). We used the ℓ_1-norm of the classification coefficients as a penalty term:

$$P_\theta = \sum_{k=1}^{K} \|\mathbf{c}_k\|_1. \quad (8)$$

This penalization is called the least-absolute-shrinkage operator (LASSO) and it makes the final classifier sparse by shrinking many classification coefficients to zero (Tibshirani, 1996). The higher the value of λ, the sparser will be the final classifier. The benefit of the sparse model is that it is easier to interpret than the model where classification coefficients are nonzero. This property is especially important in functional neuroimaging, where the goal is to extract neuroscientifically interesting information from the trained classifier (Yamashita et al., 2008). Another popular regularization method used in neuroimaging studies is the so-called elastic net (Zou and Hastie, 2005), which uses the combination of ℓ_1- and ℓ_1-norms as a penalty term. This regularization allows the selection of the correlated features in the final model. The use of the elastic net is justified in fMRI-based decoding studies, where features correspond directly to spatially correlated voxels (Ryali et al., 2010). The situation in our study is different because the features are the spectral projections of the ICs. Our goal is to find most informative ICs and investigate their spectral and spatial characteristics without too much redundancy in the visualization (see Interpretation of spectrospatial patterns section for details how we analyzed the trained classifiers). Thus, for our data, it does not seem useful or necessary to use several correlated features for the same category.

We used an optimization code based on coordinate descent (GLMNET software package[2] by Friedman et al. (2010)) for maximizing Eq. (7) using the ℓ_1-norm penalty. After learning the classifier, we classified our test data according to *maximum a posteriori* (MAP) rule to evaluate the predictive performance of the classifier. More specifically, we selected the optimal category k^* for an unknown test sample $\mathbf{x}(n)$ as:

$$k^* = \arg\max_k \{h(\boldsymbol{\theta}_k;\mathbf{x}(n)) + b_k\}. \quad (9)$$

Alternative classifiers

We constructed four additional classifiers based on logistic regression. They all utilize spectral information differently. By comparing the prediction performance of different classifiers, we can infer what type of spectral information is associated with the stimulus- or task-related processing in the brain. The classifiers are presented here in the order of increasing spectrospatial complexity, starting from the classifier that does not utilize spectral information at all and ending with the classifier that uses IC-specific detailed spectral information. Table 1 summarizes all these classifiers.

The first classifier did not utilize spectral information of the MEG signals but used as features the total energies of the ICs. More specifically, for each epoch, we computed the feature vectors $\mathbf{x}(n) \in \mathbb{R}^C$ with the elements:

$$x_i(n) = \sum_{j=1}^{F} {z_{ij}}^2(n), \quad \text{for } i = 1,2,\ldots,C, \quad (10)$$

where $z_{ij}(n)$ stands for the absolute value of the jth Fourier coefficient for ith IC. Note that although we computed the total energies from the frequency representations of the ICs, these features are not frequency-specific since they can be equally computed from the time-representations of the signals (see Parseval's Theorem e.g. in Oppenheim et al. (1999)). After feature extraction, we estimated the classification coefficients and biases by maximizing the penalized log-likelihood model of Eq. (7) with the l_1-norm penalty using the GLMNET software package. This classifier served as our baseline method when we investigated the importance of varying degrees of spectral information in our decoding task. Hence, we call this classifier "Baseline".

Second, we investigated whether the incorporation of coarse spectral information from MEG signals to the classifier design is advantageous. To this purpose, we computed the standard deviations of the power spectra of the ICs and used them as features in the logistic regression classifier. We call this classifier "Statistical". The rationale for selecting these features was to coarsely characterize the spectrum without being specific to any frequency. Also for this model, we estimated the classification coefficients by maximizing the ℓ_1-norm penalized logistic regression model using the GLMNET software package.

Third, we introduced a classifier utilizing spectrospatial information in a more detailed form. We made the assumption that each stimulus

[2] http://www-stat.stanford.edu/~tibs/glmnet-matlab.

category is associated with its own spectral signature that can be described by F spectral weights. This signature can be present in several ICs, and the contributions of each IC are captured by C classification coefficients. We used logistic regression with a bilinear kernel to build a classifier that fulfills these assumptions. A bilinear projection is given by:

$$h\left(c_{ki}, f_{kj} | i = 1, 2, \ldots, C, j = 1, 2, \ldots, F; \mathbf{Z}(n)\right) = \sum_{i=1}^{C} \sum_{j=1}^{F} c_{ki} f_{kj} z_{ij}(n) = \mathbf{c}_k^T \mathbf{Z}(n) \mathbf{f}_k, \quad (11)$$

where $\mathbf{Z}(n) \in \mathbb{R}^{C \times F}$ is now an entire matrix given in Eq. (1), and $\mathbf{c}_k \in \mathbb{R}^C$ and $\mathbf{f}_k \in \mathbb{R}^F$ denote the classification coefficients to be estimated for each category k, respectively. Now, each spectral classification coefficient depends on the category (index k) but not on the IC (index i). The index of the Fourier coefficients is denoted by j. A bilinear formulation of the logistic regression has been discussed previously in the context of BCI under the name *bilinear component analysis* (Dyrholm et al., 2007a). The learning of this classifier is feasible even from rather limited training data, because the total number of parameters to be estimated is only $K(C + F + 1)$ and not $K(CF + 1)$ due to the assumption that spectral coefficients and IC coefficients are separable. We call this classifier "Bilinear". We used a conjugate gradient method[3] by Rasmussen and Nickisch (2010) for maximizing the penalized log-likelihood, because the GLMNET software package cannot handle a bilinear kernel function. See Appendix A for details how we obtained the gradient and the objective function for this classifier.

Fourth, we used an unsupervised learning method which leads to a classifier similar in spirit to Bilinear classifier. While Bilinear classifier is based on the mathematically attractive assumption that different ICs share the same spectral characteristics for a given category, a more plausible assumption is that each IC can have unique spectral characteristics. To investigate whether the latter assumption would yield better prediction performance, we constructed a classifier for which we estimated spectral-weight vectors for each IC separately before estimating the classification coefficients of the ICs, somewhat like in Spectral LDA. We applied PCA to the short-time spectra of each IC and took the first principal directions as the estimates for spectral weights. We formed feature vectors by projecting the short-time spectra to one dimension (similar to Spectral LDA, but now we computed only one projection per ICs) and estimated the classification coefficients and biases by maximizing the ℓ_1-penalized logistic regression model of Eq. (7) using the GLMNET software package. Since we used PCA to estimate the spectral-weight vectors, we call this classifier "Spectral PCA". To show explicitly how Spectral PCA depends on the IC-specific spectral weights, we can write the kernel function in the form:

$$h\left(c_{ki} | i = 1, 2, \ldots C; \mathbf{x}(n)\right) = \sum_{i=1}^{C} \sum_{j=1}^{F} c_{ki} f_{ij} z_{ij}(n) = \mathbf{c}_k^T \mathbf{x}(n). \quad (12)$$

Here, $\mathbf{c}_k \in \mathbb{R}^C$ are classification coefficients to be estimated for category k, f_{ij} are spectral weights estimated using the PCA beforehand, $\mathbf{x}(n) \in \mathbb{R}^C$ is a feature vector, j is the index of the Fourier coefficients, and i is the index of the ICs. Note that spectral weights are not category-specific since they do not depend on the category index unlike the corresponding weights for Spectral LDA in Eq. (6) or for Bilinear in Eq. (11).

Fig. 2 illustrates the differences between Bilinear, Spectral PCA and Spectral LDA classifiers. All the classifiers assume that each category is represented by a specific combination of some of the ICs (the classifier finds these informative combinations; here, for simplicity, the found combinations were expected to be the same for all three classifiers). However, the way the spectral-weight vectors are estimated depends on the classifier. The spectral weights reflect the importance of different frequencies in the classification and are illustrated by black solid curves next to the spatial patterns. The shapes of the weight vectors depend either on the category (Bilinear), the estimated source (Spectral PCA), or both (Spectral LDA).

[3] http://www.gaussianprocess.org/gpml/code/matlab/util/minimize.m.

Cross-validation

We trained all the classifiers based on the penalized log-likelihood objective function, which involves selecting a suitable hyperparameter value λ that controls the extent of regularization. One possibility to automatically determine λ is to aim at the best decoding performance as measured by CV. Conventional CV procedures were not directly applicable to our data because of the temporal dependencies between successive time windows. However, time windows in different stimulus/rest blocks can be assumed to be rather independent due to the sharp onsets and offsets of the blocks. Thus, we can avoid the problem of temporal dependencies by using a "leave-one-block-out" CV procedure (Lemm et al., 2011). For this procedure, we used all epochs from one block for classifier validation and the rest of the epochs extracted from all the other blocks for classifier training during one CV loop. We trained the classifier and evaluated its performance using each training and validation folds for several values of λ, and selected the final value according to the highest average classification accuracy across the results.

To avoid bias, we estimated ICs and spectral-weight vectors separately for each CV data set. This procedure increased considerably the computational cost of the classifier training but could still be performed in a reasonable time. After we had estimated the hyperparameter value through CV, we trained the final classifier with this value using the entire training data set. We emphasize that for this classifier we only used data from the first session for classifier training (including hyperparameter estimation through CV) and reserved the entire second session for evaluating the performance of the classifiers. Hence, the data used in the final performance evaluation were independent from the training data.

