ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 28 (2005) 326 – 341

# Recipes for the linear analysis of EEG

Lucas C. Parra,[a,*] Clay D. Spence,[b] Adam D. Gerson,[c] and Paul Sajda[c]

[a]*Department of Biomedical Engineering, City College of New York, New York, NY 10031, USA*
[b]*Sarnoff Corporation, Princeton, NJ 08540, USA*
[c]*Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA*

Received 22 September 2004; revised 2 March 2005; accepted 5 May 2005
Available online 3 August 2005

**In this paper, we describe a simple set of "recipes" for the analysis of high spatial density EEG. We focus on a linear integration of multiple channels for extracting individual components without making any spatial or anatomical modeling assumptions, instead requiring particular statistical properties such as maximum difference, maximum power, or statistical independence. We demonstrate how corresponding algorithms, for example, linear discriminant analysis, principal component analysis and independent component analysis, can be used to remove eye-motion artifacts, extract strong evoked responses, and decompose temporally overlapping components. The general approach is shown to be consistent with the underlying physics of EEG, which specifies a linear mixing model of the underlying neural and non-neural current sources.**

*Keywords:* Electroencephalography (EEG); Linear integration; Source estimation; Generalized eigenvalue decomposition; Brain computer interfaces

## Introduction

Modern high spatial density electroencephalography (EEG) can simultaneously record surface potentials in up to 256 electrodes, promising improved spatial resolution and an increased signal-to-noise ratio (SNR). Traditional EEG analysis methods such as trial-averaging typically only considers the time course of individual channels. Thus, the increased number of sensors has created a need for tools that can analyze the time series of multiple electrodes simultaneously. The activity in multiple channels is often visualized as a topographic map across the scalp, with various methods commonly used to integrate this activity for localizing neuronal sources within the brain (Niedermeyer, 1996; Mosher et al., 1999; Michel et al., 2004; Darvas et al., 2004). Source localization represents a difficult inversion problem, as many possible current distributions can lead to the same observed EEG activity. To resolve this ambiguity, localization methods try to explain the spatio-temporal statistics of the observed data by constraining the possible source distributions in space and assuming a model of the anatomy of the head/brain. The methods presented in this paper avoid making any spatial modeling assumptions with regard to the sources or anatomy. They instead rely entirely on the statistics of the observed data and its covariation with observable stimuli and behavioral responses.

Recently, various multivariate signal processing algorithms have been proposed for EEG that linearly combine channels to generate an aggregate representation of the data that is easier to analyze (Chapman and McCrary, 1995; Koles et al., 1995; Makeig et al., 1996; Tang et al., 2002; Ramoser et al., 2000; Parra et al., 2002; Parra and Sajda, 2003; Delorme and Makeig, 2004). Specifically, denote with $\mathbf{x}(t)$ the vector of multidimensional EEG data at time $t$. A weighting vector $\mathbf{w}$ is selected that generates a one-dimensional projection $y(t)$ with,[1]

$$y(t) = \mathbf{w}^T\mathbf{x}(t) = \sum_{i=1}^{D} w_i x_i(t). \tag{1}$$

This linear projection combines the information from the multiple sensors into a single channel whose time course can be analyzed with conventional methods, i.e., temporal filtering, trial averaging, frequency power analysis, etc. The vector $\mathbf{w}$ is selected based on constraints or desired attributes of the time series $y(t)$. When compared to an individual sensor the projection $y(t)$–sometimes also called a component–can be a better estimate of neurophysiological activity, as it may have a higher signal-to-noise

* Corresponding author.
*E-mail addresses:* parra@ccny.cuny.edu (L.C. Parra), cspence@sarnoff.com (C.D. Spence), adg71@columbia.edu (A.D. Gerson), ps629@columbia.edu (P. Sajda).

**Available online on ScienceDirect (www.sciencedirect.com).**

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.05.032

[1] Throughout this paper, lower case italic characters refer to scalars, such as $y(t)$; lower case bold characters, such as $\mathbf{w}$, denote vectors, while bold upper case characters refer to matrices, as in $\mathbf{A}$. Finally, $\mathbf{w}^T$ indicates the transpose of $\mathbf{w}$.

ratio (SNR) and reduced interference from other sources. To see this, consider the following two scenarios.

### *Signal summation*

Assume that two neighboring electrodes have activities $x_1(t)$ and $x_2(t)$ which measure the surface potential associated with neural activity, $s(t)$. Each electrode also measures noise $n_1(t)$ and $n_2(t)$, respectively, originating from independent noise sources such as skin potentials, myographic activity, or inductive line noise. Since electrical potentials add linearly, we can summarize the electrodes' activities as $x_1(t) = s(t) + n_1(t)$ and $x_2(t) = s(t) + n_2(t)$. Choosing equal weights, $\mathbf{w}^T = [1,1]$, for the two neighboring electrodes in Eq. (1) results in a component, $y(t) = 2s(t) + n_1(t) + n_2(t)$. This component captures the neuronal activity, $s(t)$, with an increased signal-to-noise ratio relative to the individual electrodes (3 dB improvement in the case of independent Gaussian noise).

### *Signal subtraction*

Assume a different scenario in which an electrode measures the surface potentials $x_1(t)$ associated with the activities of two separate neuronal processes, $s_1(t)$ and $s_2(t)$, so that $x_1(t) = s_1(t) + s_2(t)$. Assume that $s_2(t)$ can be observed in isolation by an additional electrode $x_2 = s_2(t)$, whereas $s_1(t)$ cannot—i.e., $x_1(t)$ is a mixture of both sources. Choosing the weights, $\mathbf{w}^T = [1, -1]$ results in a component, $y(t) = x_1(t) - x_2(t) = s_1(t)$. This component recovers the activity of the individual source, $s_1(t)$, by subtracting the known "interfering" source, $s_2(t)$, from the measured mixture of sources.

As we will see later, one can combine these two operations to estimate sources and subtract interfering sources. The motivation for using a linear integration is that the observed surface potentials are linearly related to the current sources within the brain (see "Linear model for EEG"). The resulting component can therefore in principle be localized using any of the established source localization algorithms (for a review see (Darvas et al., 2004)). Evident is that the benefits of linear integration will depend on the proper choice of the weighting vector **w**. In fact there may be more than one useful choice depending on the desired properties of the component $y(t)$.

This paper shows how three basic criteria, namely maximum difference, maximum power, and statistical independence, can be used to select useful linear integration vectors **w**. The methods will be introduced on practical examples such as the estimation of eye-motion artifacts, the extraction of evoked responses and the decomposition of temporally overlapping components. Some of the methods and algorithms we will discuss have been previously presented (Chapman and McCrary, 1995; Makeig et al., 1996; Tang et al., 2002; Ramoser et al., 2000; Parra et al., 2002; Parra and Sajda, 2003; Delorme and Makeig, 2004). However, it is the objective of this paper to put these various algorithms into the a common framework and show that they are in fact intimately linked in terms of their relationship to the three criteria. The mathematical details, some of which are well-established and others less well-known, are recast into the present context in a series of Appendices. For instance, we show that common spatial patterns are essentially the same as independent components, and that, under some circumstances, penalized logistic regression is equivalent to a support vector machine (SVM). Computer code for the corresponding algorithms has been made available on-line.[2]

## Estimation through projections and subtraction

Before we discuss the criteria for estimating **w,** we begin by introducing a basic model of EEG generation and generalize the preceding discussion on summation and subtraction to higher dimensions. In higher dimensions, the summation and subtraction operations become a subspace projection and a subspace subtraction, respectively. In the first case, source activity is estimated by projecting the observed data onto the subspace defined by that source, while in the second case, the subspace of the source is subtracted to remove the corresponding activity. In the following sections, we formalize this process to the more general setting of multiple sources and sensors.

### *Linear model for EEG*

At the frequencies of interest (<100 Hz), tissue is often assumed to be primarily a resistive medium governed by Ohm's law and capacitive effects can be neglected.[3] Hence, the contributions of a single current source $s(t)$ to the surface potentials $\mathbf{x}(t)$ is linear in $s(t)$,

$$\mathbf{x}(t) = \mathbf{a}s(t). \tag{2}$$

The proportionality factors in vector **a** are, in general, unknown as they depend on the spatial distribution and orientation of the current source as well as the conductivity distribution of the interfacing anatomy (including brain tissue, CSF, skull, skin, electrode contact and location, etc.). For multiple sources, there is one such factor for every electrode–source pair, which we summarize in matrix **A**. Since potentials are additive, the combined effect of multiple sources summarized in vectors $\mathbf{s}(t)$ is simply,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{3}$$

The matrix **A** is often called a *forward model* relating the source activities with the observed sensor activities. Any contribution that is not described by the matrix **A** can be summarized in an additional term that we will consider as noise, $\mathbf{n}(t)$. The distinction between noise and source contributions is somewhat arbitrary, as obviously noise also has a source. In particular, we may consider genuine neurological activity as 'noise' if it forms part of the background activity that is not the subject of a particular study. Summarizing, we find the following simple linear model which is an accurate representation of the electro-physics of EEG,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \tag{4}$$

[2] MATLAB Plug-ins for EEGLAB can be downloaded from http://liinc.bme.columbia.edu/downloads.

[3] The ratio between capacitive vs. resistive conductance is independent of source location for homogeneous tissue and is estimated to be <10% at 100 Hz (Reilly, 1992). More importantly, regardless of homogeneity, this ratio is approximately constant within the frequency range of interest (Reilly, 1992) leading to a linear phase and hence a constant delay (<0.2 ms). We assume that capacitive delays introduced by the skull, electrode contacts, or skin potentials are approximately the same for all electrodes and therefore negligible.

### *Source estimation by linear projection*

Assume we have an estimate of the forward model **A**. This estimate could originate from a dipole fit of the data, or could have been generated from one of the algorithms we will describe in the following sections. We compute a linear estimate of the source activity, which we denote as

$$\hat{\mathbf{s}}(t) = \mathbf{V}^T \mathbf{x}(t). \tag{5}$$

The matrix **V** is often called a *backward model* relating the sensor activity to the originating sources. If **A** captures all sources that contribute significantly to the surface potentials, it is reasonable to choose **V** such that the noise is as small as possible. Minimizing the total noise power gives the pseudo-inverse of $\mathbf{A}^{\#}$ as the classic least mean squares estimator

$$\hat{\mathbf{V}}^T = \mathbf{A}^{\#} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T. \tag{6}$$

If one is willing to assume Gaussian noise with known spatial correlations one obtains an improved maximum likelihood estimator, which can be used to factor out the effect of EEG background activity in the source estimate (see Appendix A.1). When combining Eqs. (4)–(6) it becomes clear that $\hat{\mathbf{s}}(t)$ is a reasonable estimate for $s(t)$

$$\hat{\mathbf{s}}(t) = \mathbf{s}(t) + \mathbf{V}^T\mathbf{n}(t). \tag{7}$$

This backward model recovers the sources but also captures the portion of the noise that is collinear with the source estimates. Without additional (temporal) information, that portion of the noise is indistinguishable from the sources. When the forward model of different sources (different columns of **A**) are similar, the pseudo-inverse becomes sensitive to errors in the estimation of **A** and regularization methods may be necessary (Uusital and Ilmoniemi, 1997).

### *Interference subtraction*

In some instances, it may be desirable to minimize interference from other sources. For instance, one may wish to remove the activity associated with eye motion. Let **A** now represent the coupling of the sources $\hat{\mathbf{s}}(t)$ which are to be removed. The sources can be estimated from the sensors using (5) and estimator (6) as, $\hat{\mathbf{s}}(t) = \mathbf{A}^{\#}\mathbf{x}(t)$. Following the linear model (3) the sensor activity generated by these sources is $\mathbf{x}_{||}(t) = \mathbf{A}\hat{\mathbf{s}}(t)$. This activity can be subtracted from the sensor activity.

$$\mathbf{x}_{\perp}(t) = \mathbf{x}(t) - \mathbf{x}_{||}(t) = \left(\mathbf{I} - \mathbf{A}\mathbf{A}^{\#}\right)\mathbf{x}(t), \tag{8}$$

where **I** denotes the identity matrix. The activity $\mathbf{x}_{\perp}(t)$ is the original sensor data without the interference from sources $\hat{\mathbf{s}}(t)$. To be precise, $\mathbf{x}_{\perp}(t)$ does not contain any activity that correlates with $\hat{\mathbf{s}}(t)$. While $\mathbf{x}_{\perp}(t)$ will have an increased signal-to-noise ratio, its dimensionality is effectively reduced. Subsequent algorithms have to account for this reduced rank in the data as described in Appendix A.4. We will demonstrate this interference subtraction method for the removal of eye-motion activity in “Example: eye-motion activity”.

