NeuroImage xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Review

Joint decorrelation, a versatile tool for multichannel data analysis

Alain de Cheveigné [a,b,c,*], Lucas C. Parra [d]

[a] Laboratoire des Systèmes Perceptifs, UMR 8248, CNRS, France
[b] Département d'Etudes Cognitives, Ecole Normale Supérieure, France
[c] University College London, UK
[d] City College New York, USA

ARTICLE INFO

Article history:
Accepted 24 May 2014
Available online xxxx

Keywords:
Electroencephalography (EEG)
Magnetoencephalography (MEG)
Local field potential (LFP)
Electrocorticography (ECoG)
Optical imaging
Multielectrode array
Denoising
Noise reduction
Artifact rejection
Blind source separation (BSS)
Independent Component Analysis (ICA)
Denoising Source Separation (DSS)
Common Spatial Pattern (CSP)
Time Domain source separation (TDSEP)
Simultaneous diagonalization
Joint diagonalization
Principal Component Analysis (PCA)

ABSTRACT

We review a simple yet versatile approach for the analysis of multichannel data, focusing in particular on brain signals measured with EEG, MEG, ECoG, LFP or optical imaging. Sensors are combined linearly with weights that are chosen to provide optimal signal-to-noise ratio. Signal and noise can be variably defined to match the specific need, e.g. reproducibility over trials, frequency content, or differences between stimulus conditions. We demonstrate how the method can be used to remove power line or cardiac interference, enhance stimulus-evoked or stimulus-induced activity, isolate narrow-band cortical activity, and so on. The approach involves decorrelating both the original and filtered data by joint diagonalization of their covariance matrices. We trace its origins; offer an easy-to-understand explanation; review a range of applications; and chart failure scenarios that might lead to misleading results, in particular due to overfitting. In addition to its flexibility and effectiveness, a major appeal of the method is that it is easy to understand.

Contents

* Corresponding author at: Equipe Audition, ENS, 29 rue d'Ulm, F-75230 Paris, France.
E-mail address: Alain.de.Cheveigne@ens.fr (A. de Cheveigné).

http://dx.doi.org/10.1016/j.neuroimage.2014.05.068
1053-8119/

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

Introduction

Data are increasingly *multidimensional*. The density of electrode arrays increases exponentially (Stevenson and Kording, 2012), brain imaging techniques such as EEG (electroencephalography), MEG (magnetoencephalography) or fMRI (functional magnetic resonance imaging) involve large numbers of electrodes, sensors, or voxels, and optical imaging produces massively parallel time series of pixel values. An array offers several advantages over a single electrode. The yield is improved, as one is effectively running multiple experiments at the same time. Knowledge of the electrode geometry helps map the topography of brain sources. More importantly, the *correlation structure* helps tease apart different sources of brain activity and noise. There is a pressing need for signal processing tools to exploit the rapidly increasing number of sensors in electrophysiological data.

In some cases (e.g. intracellular recording) a sensor waveform might correspond to a single neural source. In general, however, there is mixing between sources and sensors, so that a sensor records a weighted sum of sources (Fig. 1a), while each source contributes to several sensors. This obviously complicates the interpretation of the waveforms and the topographies. Component analysis designates a family of methods that form *linear combinations* of the observed signals. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) (Hyvarinen, 2012; Hyvärinen et al., 2009) are well known, but others such as beamforming, Current Source Density (CSD), Laplacian, or differential montages used in EEG also fit this definition. Their purpose is usually to improve the signal-to-noise ratio (SNR) of the activity of interest, by canceling interference while preserving activity of interest. However they differ by the weights applied, and this begs the question as to whether there exists a "best" set of weights, and how to find it.

Fukunaga and Koontz showed in 1970 how to maximize the difference in the spectrum between two sets of data by joint diagonalization of their auto-correlation matrices (Fukunaga and Koontz, 1970; Fukunaga, 1972, 1990). The same two-step process for diagonalization was later used to identify Common Spatial Patterns (CSP) in EEG — an analysis technique now widely used in the Brain Computer Interface (BCI) community (Blankertz et al., 2008; Dornhege et al., 2006; Koles et al., 1990; Parra et al., 2005; Tangermann et al., 2011; Wang et al., 1999). The idea reoccurs in various forms in a wide range of blind and semi-blind source separation algorithms (Belouchrani et al., 1997; Blaschke et al., 2006; Cichocki, 2004; Molgedey and Schuster, 1994; Parra et al., 2005; Ramoser et al., 2000; Särelä and Valpola, 2005; Ziehe and Müller, 1998). Here we show how the basic principle, joint diagonalization, common to all these methods, in itself is a powerful tool applicable to a wide range of needs. Properly formulated, it is also very easy to understand. Our formulation follows that of Denoising Source Separation (DSS) (Särelä and Valpola, 2005), more specifically linear DSS. Our purpose is not to introduce a new method, but rather to provide a new perspective to an existing approach, in order to highlight its versatility, optimality and ease-of-use.

We will refer to the approach presented here generically as *Joint Decorrelation* (JD), because it simultaneously decorrelates the data as well as the data after filtering. This general approach subsumes prior methods such as CSP, linear DSS and other component extraction techniques. The result is to improve the signal-to-noise ratio (SNR) of the activity of interest within the data — where signal and noise are specified by a "bias filter". Depending on the choice of bias filter one can achieve a variety of common objectives in electrophysiology and imaging: e.g. reproducibility across trials, discrimination between conditions, reduction of interference, and more. Compared to other component extraction techniques, it is attractive because (a) it optimizes a specific objective, (b) components are ordered so that there is no need for post-hoc sorting and selection, (c) a wide variety of applicable objectives makes the method flexible, and (d) it is easy to implement and easy to understand. With these nice features also comes an enhanced risk of *overfitting*, that we also stress below.

The paper is organized as follows. First, we give a simple and intuitive explanation of the approach. Next, we review a series of examples to get a feeling for how it is applied and what can be achieved. Finally we review a number of *failure scenarios* to emphasize its limits and alert the user to potential pitfalls. Many useful details may be found in the appendix.