Multisubject classifier

Decoding by utilizing data simultaneously from multiple subjects could inform about the extent of across-subject similarities in the modulation of brain activity. We carried out a multisubject decoding analysis by pooling the epochs of multiple subjects from both sessions and then performing the estimation of ICs, spectral feature extraction and classifier training as described previously. To assess the decoding performance, we carried out leave-one-subject-out analysis, i.e., we assessed the classification accuracy based on the data set of one subject which was left out from the training data set. We trained and tested the classifier 9 times so that each subject was left out once of the training data set, and computed the mean decoding performance across the test results. We balanced the number of category labels between the categories separately for each subject to ensure that we had the same number of epochs per category for classifier training from each subject.

Interpretation of Spectral LDA

Interpretation of spectrospatial patterns

Spectral LDA offers a possibility to investigate how spectral characteristics of the rhythmic brain activity change according to stimulus categories. Although the initial number of features in the model was relatively high ($=256$), LASSO forced the coefficients of the useless and/or correlated features to zero, thereby considerably simplifying the interpretation of the classifiers.

We concentrated on the visualization of the ICs (both their spectral and spatial characteristics) which corresponded to the *positive* classification coefficients only. As explained earlier, it is plausible to assume

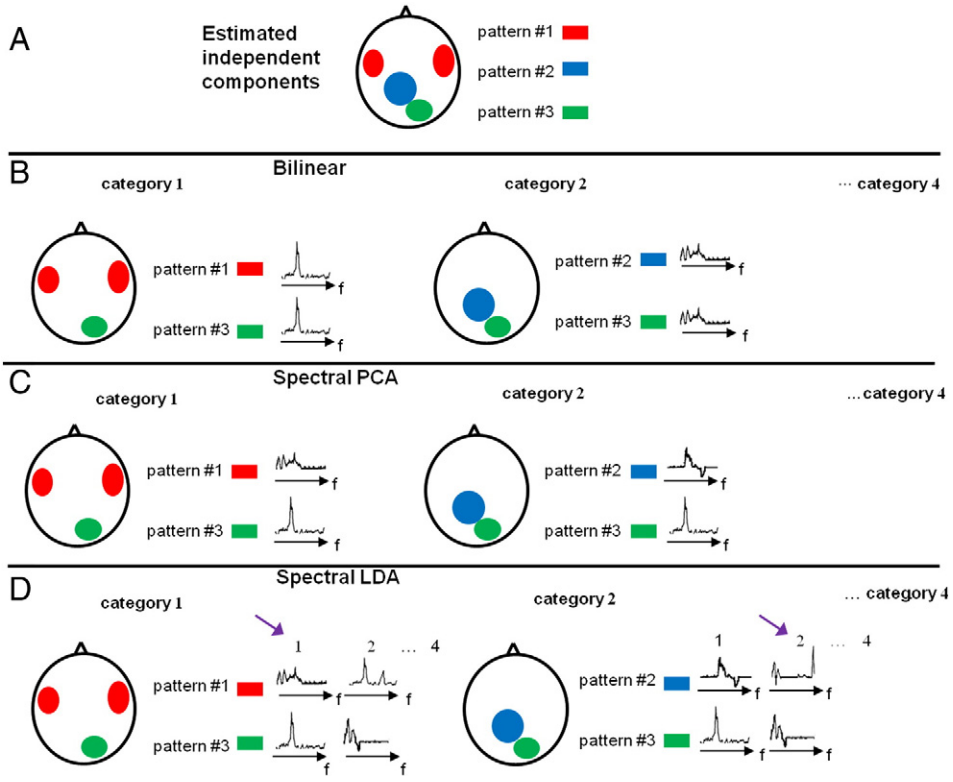

Fig. 2. Schematic illustration of the classifiers Bilinear, Spectral PCA and Spectral LDA. (A) A schematic brain with three estimated ICs. (B) Classifier Bilinear assumes the same spectral-weight vector for each IC within each category, but each category has unique weights (see the similarities and differences in the shapes of the weight vectors). (C) Spectral PCA assumes a unique weight vector for each IC but the vectors are shared across categories (see especially the shape of the vector of a shared pattern #3 that remains fixed across categories). (D) Spectral LDA is the most flexible classifier, as it assumes unique spectral-weight vectors for each IC, similar to Spectral PCA, but the spectral weights are also category-specific, similar to Bilinear: weights in the column denoted by 1 (or 2) are specific for discriminating category 1 (or 2) from the other categories. We assume that the classifier automatically captures category-specific spectral information by assigning positive classification coefficients to informative spectral projections (the two arrows emphasize the columns which are expected to be associated with high positive classification coefficients).

that a positive classification coefficient for some category is associated with those ICs and spectral weights that are estimated by discriminating the given category from the other categories (because LDA maximizes the class separation in the projected space by definition). On the other hand, because the decrease in the value of this same projection means that the given category becomes less probable with respect to other categories, negative classification coefficients do not have interesting interpretation in this model.

The spectral weights themselves can be either positive or negative. A positive value of a certain frequency bin means that an increase in the power at this frequency band makes the category of the corresponding classification coefficient more probable. Similarly, a negative value means that a decrease in the power at this frequency increases the probability of the category. To make the interpretation of the results meaningful, we normalized all classification coefficients and spectral weights by the standard deviation of the input data corresponding to each coefficient.

Across-subject cluster analysis

In Spectral LDA, each classification coefficient is associated with one spectrospatial pattern given by the corresponding IC. It is convenient to visualize this pattern using three adjacent plots: one for the spatial pattern, one for the spectral-weight vector, and one for the real spectrum. If the classifier contains several positive classification coefficients, it is obvious that the interpretation of the findings becomes complicated due to the high number of plots. For some subjects and categories, the number of positive classification coefficients was so low that the findings were relatively easy to interpret. However, for other subjects and categories, the number of positive classification coefficients was higher (e.g. more than 10), making the interpretation of the findings more difficult. One possible strategy to simplify interpretations is to investigate features associated with the highest classification coefficients only. The drawback of this approach is that it is not always obvious that high classification coefficients are related to the most interesting findings. For instance, it is possible that some of the coefficients with high magnitude are related to suppression of noise whereas some coefficients with a low magnitude are related to neuroscientifically interesting phenomena (Blankertz et al., 2011). Perhaps a more reliable strategy for simplifying the interpretation is to visualize the most consistent features across subjects. To enhance such visualization, we designed a clustering procedure to find similar features across subjects for each category. The procedure consisted of the following steps:

1. *Construction of the similarity matrix*:
   We first identified the spectral-weight vectors associated with positive classification coefficients and pooled them across the classifiers of the subjects (separately for each category). Then, for each category k, we formed a similarity matrix by computing the pairwise similarities between the weight vectors based on a cross-correlation sequence, that is, cross-correlations computed when one of the spectra is shifted towards lower and higher frequencies. We computed the cross-correlations to account for individual variability in peak frequencies in the rhythmic oscillatory activity. For each vector pair $\mathbf{f}_m, \mathbf{f}_n$ (the subscripts of the category and IC are omitted here for convenience), we computed the cross-correlations across the frequency

range [−2.5 Hz 2.5 Hz], and normalized the values so that the auto-correlations at zero lag were identically 1.0. We selected the maximum value as the similarity value and set all the negative cross-correlations to zero because we were interested in positive correlations. Hence, the elements of the similarity matrix were numbers between zero and one given by:

$$h_{sim}(m,n) = \max\{0, \max[\mathrm{xcorr}(\mathbf{f}_m, \mathbf{f}_n)]\}, \quad (13)$$

where xcorr() denotes the normalized cross-correlation sequence between the two vectors as described above.

2. *Adding spatial constrains to clustering*:
We required that only spatially similar ICs can be clustered together. To this aim, we computed a spatial binary similarity matrix by thresholding the magnitudes of the spatial patterns of the ICs (corresponding to the pooled spectral-weight vectors) and investigated pairwise whether the thresholded patterns overlapped. In this matrix, the value one denoted overlap between the patterns (at least in one channel) and the value zero meant that the patterns did not overlap. We also accounted for the hemispheric symmetry of the brain by flipping one pattern from each pair across the midline into the opposite hemisphere and investigated a possible overlap with another pattern (and gave a value 1 also in the case of symmetric "overlap"). As a result, we obtained a binary matrix with elements $b(m, n)$ denoting the pairs of spatially similar ICs. We then weighted the similarity matrix of the LDA weight vectors with this matrix and transformed the resulting similarity matrix to a dissimilarity matrix. Hence, the elements of the final dissimilarity matrix used for clustering were given by:

$$h(m,n) = 1 - h_{sim}(m,n)b(m,n). \quad (14)$$

We constructed one dissimilarity matrix for each category.

3. *Clustering and post-processing*:
We clustered the data based on the obtained dissimilarity matrices using the average-linkage agglomerative hierarchical clustering algorithm (Hastie et al., 2009).[4] A clustering cutoff value and a threshold for the spatial patterns were manually adjusted so that the spectrospatial characteristics of different clusters became visible. After clustering, it was possible that clusters contained more than one pattern from single subjects. To simplify interpretation, in each cluster we retained only the pattern corresponding to the highest classification coefficient for each subject. Hence, after post-processing, the maximum number of spectrospatial patterns in each cluster was nine: one pattern per subject.

4. *Visualization*:
We visualized the spectral-weight vectors and spatial patterns within the clusters on top of each other using a distinct color for each subject. To facilitate the overall interpretation of the findings, we sorted the clusters according to their size from the largest to the smallest.