### *Forward model estimate*

The previous discussion assumed that a forward model **A** of some current sources is available. The emphasis of this paper lies on generating a component, $y(t) = \mathbf{w}^T\mathbf{x}(t)$, where $y(t)$ has some desired temporal properties. For instance, we will ask for a projection of the data which is maximally different between two experimental conditions. This section determines how to relate that component activity $y(t)$ to a current source $\hat{s}_y(t)$ and the corresponding estimated forward model $\hat{\mathbf{a}}_y$—one column of the matrix **A**.

We define $\hat{s}_y(t)$ as the source of all the activity in $\mathbf{x}(t)$ which correlates with $y(t)$ during a set of times $t = \{t_1,\ldots,t_N\}$. The corresponding forward model $\hat{\mathbf{a}}_y$ can be found by linearly predicting $\mathbf{x}(t)$ from $y(t)$, as this will capture the linear correlation between sources and observations. The answer is given by the optimal linear predictor as shown in Appendix A.2. Arranging the times of interest along a row, for $\mathbf{y} = [y(t_1),\ldots,y(t_N)]$, and $\mathbf{X} = [\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)]$, the result derived in the Appendix can be written for zero-mean $\mathbf{x}(t)$ and $y(t)$ as

$$\hat{\mathbf{a}}_y = \mathbf{X}\mathbf{y}^T\left(\mathbf{y}\mathbf{y}^T\right)^{-1}. \tag{9}$$

Note that this estimate is a normalized correlation between the component $y(t)$ and the data $\mathbf{x}(t)$. In Appendix A.2, it is shown that the corresponding source $\hat{s}_y(t)$ captures all the activity in the data that correlates with the component $y(t)$. This forward model $\hat{\mathbf{a}}_y$ provides a spatial distribution of correlated activity which we term a *scalp projection*.

## Maximum difference

### *Evoked response difference*

A common question in EEG analysis is how the observed EEG activity differs under two different experimental conditions, such as differing stimuli or behavioral responses of the subject. The conventional approach is to average the activity over repeated presentations (or trials) and compute the difference of these trial averages for the two conditions. Typically, the onset of a stimulus or the time of a behavioral response are marked in the data by corresponding event triggers, i.e., the time instances at which a specific stimulus or response occurred. Denoting $t_1$ and $t_2$ as the times of event triggers associated with conditions 1 and 2, respectively, the traditional evoked response difference is given by

$$\Delta\mathbf{x}(\tau) = \frac{1}{N_1}\sum_{t_1}\mathbf{x}(t_1+\tau) - \frac{1}{N_2}\sum_{t_2}\mathbf{x}(t_2+\tau). \tag{10}$$

At any given point in time $\tau$, relative to the event trigger, $\Delta\mathbf{x}(\tau)$ specifies the difference activity in each electrode. This activity is referred to as the trial-averaged evoked response. We note that the evoked response integrated over a window of interest[4],

$$\overline{\Delta\mathbf{x}} = \sum_{\tau}\Delta\mathbf{x}(\tau), \tag{11}$$

can itself be seen as specific forward model estimate $\hat{\mathbf{a}}_y$. Assume that there is a component $y(t)$ that takes on different constant values $y(t_1+\tau) = y_1$ and $y(t_2+\tau) = y_2$ during conditions 1 and 2,

[4] With the $\sum_{t_1}$ or $\sum_{\tau}$ we denote, in this paper, the sum over all the possible values of variables $t_1$ or $\tau$, respectively. In general, the sum over $t_1$ or $t_2$ will therefore refer to a trial-average while the sum over $\tau$ will refer to an average over a window of time relative to an event.

respectively. We find that the corresponding forward model $\hat{\mathbf{a}}_y$ after mean subtraction is given according to Eq. (9) by

$$\hat{\mathbf{a}}_y = \frac{\overline{\Delta \mathbf{x}}}{(N_1 + N_2)(y_1 - y_2)}. \tag{12}$$

The classic evoked response difference therefore can be interpreted itself as a forward model (up to a constant factor). Specifically, it is the forward model of the current sources that correlate with a component which takes on different constant values during the two different conditions. Any localization method that uses the average evoked response difference rather than the original responses can be thought of as localizing such a current source.

*Example: eye-motion activity*

As an example of this difference method, consider the estimation of activity associated with eye motion. Vertical and horizontal eye motion induces in the electro-oculogram (EOG) and frontal EEG electrodes a significant deviation from baseline. Interestingly, the direction and magnitude of the deviation are indicative of the orientation and degree of eye motion. The time course and a corresponding scatter plot for different motions are shown in Figs. 1 and 2. This data was collected during a simple eye-tracking task in which the subject follows a cross on the screen. The cross jumps repeatedly from left to right or from top to bottom of the screen. The corresponding cross positions are marked in the data and can be used as truth data to indicate eye position. In Eq. (10), times $t_1$ and $t_2$ indicate the position of the cross on the screen, left/right (or up/down). Times $\tau$ extend over the duration in which the cross is holding a position. Horizontal and vertical eye-motion components can be estimated individually using (12), i.e., the estimate $\hat{\mathbf{a}}_\mathrm{h}$ is given by the difference between the mean activity on left and right position and $\hat{\mathbf{a}}_\mathrm{v}$ by the difference activity between the upper and lower positions. The scatter plots in Fig. 1, center and right, show that this subject did not follow the cross perfectly, yet the difference estimate, shown as solid line, captures the main orientation of the motion. In the "Maximum power" Section, we will also present a method for estimating the activity associated with eye blinks. The corresponding scalp projection – computed with Eq. (16) – is shown on the left of Fig. 1 and is denoted here as $\hat{\mathbf{a}}_\mathrm{b}$.

To demonstrate the estimation and subtraction of interference activity, we consider all three types of eye-motion artifacts: horizontal and vertical motion as well as eye blinks. We combine the three sources into a vector $\hat{\mathbf{s}}(t) = [\hat{s}_\mathrm{b}(t), \hat{s}_\mathrm{h}(t), \hat{s}_\mathrm{v}(t)]^T$, and the corresponding scalp projections into a matrix $\hat{\mathbf{A}}_\mathrm{eye} = [\hat{\mathbf{a}}_\mathrm{b}, \hat{\mathbf{a}}_\mathrm{h}, \hat{\mathbf{a}}_\mathrm{v}]$.

According to Eqs. (5) and (6), the estimate for the eye activity is $\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{\#}_\mathrm{eye}\mathbf{x}(t)$, shown in Fig. 2, left. Note that there is no cross-talk between the estimated sources despite the fact that all the sources of eye motion contribute to the interference in any one electrode, e.g., the VEOG-electrode shown on Fig. 2, bottom left, responds to both eye blinks and vertical eye motion. Referring back to Fig. 1, we also see this in the scalp plots of $\hat{\mathbf{a}}_\mathrm{b}$ and $\hat{\mathbf{a}}_\mathrm{v}$, with the activity of these two components being quite similar across the electrodes. Increasing the number of EOG electrodes results in a better estimate of the eye activity. At a minimum, one should include one electrode below the eye such as VEOG− to differentiate between eye blinks and vertical eye motion which differ in their polarity for this electrode.

Since eye-motion activity is often considered an artifact, we remove it by computing $\mathbf{x}_\perp(t)$ using Eq. (8). The activity before

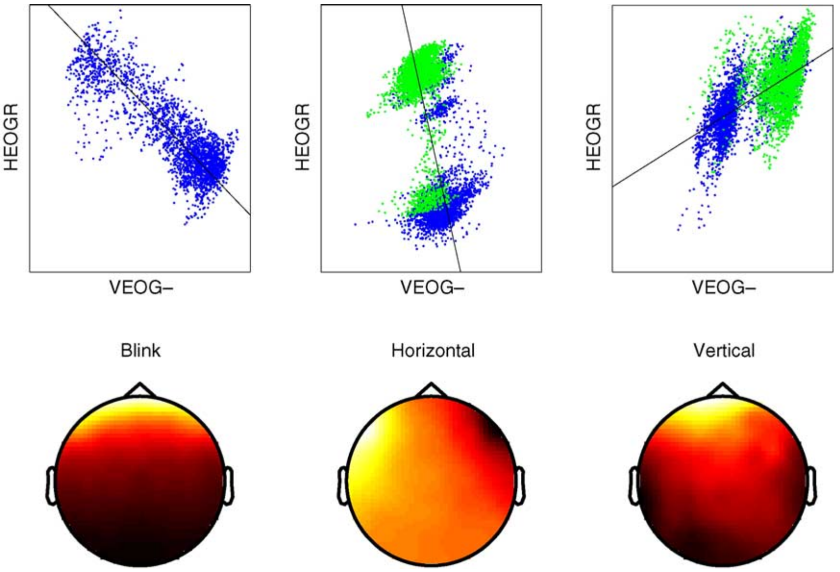

Fig. 1. Removing eye-motion artifacts using maximum power (left) and maximum difference (center, right) criteria. Top: scatter plot of two EOG electrodes during eye blinks (left), horizontal (center), and vertical (right) eye motion. Each point in the scatter plot represents one sample in time. Dark points (blue) represent samples where the fixation cross is in the left or lower screen position, and light points (green) represent samples where the fixation cross is in the right or upper screen position. The black lines indicate the scalp projection orientation $\hat{\mathbf{a}}_y$ in these two EOG coordinates estimated with principal components (left) and mean difference (center, right). The HEOGR electrode is placed next to the right eye and VEOG− below one of the two eyes. Bottom: dorsal view of the estimated scalp projection $\hat{\mathbf{a}}_y$ of sources for the three types of eye motion as shown in the top part of this figure. A "hot" colormap is used in all cases (also all figures in this paper). White (brighter) represents strong positive correlation of the source with the corresponding sensor while dark red (darker) indicates strong negative correlation.

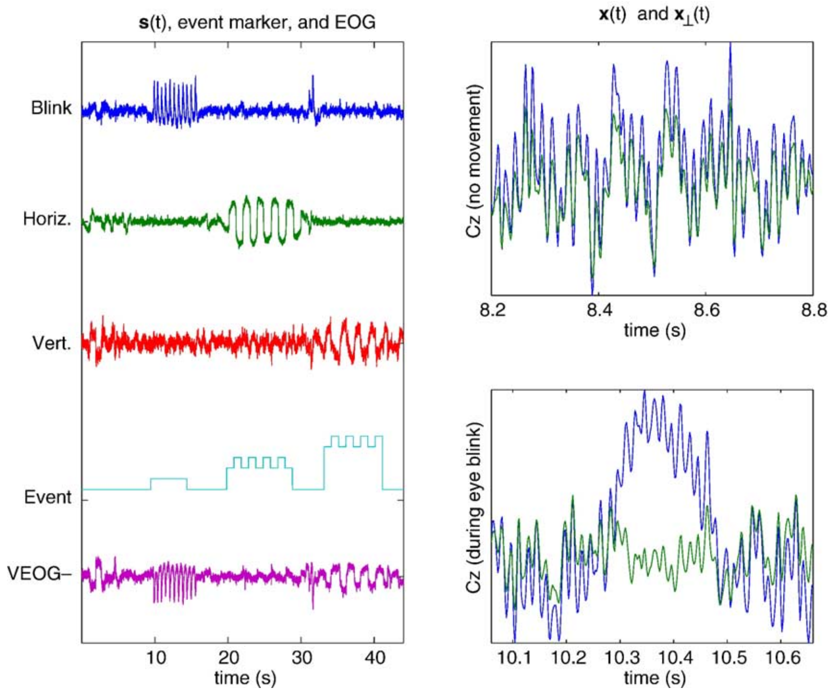

Fig. 2. Source estimates and subtraction demonstrated for eye-motion activity. Left: first three traces show the time courses of the estimated sources $\hat{\mathbf{s}}_y$. The fourth trace indicates the position of the fixation cross during this 45 s recording. The subject is instructed to blink repeatedly at the first position (9–15 s) and to follow the cross on the left/right positions (20–28 s) and up/down positions (34–42 s). Raw VEOG data is shown at the bottom as an example of cross-talk prior to estimation. Right: 600 ms section for Cz electrode (mid-line center) is shown during a period of immobility (top) and during an eye blink (bottom). The two traces in each panel show the activity before, $x(t)$, and after subtraction $x_{\perp}(t)$. Notice that $x_{\perp}(t)$ follows $x(t)$ during immobility but deviates during the blink indicating that the eye-blink activity has been removed without affecting the background EEG.

and after subtraction is shown for one electrode (Cz) in Fig. 2, right. Note that this data was sampled at 500 Hz and the 2 s mean around each sample was subtracted (zero-phase moving-average high-pass filter with 0.5 Hz cutoff). This mean subtraction or baseline subtraction is customary in EEG to remove DC drifts which are of little neurophysiological significance. In addition, most estimation methods are sensitive to such baseline drifts.