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

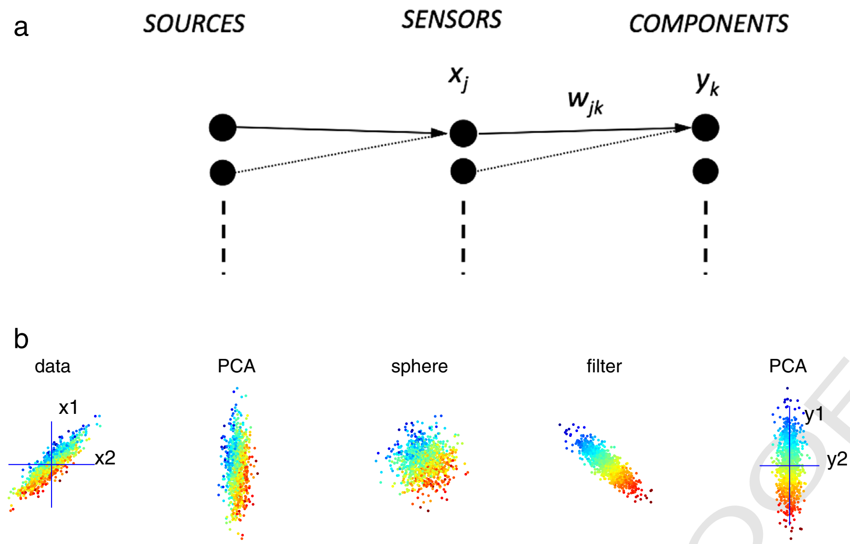


Fig. 1. Signal model and principle of the JD method. (a) Each component signal (right) is the weighted sum of sensor or electrode signals (center), themselves weighted sums of neural sources (left). (b) Illustration of the JD procedure. A signal of interest (coded as color) is embedded with noise in sensor signals x_1 and x_2 (left). PCA decorrelates the data and finds the direction of maximum power (second to left). Scaling renders the data set spherical (center). The "bias filter" enhances the direction that captures the signal of interest while reducing directions with noise (second to right). The final PCA aligns these directions with the final component axes (right).

The joint decorrelation method

Our goal is to combine sensor signals so as to obtain component signals with maximal signal-to-noise ratio. The word "sensor" here designates an individual electrode, MEG sensor, pixel, or voxel. The sensor signals are arranged as columns of a matrix $\mathbf{X} = [x_{tj}]$, where t is time. If the data are made up of multiple trials these are concatenated in time. The J time series x_{tj} of sensor values will be combined linearly to produce K component signals y_{tk} (Fig. 1a):

$$y_{tk} = \sum_{j=1}^{J} x_{tj} w_{jk}, \tag{1}$$

where w_{jk} are weights that will be optimized. In matrix notation, $\mathbf{Y} = \mathbf{XW}$, where $\mathbf{W}$ is the analysis matrix of dimensions $J \times K$, which converts from sensors to components. Component analysis algorithms often assume $K = J$, but we will allow $K \leq J$ to focus on a subset of the components, or handle the case of data of deficient rank.

The sensor signals themselves might be a linear superposition (mixture) of multiple sources of brain activity, noise such as eye blinks and muscle artifacts, power line interference, sensor noise and so on. Ideally, we would like each component to reflect an individual source of neural activity, with the analysis matrix $\mathbf{W}$ serving as an un-mixing matrix that reverses the effects of source-to-sensor mixing. However, brain sources vastly outnumber sensors so this unmixing will not be possible in a strict sense. Instead it is fruitful to see the analysis as a tool to find the "best angle" to view the data, maximizing the SNR for activity of interest.

A noisy signal can often be enhanced by averaging over trials (to enhance trial-locked activity), or applying a filter (to suppress frequency regions dominated by noise), or simply by selecting a temporal interval of higher SNR. These operations can all be formalized as left-multiplication of the data by a matrix $\mathbf{L}$ that we will call "bias filter". JD leverages the selectivity of this filter to find optimal weights for Eq. (1). We restrict ourselves to linear filters which have a number of advantages as discussed in Appendix 1. Non-linear filtering is discussed in Särelä and Valpola (2005).

The JD algorithm is simple. Given a set of sensor or electrode signals $\mathbf{X}$, the analysis matrix $\mathbf{W}$ is found by the following steps:

1. PCA applied to $\mathbf{X}$ produces a rotation matrix $\mathbf{P}$ that orthogonalizes the data, so that columns of $\mathbf{XP}$ are mutually uncorrelated in time.
2. Normalization of $\mathbf{XP}$ produces a diagonal matrix $\mathbf{N}$ that renders the data set "spherical" (unit power in all directions).
3. The bias filter $\mathbf{L}$ applied to $\mathbf{XPN}$ enhances power along relevant directions while reducing power in noise directions.
4. PCA applied to the filtered data $\mathbf{LXPN}$ produces a rotation matrix $\mathbf{Q}$ that aligns the relevant power with the final component axes.

The algorithm is defined more precisely in Appendix 1. The analysis matrix is obtained as $\mathbf{W} = \mathbf{PNQ}$, which transforms the raw observations $\mathbf{X} = [x_{tj}]$ into the components $\mathbf{Y} = [y_{tk}]$. The first component signal $[y_{t1}]$ is the linear combination with the highest possible score, where score is defined as the ratio of power in the bias-filtered data relative to the raw data. The second component signal $[y_{t2}]$ is uncorrelated to the first and has the next highest score, and so on. If the bias filter enhances the signal of interest and reduces noise, this process produces components sorted by SNR, and indeed in some cases JD is guaranteed to generate components with *optimal* SNR (see Appendix 3).

The principle is illustrated in Fig. 1b. The raw observations x_1 and x_2 covary with a signal of interest (coded as color) along some direction that does not coincide with either of the observed dimensions (left). That direction is also not co-linear with directions of maximum or minimum power, so PCA cannot isolate it (second to left). However, rotation and scaling remove the influence of correlation between sensors so that the data set is now "spherical" (center). The bias filter then emphasizes the power of the signal of interest relative to irrelevant directions (second to right). The second PCA aligns these signal directions with the component axes (right), thus producing a component that is maximally sensitive to the signal of interest. Intuitively, JD can be understood as a form of principal component analysis that maximizes the *power-ratio* between filtered and raw signal, and not just power as in conventional PCA.