Results

Classification performance

Fig. 3 presents the mean classification accuracy of the classifiers across subjects. The mean accuracy is shown together with the standard error of mean (SEM). We report results obtained with 4-s window length and 2/3 window overlap, since these parameters yielded the highest performance across all classifiers.[5] The mean classification accuracy for all classifiers was well above the chance level (0.25). Spectral

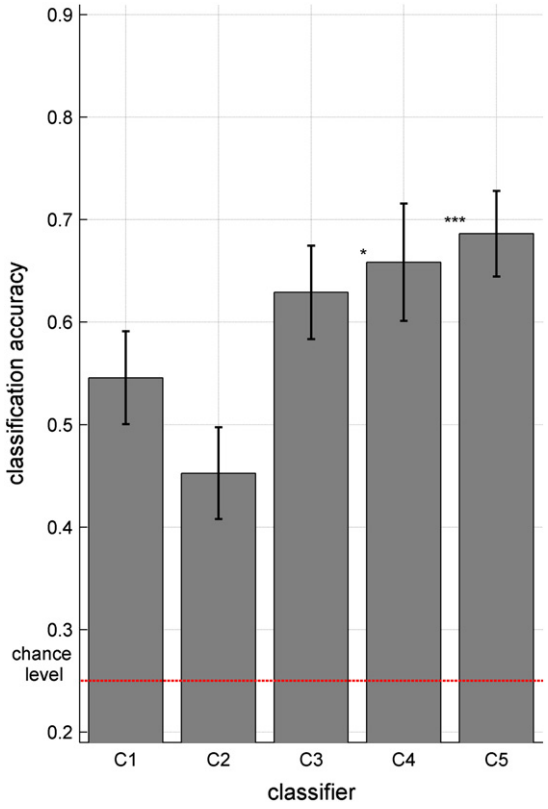

Fig. 3. Mean (SEM) classification accuracy across the subjects for the five classifiers: *C1* = Baseline, *C2* = Statistical, *C3* = Bilinear, *C4* = Spectral PCA, and *C5* = Spectral LDA. The shown significance levels (marked by asterisks) refer to the comparison of the classifiers against Baseline.

LDA provided the best performance with the mean accuracy of 0.686 (i.e., 68.6% of the epochs were correctly classified for the test data). The result was significantly higher (at $\alpha = 0.001$, Bonferroni corrected) compared with that of Baseline (mean performance 0.546; paired *t*-test; $p = 0.0002$). Also the result of Spectral PCA (mean performance 0.659) was statistically significant (at $\alpha = 0.05$, Bonferroni corrected) compared with that of Baseline ($p = 0.0076$). The corresponding result of Bilinear (mean performance 0.629) was not statistically significant after the Bonferroni correction ($p = 0.0188$). The result of Spectral LDA was significantly higher (at $\alpha = 0.05$, Bonferroni corrected) in the comparison against Bilinear ($p = 0.0024$) but not in the comparison against Spectral PCA ($p = 0.3198$). The classifier Statistical performed considerably worse (mean performance 0.453) than any other classifier.

Fig. 4 shows the subject-wise classification results (ordered from the highest to the lowest based on the classification accuracies of Spectral LDA). The classification accuracy varied considerably between individuals for all classifiers. Note that Spectral LDA yielded relatively good performance for all subjects (the results varied between 0.485 and 0.838) whereas Spectral PCA yielded very good results for some subjects but relatively poor results for some others (the results varied between 0.382 and 0.882).

To test whether Spectral LDA would also work at single-individual level, we compared its classification accuracy with the performance of a random classifier which assigns the data points to the classes randomly, with equal probabilities. The results of Spectral LDA were significantly above the chance level (at $\alpha = 0.001$, Bonferroni corrected) for all subjects ($p < 8 \times 10^{-6}$ for all the subjects); see Pereira et al. (2009) for details of the test.

Table 2 presents the confusion matrices of the classifiers, with the rows denoting the true and the columns the estimated categories. The Spectral LDA provided the highest mean classification accuracies for the categories "auditory" (0.575) and "visual" (0.830). Bilinear yielded

[4] We used the implementation from the Statistics Toolbox of Matlab.

[5] This window parameter combination did not yield the highest accuracy for Spectral LDA, but it was used to avoid bias in the comparison of the results.

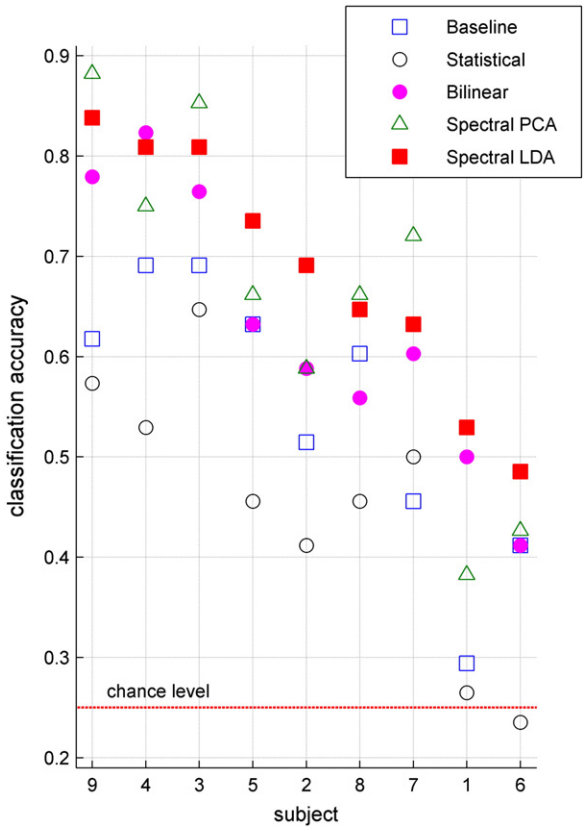

Fig. 4. Accuracies of the tested classifiers for individual subjects. The subjects are ordered according to the individual mean classification accuracy.

the best mean accuracy for the category "tactile" (0.804). Both Spectral LDA and PCA provided the highest mean accuracy for the category "rest" (0.549). For classifiers Baseline, Statistical, and Bilinear, the most difficult category to decode was "auditory", and epochs from this category were often classified either as "tactile" or "rest". For Bilinear, Spectral PCA and Spectral LDA, the most difficult category to decode was "rest", which was most often confused with the "auditory" category. The "auditory" category was difficult to decode also with these classifiers.

Table 2
Average confusion matrices across subjects for all classifiers. Rows denote the true and columns the estimated category. The reported values are mean classification accuracies across subjects, and the corresponding standard errors are given inside parentheses. The correct classification results are shown in bold.

	Auditory	Visual	Tactile	Rest
Baseline				
Auditory	**.477 (.124)**	.092 (.049)	.157 (.052)	.275 (.106)
Visual	.092 (.037)	**.654 (.097)**	.150 (.049)	.105 (.069)
Tactile	.209 (.073)	.059 (.035)	**.529 (.064)**	.203 (.075)
Rest	.235 (.067)	.124 (.048)	.118 (.037)	**.523 (.116)**
Statistical				
Auditory	**.379 (.107)**	.118 (.071)	.288 (.081)	.216 (.071)
Visual	.118 (.040)	**.444 (.100)**	.294 (.105)	.144 (.038)
Tactile	.222 (.054)	.111 (.055)	**.490 (.068)**	.177 (.040)
Rest	.163 (.051)	.157 (.051)	.183 (.075)	**.497 (.067)**
Bilinear				
Auditory	**.490 (.077)**	.092 (.037)	.216 (.045)	.203 (.049)
Visual	.105 (.036)	**.752 (.067)**	.065 (.025)	.078 (.035)
Tactile	.085 (.026)	.052 (.029)	**.804 (.049)**	.059 (.022)
Rest	.255 (.056)	.137 (.028)	.137 (.056)	**.471 (.070)**
Spectral PCA				
Auditory	**.556 (.110)**	.137 (.070)	.150 (.048)	.157 (.056)
Visual	.072 (.027)	**.804 (.064)**	.059 (.033)	.065 (.030)
Tactile	.118 (.046)	.078 (.029)	**.726 (.090)**	.078 (.033)
Rest	.261 (.066)	.059 (.020)	.131 (.039)	**.549 (.064)**
Spectral LDA				
Auditory	**.575 (.078)**	.059 (.024)	.137 (.045)	.229 (.052)
Visual	.098 (.045)	**.830 (.045)**	.033 (.014)	.039 (.022)
Tactile	.105 (.036)	.033 (.017)	**.791 (.061)**	.072 (.045)
Rest	.222 (.046)	.131 (.040)	.098 (.037)	**.549 (.065)**

Interpretation of spectrospatial patterns

Figs. 5 and 6 show two examples of spectrospatial patterns learned by the Spectral LDA (the confusion matrix of the results of this subject is shown in Table 3). Fig. 5 (Subject 4, visual category) is an example of an extremely sparse solution, where only two classification coefficients in the final classifier were nonzero for the given category. The spectrospatial pattern associated with the higher classification coefficient (in the first row) shows that the suppression (reflected by the negative sign of the spectral weights) of the 10-Hz power in the occipital cortex increased the probability of the visual category. The phenomenon can be verified from the real spectrum on the right: the 10-Hz activity was strongly present when the visual stimulus was absent but suppressed when the stimulus was present. The finding can be related to the classic alpha rhythm originated in the posterior cortex and known to be suppressed during visual processing or attention (Hari and Salmelin, 1997).

The second pattern shows that the increase of the 12-Hz power in the Rolandic areas made the visual category more probable. This pattern may not be directly related to the processing of the visual input but rather to the absence of processing tactile stimuli. It likely reflects the well-known Rolandic mu rhythm that is known to be suppressed during tactile stimulation and sensorimotor activity (Hari and Salmelin, 1997). Note that the classification coefficient of the "visual pattern" is much larger (2.52) than that of the "Rolandic pattern" (0.18), indicating that the suppression of the occipital alpha was a by far more discriminative feature. Although the finding may seem obvious, we want to point out that the classifier found it automatically from high-dimensional MEG recordings without any assumptions about the spatial location of the feature, and the frequency-band of interest was only coarsely specified.