The subtraction method proposed here is similar to the standard regression approach where a single (or a few) EOG electrodes are used as a reference signal that is subtracted in proportion to their contribution to each EEG channel (Croft and Barry, 2000). The difference lies in that we use all EEG and EOG channels to construct an optimal estimate $y(t)$ of the corresponding activity as reference for subtraction. Many researchers have proposed to use independent component analysis (ICA) to estimate eye-motion activity. ICA also combines all electrodes during the estimation, however, we found the proposed method to be more reliable in practice requiring significantly less data. This is not surprising given that the present method exploits timing information on eye motion, whereas ICA aims to compute the activity "blindly".

### *Maximum magnitude difference*

The difference activity discussed above can itself be thought of as a forward model of a component that takes on different constant values for two different conditions. However, the resulting source estimator does not necessarily give a source estimate that differs maximally between the two conditions.

To capture all the changes between the two experimental conditions, we use a projection vector $\mathbf{w}$ with a component $y(t)$ that best separates the two conditions, i.e., $y(t)$ acts as a discriminator. There are a variety of optimal solutions to this problem, each assuming a different distribution of the data in the two conditions (Duda et al., 2001). If we assume that activity during $t_1 + \tau$ and $t_2 + \tau$ can be modeled as two Gaussians having identical covariance matrices $\mathbf{R}$ but differing in their means, the optimal maximum likelihood linear discriminator is given by

$$\mathbf{w}_{\mathrm{ml}} = \mathbf{R}^{-1}\overline{\Delta\mathbf{x}}. \tag{13}$$

This approach gives the same result as the evoked response difference activity but, in addition, accounts for correlated activity in the sensors. We see that if the sensors are strongly correlated, which is in fact often the case for high density EEG, then the evoked response difference may not necessarily represent the component of the EEG that optimally captures changes between the two conditions. The assumption of Gaussian data with equal covariances during both conditions may be too restrictive. Instead, one can use the Fisher linear discriminant (FLD) which is given for two classes by

$$\mathbf{w}_{\mathrm{fld}} = (\mathbf{R}_1 + \mathbf{R}_2)^{-1}\overline{\Delta\mathbf{x}}. \tag{14}$$

$\mathbf{R}_1 + \mathbf{R}_2$ is the within-class covariance and $\mathbf{R}_1$ and $\mathbf{R}_2$ are the covariances for each of the two conditions: $\mathbf{R}_i = \sum_{t_i}\sum_{\tau}(\mathbf{x}(t_i+\tau) - \overline{\mathbf{x}}_i)(\mathbf{x}(t_i+\tau) - \overline{\mathbf{x}}_i)^T$. The FLD maximizes the signal-to-noise ratio (SNR) defined here as difference of the means divided by the within-class standard deviation, also known as the $F$ statistic. Increased SNR of the evoked response is the main motivation for using discrimination methods since this can facilitate single trial analysis that may otherwise not be possible

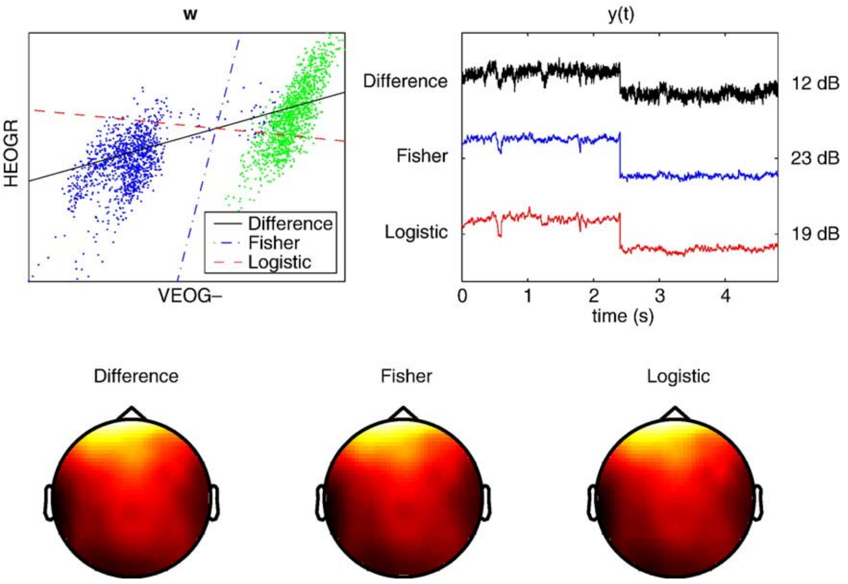

Fig. 3. Comparison of linear discriminant methods. Components were extracted for EEG/EOG electrodes using evoked response difference ($\mathbf{w}_{\text{erd}} \propto \overline{\Delta \mathbf{x}}$), Fisher linear discriminant ($\mathbf{w}_{\text{fld}}$), and penalized logistic regression ($\mathbf{w}_{\text{plr}}$). Top left: values for two out of a total of 64 electrodes are shown. The scatter plot shows samples for EOG electrodes for two different eye positions (fixation to lower or upper position on the screen). The lines indicate the projection vectors $\mathbf{w}$. Top right: comparison of the resulting discriminant components $y(t)$. The first and second halves of the data show the activity for upper and lower positions, respectively. The numbers on the far right indicate the SNR of the recovered components. For comparison, the raw VEOG-electrode has an SNR of 16 dB for this data. Bottom: the three scalp projections $\mathbf{a}_y$ have been estimated with Eq. (9) for each component $y(t)$ separately. Since the main variability in $y(t)$ is similar for the three different results, we obtain similar scalp projections (despite the significant differences in the projection vectors $\mathbf{w}$).

(Parra et al., 2002, 2003; Gerson et al., 2005; Blankertz et al., 2003; Müller et al., 2003).

A problem with this solution is that the inverse covariance of the data is quite sensitive to outliers. In EEG, outliers are inevitable. We therefore use a linear discrimination method that is less sensitive to outliers and instead focuses on the samples at the discrimination boundary. Two candidates for this problem are logistic regression (LR) and linear support vector machines (SVM). Their relationship and relative merits are discussed in Appendix B.2. In general, there are no closed-form solutions for the corresponding optimization criteria of these techniques. In Appendix B.1, we present an algorithm for logistic regression based on iteratively re-weighted least-squares (IRLS), which can also compute an approximate support vector solution when including a regularization term (see Appendix B.2). The logistic regression solution $\mathbf{w}_{\text{lr}}$ equals $\mathbf{w}_{\text{ml}}$ of Eq. (13) if the assumptions of the Gaussian classifier are satisfied. Logistic regression is therefore a generalization of the Gaussian classifier. In practice, we use penalized logistic regression (PLR) as it combines the benefits of support vectors (robust to over-training for well-separated classes) and logistic regression (robust to outliers and good performance for poorly separated distributions).

### *SNR and robustness comparison for linear discriminants*

A comparison of the evoked response difference, the FLD, and the PLR is provided in Fig. 3. For the evoked response difference, we use the source estimator $\mathbf{v}$ as component projector, $\mathbf{w}_{\text{erd}} = \mathbf{v} = \mathbf{a}^{\#T} = \overline{\Delta \mathbf{x}} \,/\, \| \overline{\Delta \mathbf{x}} \|^2$ following Eqs. (6) and (12). The component estimates $y(t)$ for the FLD and the PLR are better separated resulting in a higher SNR. As expected, the SNR is highest for the FLD. In this example, an effort has been made to include activity during fixation only.[5] Nevertheless, a few mislabeled samples during a saccade are included in one of the classes. As a result, the within-class covariance is not well-estimated and the FLD does not capture the difference orientation between the two conditions for the electrodes shown. The evoked response difference maintains its orientation even if a large fraction of the data is mislabeled. It is therefore the most robust, but unfortunately, it has the lowest SNR. Finally, PLR gives high SNR and remains robust to outliers. A quantitative measure of robustness and its statistical significance will be presented in the following section.

### *Example: Identifying neural sources of visual target detection*

We consider the problem of identifying neural sources associated with the detection of infrequent visual targets during a rapid serial visual presentation task (RSVP), a paradigm which has been used to study the speed of visual processing (Thorpe et al., 1996). In this experiment, subjects are shown sequences of images which consist of natural images with people (targets) or natural images without people (distractors). A trial consists of a sequence of 50 images, each image flashed for 50 ms. Subjects are instructed to press a button at the end of the trial if one of the images was a target. Having subjects respond at the end of the trial reduces potential confounds from motor activity. In each trial, there is a

[5] We have chosen, once again, to show data for eye motion, as the difference between the two conditions is strong and can be visualized in only two dimensions. Another, perhaps more interesting, example will be shown in Section on visual target detection.

50% chance that one (and only one) of the images in the sequence is a target image. Thus, target images are rare relative to distractor images.

We aim to find a weight matrix, **w**, which can discriminate, single-trial, between the two conditions of targets vs. distractors. We argue that the activity that discriminates between the two conditions represents the neurological processing associated with the task of visual target detection. $y(t)$ represents the projection of the data that is maximally discriminating for the task. We refer to this activity, therefore, as a task-related component. Studies using trial-averaged responses have shown that such "oddball" experiments having infrequent targets amongst distractors elicit late positive complexes (LPC) with components between 300 and 600 ms post-stimulus (Makeig et al., 1999). In particular, such paradigms produced a strong P300 ERP (Friedman et al., 2001). We therefore consider $y(t)$ for a temporal window in the neighborhood of the reported LPC time course (480–530 ms) and compare this to conventional trial-averaging results.

Fig. 4 shows results, including a projected component of the PLR discriminator. The discriminator is a 64-dimensional hyperplane (64 electrodes) and this single 2D representation shows some of the separation between the two conditions. The discriminator is trained by considering all data points in the temporal window as being independent samples. We can assess the performance of the discriminator using leave-one-out (LOO) receiver operator characteristic (ROC) analysis (Green and Swets, 1966), reporting the area under the ROC curve (Az). Chance performance corresponds to Az = 0.5, and perfect classification to Az = 1.0. The extracted activity $y(t)$ can be regarded as significant only if the Az value is above a significance threshold (a threshold with significance level of $p = 1/N$ can be estimated as the maximum Az value obtained from $N$ repetitions of the training and testing procedure using random class labels (Edington, 1980), a method that is also called 'bootstrapping'). Because of noise, we compute Az by averaging the output of the discriminator within a trial—i.e., we compute the LOO Az over $\bar{y}(t) = \sum_{\tau} y(t + \tau)$. The results are a reasonable estimate of single-trial performance since the averaging of outputs occurs within a trial.

For the discriminator trained during this temporal window, we find Az = 0.87 ($P < 0.001$), indicating a robust discrimination in the full space of the sensors. Also shown in Fig. 4 is the scalp projection $\hat{\mathbf{a}}_y$ of the discriminating task-related component. The topography is very consistent with what has been reported for LPC (Makeig et al., 1999) and trial-averaged P300. This approach can be extended for recovering task-relevant components across different temporal windows to analyze the spatio-temporal changes in neural processing, an example of which will be presented in the methods comparison section.