The choice of bias filter $\mathbf{L}$ depends upon the task, i.e. what should be considered signal and what is noise (see Appendix 2 and examples below). Different filters may be applied to the same data to emphasize different aspects of the data. While the filter $\mathbf{L}$ is involved in determining

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

the projection matrix $\mathbf{W} = \mathbf{PNQ}$, the resulting component signals $\mathbf{Y} = \mathbf{WX}$ are not filtered by $\mathbf{L}$. Of course, it is possible to *also* include filtering, i.e. calculate $\mathbf{Y'} = \mathbf{LXW}$.

From a practical point of view, the matrix $\mathbf{W}$ is calculated on the basis of two covariance matrices: $\mathbf{C}_0$, covariance of the raw data $\mathbf{X}$, and $\mathbf{C}_1$, covariance of the filtered data $\mathbf{LX}$. Once the components are obtained, they may be interpreted directly (as statistics derived from the data), or *projected back* into sensor space, or *projected out* to obtain denoised data (see Appendix 4 for a precise definition of these notions). The following examples show how these ideas can be applied to actual data.

Examples

The following tasks are typical of electrophysiology. JD solves the problem in each case with a bias filter tailored to the task. In some cases it is applied repeatedly with different bias filters. Details may be found in Appendix 6.

Power line noise

The aim here is to identify a subspace dominated by "line noise" (50 or 60 Hz and harmonics), and project it out of the data. This is a common problem in animal and human electrophysiology; ideally it is avoided by appropriate equipment design and shielding, but there are situations where these precautions are not fully effective. If "reference channels" are available, that pick up environmental noise but no brain activity, the noise can be removed by regression (de Cheveigné and Simon, 2007) However in the general case, the interference is intimately mixed with brain activity at all sensors. As an illustration, Fig. 2a shows the power spectrum of an MEG data set. Power at 50 Hz and harmonics is prominent, accounting for 38% of the power in these data.

JD was applied using a bias filter with a comb-shaped transfer function, with peaks at 50 Hz and harmonics, and zeros elsewhere, producing a set of orthogonal components. The power-ratio score (filter output to input) is plotted in Fig. 2c, showing that the first components are strongly dominated by 50 Hz and harmonics. The first 20 components (out of 274) were projected out of the data (see Appendix 6) to obtain clean, noise-free data. At frequencies other than 50 Hz and harmonics, the power spectrum of the clean data (Fig. 2b, red) is similar to that of the raw data (Fig. 2a). The spectrum level of the noise (part removed) is much lower [compare Figs. 2(a) and (b, green)], implying that the impact of denoising on brain activity must be minimal. This example shows how JD can be used to suppress environmental noise.

Stimulus-evoked activity

The aim here is to improve SNR by finding the subspace that is most repeatable across trials. MEG data were obtained in response to repeated visual stimulation. The stimulus appeared 2.5 s from the onset of each 5 s trial (see Appendix 6 for more details). Data were submitted to JD using as a bias filter the average over 30 trials. To be precise, the matrix $\mathbf{C}_0$ (see above) was the covariance matrix of the raw data, and the matrix $\mathbf{C}_1$ was the covariance matrix of the data averaged over trials. In this case the optimality criterion is the power of the mean divided by total power, which implies that the first component is characterized by the strongest possible mean effect relative to overall variability. Fig. 3a shows the power-ratio score for each JD component. The gray band shows the 5–95% interval for that statistic based on surrogate data (see Overfitting and circularity section). Fig. 3c shows the waveforms

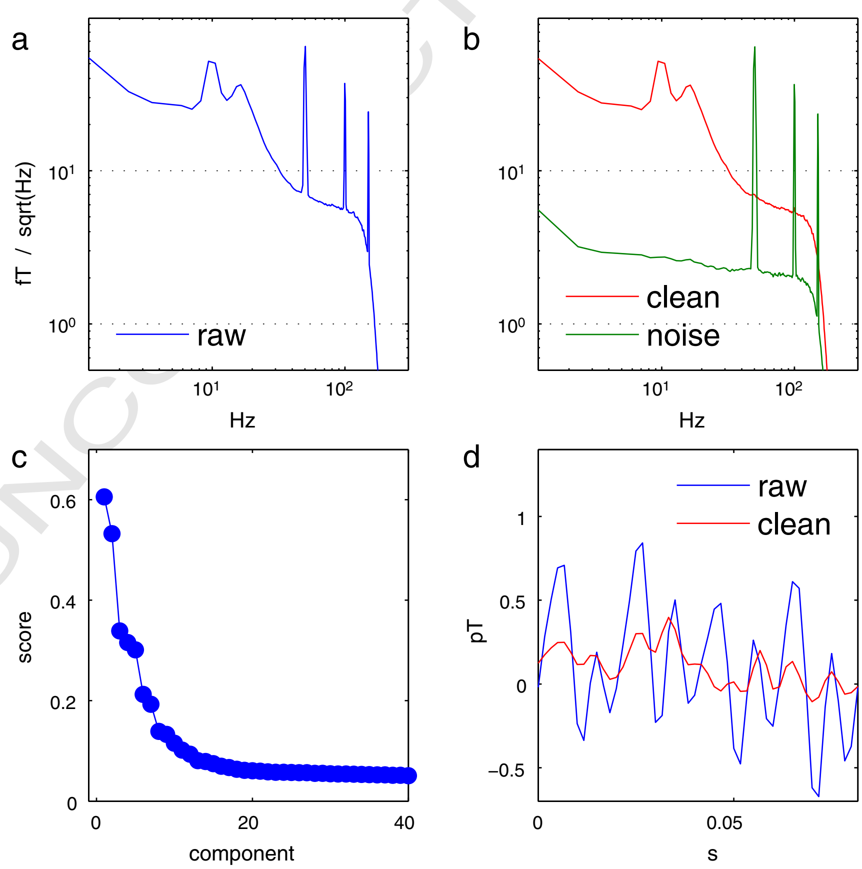


Fig. 2. Removing power line interference from MEG data. (a) Power spectral density averaged over sensors. (b) Red: power spectral density after removal of interference, green: power spectral density of noise. (c) Power-ratio scores for the first 40 components. (d) Time course of one particular channel before (blue) and after (red) noise removal.