Fig. 6 shows the corresponding results for the category "rest" for the same subject. Overall, these results are more difficult to interpret than the plots of Fig. 5, which is not surprising given the much lower classification accuracy of "rest" (0.647) than "visual" (0.941) in this subject. The first pattern shows increased occipital alpha, likely due to decreased visual processing and attention during the rest periods. Note that in contrast to the plot in Fig. 5, the positive sign of the spectral-weight vector now suggests increased 10-Hz activity. The increase in the power can be verified from the true spectrum on the right. An interesting component is IC #4 in which the 10-Hz activity is decreased during "rest" compared with stimulation; it is somewhat puzzling that the spatial pattern overlaps strongly with the pattern of IC #1. From the viewpoint of decoding methodology, particularly interesting are ICs #5 and #6. Again, the decoder finds that decreases in rhythmic activity are predictive to "rest"; however, in these components the differences in the spectra (on the right) are not clear, and the decoding methodology may be necessary to find this connection. Regarding the remaining components, ICs #2 and #7 seem to be similar to #1 in the sense of connecting "rest" with increased rhythmic activity in sensory cortices, but now close to Rolandic (#2) and temporal (#7) areas. IC #3 is presumably an artifact.

Figs. 7–10 show across-subject clustering results for the Spectral LDA for the categories "auditory", "visual", "tactile" and "rest", respectively. Three largest clusters are shown for each category. The most interesting finding for the auditory category (Fig. 7) is the second cluster, which indicates suppression of the band power (at 10–14 Hz, depending on the subject) in the left temporal cortex that would agree spatially with the generation site of the temporal-lobe tau rhythm which however has been reported to occur at slightly lower frequencies

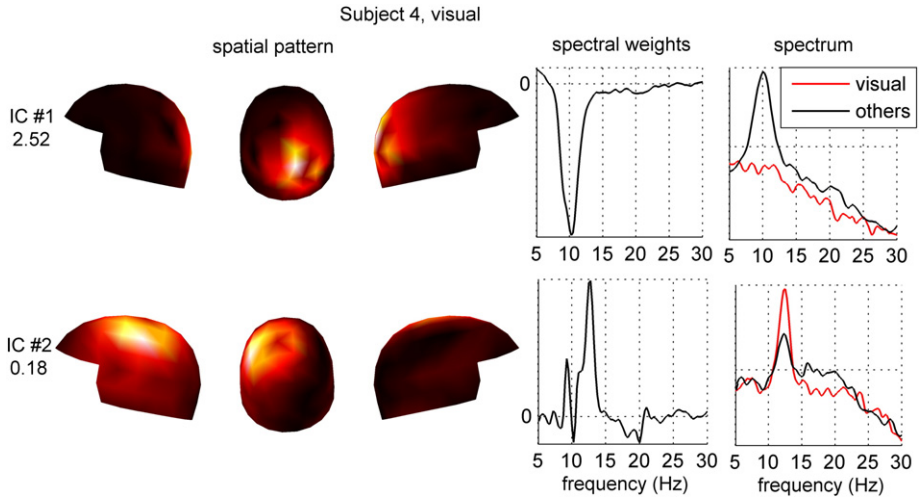

Fig. 5. Spectrospatial patterns (ICs and their associated LDA weight vectors) found for Subject 4 for the category "auditory". Rows denote all the found patterns in the order of decreasing classification coefficients (the values of the classification coefficients are shown on the left), and columns denote the spectrospatial characteristics of each pattern: the spatial pattern of the IC on the sensor helmet viewed from left, back, and right (leftmost column), the LDA weight vector associated with the IC (middle column), and the "real" spectra of the IC at a logarithmic scale (rightmost column). In the rightmost column, the black color of the real spectrum denotes the average spectra computed across the epochs belonging to the category of interest, and the red color indicates the corresponding spectrum computed across all the other categories. See Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

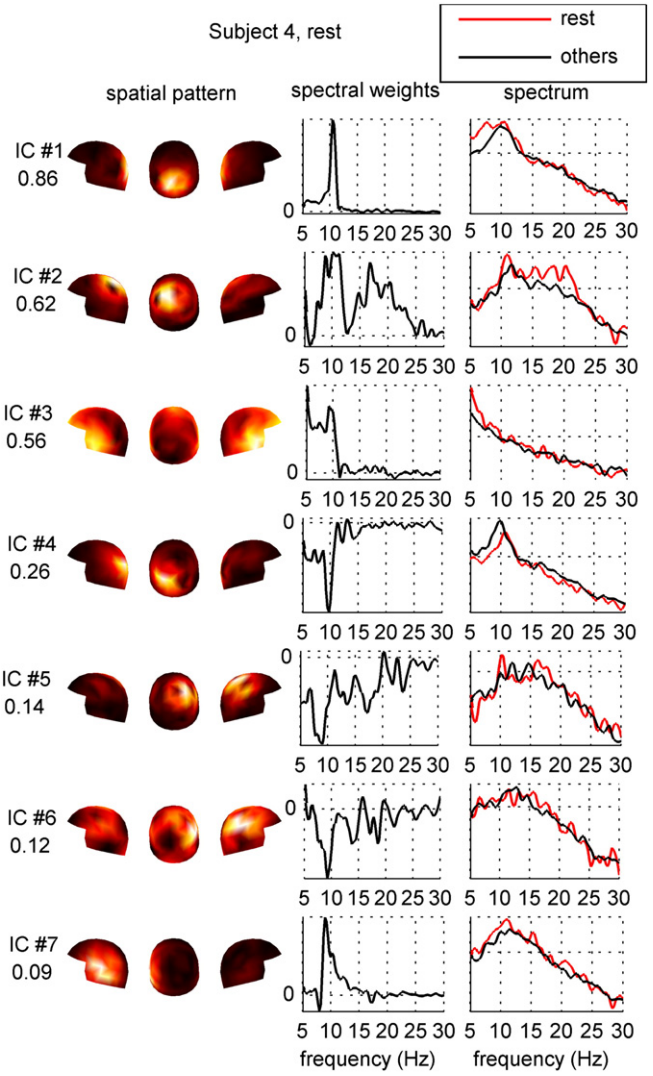

Fig. 6. Spectrospatial patterns found for Subject 4 for the category "rest". See the caption of Fig. 5 for the meaning of the plots and Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

of 8–10 Hz (Lehtelä et al., 1997). The first and the third clusters seem to show increases in the Rolandic mu and occipital alpha rhythms, respectively. The signs of the spectral weights are plausible because the tactile and visual stimuli were not present.

The first cluster of the visual category (Fig. 8) indicates strong suppression of the occipital alpha rhythm, suggesting involvement of visual processing. The second cluster was also located in the posterior cortex, with increase in power in the 8–10 Hz band, implying that alpha rhythms of different center frequencies may behave functionally differently. The third cluster, with weak (0.09) weight overlaps with Rolandic areas and suggests suppression of the two main frequency components of the mu rhythm.

The clusters of the tactile category (Fig. 9) were all located around the Rolandic areas. The first cluster showed suppression at around 10–13 Hz as well as at 19–22 Hz (however, one subject showed suppression already at 7 and 13 Hz). These findings could reflect suppression of the Rolandic mu rhythm due to the tactile stimuli, here with considerably stronger weights than in the lowest panel of Fig. 8. Also the second and the third cluster indicated suppressed activity in two distinct frequency bands, but the frequencies were lower than in the first cluster (the approximate frequency bands were 6–10 Hz and 14–17 Hz). Because the subjects of the second and third cluster were different from those of the first cluster, also these findings may reflect Rolandic mu but with notable intersubject variation in the characteristic frequencies. In any case, most of these findings agree with the current literature because the mu rhythm typically contains both 8–13 Hz and 15–25 Hz frequency bands (Hari and Salmelin, 1997).

The largest cluster of the category "rest" (Fig. 10) showed variable spectral characteristics across a wide frequency band (5–25 Hz) in the parietal cortex, mainly indicating increased oscillatory activity. This result may be related to increased Rolandic mu due to the absence of tactile stimuli, or it might reflect spontaneous variations not related

Table 3
The confusion matrix of the subject 4 whose spectrospatial features are shown in Figs. 5 and 6. Rows denote the true and columns the estimated category memberships. The correct classification results are shown in bold.