## Maximum power

Often times, the main source of temporal variability in the EEG data is well known. A straightforward example is the activity of frontal electrodes associated with eye blinks. The activity is considered an artifact and sections of data containing eye blinks–typically identified by visual inspection–are excluded from further analysis. Eye blinks cause a strong deflection of the EEG signal from baseline lasting 50–100 ms. Relative to a common reference the potentials are negative for electrodes placed below the eyes and positive for electrodes above. The relative magnitudes depend on the location and the coupling of the electrode (see Fig. 2). During those 100 ms, the activity associated with the eye blink is typically 20 dB above neuronal background activity. To estimate this activity, it is therefore

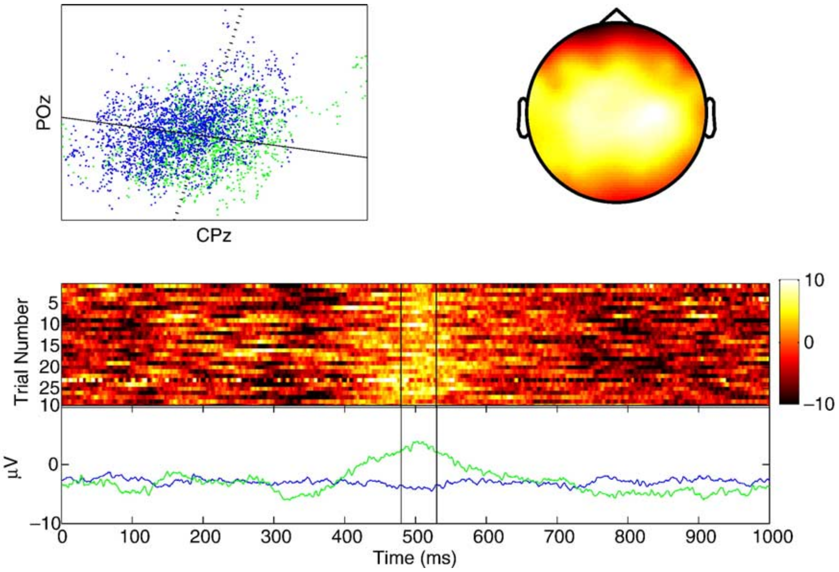

Fig. 4. Identifying task-relevant components in visual target detection using the maximum magnitude difference method. Top left: 2D representation (electrodes CPz and POz) of discrimination vector **w** (black line), determined using penalized logistic regression, illustrating the orientation of maximum separation between target (light/green) and distractor (dark/blue) trials. Top right: scalp projection of corresponding forward model $\hat{\mathbf{a}}_y$. The color code indicates positive correlation (bright) of the sensor readings with the component and negative correlation (dark). Bottom: component activity $y(t)$ across trials locked to the time of stimulus presentation at 0 ms, and trial average of the component activity for target (green) and distractor (blue) trials. Estimation of this component was done using the time window outlined with black lines.

reasonable to find a component that captures the maximum power in the signal during and around the time of eye blinks:

$$\mathbf{w}_{\mathrm{pc}} = \arg\max_{\mathbf{w}, \|\mathbf{w}\|=1} \sum_t y^2(t) = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{R} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}. \tag{15}$$

$\mathbf{R} = \sum_t \mathbf{x}(t)\mathbf{x}^T(t)$ represents the covariance of the $\mathbf{x}(t)$ estimated during the appropriate times. The concept of maximum power or variance dates back to the *varimax* algorithm (Kaiser, 1958). The constraint, $\|\mathbf{w}\| = 1$ is required since, otherwise, the power of the component can be arbitrarily maximized by changing the scale of $\mathbf{w}$. By normalizing with $\|\mathbf{w}\|^2$, the second optimization criterion becomes insensitive to the scale. At the solution of (15), the ratio becomes the maximum eigenvalue of $\mathbf{R}$ and $\mathbf{w}_{\mathrm{pc}}$ the corresponding eigenvector (see Appendix C.1). The result of this maximization is also known as the principal component of the samples $\mathbf{x}(t)$. There are many algorithms which can find the principal component. One of the simplest is the Power method (Reiter, 1990), which consists of iterating the update Equation, $\mathbf{w} \leftarrow \mathbf{R}\mathbf{w}/\|\mathbf{w}\|$, until convergence starting with any non-zero initial guess for $\mathbf{w}$.

Note that at convergence $\mathbf{w}$ satisfies $\mathbf{R}\mathbf{w} = \|\mathbf{w}\|\mathbf{w}$. Using Eq. (29), we find therefore that the corresponding forward model of the principal component is simply,

$$\hat{\mathbf{a}}_{\mathrm{pc}} = \mathbf{R}\mathbf{w}_{\mathrm{pc}}\left(\mathbf{w}_{\mathrm{pc}}^T \mathbf{R} \mathbf{w}_{\mathrm{pc}}\right)^{-1} = \frac{\mathbf{w}_{\mathrm{pc}}}{\|\mathbf{w}_{\mathrm{pc}}\|^2}. \tag{16}$$

Crucial for this operation to give the desired projection vector is that the sample times $t$ selected for this analysis cover a range of data where the activity of interest is the main source of variability. Other sources of variability should be excluded. For instance, a common source of variability in EEG is baseline drift, which should be subtracted to prevent interference in the maximum power estimation of eye blinks. We typically subtract baseline shifts with a high-pass filter having a cutoff frequency of 0.5 Hz prior to this analysis.

### *Example: removing artifacts and noise (continued)*

We have already presented how the maximum power criterion can be used to remove eye-blink activity in the EEG (see Figs. 1 and 2). Note that the same process can also be used to compute projections that estimate other sources of artifacts such as myographic activity, though these are in general more difficult to eliminate than eye motion. For a cooperating subject, a variety of artifacts can be estimated automatically by designing a protocol in which the subject is instructed to generate the corresponding muscle artifacts at predetermined times.

### *Maximum power-ratio*

Before introducing this next criterion, we point out a basic difference between traditional evoked responses and EEG power spectrum analysis. To compute the evoked response, raw EEG signals are averaged over trials. This increases the SNR of low frequency activity that occurs phase-locked to the event timing. High-frequency activity is typically lost due to limited repeatability of precise phase or limited temporal accuracy of event timings. Activity that is not time-locked is lost. On the other hand, power spectral analysis first computes the power spectrum and only then averages over trials, thus capturing all frequency bands in the average. The approach presented in "Maximum magnitude difference", for extracting components that maximally discriminate in *magnitude*, will have the same limitation as trial averaging and will therefore not recover high frequency information. An alternative method is to extract a component that is maximally different in *power* for two different sets of times.

In this method, we find a projection of the data that has minimal power during times $t_1 + \tau$ and maximal power during $t_2 + \tau$, or equivalently a maximum power-ratio

$$\mathbf{w}_{\mathrm{ge}} = \arg\max_{\mathbf{w}, \|\mathbf{w}\|=1} \frac{\sum_{t_2} \sum_{\tau} y^2(t_2 + \tau)}{\sum_{t_1} \sum_{\tau} y^2(t_1 + \tau)} \tag{17}$$

$$= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{R}_2 \mathbf{w}}{\mathbf{w}^T \mathbf{R}_1 \mathbf{w}}. \tag{18}$$

The constraint $\|\mathbf{w}\| = 1$ is again required to normalize for an arbitrary scaling factor. Notice the similarity with condition (15) which results in the first principal component. In fact, while the solution for (15) is given by the eigenvector for $\mathbf{R}$ with maximum eigenvalue, the solution for (18) is given by the generalized eigenvector of $\mathbf{R}_1^{-1}\mathbf{R}_2$ with maximum eigenvalue (see Appendix C.1). Furthermore, notice that the eigenvector with minimum eigenvalue minimizes rather than maximizes the power ratio in (18). An illustrative example for artificial data is shown in Fig. 5.

A limitation of this approach is the sensitivity of the covariance matrix to outliers, given that the covariance to be minimized is in the denominator of (18). For instance, we find that the largest (smallest) component often captures myographic activity if the data is not carefully selected to avoid sections with motion artifacts. Therefore, in practice, it has been suggested (Ramoser et al., 2000) to consider not just the maximum and minimum eigenvectors but several of the eigenvectors corresponding to the largest and lowest eigenvalue ratios. The powers in these projections have been used to discriminate between two distinct task conditions. The recovered components have been termed common spatial patterns (Koles, 1997; Ramoser et al., 2000).

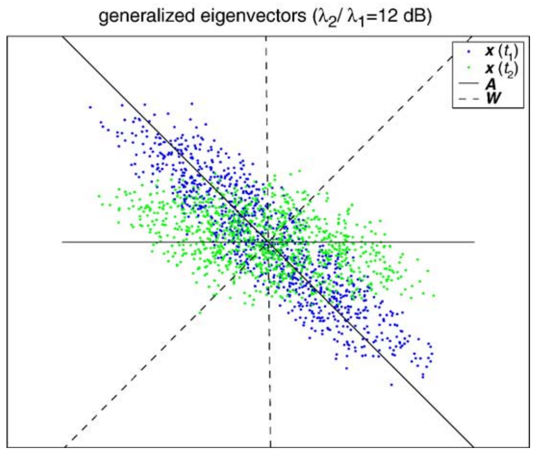

Fig. 5. Generalized eigenvalues and independent components. Dark and light dots indicate samples with covariance matrix $\mathbf{R}_1$ and $\mathbf{R}_2$. Dashed lines indicate the projection vectors $\mathbf{w}_{\mathrm{ge}}$ that generate the maximum and minimum power-ratio for projected component $y(t)$ on all samples. Solid lines indicate the columns of the corresponding $\hat{\mathbf{A}}_y$.

*Example: evoked response synchronization and desynchronization*

It is well established that in preparation for movement, ongoing alpha activity (8–12 Hz oscillation) over the motor cortex is reduced. The activity recovers slowly after movement. This type of increase and decrease of oscillatory activity in response to an event is known as event-related synchronization and desynchronization (ERS/ERD) and has been reported for different frequency bands and different types of events (Pfurtscheller and Lopes da Silva, 1999). Though a general topographic distribution for the activity of interest is known, the spatial distribution and precise frequency bands are quite diverse across subjects. The maximum power-ratio method can be used to adaptively extract ERD and ERS. To this end, the EEG activity is bandpass filtered in the frequency band of interest prior to applying the optimization criteria (18). An example is shown in Fig. 6.

## Statistical independence

The methods presented thus far all use explicit timing information of observable events. This timing information is used to extract the EEG activity that coincides with those events (potentially with some time lag). In some instances, however, one may be interested in EEG activity associated with neuronal activity for which no observable events occur, e.g., varying sleep stages, covert shifts of attention, mental imagery, etc. A promising approach in that case has been to decompose the EEG signal into multiple components $y_i(t)$ with independent time courses. The rationale for this has been that the central nervous system is engaged in a multitude of separate processes, some of which are interdependent while others are independent. One reasonable goal is to combine all co-varying EEG activity into a single component, while activity that takes on independent time courses should be grouped into separate components. With linear methods, at most $D$ independent sources can be recovered, where $D$ is the number of sensors. The goal is to identify $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_D]$ such that the components $\mathbf{y}(t) = [y_1(t),\ldots,y_D(t)]$

$$\mathbf{y}(t) = \mathbf{W}^T\mathbf{x}(t) \tag{19}$$

are statistically independent. There are a number of measures that can be used to assess whether two time sequences are statistically independent, each leading to a different optimality condition for determining $\mathbf{W}$. We avoid going into much detail except to note that a necessary, though not sufficient, condition is that independent sequences have a diagonal covariance matrix. Interestingly, this diagonalization criterion when applied to $\mathbf{R}_1$ and $\mathbf{R}_2$ leads to the same generalized eigenvalue problem as above (Parra and Sajda, 2003). In fact, the solution shown in Fig. 5 are the independent components that span the data.