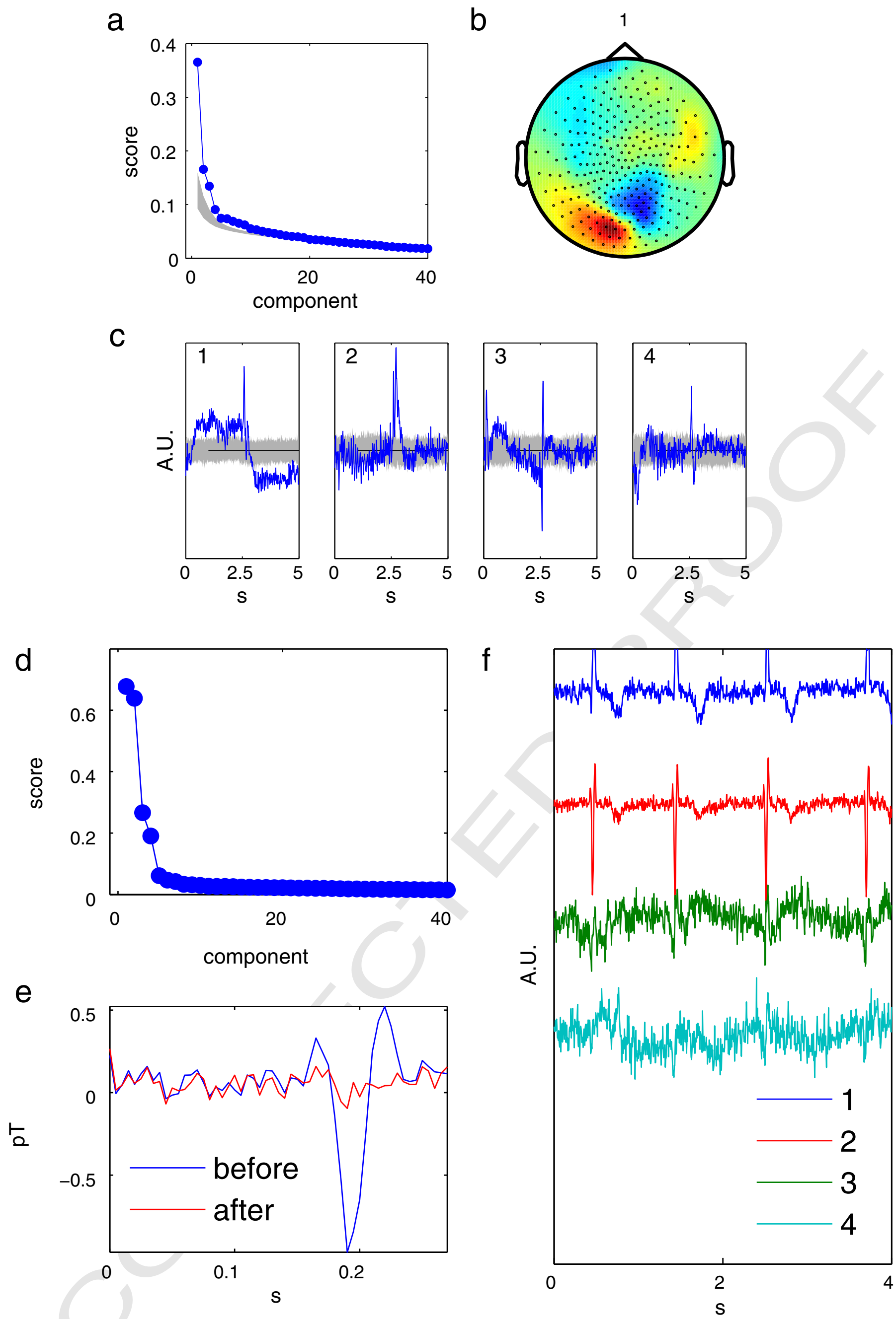


Fig. 3. (a–c) Isolating repeatable components in MEG data. (a) Power-ratio score for each the first 40 components. The gray band indicates the 5 and 95 percentiles of this statistic calculated from surrogate data (see Overfitting and circularity section). (b) Spatial pattern associated with the first JD component. (c) Mean (blue) and ±2 standard deviations of a bootstrap resampling of the mean (gray) of the first 6 JD components. (d–f) Cardiac activity in MEG data. (d) Power-ratio score of the first 40 components. (e) Sample of the time course of one particular sensor before (blue) and after (red) removal of cardiac components. (f) Time courses of first four JD components (offset vertically for clarity).

of the first 4 components. The blue line represents the average over trials, and the gray band ±two standard deviations of a bootstrap resampling of the mean (Efron and Tibshirani, 1993). Fig. 3b shows the topography of the first JD component, calculated as the cross-correlation coefficient between that component and the signal at each sensor (Haufe et al., 2014; Parra et al., 2005). The first component is *the most repeatable linear combination* of sensor signals; the first K components span a "most repeatable subspace" of dimension K. One or more components may be projected back into sensor space to obtain "clean data" (de Cheveigné and Simon, 2008a).

Cardiac artifacts

The aim here is to identify a subspace dominated by electric or magnetic fields originating from the cardiac muscle, or indirect effects of changes in blood pressure or flow, and project it out of the data. If an electrocardiogram (ECG) channel is available, that signal may be regressed out of the data, but the improvement is often limited by differences in shape between the ECG and the artifacts, for example due to different degrees of distortion along different pathways. An alternative strategy is to use the ECG to define epochs corresponding to

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

cardiac cycles, and apply JD as described above for evoked activity, to find a subspace that maximizes the power of the mean cardiac signal versus total power. Fig. 3d shows the score for each component, and Fig. 3f the waveforms of the first four components. These are clearly locked to the cardiac rhythm. These components were then projected out of the data to obtain "clean" data. Fig. 3e compares the signal from one sensor before and after removal.