Subject 4	Auditory	Visual	Tactile	Rest
Auditory	**.941**	0	0.059	0
Visual	0	**.941**	0.059	0
Tactile	0.177	0.118	**.706**	0
Rest	0.177	0.117	0.059	**.647**

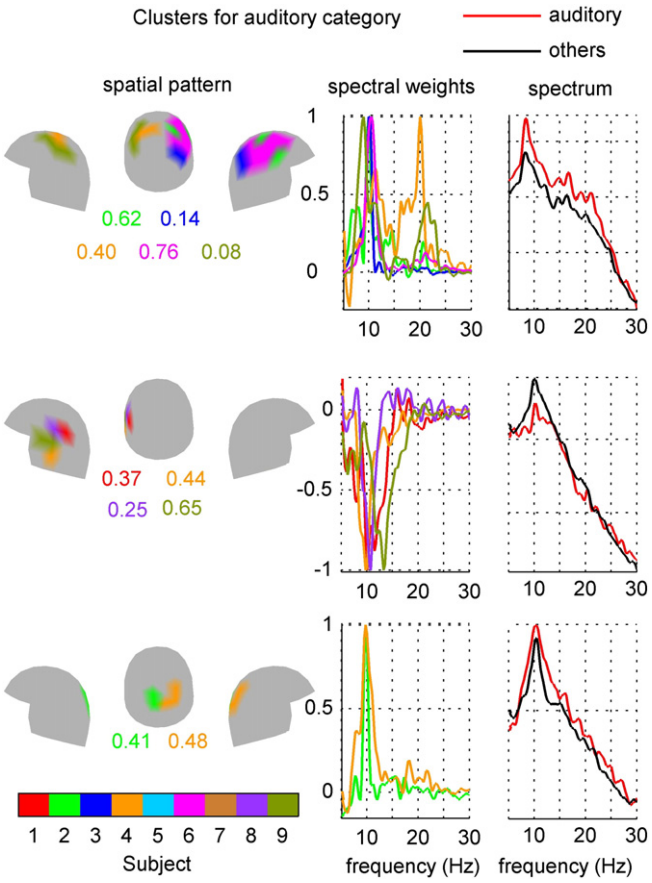

Fig. 7. Across-subject clustering results for the auditory category. Rows denote clusters and columns denote the spectrospatial characteristics of each cluster: the spatial patterns of the ICs on the sensor helmet viewed from left, back, and right (leftmost column), the spectral-coefficient vectors associated with the ICs (middle column), and the "real" spectra of the ICs at a logarithmic scale (rightmost column). Colors of the spatial patterns and spectral-coefficient vectors correspond to different subjects according to the color map (the values of the classification coefficients are shown below the helmets using the same color code). In the rightmost column, the black color of the real spectra denotes the average spectra computed across the epochs (and subjects) belonging to the category (and cluster) of interest, and the red lines show the corresponding spectra computed across all the other categories. See Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

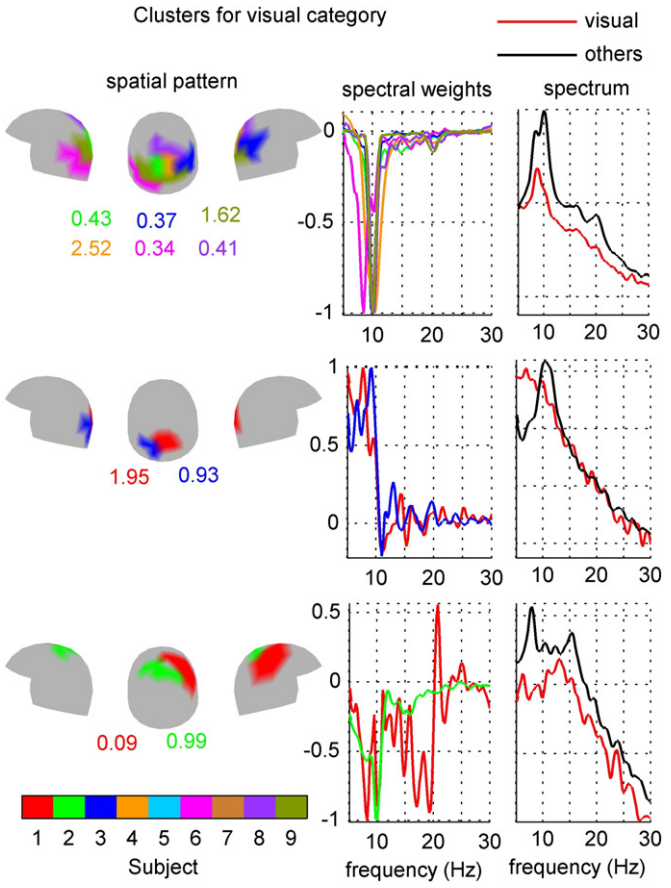

Fig. 8. Across-subject clustering results for the visual category. See the caption of Fig. 7 for the meaning of the plots and Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

to tasks or stimuli. The second cluster could be related to increased occipital alpha power due to decreased visual attention. The origin of the pattern in the third cluster is unclear. Note that the sign is negative, i.e. the decoder found suppression specifically related to "rest", and thus it is possible that the results reflect increased activation (and decreased rhythmic activity) in the Rolandic cortex.

Multisubject classifier

The mean classification accuracy across subjects for the multisubject Spectral LDA classifier was 0.46 ($SEM = 0.03$). The results of individual subjects were all statistically significantly higher than the results of a random classifier (at $\alpha = 0.05$, Bonferroni corrected; binomial test; the p-values were between 2.66×10^{-10} and 0.0064), but the result was significantly worse (alpha = 0.001) than that obtained by the subject-specific classifiers (paired t-test, $p = 0.00013$). Table 4 presents the confusion matrix of the classification results. Similar to single-subject classifiers, the categories "visual" and "tactile" were the easiest categories to decode. The classification of epochs belonging to the "rest" category was completely random.

Assessment of Spectral LDA design

We evaluated the contribution of the key design steps for Spectral LDA, including the selection of the optimal windowing parameters, the use of spatial filtering using ICA, the use of regularization in the estimation of the spectral-weight vectors, and the effect of frequency band on the decoding performance. The evaluation of these steps is important especially for future studies.

Effect of windowing parameters

We first evaluated the effect of window length and window step size (which controls window overlap) on the classification performance. While computing the results for different window lengths, we did not change the window step size so that the number of windows for classifier training was kept constant. In addition, we always evaluated the test performance based on non-overlapping windows. Fig. 11A shows that window length had notable effect on the classification results: the 1-s time window yielded a mean accuracy of 0.553 which with the 4-s time window improved to 0.686 (change statistically significant at $\alpha = 0.001$; paired t-test; $p = 0.00005$).

Increasing the window length from 4 to 5 s did not any more change significantly the classification accuracy. It is possible that the 5-s window was suboptimal because the number of independent epochs used in the classifier training decreased as the window length was increased. Another possible explanation is that the non-stationarity of the time-series within the category-blocks started to degrade the classification performance.

We examined both these possibilities in more detail. First, we found that even when exactly the same number of independent training epochs was used, the classifier trained on 5-s epochs did not outperform

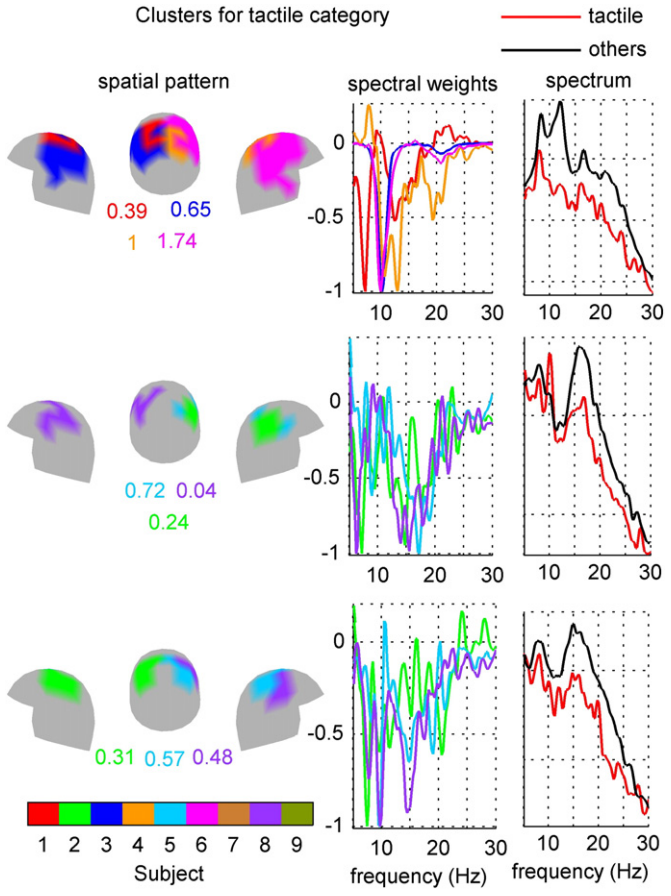

Fig. 9. Across-subject clustering results for the tactile category. See the caption of Fig. 7 for the meaning of the plots and Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

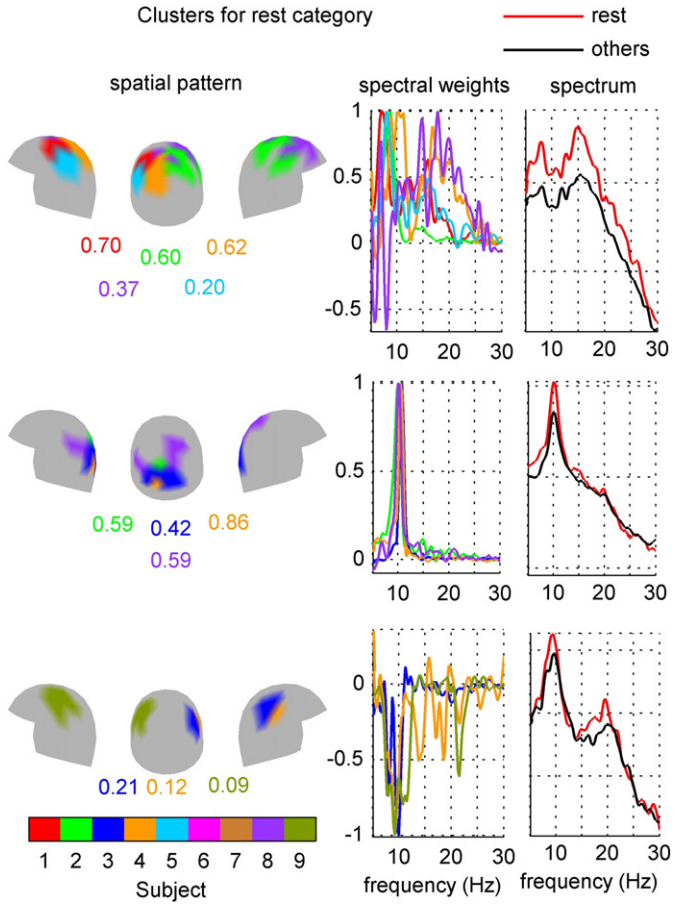

Fig. 10. Across-subject clustering results for the rest category. See the caption of Fig. 7 for the meaning of the plots and Interpretation of spectrospatial patterns section on how to interpret the model coefficients.

the classifier trained on 4-s epochs (mean performances 0.662 and 0.680, respectively). Second, to find out the effect of non-stationarity of the MEG signals, we trained one classifier with epochs appearing immediately (0–3 s) after the onset of each stimulus/rest block and another with epochs appearing later (3–6 s after the onset of each block). We used 3-s time-windows because the duration of the shortest blocks was 6 s and we wanted to ensure that we had exactly the same number of epochs for training both classifiers and that at least one epoch was extracted from each block. The accuracies did not differ (mean performances 0.493 and 0.470 for classifiers using early and late information, respectively; paired t-test; $p = 0.19$).