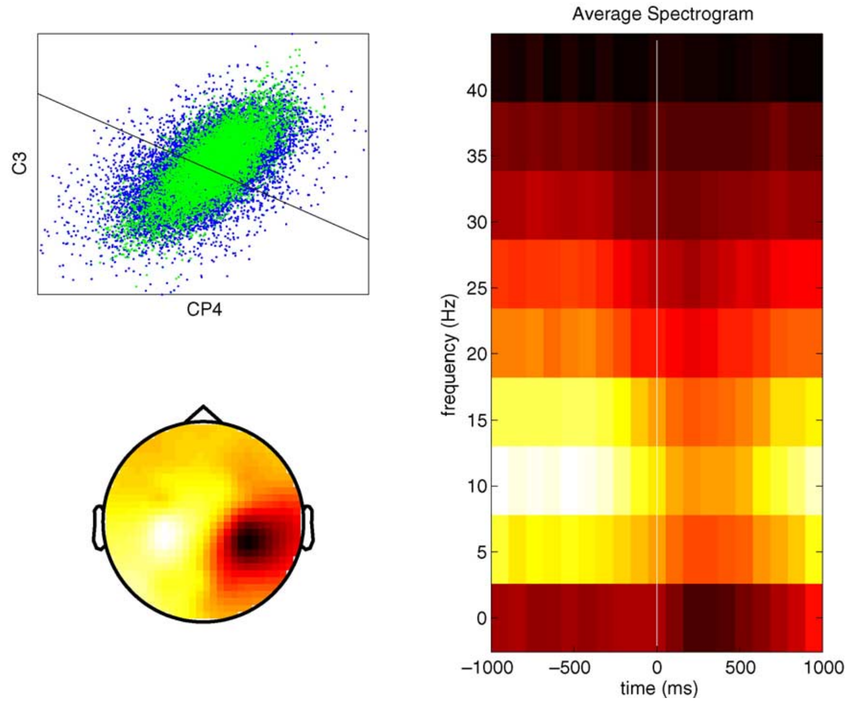

Fig. 6. ERD/ERS with generalized eigenvectors. In this experiment the subject responds to a visual stimulus with a button press. Prior to the maximum power-ratio analysis, all EEG channels are bandpass filtered between 5–40 Hz. The covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ are computed in a window 200 ms before ($\mathbf{R}_1$) and 200 ms after ($\mathbf{R}_2$) the button press. Top left: Scatter plot of the corresponding activity for two of the 64 EEG sensors. The solid line indicates the orientation, $\mathbf{w}_{\text{ge}}$, along with the two distributions having a maximum power (variance) ratio, estimated using generalized eigenvalues. Bottom left: Estimated forward model $\hat{\mathbf{a}}_y$ corresponding to $\mathbf{w}_{\text{ge}}$. Clear is that the source activity originates over motor areas (it is maximal over C3 and CP4) and has opposite sign (180 phase delay) between the hemispheres. Right: Spectrogram computed over $\tau$ for component $y(t + \tau)$ and averaged over $t$, which is approximately 300 button press events for 1 s before and after the button press (indicated with a vertical white line). Alpha band activity (maximal at 12 Hz for this subject) decreases (de-synchronizes) for about 500 ms after the button push.

Finally, we note that for any complete decomposition as in independent component analysis (ICA), where one recovers as many sources as there are sensors, the corresponding forward model as given in (29) simplifies as shown in Appendix A.3 to

$$\hat{\mathbf{A}}_y = \mathbf{W}^{-T}. \tag{20}$$

This is the expression conventionally used for the forward model in ICA. In practice a variety of algorithms have been used to compute independent components in EEG. They often give equivalent results (Parra and Sajda, 2003). Most ICA algorithms are based on simultaneous diagonalization of multiple (more than two) cross-moments such as JADE (Cardoso and Souloumiac, 1993, 1996), SOBI (Belouchrani et al., 1997), or multiple decorrelations (Parra and Spence, 2000; Parra and Sajda, 2003; Pham and Cardoso, 2001; Ziehe et al., 2004). An exception is the popular *infomax* ICA (Bell and Sejnowski, 1995), which is based on a maximum entropy criterion.

## A comparison of methods: recovering sources during a complex visual detection task

In this section, we compare three different approaches described above, namely statistical independence, maximum power-ratio, and maximum magnitude difference, for recovering source activity during a visual detection task. The corresponding algorithms are independent component analysis (ICA), generalized eigenvalue decomposition (GEVD), and penalized logistic regression (PLR). Specifically, we focus on an experiment involving an interactive video game which results in many challenges in terms of removing motion and eye-blink artifacts, detecting event-related neural activity and removing confounding neural and non-neural sources.

The experiment has subjects play a video game where they use a keypad to move a "ship" in order to avoid incoming torpedoes from a fleet of submarines. While subjects avoid incoming torpedoes, they are instructed to also detect infrequent task-relevant visual changes (counting changes in the color of submarines on the display). The nature of the task, with rather infrequent secondary task events, predicts N2 and P3b ERP activity (Friedman et al., 2001) locked to detected visual changes.

### *Data acquisition and preprocessing*

A 64-channel EEG system with a sampling rate of 1000 Hz was used for these recordings. Eye-blink and eye-movement activities were recorded separately so that these artifacts could be removed from the EEG using the maximum power method, described in "Maximum power". Prior to the analysis, the data was filtered for low frequency drift, 60 Hz noise, and 120 Hz harmonic noise.

Events were recorded and data was locked to visual "color change" events. Subjects were required to respond with a button press if they detected a visual change. A unique button was assigned for detecting submarine color changes, while other buttons were used for moving the player's ship and avoiding torpedoes. All button responses were made right handed. Trials were given a binary label based on whether or not the subject detected a change (pressed the detection button). Changes that were not detected by the subject were not considered in this analysis.

### *Recovered sources*

Figs. 7–9 show results for sources recovered using statistical independence, maximum power-ratio, and maximum magnitude difference, respectively. For the independence criterion, the data are decomposed using the infomax ICA algorithms provided in EEGLAB (Delorme and Makeig, 2004). For the maximum power-ratio criterion, we use generalized eigenvalue decomposition (GEVD) (Parra and Sajda, 2003), and for the maximum magnitude difference, we use penalized logistic regression (PLR) (Parra et al., 2002). For ICA and GEVD, we plot the scalp projections for the 10 components, locked to the onset of detected visual change, which capture the largest fraction of the variance in the data. Note that there is a sign and scaling difference between the scalp projections for ICA and GEVD which results from the fact that neither method can uniquely recover the sign and scale for these decompositions. Nonetheless, the two decompositions show some similarity in the spatial distributions of scalp activity. For example, the 8th ICA component and the 1st GEVD component show very similar scalp projections and corresponding time courses (absent the sign) which is consistent with the expected N2 and P3b ERPs. The results for the two methods vary for three reasons: (1) the specific

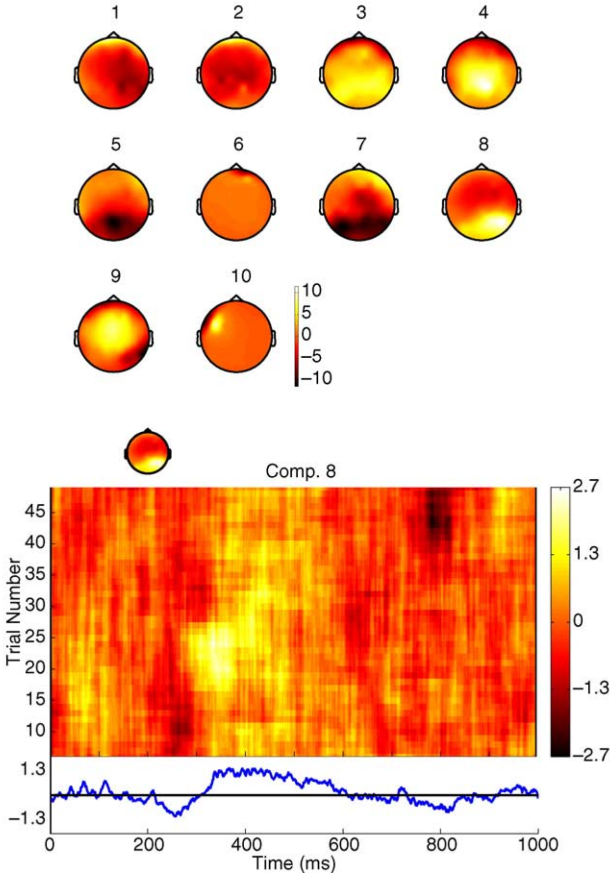

Fig. 7. Sources recovered using ICA. (Top) Scalp projections for first 10 components recovered using infomax ICA in response to detected visual (color) change. Ordering of the scalp projections is based on the fraction of variance captured by each component. (Bottom) Time course of component 8 recovered using ICA. The plot at the bottom shows the average across trials. Note negativity beginning around 200 ms and positivity around 300 ms, consistent with N2 and P3b.

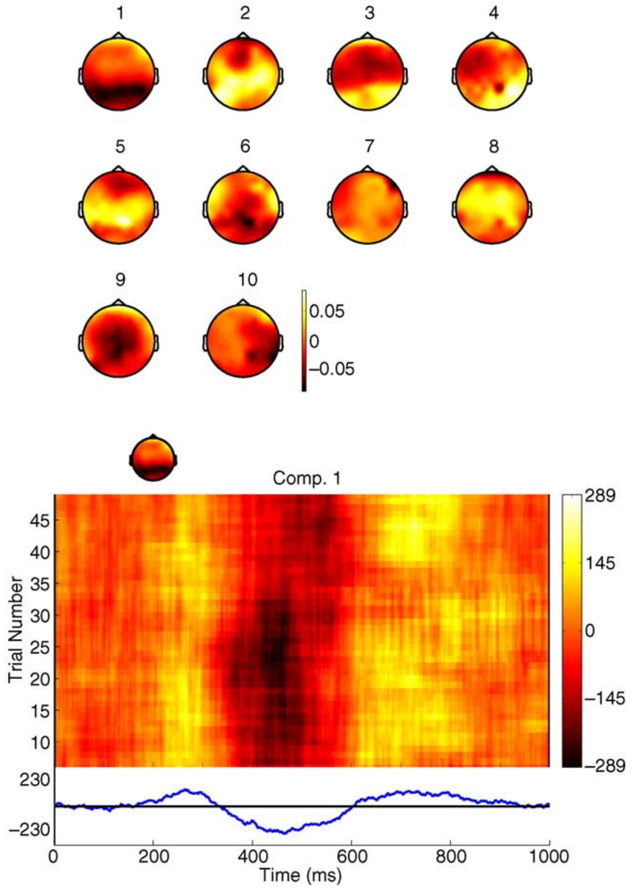

Fig. 8. Sources recovered using GEVD. (Top) Scalp projections for first 10 components with ordering based on the fraction of variance captured by each component. (Bottom) Time course for component recovered using GEVD. Note positivity beginning around 200 ms and negativity around 300 ms.

criterion for statistical independence used in the infomax ICA algorithm is different from the diagonalization criterion discussed above, (2) the infomax algorithm is sensitive to initialization, and (3) The GEVD solution is sensitive to outliers.

For the maximum magnitude difference criterion, with sources recovered using PLR, a single component is recovered which may be interpreted as the neural activity associated with the subject's detection of visual change. In our analysis, we first apply PLR during a 100 ms wide training window starting at 400 ms after the onset of the detected visual change. Within this event-locked window, the PLR model is trained to distinguish activity following a visual change that was detected versus a change that was not detected. The resulting recovered component is displayed in Fig. 9 with the inset showing the corresponding scalp projection. From the PLR result, we see that this component is similar to the 8th component recovered via ICA and the 1st component in GEVD.

We can use the maximum difference criterion to investigate the dynamics of the task-relevant activity, which represents neurological processing associated with change detection. In this case, we train a PLR model to discriminate activity using different temporal windows. For example, using a 100 ms window for $\tau$, we compute the maximally discriminating components $y(t + \tau)$ for the window shifting from 1000 ms pre-stimulus to 1000 ms post-stimulus. In this manner we compute the task-relevant component during varying time intervals relative to stimulus presentation, giving us the progression of the task-relevant component.

We plot the results of these discriminators using a leave-one-out ROC. Fig. 9 illustrates the results with the abscissa showing the time index relative to when the change occurs (locked at 0 ms), and ordinate showing the Az. Also shown are the scalp projections for different discriminators as they are trained on the different windows. Clear is that the PLR estimate recovers different discriminating activity, depending on the timing of the training window and that, in this case, there is a systematic change of discriminating activity from posterior to central regions as the window progresses from 300–800 ms post-stimulus. This changing spatial distribution of activity is consistent with a P300 source followed by a strong motor response that occurs roughly at 650 ms post-stimulus (average response time is 688 ms for this subject).