Narrowband cortical activity

The aim here is to improve the SNR of oscillatory activity. Narrowband oscillations are observed in deep electrode recordings in many parts of the brain (Buzsáki, 2006), but in EEG and surface recordings they are more elusive, often obscured by other activity. Time–frequency analysis, or filtering, may be used to improve signal-to-noise ratio, but there is a concern that filter ringing may masquerade as oscillations and complicate the interpretation of the data (Yeung et al., 2004). Component analysis offers an alternative with potentially less artifacts (but see caveats later on).

The same data, after removal of 50 Hz components, were submitted to JD using a narrowband bias filter centered on 10 Hz. Fig. 4a shows the power-ratio score (bias filter output to input) for each JD component. Fig. 4b displays a raster plot of the power spectra of the first 20 components, showing that they are indeed dominated by 10 Hz power. Approximately 60% of the first component's power is within the spectral region defined by the bias filter. Fig. 4d compares the power spectra of this component (red) to that of the sensor most dominated by 10 Hz (green), and Fig. 4c shows a sample of its time course, which is shaped as a spindle-shaped oscillatory burst. This oscillatory shape is not the result of, or distorted by, filter ringing (the bias filter used to identify spatial components with maximal SNR is *not* included in the sensor-to-component transform). This is in contrast to time–frequency analysis for which the time course is smeared by convolution with the analysis filter. The analysis thus appears to have uncovered genuine oscillatory activity. The topography associated with the first component is shown in Fig. 4e. Fig. 4b shows that more than one component is dominated by alpha, suggesting multiple sources with different time courses and spatial extent. Note that it is unlikely that these JD components map to individual neural sources, instead they collectively define a *signal subspace* within which the measurable alpha activity is concentrated.

The same analysis can be repeated with other bias filter frequencies, to search for other narrowband activity. Looking closely at Fig. 4b, the 7th component seems closer to 12 Hz than 10 Hz, and in Fig. 2b there was also some hint of power near 16 Hz. Applying JD with a bandpass bias filter centered on 12 Hz or 16 Hz isolates narrowband components at those frequencies (Fig. 4f), and a wider bias filter centered on 30 Hz isolates a source of activity within the lower gamma band, with a narrowly localized quadri-polar topography (Fig. 4e). The topographies of the other three components are dipolar, roughly consistent with a current dipole source oriented parallel to the surface of the head. Varying the bias filter frequency systematically did not reveal any other narrowband components (which does not mean that none exist, see de Cheveigné, 2012; Duncan et al., 2009). These examples show how JD can be used to isolate neural activity with specific spectral characteristics (see Nikulin et al., 2011 for a similar method).

Event-related desynchronization (ERD)

Visual and other perceptual stimuli may produce an increase or decrease in power in certain frequency bands, referred to as event-related synchronization or desynchronization (ERS/ERD). This is usually revealed by time–frequency analysis that serves both to improve the SNR of the

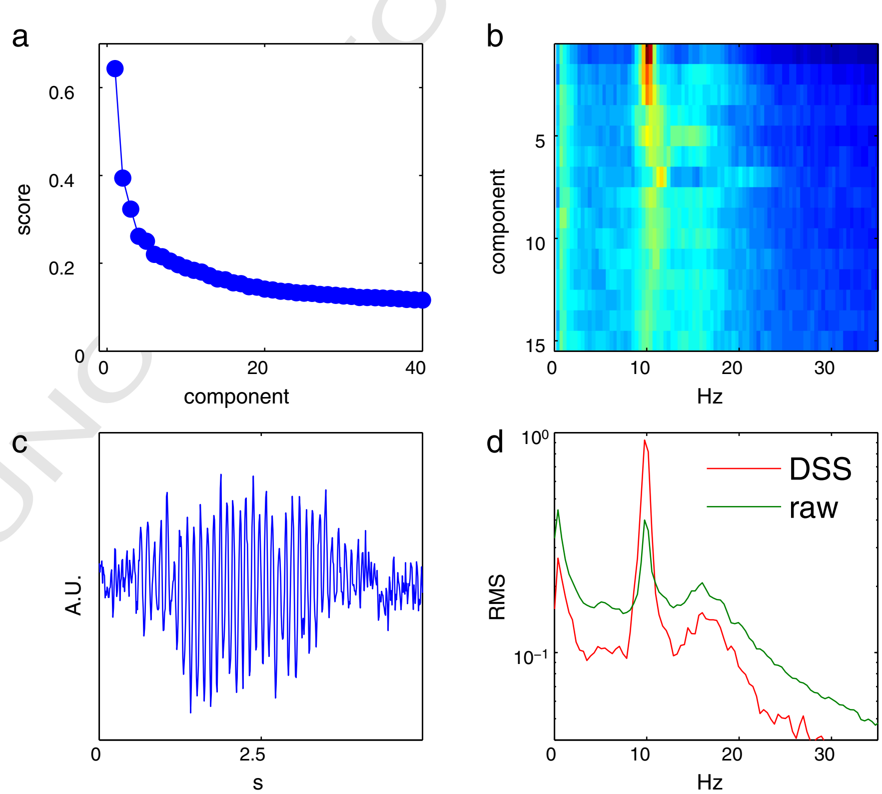


Fig. 4. Narrowband activity in MEG data. (a) Power-ratio score for the first 40 components, for a bias filter centered on 10 Hz. (b) Power spectra of the first 20 JD components. Each line represents the power spectrum of a component coded as color. (c) Sample of the time course of the first JD component. (d) Power spectra of the first JD component (red) and the sensor most strongly dominated by 10 Hz power (green). (e) Topographies of first JD components for bias filters centered on 10, 12, 16 Hz and 30 Hz. (f): Power spectra of these components.