Fig. 11B shows the effect of time-window overlap on classification performance. The maximum performance, with mean accuracy of 0.711, was attained with 2/3 overlap; this performance tended to be higher than our original mean accuracy of 0.686 ($\alpha = 0.05$; paired t-test, $p = 0.055$). The tendency in the results is that the classification accuracy first increases and then decreases as a function of the window overlap. Thus it seems that some time-window overlap is useful.

Effect of spatial filtering

Fig. 12A shows the impact of ICA and the number of estimated ICs on the decoding performance. Without spatial filtering, i.e., when we trained and tested the classifiers for each subject directly based on preprocessed signals from the 204 gradiometer channels, the mean accuracy was 0.619 (bar on the left). The use of ICA (training and testing the classifier after estimating 64 ICs from 204 gradiometer signals) improved the mean classification accuracy to 0.686 ($\alpha = 0.05$; paired t-test; $p = 0.014$).

Fig. 12B shows the classification results as a function of the number of estimated ICs. The mean classification accuracy tended to increase together with the number of ICs until 40 ICs were reached (the accuracy was significantly higher with 40 than 10 ICs at $\alpha = 0.01$; paired t-test; $p = 0.0012$). Above 40 ICs, the results fluctuated more but the decoding performance remained high (the highest mean performance 0.701 was obtained using 60 ICs; this result was not significantly higher than the one based on 40 ICs; paired t-test; $p = 0.364$). Note that although the number of gradiometer channels was 204, the effective dimension of the data after the SSS preprocessing was 64, rendering the estimation of a higher number of ICs meaningless.

Effect of regularization

The estimation of spectral-weight vectors in our study was based on a heavily regularized version of LDA, where the covariance matrix was effectively assumed spherical. When spectral-weight vectors were estimated using the LDA without regularization, the mean classification accuracy was 0.472. Our original result (the mean classification accuracy

Table 4
The mean (SEM) confusion matrix of the leave-one-subject-out analysis carried out with the Spectral LDA. Rows denote the true and columns the estimated category memberships. The correct classification results are shown in bold.

All subjects	Auditory	Visual	Tactile	Rest
Auditory	**.439 (0.047)**	0.216 (0.056)	0.164 (0.036)	0.199 (0.036)
Visual	0.193 (0.050)	**.567 (0.079)**	0.135 (0.037)	0.123 (0.037)
Tactile	0.140 (0.047)	0.076 (0.022)	**.632 (0.096)**	0.170 (0.061)
Rest	0.316 (0.043)	0.257 (0.053)	0.205 (0.032)	**.240 (0.046)**

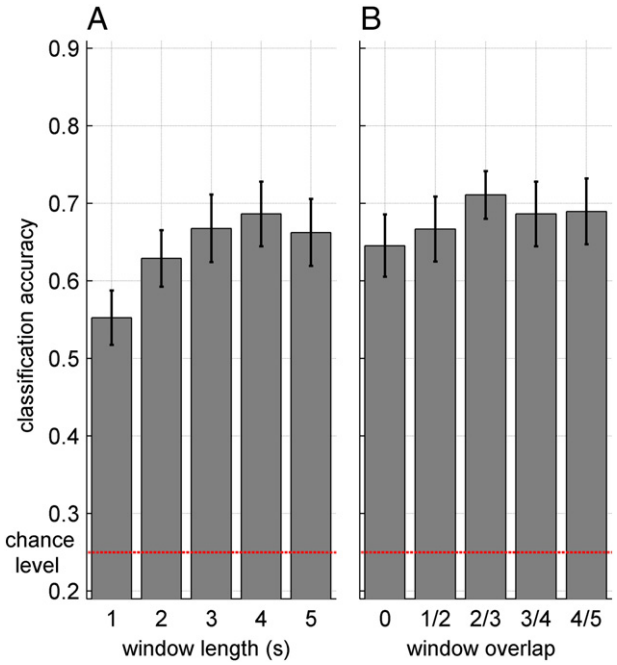

Fig. 11. The effect of window parameters on the classification performance of Spectral LDA: (A) the effect of window length, and (B) the effect of window step size (presented as fractions of the overlap between two successive windows). For instance, window overlap = 0 means that successive windows did not overlap at all (but are adjacent to each other), and window overlap = 1/2 indicates 50% overlap between the successive windows.

0.686) was significantly better than the unregularized one at $\alpha = 0.001$ (paired t-test; $p = 0.00026$). The poor performance of the unregularized LDA was not surprising, because the dimension of the short-time spectra was roughly equal to the number epochs, making the estimation of the inverse of the covariance matrix unstable.

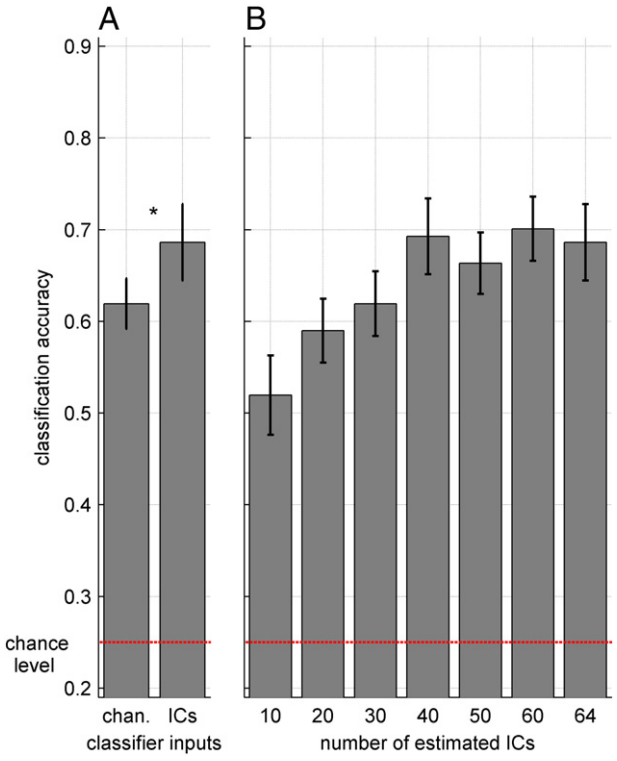

Fig. 12. The effect of ICA on the classification performance of Spectral LDA. A) mean classification accuracies without (a left bar labeled as "chan.") and with (a right bar labeled as "ICs") applying ICA to the 204 gradiometer signals. B) The effect of the estimated number of ICs from 10 to 64 on the classification accuracy.

Effect of frequency range

We limited our analysis to frequencies below 30 Hz because we assumed that higher frequencies might not be useful in our decoding task due to their low SNR. Because the limit was arbitrary, we tested whether the decoding accuracy would improve if we extended the frequency range of interest to contain the gamma-band (frequencies up to 80 Hz). The classification accuracy based on the 5–80-Hz range (mean performance 0.662) was not higher than based on our original 5–30-Hz range (mean performance 0.686). The visual inspection of the LDA weight vectors verified that category-discriminative information was dominantly present below 30 Hz (the values of the weights above 30 Hz were zero or close to zero).

Discussion

In this paper, we introduced a novel general-purpose MEG decoding method, Spectral LDA, for the investigation of whole-brain rhythmic activity during distinct conditions. We assessed the usefulness of the method in a naturalistic setup comprising visual, auditory and tactile stimuli. Spectral LDA assumes that distinct brain activations (sources) are characterized by unique spectral patterns of cortical rhythms. Spectral LDA also assumes that the spectral patterns may vary according to the brain state. Although these assumptions are rather obvious as far as brain function is concerned (see e.g. Singer, 1993; Buzsáki and Draguhn, 2004), decoding of different brain states and different applied stimuli from the very noisy and relatively short single-trial MEG traces is far from trivial. However, Spectral LDA performed very well in this difficult task and was superior to three out of four classifiers based on more restrictive assumptions concerning spectrospatial information in the data (Figs. 3, 4; these three classifiers were Baseline, Statistical, and Bilinear). The better performance of Spectral LDA compared with Baseline implies the usefulness of the spectral content of MEG signals in the given decoding task. In addition, detailed spectral information was more useful than unspecific spectral information (as Spectral LDA performed better than Statistical). The results also indicate that it is wise to estimate spectral signatures separately for distinct ICs instead of estimating a common spectral feature for each category and using that with each IC (Spectral LDA was better than Bilinear). Spectral PCA provided comparable classification accuracy with Spectral LDA, but the visualization of the final results is more meaningful with Spectral LDA.