In summary, ICA may be useful when no timing information is available, for instance when the variability across trials is very large. The difficulty with ICA is that the results are subject to interpretation. The same is true for CSP which, in addition, is sensitive to outliers. The maximum power method may be most

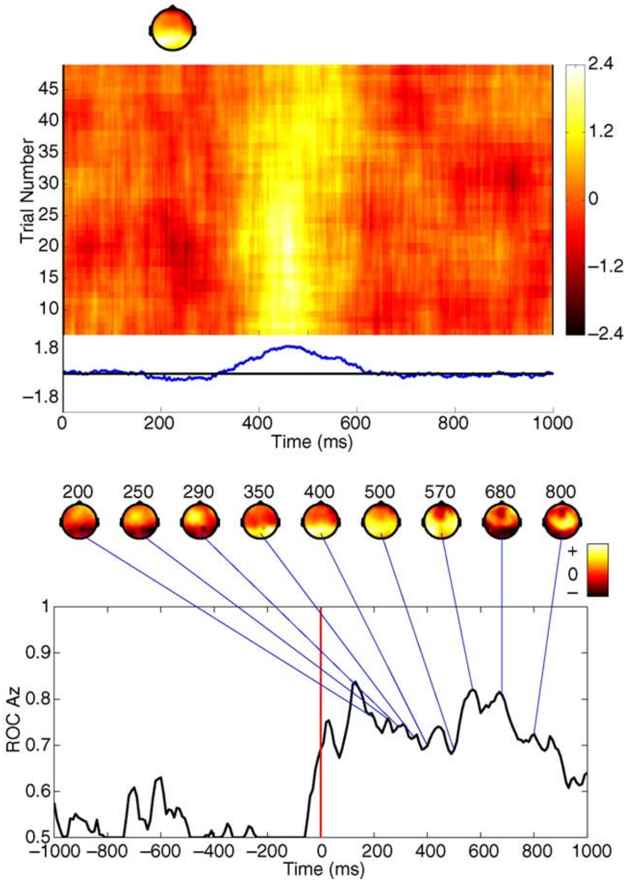

Fig. 9. Sources recovered using PLR show progression of task-related activity. (Top) Scalp projections and level of discrimination (plotted as Az) computed using PLR trained for different temporal windows. (Bottom) Time course and scalp projection of discriminating source recovered via training in a 100 ms window starting at 400 ms. Onset time of training window is indicated at the top of each scalp projection. This analysis shows the progression of task related activity from frontal to parietal areas.

useful to identify motion artifact activity. The maximum power-ratio method is useful when trying to identify non-time locked activity. One must be careful, however to exclude any periods with motion artifacts or myographic activity. Preferably, the data should be band-pass filtered in the range of interest to further reduce the effect of potential artifacts. When precise timing information is available PLR gives the most robust results and high SNR for the case of comparing the evoked response between two experimental conditions.

## Conclusion

In this paper, we have described three basic criteria for recovering physically meaningful neural and non-neural sources via a linear transformation of the EEG activity. These criteria assume different characteristics and/or statistical properties of the recovered source signals, though all assume an underlying linear (Ohmic) model between the source currents and scalp potentials. An advantage of the linear methods proposed here is that they are, in general, computationally inexpensive and can be implemented on-line and in real time, making them useful for a variety of applications including brain–computer interfaces and cognitively adaptive information displays. For example, we currently use the subspace subtraction and discrimination methods in a real-time system (Parra et al., 2003). The eye-motion activity is estimated after a calibration run and subspace subtraction is then applied in real time. The linear discriminator is trained on-line using the Fisher linear discriminator.

In this paper, we have touched on robustness issues in detail only for the discrimination methods for which we used ROC analysis and leave-one-out cross-validation to determine robustness and bootstrapping to determine statistical significance. Future work should address robustness issues in particular for the maximum power-ratio method borrowing perhaps on the statistical methods recently introduced for MEG source localization (Darvas et al., 2004, 2005; Pantazis et al., in press). Finally, we point out that all methods described here are equally applicable to other functional imaging techniques governed by linear imaging equations such as magneto-encephalogy (MEG), and functional near-infrared imaging (fNIR) where the measured optical signal is a linear combination of the modulatory effect of the hemodynamic response on the IR light.

## Acknowledgments

The authors would like to thank An Luo for help with the data collection and analysis. We also thank Xiang Zhou for the help with some of the programming. This work was supported by grants from DARPA, under the Augmented Cognition Program, and the NIH (EB004730).

## Appendix A

Much of what is discussed in the first four Appendices is based on standard linear algebra. A good reference for this material is Strang (2003). To make the relationship with the current material explicit, we reproduce the standard optimality criteria and results in the notation of this paper.

### A.1. Minimum noise source estimate

For a given forward model **A**, we ask 'what is the projection **V** that gives the best estimate of the source activities?'. If we assume that the sources described by the columns in **A** capture most of the power in the signal, it is reasonable to require that the noise account for as little power in the data as possible. The best linear source estimator thus can be obtained by assuming minimal noise power:

$$\hat{\mathbf{V}}^T = \arg\min_{\mathbf{V}} \sum_t \|\mathbf{n}(t)\|^2 \tag{21}$$

$$= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T = \mathbf{A}^{\#}. \tag{22}$$

This is the conventional linear least-squares solution, and it is independent of the data $\mathbf{x}(t)$. If the noise is spatially correlated and we know its covariance matrix, this estimate can be improved. By assuming temporally uncorrelated (white) Gaussian noise with covariance $\mathbf{R}_n$ the maximum likelihood solution to this problem becomes

$$\hat{\mathbf{V}}^T = \arg\min_{\mathbf{V}} \sum_t \mathbf{n}^T(t)\mathbf{R}_n^{-1}\mathbf{n}(t) \tag{23}$$

$$= \left(\mathbf{A}^T\mathbf{R}_n^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{R}_n^{-1}. \tag{24}$$

For spatially uncorrelated noise, $\mathbf{R}_n = \mathbf{I}$, we recover the pseudo-inverse $\mathbf{A}^{\#}$. The relation (7) between a source and its estimate apply equally for this more general estimator $\mathbf{V}^T$.

In practice, the noise covariance can be estimated during resting periods to capture the background EEG where the activity of interest is not present. This method can be used to factor out the effect of background EEG activity in the component estimate. In the case that this noise is temporally correlated, it should be whitened prior to computing $\mathbf{R}_n$, for instance, by applying a moving average filter with the linear prediction coefficients of the noise (Hayes, 1996).

### A.2. Minimum noise forward model estimate

For a given component $y(t)$, we ask 'what is the best forward model $\hat{\mathbf{A}}$ that explains most of the observed data during times $t = t_1,\ldots,t_N$?'. To keep the derivation general we consider a set of components summarized in column vector $\mathbf{y}(t)$. Denote the linear prediction of $\mathbf{x}(t)$ from $\mathbf{y}(t)$ with $\mathbf{x}_{||}(t) = \mathbf{A}\mathbf{y}(t)$. We minimize the distance between the prediction and the data,

$$\hat{\mathbf{A}}_y = \arg\min_{\mathbf{A}} \sum_t \|\mathbf{x}(t) - \mathbf{x}_{||}(t)\|^2 \tag{25}$$

$$= \mathbf{X}\mathbf{Y}^T\left(\mathbf{Y}\mathbf{Y}^T\right)^{-1}. \tag{26}$$

In the last line, we have arranged the times of interest in the columns of $\mathbf{Y} = [\mathbf{y}(t_1),\ldots,\mathbf{y}(t_N)]$ and $\mathbf{X} = [\mathbf{x}(t_1),\ldots,\mathbf{x}(t_N)]$.

We note that the corresponding source estimate $\hat{\mathbf{S}}_y = \hat{\mathbf{A}}_y^{\#}\mathbf{X}$ captures all the activity in $\mathbf{X}$ that is correlated with $\mathbf{Y}$. To see this consider the correlation of $\mathbf{X}_\perp$ with $\mathbf{Y}$

$$\mathbf{X}_\perp\mathbf{Y}^T = \left(\mathbf{I} - \hat{\mathbf{A}}_y\hat{\mathbf{A}}_y^{\#}\right)\mathbf{X}\mathbf{Y}^T = 0. \tag{27}$$

The second equality follows after inserting (22) and (26) and some simplification. The same can be shown for the source estimator (24) if we define more generally, $\mathbf{X}_\perp(t) = (\mathbf{I} - \hat{\mathbf{A}}_y\hat{\mathbf{V}}_y^T)$, where the estimator $\hat{\mathbf{V}}_y^T$ is determined from $\hat{\mathbf{A}}_y$ and some noise covariance $\mathbf{R}_n$ using (24). This means that after subtracting the contribution of $\hat{\mathbf{S}}_y(t)$, there is no correlated activity left in $\mathbf{X}$, as was intended.

Finally, we note that the result of (26) is sensitive to constant offsets in the electrode potentials $\mathbf{x}(t)$ and components $\mathbf{y}(t)$. Constant offsets in the potentials originate from constant current sources and are usually of little interest in neurophysiology. To segregate the activity of constant sources from that of variable sources, one should always include in $\mathbf{Y}$ a constant current component $y(t) = 1$, i.e., augment if necessary, $\mathbf{Y}^T \leftarrow [\mathbf{Y}^T, \mathbf{1}]$, with an $N$-dimensional column vector $\mathbf{1} = [1, 1,\ldots, 1]^T$. The additional (last) column of $\hat{\mathbf{A}}_y$ captures the forward model of any observed DC activity. If $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are zero mean, this step can be omitted as in (9).

### *A.3. Relation between* $\mathbf{V}$*,* $\mathbf{w}$ *and* $\hat{\mathbf{A}}_y$

In general, $\mathbf{A}$ is the forward model while $\mathbf{w}$ and $\mathbf{V}$ can be thought of as backward models. The difference between $\mathbf{w}$ and $\mathbf{V}$ is that $\mathbf{w}$ generates projections with specific optimality conditions, while $\mathbf{V}$ is an estimator that recovers sources for a given set of forward models.

The model $\hat{\mathbf{A}}_y$ associated with one or more $\mathbf{w}$ can be computed using Eq. (9) or (26), and estimator $\mathbf{V}$ can be computed for any $\hat{\mathbf{A}}$ with Eq. (22) or (24). In the case of $\hat{\mathbf{A}}_y$ the estimate $\hat{\mathbf{s}}_y(t)$ in (5) is the source of all the activity in the data $\mathbf{x}(t)$ that correlates with components $\mathbf{y}(t)$.

We establish now some useful relationships between $\hat{\mathbf{A}}_y$, $\mathbf{w}$ and $\mathbf{V}$. In multiple dimensions and with the notation of the previous section, we can write $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where each row of $\mathbf{W}$ represents the projection operator for each component in column vector $\mathbf{y}(t)$. Denote the correlation of the sensors with

$$\mathbf{R} = \sum_t \mathbf{x}(t)\mathbf{x}^T(t) = \mathbf{X}\mathbf{X}^T. \tag{28}$$

Eq. (26) can then be rewritten as

$$\hat{\mathbf{A}}_y = \mathbf{X}\mathbf{X}^T\mathbf{W}\left(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}\right)^{-1} = \mathbf{R}\mathbf{W}\left(\mathbf{W}^T\mathbf{R}\mathbf{W}\right)^{-1}. \tag{29}$$

When $\mathbf{W}$ and $\mathbf{R}$ are full rank, this equation simplifies to Eq. (20) in the main text. Furthermore, in the case of uncorrelated data, $\mathbf{R} = \mathbf{I}$, Eq. (29) simplifies to $\hat{\mathbf{A}}_y = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1} = \mathbf{W}^{\#T}$, regardless of the rank of $\mathbf{W}$. Therefore

$$\mathbf{V} = \hat{\mathbf{A}}_y^{\#T} = \left(\mathbf{W}^{\#T}\right)^{\#T} = \mathbf{W}. \tag{30}$$

Thus, the projection operator that recovers the sources $\hat{\mathbf{s}}_y(t)$ associated with components $\mathbf{y}(t)$ is identical to the projection operator that generates components $\mathbf{y}(t)$, and therefore, $\hat{\mathbf{s}}_y(t) = \mathbf{y}(t)$. If we choose estimator (24) instead of the pseudo-inverse (22) the same can be shown for the case that the sensor data and the noise data have the same spatial correlations, i.e., $\mathbf{R} = \mathbf{R}_n$.