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

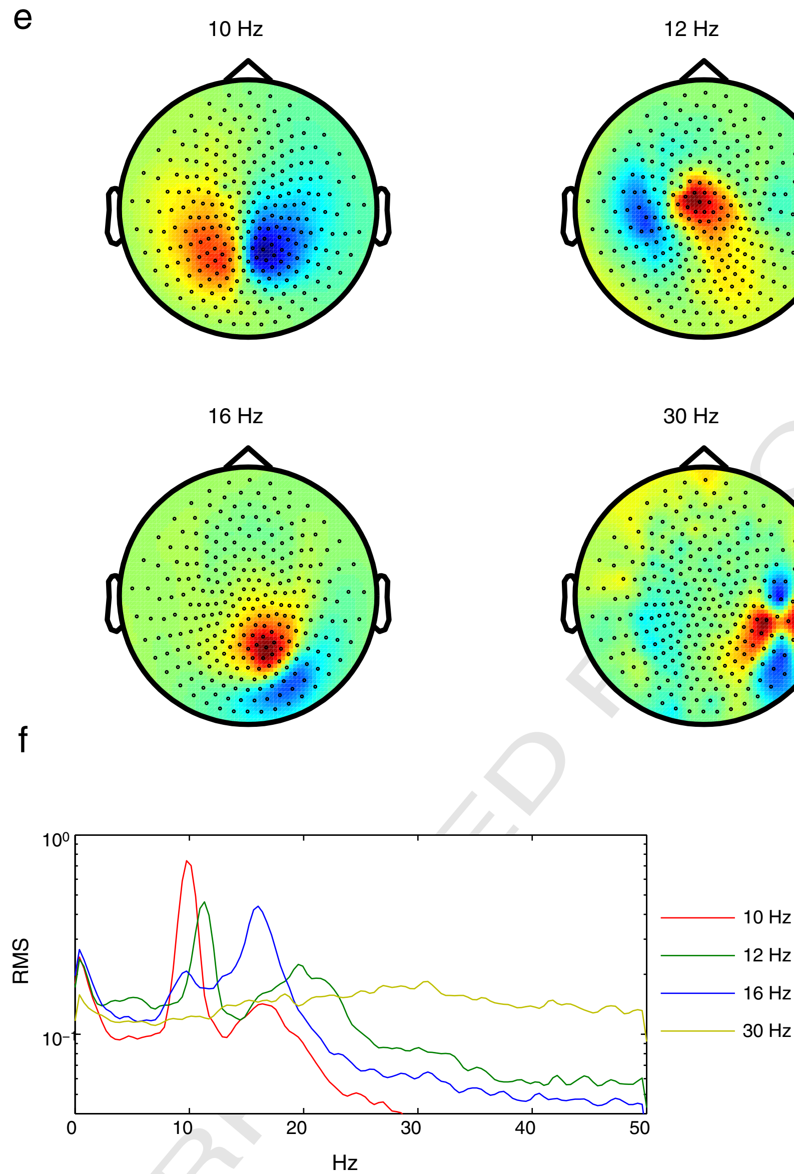


Fig. 4 (*continued*).

effect, and to display its time course. However, time–frequency analysis is subject to temporal smearing, and furthermore a weak ERS/ERD source might be masked by other sources within the same frequency band.

Using the same MEG data as before (visual stimulation), JD was applied using a bias filter that set to zero all samples beyond the onset of stimulation (2.5 s from trial onset), within each trial. This will maximize the power-ratio between the two intervals and thus capture ERD/ERS as proposed in Parra et al. (2005). More precisely, $\mathbf{C}_0$ was calculated as the covariance matrix of data in the 0–5 s interval, and $\mathbf{C}_1$ as the covariance matrix of data in the 0–2.5 s interval (see Appendix 2). Fig. 5a shows the power-ratio between interval 0–2.5 s and interval 2.5–5 s. The power of the first component was almost two times greater in the first than in the second interval. Its topography, and a raster plot of individual trials, are shown in Figs. 5b and c respectively. Fig. 5d shows the spectrogram of the first 4 ERD components. This spectrogram is dominated by power in the 10–16 Hz region, suggesting that the ERD activity is partly included within the subspace of alpha activity found by the previous analysis.

Two conditions, repeated trials

The aim here is to optimize the SNR of brain activity that differs between two different experimental conditions, each of which involves repeated trials. We are interested in activity that is reproducible over trials *and* distinct between conditions. As two criteria are involved, we expect the solution to be within the intersection of two subspaces, each one optimal for one of the criteria. Accordingly we apply JD twice, first to identify a signal subspace that favors reproducibility, and next to find the directions in that subspace that optimize the effect of condition. To illustrate this we use MEG data from a study that recorded responses to visual (V) or combined auditory and visual (AV) stimuli (Molloy et al., in preparation), presented randomly interleaved. Subjects performed a demanding task involving the visual stimulus only, and did not attend to the auditory stimulus present on half the trials. Accordingly, visual and task-related correlates were strong in the MEG data, and there was little evidence of any auditory activity within the raw sensor waveforms. Q7