The investigation of the trained classifiers showed that the Spectral LDA can provide neuroscientifically relevant information about state-dependent changes of rhythmic brain activity. To make the investigation of the spectrospatial features across subjects easier, we developed a clustering method that took into account some functional and anatomical differences across individuals. The cluster analysis revealed that many subjects shared similar spectrospatial features (Figs. 7–10). The spatial patterns of the dominant clusters agreed with functionally meaningful brain areas. Naturally, the verification of different findings will require additional studies with new data sets collected from a larger number of subjects. In any case, already these findings are neurophysiologically encouraging, because the method seeks patterns in an exploratory manner based on minimal *a priori* assumptions concerning the brain areas and frequencies of interest.

The decoding performance of the classifier trained with the data of several subjects and tested with a subject not included in the training data showed that Spectral LDA was capable of utilizing common spectrospatial features across subjects. However, the multisubject classifier was not competitive against the single-subject classifiers, suggesting notable interindividual variation in the rhythmic brain activity. Across-subject clustering results supported this view: cluster

analysis revealed similar spectrospatial characteristics across subjects, but each cluster contained data only from a subset of subjects.

The choice of the time-window length is a key design parameter in the type of analysis presented here because it determines the spectral and temporal resolution of the analysis. We found that the window length had a notable effect on the classification performance, the 4-s time window yielding the best results (although not significantly different from those with a 5-s window). Non-stationarity of the MEG time-series possibly explained the lack of further improvement of the classification performance with longer than 4-s time windows. Using at least some time-window overlap turned out to be beneficial (the best results were obtained with 2/3 overlap), but the selection of the amount of overlap was not critical.

Because the signal captured by each gradiometer sensor is a mixture of activity from different neuronal sources, we assumed that linear unmixing of the 204 gradiometer signals using ICA would map the data closer to the sources and thus improve decoding performance. To investigate the usefulness of ICA in our decoding task, we trained Spectral LDA also using the 204 gradiometer signals as inputs and compared the results with those obtained using ICA. The direct decoding based on the gradiometer signals degraded the classification accuracy significantly, showing that spatial filtering is an important part of the classifier design. In our original design, we estimated 64 ICs, corresponding to the dimension of the data after preprocessing using the SSS method. The decoding performance using a much smaller number of ICs (10–30) was considerably worse, implying that discriminative information was distributed across many ICs. On the other hand, once a sufficiently high number of ICs was estimated, the decoding results remained high even when more ICs were estimated. This result was expected because of the efficient regularization with LASSO which was capable of removing redundant and uninformative components. Thus, the selection of the number of ICs is not critical as long as the number is high enough. One benefit of ICA is that it provides an easily interpretable model. However, in the future we will also assess the suitability of other spatial filters, such as those based on common spatial pattern (CSP) (Blankertz et al., 2008), for Spectral LDA.

A major challenge in MEG-based decoding of spontaneous activity is the variability of the statistical properties of the MEG signals between sessions and subjects. The session-to-session variability can be partially alleviated by using appropriate regularization to avoid overlearning of the classifier. In Spectral LDA design, we applied regularization at two stages: first as a part of LDA in the estimation of spectral-weight vectors, and then in the training of the logistic regression classifier using the sparsity-enforcing ℓ_1-norm penalty. We used a heavily regularized version of LDA to estimate spectral weights for each IC by assuming that the covariance matrices of the class-conditional distributions of the estimated short-time spectra were spherical. Even though this assumption may not be realistic, the regularization improved the classification performance significantly, suggesting that the prevention of overlearning is a critical requirement for successful classification when the training and test data sets come from independent sessions. The use of the sparsity-enforcing penalty in the final classifier training was important not only in the prevention of overlearning, but also for automatic selection of important features in the final model.

In the current study, we limited the frequency range of interest to 5–30 Hz. Extending this range to the gamma band (up to 80 Hz) did not improve classification, possibly because the highest SNR for MEG signals and their changes due to sensory stimuli occurs below the gamma band.

Besides Spectral LDA, also two other classifiers utilizing detailed spectral information (Spectral PCA and Bilinear) performed well in this study. Spectral PCA yielded even higher classification accuracy for some individual subjects than Spectral LDA, suggesting that the unsupervised estimation of the spectral weight-vectors may, in some cases, be less prone to overfitting compared with the supervised estimation. On the other hand, poor results of Spectral PCA for other subjects indicate that the directions of the maximal variances in the short-time spectra of the ICs could not always capture category-discriminative information. The classification performance of Bilinear was inferior to Spectral LDA and PCA but nevertheless also this classifier performed relatively well. In the current study, we used a bilinear model imposing rank $= 1$ constraint on the fully parameterized model. This choice was well-motivated because we wanted to test a classifier which assumes that spectral characteristics in the brain are category-specific but not IC-specific. The use of constraints with rank > 1, similar as in Dyrholm et al. (2007a), is an important topic for future research.

One limitation in the current study was the relatively short duration of the 12-min sessions, which provided only a limited number of (independent) epochs per category for training the classifier. In the future, it would be important to perform similar experiments with longer recordings to make the training of the classifier more accurate. It would also be useful to analyze recordings containing a more diverse set of different stimulus blocks compared with the current experiments. In this case, one could try to decode even a higher number of categories. In our current data, the categories “visual” and “auditory” consisted of different subcategories. However, we did not intend to decode these categories because the limited amount of the available data.

We would like to emphasize that the proposed method is not limited to the type of experiments presented here, but can be applied to any MEG or EEG decoding problem. High classification performance and meaningful spectrospatial patterns provided by the classifier indicate that the analysis of MEG signals using Spectral LDA can be an interesting alternative to fMRI-based decoding studies, in which detailed spectral information cannot be utilized due to the poor temporal resolution of the fMRI. We provide all the implemented methods as a free Matlab toolbox[6] and hope them to be useful in advancing neuroscience.

Acknowledgments

This work was supported financially by the Academy of Finland (National Centers of Excellence Programme, Computational Sciences Program, research grant), European Research Council (Advanced Grant #232946), and Aalto University's aivoAALTO project.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication.

Appendix A. Multicategory bilinear logistic regression

Here, we present the objective function and the gradient for the bilinear logistic regression classifier. Gradient for the classifier with a bilinear kernel has been presented earlier (Dyrholm et al., 2007a), but the analysis was restricted to two-category classification. Although it is possible to solve multicategory problems using multiple binary classifiers (e.g. by using one-versus-rest discriminations), the multicategory extension is useful because it allows optimal assessment of performance based on a single model rather than on constructs of several two-category classifiers.

We used ℓ_2-norm of the classification coefficients as a penalty term because it is a differentiable function and therefore suitable for gradient-based optimization. The use of this penalty did not force the classification coefficients corresponding to distinct ICs to zero as was the case with our other classifiers which were optimized using the linear kernel function and ℓ_1-norm penalty. However, this was not

[6] http://www.cs.helsinki.fi/group/neuroinf/code/spedebox.

a major concern because we were solely interested in the prediction performance (and not interpretability) of the bilinear classifier in this study.

We found the gradients separately for the log-likelihood and the penalty terms of the penalized objective function $J_\theta = L_\theta - \lambda P_\theta$, and then computed the final gradient as $\nabla J_\theta = \nabla L_\theta - \lambda \nabla P_\theta$. The components of the gradient of the penalty term are:

$$\frac{\partial P_\theta}{\partial c_{ki}} = \frac{\partial\left(\sum_{k=1}^{K}\|\mathbf{c}_k\|_2^2\right)}{\partial c_{ki}} = 2c_{ki}. \tag{A.1}$$

For the bilinear classifier, observations are matrices $\mathbf{Z}(n)$. Under the assumptions that $\mathbf{Z}(n)$ are independent and category labels y_{nk} follow a multinomial distribution with a parameter $p_k(\mathbf{Z}(n))$, the log-likelihood of the symmetric logistic regression model is:

$$\begin{aligned} L_\theta &= \frac{1}{N}\log\prod_{n=1}^{N}\prod_{k=1}^{K} p_k^{y_{nk}}(\mathbf{Z}(n)) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\log\left(\frac{e^{h(\boldsymbol{\theta}_k;\mathbf{Z}(n))+b_k}}{\sum_{j=1}^{K}e^{h(\boldsymbol{\theta}_j;\mathbf{Z}(n))+b_j}}\right)^{y_{nk}} \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}y_{nk}\left(h(\boldsymbol{\theta}_k;\mathbf{Z}(n))+b_j\right) - \left(\sum_{k=1}^{K}y_{nk}\right)\log\sum_{j=1}^{K}e^{h(\boldsymbol{\theta}_j;\mathbf{Z}(n))+b_j}\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}y_{nk}\left(h(\boldsymbol{\theta}_k;\mathbf{Z}(n))+b_j\right) - \log\sum_{j=1}^{K}e^{h(\boldsymbol{\theta}_j;\mathbf{Z}(n))+b_j}\right), \end{aligned} \tag{A.2}$$

where we have used knowledge that $\sum_{k=1}^{K} y_{nk} = 1$. With bilinear kernel function, Eq. (A.2) becomes:

$$L_\theta = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}y_{nk}\left(\mathbf{c}_k^T\mathbf{Z}(n)\mathbf{f}_k + b_k\right) - \log\sum_{j=1}^{k}e^{\mathbf{c}_j^T\mathbf{Z}(n)\mathbf{f}_j+b_j}\right), \tag{A.3}$$

and the elements of the gradient are partial derivatives of this function with respect to parameters $\mathbf{c}_k$, $\mathbf{f}_k$, and b_k. These are given by:

$$\begin{aligned} \frac{\partial L_\theta}{\partial c_{ki}} &= \frac{1}{N}\sum_{n=1}^{N}\left(y_{nk}\sum_{j=1}^{F}f_{kj}z_{ij}(n) - \frac{e^{\mathbf{c}_k^T\mathbf{Z}(n)\mathbf{f}_k+b_k}}{\sum_{j=1}^{K}e^{\mathbf{c}_j^T\mathbf{Z}(n)\mathbf{f}_j+b_j}}\sum_{j=1}^{F}f_{kj}z_{ij}(n)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{F}f_{kj}z_{ij}(n)(y_{nk} - p_k(\mathbf{Z}(n))) \end{aligned} \tag{A.4}$$

$$\begin{aligned} \frac{\partial L_\theta}{\partial f_{kj}} &= \frac{1}{N}\sum_{n=1}^{N}\left(y_{nk}\sum_{i=1}^{C}c_{ki}z_{ij}(n) - \frac{e^{\mathbf{c}_k^T\mathbf{Z}(n)\mathbf{f}_k+b_k}}{\sum_{j=1}^{K}e^{\mathbf{c}_j^T\mathbf{Z}(n)\mathbf{f}_j+b_j}}\sum_{i=1}^{C}c_{ki}z_{ij}(n)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C}c_{ki}z_{ij}(n)(y_{nk} - p_k(\mathbf{Z}(n))) \end{aligned} \tag{A.5}$$

$$\frac{\partial L_\theta}{\partial b_k} = \frac{1}{N}\sum_{n=1}^{N}(y_{nk} - p_k(\mathbf{Z}(n))), \tag{A.6}$$

where $z_{ij}(n)$ are the elements of the matrix $\mathbf{Z}(n)$.