### *A.4. Reduced rank space*

In "Interference subtraction", we argue that noise artifacts can be removed by subspace projection generating an artifact free representation of the original data $\mathbf{x}_\perp(t)$. This data has a reduced rank which may be problematic for subsequent algorithms that estimate various optimal projections $\mathbf{W}$. For instance, all the covariance matrices that are discussed in this paper, when computed with $\mathbf{x}_\perp(t)$ rather than the original $\mathbf{x}(t)$, will be rank deficient.[6]

An alternative approach, which we use in our implementation of the PLR, is to generate a representation of the data with reduced dimensionality but full rank. This is easily accomplished using the normalized null space of $\mathbf{A}$ which we denote here as $\mathbf{A}_\perp$ such that $\mathbf{A}^T\mathbf{A}_\perp = 0$ and $\mathbf{A}_\perp^T\mathbf{A}_\perp = \mathbf{I}$. The representation of the data with reduced dimensions is then

$$\mathbf{z}(t) = \mathbf{A}_\perp^{\#}\mathbf{x}(t) = \mathbf{A}_\perp^T\mathbf{x}(t). \tag{31}$$

It is easy to show that $\mathbf{x}_\perp(t)$, which is the corresponding interference free activity in sensor space as defined in (8), is given by

$$\mathbf{x}_\perp(t) = \mathbf{A}_\perp\mathbf{z}(t). \tag{32}$$

Any $\mathbf{W}$ that has been computed using $\mathbf{z}$ instead of $\mathbf{x}$ corresponds to $\mathbf{W}\mathbf{A}_\perp^T$ when computed for the original sensor space $\mathbf{x}$. In particular use, $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t) = \mathbf{W}\mathbf{A}_\perp^T\mathbf{x}(t)$, prior to computing the forward model $\hat{\mathbf{A}}_y$ in (26).

## Appendix B

### *B.1. Logistic regression*

The assumption in logistic regression is that the data when projected onto coordinate, $y = \mathbf{w}^T\mathbf{x} + b$, is distributed according to a logistic function, i.e., the likelihood that sample $x$ belongs to the class of positive examples, $c = +1$, follows

$$p(c = +1|\mathbf{x}) = f(y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}. \tag{33}$$

This likelihood is parametrized by the projection vector $\mathbf{w}$ and bias $b$. The likelihood for negative examples, $p(c = -1\,|\,\mathbf{x}) = 1 - f(x)$, is also a logistic (see Fig. 10). The hyper-plane orthogonal to $\mathbf{w}$ and shifted by $b$ from the origin divides the two classes. This model is accurate assuming that the data for each class is distributed according to a Gaussian, Bernoulli, Poisson, or any other distribution within the exponential family. The method is called logistic regression because the regression orientation $\mathbf{w}$ is adjusted such that the data matches the logistic model distribution

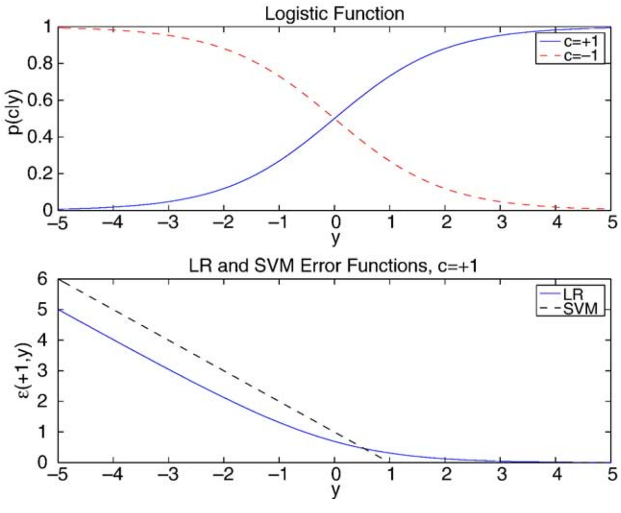

Fig. 10. Top: logistic function. Bottom: cross-entropy error $\epsilon_{\mathrm{lr}}$ and effective SVM error function $\epsilon_{\mathrm{svm}}$ for positive examples.

[6] Principal component algorithms and generalized eigenvalue routines will remain unaffected by this, simply the lower eigenvalues have to be discarded. The inverse for the Fisher linear discriminant should be replaced by the pseudo-inverse.

(33). This can be accomplished by maximizing the likelihood of the data with respect to the model parameters. Given i.i.d. samples $\mathbf{x}_t$, $c_t$, the negative log-likelihood of the data is

$$L(\mathbf{w}, b) = -\sum_t \log p(c_t|y_t). \tag{34}$$

The maximum likelihood solution for the parameters $\mathbf{w}$ is

$$\mathbf{w}_{\mathrm{lr}} = \arg\min_{\mathbf{w}} \; L(\mathbf{w}, b). \tag{35}$$

There is no closed-form solution to this optimization problem. However, the maximum can be computed using a fast algorithm based on iteratively reweighted least squares (IRLS) (McCullagh and Nelder, 1989). It is a type of Newton–Raphson gradient descent algorithm called the Fisher Scoring method where the Hessian is given by the Fisher information matrix (FIM):

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - E\left[\frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w} \partial w^T}\right]^{-1} \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}. \tag{36}$$

The algorithm is summarized in Table 1 with $\lambda = 0$.

For perfectly separable classes the transition between classes can be arbitrarily sharp and the norm $||\mathbf{w}||$ grows without bounds. We can limit the slope of the transition between classes by adding a penalty term $\frac{\lambda}{2}||\mathbf{w}||^2$ to the cost function (34), which is equivalent to assuming a priori that $\mathbf{w}$ is Gaussian distributed with zero-mean. The corresponding algorithm is called penalized logistic regression (PLR). The terms with $\lambda$ in Table 1 implement this additional penalty.

### B.2. Support vector machines and penalized logistic regression

The optimality criteria for support vectors machines (SVMs) differs, on the surface, from logistic regression: the projection vector $\mathbf{w}$ is chosen so that the region around the dividing hyperplane that is free of samples is as wide as possible. This region is called the "margin", and so this is called a "maximum margin" criterion. Since the margin is as wide as possible, there will be samples on its boundaries that therefore define the margin. These are the so-called "support vectors". If the problem is not separable there is no such margin. In this case, some samples are allowed to violate the margin, but with a penalty that is linear in the amount of the violation. The optimization problem for SVMs is a quadratic programming (QP) problem, so there is one optimum, and the algorithms to solve it are relatively fast (Burges, 1998; Müller et al., 2001). Here, we highlight the similarity between support vector machines and penalized logistic regression.

Table 1
Penalized logistic regression using IRLS

```
Initialize $\mathbf{w} = 0$
Repeat these 4 steps until $\mathbf{w}$ converges
$\mathbf{p} \leftarrow f(\mathbf{w}^T * \mathbf{X})$
$\mathbf{g} \leftarrow \mathbf{X} * (\mathbf{d} - \mathbf{p}) - \mathbf{\Lambda} * \mathbf{w}$
$\mathbf{H} \leftarrow \mathbf{X} * \mathrm{diag}(\mathbf{p}.* (1 - \mathbf{p})) * \mathbf{X}^T + \mathbf{\Lambda} * \mathbf{I}$
$\mathbf{w} \leftarrow \mathbf{w} + \mathbf{H}^{-1} * \mathbf{g}$
```

$f()$ is applied element-wise, * is the inner product, .* is the element-wise product, $\mathbf{d}$ is a row vector containing $\{1, 0\}$ class labels for each sample, and diag() converts a vector into a diagonal matrix. Column vector $\mathbf{g}$ represents the gradient, and $\mathbf{H}$ is the FIM in Eq. (36). In this implementation, the bias term was merged with vector $\mathbf{w}$ by substituting $[\mathbf{w}, b] \rightarrow \mathbf{w}$, $[\mathbf{x}, 1] \rightarrow \mathbf{x}$, and $\mathbf{\Lambda} \rightarrow \lambda \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$.

The conditions on the support vectors can be derived from a cost function that is quite similar to penalized logistic regression,

$$L(\mathbf{w}, b) = \sum_t \epsilon(c_t, y_t) + \frac{\lambda}{2}||\mathbf{w}||^2. \tag{37}$$

In fact, we recover PLR with

$$\epsilon_{\mathrm{lr}}(c, y) = -\log p(c|y) = -d \log f(y) - (1-d) \log(1 - f(y)), \tag{38}$$

where $d = (c + 1)/2$. This expression is also called the cross-entropy error.

The gradient of (37) with respect to $\mathbf{w}$ must be zero at a minimum of $L$, which implies that

$$\mathbf{w} = -\frac{1}{\lambda} \sum_t \frac{\partial \epsilon}{\partial y}(c_t, y_t)\mathbf{x}_t \tag{39}$$

at a minimum. If we define $\alpha_t = -c_t \lambda \partial \epsilon(c_t, y_t)/\partial y$, then $\mathbf{w} = \sum_t \alpha_t c_t \mathbf{x}_t$, which is the expression normally obtained for a linear SVM (Burges, 1998).

To carry the analogy with SVMs further, let us require $\alpha_t \geq 0$ and $\sum_t \alpha_t c_t = 0$. The first of these requirements is then

$$-c_t \frac{\partial \epsilon}{\partial y}(y_t, c_t) \geq 0. \tag{40}$$

This is the plausible condition that for positive examples, the error function does not increase with $y$ and for negative examples it does not decrease with $y$. An important example is $\epsilon_{\mathrm{lr}}$, so this applies to PLR discriminators. The second condition, $\sum_t \alpha_t c_t = 0$, becomes $\sum_t \partial \epsilon / \partial_y (y_t, c_t) = 0$: For both SVMs and the other linear classifiers, we are considering, this just expresses the condition for the objective function to be at an extremum with respect to $b$. Thus, for a linear discriminator, the parameters that optimize most reasonable error functions plus a weight-decay term can be expressed in exactly the same way as the SVM expression for the weight vector.

Furthermore, we can express the SVM optimization problem in the same terms, with the error function

$$\epsilon_{\mathrm{svm}}(y, c) = \begin{cases} 1 - cy & \text{if } cy < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

To show this, note that $\epsilon_{\mathrm{svm}}$ is continuous but not differentiable at $cy = 1$. A standard trick for optimizing a piece-wise differentiable function is to introduce an extra variable to the optimization problem, say $\xi$, replace the piece-wise differentiable function with this new variable, and add the constraints that this variable must be greater than each of the differentiable pieces. In our case, we introduce a new variable $\xi_t$ for each example in the training set. Each occurrence of the error function in Eq. (37) is then replaced with the corresponding new variable. If we let $\lambda = 1/C$ and multiply $L$ by $C$, we get

$$L(w, b) = \frac{1}{2}||\mathbf{w}||^2 + C \sum_t \xi_t, \tag{42}$$

which we minimize subject to the constraints

$$c_t(\mathbf{w}^T \mathbf{x}_t + b) - 1 + \xi_t \geq 0, \text{ and} \tag{43}$$

$$\xi_t \geq 0. \tag{44}$$

This is, of course, exactly the SVM optimization problem, with slack variables introduced to allow some data points to violate the margin conditions, with a penalty.

To compare $\epsilon_{\text{svm}}$ with $\epsilon_{\text{lr}}$, consider the behavior of $\epsilon_{\text{lr}}(c, y)$ for $c = +1$ at the extremes of $y$. For large positive $y$, this is $\epsilon_{\text{lr}}(+1, y) = \log(1 + e^{-y}) \approx e^{-y}$, so that it approaches zero rapidly. For large negative $y$, this is $\epsilon_{\text{lr}}(+1, y) = -y + \log(1 + e^{y}) \approx -y$: Thus, it increases linearly as $y$ decreases. This similarity with $\epsilon_{\text{svm}}(+1, y)$ is clear (see Fig. 10). The behavior for negative examples is simply reflected about $y = 0$.

Given the similarity between SVMs and PLR discriminators, we might expect the performance of discriminators trained with the two methods to be similar. The main difference lies in the transition region, near $y = 0$, so the SVM and PLR solutions will be similar if the classes are well separated and the penalty constant $\lambda$ is not too large. For a PLR classifier, few if any of the $\alpha$ will be zero, but those that would be zero in the SVM solution will be very small in the PLR.