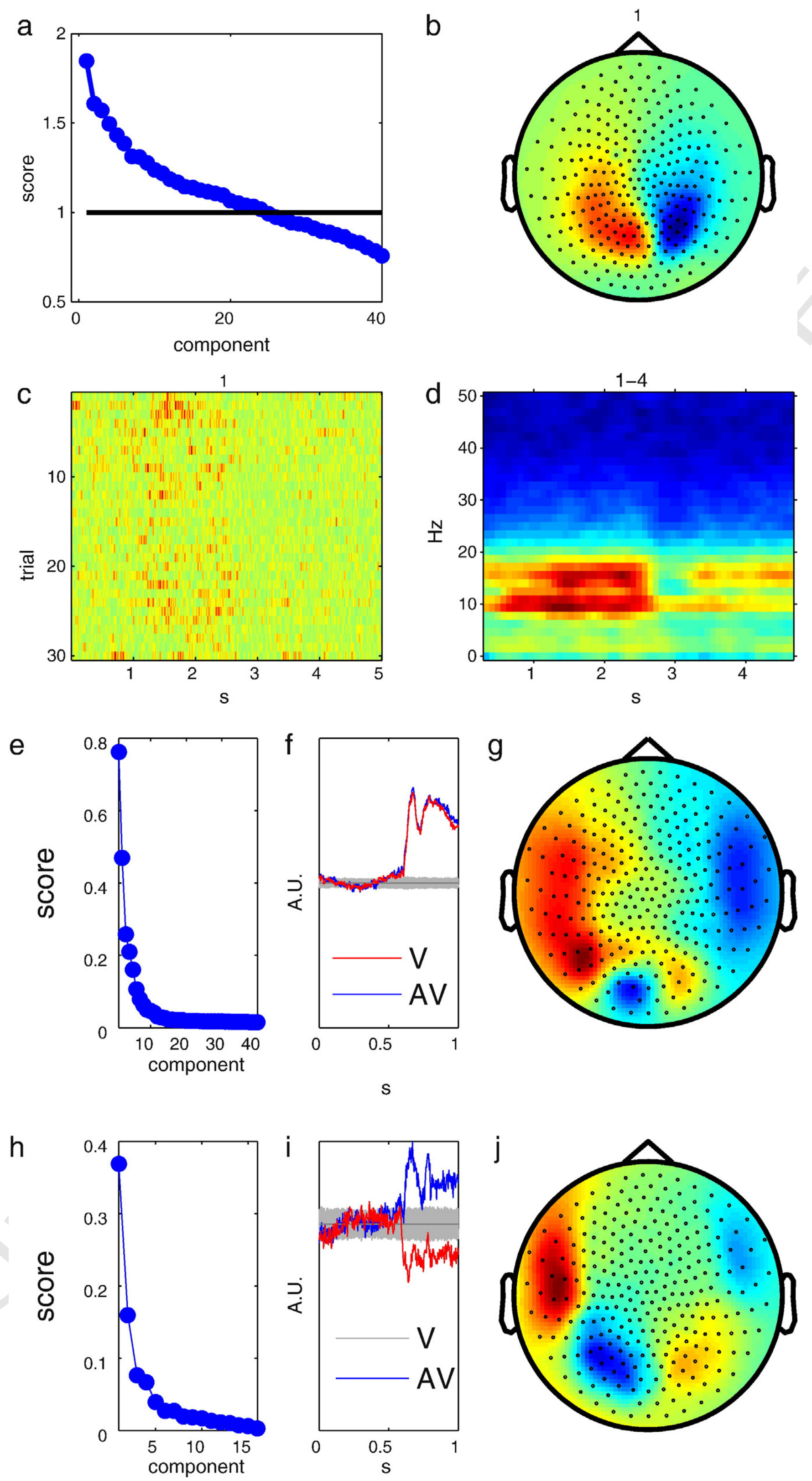


Fig. 5. Isolating interval-specific responses. (a–d) Event-related desynchronization (ERD). (a) Ratio of power in the 0–2.5 s interval relative to the 2.5–5 s interval for the first 40 JD components. (b) Topography of the first component. (c) Raster plot of individual trials for the first component, showing a drop in power after approximately 2.5 s for most trials. (d) Spectrogram of the first four JD components averaged over trials. (e–g) Stimulus-evoked response to repeated visual or audio-visual stimuli. (e) Power-ratio score of the first 40 components. (f) Time course of the first component in response to a visual (red) or audio-visual (blue) stimulus. (g) Topography of the first component. (h–j) Components that differ between visual and audio-visual stimulation. (h) Power-ratio score of all components. (i) Time course of first component in response to visual (red) stimulation and audio-visual (blue) stimulation. (j) Topography of the first component.

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

JD was first applied to isolate a subspace of component signals that responded reproducibly to both stimuli (V and AV). Matrix $\mathbf{C}_0$ was the covariance matrix of the entire data, and matrix $\mathbf{C}_1$ the sum of covariance matrices of trial-averaged data for the V and AV conditions. Fig. 5e shows the power-ratio score for the first 40 components, and Figs. 5f and g show the time course and topography of the first component, respectively. The time course of this component is very similar for V and AV (compare red and blue in Fig. 5f), and the same was true for subsequent components (not shown). There was no obvious sign of an auditory response in any of these components.

In a second stage, JD was applied to a selected subset of components ($K = 16$) from the first stage, using as matrix $\mathbf{C}_0$ the covariance matrix of this subset, and as matrix $\mathbf{C}_1$ the covariance matrix of the difference between averages over trials for the V and AV conditions. Fig. 5h shows the power-ratio score, and Figs. 5i and j show the time course and topography of the first component, respectively. The time course of this component differs clearly between V and AV (compare red and blue in Fig. 5f), and its dipolar topography is consistent with activity in the auditory cortex. Without this two-stage analysis this activity would have been invisible. This example shows how JD can extract an extremely weak source of condition-specific, stimulus-evoked activity from a competing background.

Additional examples

These examples involve a wider range of data types and tasks. A first additional example involves electrocorticogram (ECoG) data recorded from a 128-channel surface array on the cortex of a monkey (NeuroTycho project, http://www.neurotycho.org/), at the transition between awake and anesthetized state. The processing goal is to characterize brain activity affected by anesthesia. After dimensionality reduction ($N = 22$), JD was used to contrast the power after injection relative to the power before injection (as in the ERD example above). Fig. 6a shows the post/pre power ratio for each component (bottom left), together with the power of each component as a function of time (top), and the RMS (root mean square) of the topographies associated