References

Anemüller, J., Sejnowski, T.J., Makeig, S., 2003. Complex independent component analysis of frequency-domain electroencephalographic data. Neural Networks 16, 1311–1323.

Bahramisharif, A., Van Gerven, M., Heskes, T., Jensen, O., 2010. Covert attention allows for continuous control of brain–computer interfaces. Eur. J. Neurosci. 31, 1501–1508.

Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., Garnero, L., 2007. Classification methods for ongoing EEG and MEG signals. Biol. Res. 40, 415–437.

Bingham, E., Hyvärinen, A., 2000. A fast fixed-point algorithm for independent component analysis of complex valued signals. Int. J. Neural Syst. 10, 1–8.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.R., 2008. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process. Mag. 25, 41–56.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.R.R., 2011. Single-trial analysis and classification of ERP components — a tutorial. NeuroImage 56, 814–825.

Buzsáki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304, 1926–1929.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44, 112–122.

Chan, A.M., Halgren, E., Marinkovic, K., Cash, S.S., 2011. Decoding word and category-specific spatiotemporal representations from MEG and EEG. NeuroImage 54, 3028–3039.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

de Brecht, M., Yamagishi, N., 2012. Combining sparseness and smoothness improves classification accuracy and interpretability. NeuroImage 60, 1550–1561.

De Martino, F., Valente, G., Staeren, N.l., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.

Dyrholm, M., Christoforou, C., Parra, L.C., 2007a. Bilinear discriminant component analysis. J. Mach. Learn. Res. 8, 1097–1111.

Dyrholm, M., Makeig, S., Hansen, L.K., 2007b. Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG. Neural Comput. 19, 934–955.

Friedman, J.H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Hari, R., Kujala, M.V., 2009. Brain basis of human social interaction: from concepts to brain imaging. Physiol. Rev. 89, 453–479.

Hari, R., Salmelin, R., 1997. Human cortical oscillations: a neuromagnetic view through the skull. Trends Neurosci. 20, 44–49.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640.

Hasson, U., Furman, O., Clark, D., Dudai, Y., Davachi, L., 2008. Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. Neuron 57, 452–462.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Springer, New York.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nature Rev. Neurosci. 7, 523–534.

Hejnar, M.P., Kiehl, K.A., Calhoun, V.D., 2007. Interparticipant correlations: a model free fMRI analysis technique. Hum. Brain Mapp. 28, 860–867.

Huttunen, H., Manninen, T., Kauppi, J.P., Tohka, J., 2012. Mind reading with regularized multinomial logistic regression. Mach. Vision Appl. 2012, 1–15.

Hyvärinen, A., Ramkumar, P., Parkkonen, L., Hari, R., 2010. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. NeuroImage 49, 257–271.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature Neurosci. 8, 679–685.

Kauppi, J.P., Jääskeläinen, I.P., Sams, M., Tohka, J., 2010. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front. Neuroinform. 4, 5 (10 pp.).

Kauppi, J.P., Huttunen, H., Korkala, H., Jääskeläinen, I.P., Sams, M., Tohka, J., 2011. Face prediction from fMRI data during movie stimulus: strategies for feature selection. Artif Neural Networks and Mach Learn — ICANN. Springer, pp. 189–196.

Klami, A., Ramkumar, P., Virtanen, S., Parkkonen, L., Hari, R., Kaski, S., 2011. ICANN/PASCAL2 challenge: MEG mind reading — overview and results. Proceedings of ICANN/PASCAL2 Challenge: MEG Mind Reading. Aalto University Publication series SCIENCE + TECHNOLOGY 29/2011 Espoo, Finland, pp. 3–19.

Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. NeuroImage 56, 411–421.

Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012. Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front. Hum. Neurosci. 6, 233 (14 pp.).

Lehtelä, L., Salmelin, R., Hari, R., 1997. Evidence for reactive magnetic 10-Hz rhythm in the human auditory cortex. Neurosci. Lett. 222, 111–114.

Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56, 387–399.

Liu, G., Huang, G., Meng, J., Zhu, X., 2010. A frequency-weighted method combined with common spatial patterns for electroencephalogram classification in brain–computer interface. Biomed. Signal Process. 5, 174–180.

Malinen, S., Hlushchuk, Y., Hari, R., 2007. Towards natural stimulation in fMRI — issues of data analysis. NeuroImage 35, 131–139.

Mellinger, J., Schalk, G., Braun, C., Preissl, H., Rosenstiel, W., Birbaumer, N., Kübler, A., 2007. An MEG-based brain–computer interface (BCI). NeuroImage 36, 581–593.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M.A., Newman, S.D., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57, 145–175.

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., Lakany, H., 2011. EEG decoding of semantic category reveals distributed representations for single concepts. Brain Lang. 117, 12–22.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., et al., 1999. Discrete-Time Signal Processing, vol. 5. Prentice Hall, Upper Saddle River.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209.

Pfurtscheller, G., Neuper, C., 2001. Motor imagery and direct brain–computer communication. IEEE Proc. 89, 1123–1134.

Ramkumar, P., Parkkonen, L., Hari, R., Hyvärinen, A., 2012. Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. Hum. Brain Mapp. 33, 1648–1662.

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., Parkkonen, L., 2013. Feature-specific information processing precedes concerted activation in human visual cortex. J. Neurosci. 33, 7691–7699.
Rasmussen, C.E., Nickisch, H., 2010. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. 11, 3011–3015.
Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C., 2012. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recogn. 45, 2085–2100.
Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D., 2011. Decoding brain states from fMRI connectivity graphs. NeuroImage 56, 616–626.
Rieger, J.W., Reichert, C., Gegenfurtner, K.R., Noesselt, T., Braun, C., Heinze, H.J.J., Kruse, R., Hinrichs, H., 2008. Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. NeuroImage 42, 1056–1068.
Ryali, S., Supekar, K., Abrams, D.A., Menon, V., 2010. Sparse logistic regression for whole brain classification of fMRI data. NeuroImage 51, 752–764.
Ryali, S., Chen, T., Supekar, K., Menon, V., 2012. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. NeuroImage 59, 3852–3861.
Santana, R., Bielza, C., Larrañaga, P., 2012. Regularized logistic regression and multiobjective variable selection for classifying MEG data. Biol. Cybern. 106, 389–405.
Simanova, I., van Gerven, M., Oostenveld, R., Hagoort, P., 2010. Identifying object categories from event-related EEG: toward decoding of conceptual representations. PloS One 5, e14465 (12 pp.).
Singer, W., 1993. Synchronization of cortical activity and its putative role in information processing and learning. Annu. Rev. Physiol. 55, 349–374.
Spiers, H., Maguire, E., 2007. Decoding human brain activity during real-world experiences. Trends Cogn. Sci. 11, 356–365.
Suk, H.I., Lee, S.W., 2013. A novel Bayesian framework for discriminative feature extraction in brain–computer interfaces. IEEE Trans. Pattern Anal. Mach. Intell. 35, 286–299.
Taulu, S., Kajola, M., 2005. Presentation of electromagnetic multichannel data: The signal space separation method. J. Appl. Phys. 97, 124905 (10 pp.).
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Series B (Stat. Meth.) 58, 267–288.
Tomioka, R., Müller, K.R., 2010. A regularized discriminative framework for EEG analysis with application to brain–computer interface. NeuroImage 49, 415–432.
Tong, F., Pratte, M.S., 2012. Decoding patterns of human brain activity. Annu. Rev. Psychol. 63, 483–509.
van Gerven, M., Jensen, O., 2009. Attention modulations of posterior alpha as a control signal for two-dimensional brain–computer interfaces. J. Neurosci. Methods 179, 78–84.
van Gerven, M., Hesse, C., Jensen, O., Heskes, T., 2009. Interpreting single trial data using groupwise regularisation. NeuroImage 46, 665–676.
van Gerven, M.A., Cseke, B., De Lange, F.P., Heskes, T., 2010. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. NeuroImage 50, 150–161.
Wolf, I., Dziobek, I., Heekeren, H.R., 2010. Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. NeuroImage 49, 894–904.
Yamashita, O., Sato, M.A., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42, 1414–1429.
Zhdanov, A., Hendler, T., Ungerleider, L., Intrator, N., 2007. Inferring functional brain states using temporal evolution of regularized classifiers. Comput. Intell. Neurosci. 52609 (8 pp.).
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B (Stat. Meth.) 67, 301–320.