Conversely, the two classifiers will be different if there are many examples that fall in the margin or non-asymptotic region, i.e., when the two classes do not separate well. In our experience, SVMs fail under these conditions, misclassifying many examples, whereas PLR gives reasonable answers. We can explain this observation as the breakdown of the intuitive idea of a margin when most examples fall within it.

Finally, we note that the same arguments can be applied to non-linear SVM and non-linear Logistic Regression, which is presented in greater detail in Wahba (1999) and Girosi (1998) along with the linear case.

## Appendix C

### C.1. Generalized eigenvalue problem

Here, we show the relationship of the generalized eigenvalue problem to the maximum power criterion and the maximum power-ratio criterion.

At its optimum, the gradient of the power ratio must vanish. Defining $\lambda$ as the power ratio

$$\lambda = \frac{\mathbf{w}^T\mathbf{R}_2\mathbf{w}}{\mathbf{w}^T\mathbf{R}_1\mathbf{w}}, \tag{45}$$

we can write this condition as

$$0 = \frac{\partial\lambda}{\partial\mathbf{w}^T} = \frac{\mathbf{R}_2\mathbf{w} - \lambda\mathbf{R}_1\mathbf{w}}{\mathbf{w}^T\mathbf{R}_1\mathbf{w}}, \tag{46}$$

and therefore

$$\mathbf{R}_2\mathbf{w} = \lambda\mathbf{R}_1\mathbf{w}, \tag{47}$$

which is a generalized eigenvalue equation with $\lambda$ as the eigenvalue. Of the possible solutions, which can be found with standard numerical methods, we choose the one with the largest eigenvalue, as this is equivalent to the maximum power-ratio according to the definition (45).

The same argument applies for the maximum power criterion of Eq. (15) if we set $\mathbf{R}_1 = \mathbf{I}$, in which case (47) becomes a conventional eigenvalue equation and the maximum eigenvalue corresponds to the maximum power. The corresponding eigenvector is also called the first principal component.

### C.2. Generalized eigenvectors are commons spatial patterns

Koles introduced to EEG analysis a decomposition that diagonalizes the covariance matrices during two experimental conditions which they termed *common spatial patterns* (CSP) (Koles, 1997; Ramoser et al., 2000). We review here the fact that such simultaneous diagonalization is equivalent to the generalized eigenvectors of $\mathbf{R}_1$ and $\mathbf{R}_2$ (Fukunaga, 1990), and according to Parra and Sajda (2003) are therefore an estimate of the independent components of the data. The CSP components are computed as follows (Ramoser et al., 2000):

(1) Diagonalize and scale matrix $(\mathbf{R}_1 + \mathbf{R}_2)$ by finding an orthogonal matrix $\mathbf{P}$ such that the covariance becomes identity (a process known as "sphering")

$$\mathbf{P}^T(\mathbf{R}_1 + \mathbf{R}_2)\mathbf{P} = \mathbf{I}. \tag{48}$$

(2) Diagonalize matrix $(\mathbf{P}^T\mathbf{R}_1\mathbf{P})$ by finding an orthonormal rotation matrix $\mathbf{R}$ such that

$$\mathbf{R}^T\left(\mathbf{P}^T\mathbf{R}_1\mathbf{P}\right)\mathbf{R} = \mathbf{D}, \tag{49}$$

with diagonal $\mathbf{D}$. The CSP components are then

$$\mathbf{W}_{\text{csp}} = \mathbf{PR}. \tag{50}$$

It is easy to show that $\mathbf{W}_{\text{csp}}$ satisfies the following two diagonalization conditions

$$\mathbf{W}^T\mathbf{R}_1\mathbf{W} = \mathbf{D}_1, \tag{51}$$

$$\mathbf{W}^T\mathbf{R}_2\mathbf{W} = \mathbf{D}_2, \tag{52}$$

with some diagonal matrices $\mathbf{D}_1$, and $\mathbf{D}_2$. Eq. (51) with, $\mathbf{D}_1 = \mathbf{D}$, follows from (49) and definition (50). Eq. (52) with, $\mathbf{D}_2 = \mathbf{I} - \mathbf{D}$, follows from (48)–(50).

Another matrix that satisfies these diagonalization conditions (with different diagonal values) can be found by combining the two equations: Right-multiply (51) with $\mathbf{W}^{-1}\mathbf{R}_1^{-1}$ leading to $\mathbf{W}^T = \mathbf{D}_1\mathbf{W}^{-1}\,\mathbf{R}_1^{-1}$; insert this into (52), and left-multiply with $\mathbf{R}_1\mathbf{W}\mathbf{D}_1^{-1}$, which gives

$$\mathbf{R}_2\mathbf{W} = \mathbf{R}_1\mathbf{W}\Lambda, \tag{53}$$

with $\mathbf{\Lambda} = \mathbf{D}_1^{-1}\mathbf{D}_2$. This is a standard generalized eigenvalue equation where all the eigenvalues are written as a diagonal matrix $\mathbf{\Lambda}$.

We denote the solution for the corresponding eigenvectors as $\mathbf{W}_{\text{gev}}$. We find therefore that both $\mathbf{W}_{\text{csp}}$ and $\mathbf{W}_{\text{gev}}$ satisfy the diagonalization conditions (51) and (52). Indeed, they are related by a simple diagonal scaling matrix $\mathbf{S}$

$$\mathbf{W}_{\text{gev}} = \mathbf{S}^{1/2}\mathbf{W}_{\text{csp}}, \tag{54}$$

which we identify as $\mathbf{S} = \mathbf{D}_1 + \mathbf{D}_2 = \mathbf{W}_{\text{gev}}^T(\mathbf{R}_1 + \mathbf{R}_2)\mathbf{W}_{\text{gev}}$ after inserting (54) into (51) and (52), eliminating $\mathbf{D}$, and solving for $\mathbf{S}$.

In summary, we have shown that common spatial patterns are essentially the generalized eigenvectors of covariance matrices estimated at different times. This, in turn, is known to be an estimate of the independent components of the data under the assumption of non-stationary sources (Parra and Sajda, 2003).

## References

Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159.

Belouchrani, A., Meraim, A., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique based on second order statistics. IEEE Trans. Signal Process. 45 (2), 434–444.

Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K.-R., Kunzmann, V., Losch, F., Curio, G., 2003. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trail EEG analysis. IEEE Trans. Neural Syst. Rehabil. Eng. 11 (2), 127–131.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2 (2), 121–167.

Cardoso, J.-F., Souloumiac, A., 1993. Blind beamforming for non Gaussian signals. IEE Proc.-F 140 (6), 362–370.

Cardoso, J.-F., Souloumiac, A., 1996. Jacobi angles for simultaneous diagonalization. SIAM J. Matrix Anal. Appl. 17 (1), 161–164.

Chapman, R., McCrary, J., 1995. EP component identification and measurement by principal components analysis. Brain Cogn. 27 (3), 288–310 (review, Erratum in: Brain Cogn. 1995 Aug; 28 (3) 342).

Croft, R., Barry, R., 2000. Removal of ocular artifact from the EEG: a review. Neurophysiol. Clin. 30 (1), 5–19 (Feb).

Darvas, F., Pantazis, D., Kucukaltun-Yildirim, E., Leahy, R., 2004. Mapping human brain function with MEG and EEG: methods and validation. NeuroImage 23 (1), S289–S299.

Darvas, F., Rautianinen, M., Pantazis, D., Baillet, S., Benali, H., Mosher, J., Garnero, L., Leahy, R., 2005 (Apr 1). Investigations of dipole localization accuracy in MEG using the bootstrap. NeuroImage 25 (2), 355–368.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, 2nd ed. John Wiley & Sons.

Edington, E., 1980. Randomization Tests. Marcel Dekker, New York.

Friedman, D., Cycowics, Y., Gaeta, H., 2001. The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci. Biobehav. Rev. 25, 355–373.

Fukunaga, K., 1990. Introduction to statistical pattern recognition, Computer Science and Scientific Computing Series, 2nd ed. Academic Press.

Gerson, A., Parra, L., Sajda, P., 2005. Cortical origins of response time variability during rapid discrimination of visual objects. NeuroImage 28, 342–353 (this issue).

Girosi, F., 1998. An equivalence between sparse approximation and support vector machines. Neural Comput. 10 (6), 1455–1480.

Green, D., Swets, J., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.

Hayes, M.H., 1996. Statistical digital signal processing and modeling. John Wiley & Sons, Inc.

Kaiser, H.F., 1958. The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200.

Koles, Z., 1997. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. Electroencephalogr. Clin. Neurophysiol. 79 (6), 440–447 (December).

Koles, Z., J.C., L., Soong, A., 1995. Spatio-temporal decomposition of the EEG: a general approach to the isolation and localization of sources. Electroencephalogr. Clin. Neurophysiol. 95 (4), 219–230.

Makeig, S., Bell, A., Jung, T., Sejnowski, T., 1996. Independent component analysis of electroencephalographic data. Advances in Neural Information Processing Systems, vol. 8. MIT Press, pp. 145–151.

Makeig, S., Westerfield, M., Jung, T.-P., Covington, J., Townsend, J., Sejnowski, T., Courchesne, E., 1999. Independent components of the late positive response complex in a visual spatial attention task. J. Neurosci. 19, 2665–2680.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Michel, C., Murray, M., Lantz, G., Gonzalez, S., Spinelli, L., Grave De Peralta, R., 2004. EEG source imaging. Clin. Neurophysiol. 115 (10), 2195–2222.

Mosher, J.S.B., Leahy, R., 1999. EEG source localization and imaging using multiple signal classification approaches. J. Clin. Neurophysiol. 16 (3), 225–238.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Net. 12 (2), 181–201.

Müller, K.-R., Anderson, C.W., Birch, G.E., 2003. Linear and non-linear methods for brain-computer interfaces. IEEE Trans. Rehabil. Eng. 11 (2), 165–169.

Niedermeyer, E., 1996. Dipole theory and electroencephalography. Clin. Electroencephalogr. 27 (3), 121–131.

Pantazis, D., Nichols, T., Baillet, S., Leahy, R., in press. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. NeuroImage.

Parra, L.C., Sajda, P., 2003. Blind source separation via generalized eigenvalue decomposition. J. Mach. Learn. Res. 4, 1261–1269.

Parra, L.C., Spence, C.D., 2000. Convolutive blind source separation of non-stationary sources. IEEE Trans. Speech Audio Process., 320–327 (May).

Parra, L.C., Alvino, C., Tang, A., Pearlmutter, B., Young, N., Osman, A., Sajda, P., 2002. Linear spatial integration for single-trial detection in encephalography. NeuroImage 17, 223–230.

Parra, L.C., Spence, C.D., Gerson, A., Sajda, P., 2003. Response error correction—A demonstration of improved human–computer performance using real-time EEG monitoring. IEEE Trans. Neural Syst. Rehabil. Eng. 11 (2), 173–177.

Pfurtscheller, G., Lopes da Silva, F.H., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110 (11), 1842–1857 (November) (review).

Pham, D.-T., Cardoso, J.-F., 2001. Blind separation of instantaneous mixtures of non stationary sources. IEEE Trans. Signal Process. 49 (9), 1848–1937.

Ramoser, H., Mueller-Gerking, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehabil. Eng. 8 (4), 441–446.

Reilly, J., 1992. Applied Bioelectricity. Springer.

Reiter, C., 1990. Easy algorithms for finding eigenvalues. Math. Mag. 63 (3), 173–178 (June) (in Notes).

Strang, G., 2003. Introduction to Linear Algebra. Wellesley-Cambridge Press.

Tang, A.C., Pearlmutter, B.A., Malaszenko, N.A., Phung, D.B., Reeb, B.C., 2002. Independent components of magnetoencephalography: localization. Neural Comput. 14 (8), 1827–1858.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Uusital, M.A., Ilmoniemi, R.J., 1997. Signal-space projection method for separating MEG or EEG into components. Med. Biol. Eng. Comput. 35 (2), 135–140 (March).

Wahba, G., 1999. Support Vector Machines, reproducing kernel Hilbert spaces and the randomized GACV. In: Schölkopf, B., Burgess, C., Smola, A.J. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT Press, pp. 69–88.

Ziehe, A., Laskov, P., Nolte, G., Müller, K.-R., 2004. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. J. Machine Learning Res. 5, 777–800.