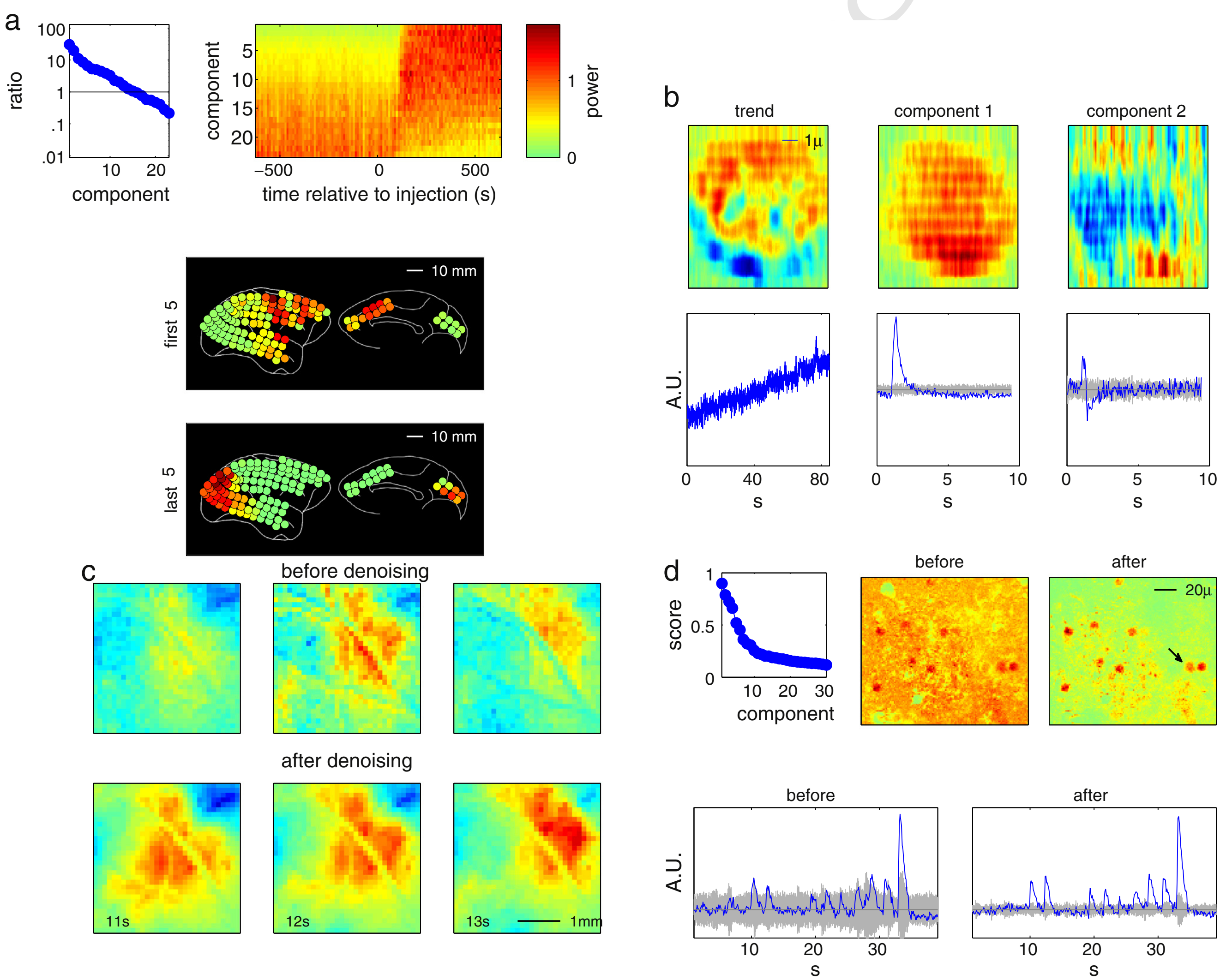


Fig. 6. ECoG (electrocorticography) and optical imaging data. (a) ECoG of monkey showing the effect of injecting a dose of anesthetic. The upper left plot shows the power-ratio of the post-injection interval relative to the pre-injection interval. The upper right plot shows the time course of the power of individual components, coded as color. The lower plots show the RMS average of the topographies associated with the first 5 components (most active after injection) and last 5 components (most active before injection). (b) Two-photon calcium imaging of the base of a cochlear hair cell. Bottom left: time course of a linear trend component. Top left: associated topography (the hair cell fills most of the plot). Bottom middle and right: stimulus-evoked responses averaged over repetitions (blue) and 2 s.d. of a bootstrap resampling of the mean (gray band). Top middle and right: associated topographies. (c) Intrinsic optical imaging in ferret auditory cortex in response to tone sweeps. Data were processed to suppress non-repeatable noise (e.g. speckle and blood-flow related), revealing a gradual shift of activity from lower left to upper right. (d) Two-photon calcium imaging of mouse auditory cortex. Top: power-ratio score (left), topography of the RMS power before and after denoising (middle and right). Bottom: time course of one individual neuron (arrow in top right) before (left) and after (right) denoising. Blue is the mean over 10 trials, gray is ±2 standard deviations of a bootstrap resampling of the mean.

with the first 5 and last 5 components (bottom right). Brain activity is radically changed by anesthesia: components active in the awake state are shut down, whereas hitherto silent components become active. Few components maintain a constant level of activity throughout the recording.

In a second example the aim was to improve the SNR of calcium signals recorded using two-photon microscopy in a mouse cochlear inner hair cell (Culley and Ashmore, 2010, in preparation). A fluorescent probe was introduced through a patch pipette that was also used to depolarize the cell for 100 ms, opening channels in the cell membrane to increase the intracellular calcium. This was repeated 9 times. JD was applied twice in succession, each time with a different bias filter. First, a linear trend was isolated using a bias filter that emphasized the difference between trial means and global mean. The topography and time course of the first component are plotted in Fig. 6b, left. This component was then projected out of the data, to reduce the dominance of the linear trend, and JD was applied again, this time to extract the stimulus-evoked activity. The time course and topography of the first two components are plotted in Fig. 6b, center and right. These patterns suggest a gradual change in calcium level gradient across the cell, superimposed on a phasic response to stimulation. The presence of more than one reproducible component suggests that the stimulus-evoked response was not perfectly synchronous across the imaging field.

The third example involves intrinsic optical imaging of the auditory cortex of a ferret in response to pure tone sweeps (Nelken et al., 2008). Each sweep (100 to 3200 Hz within 14 s) was repeated nine times. JD was used to find linear combinations of pixel time series that were most repeatable across repetitions. The four most repeatable JD components were projected back to form "clean" data. Fig. 6c shows responses sampled during the last 4 s of the sweep before (upper row) and after (lower row) denoising. Here, non-repeatable components, such as bloodflow-related, are attenuated making more salient the gradual shift of activity across the cortex (from upper left to lower right).

The fourth example involves two-photon calcium imaging of the auditory cortex of mouse in response to repeated stimulation by a sequence of 17 pure tone pips of different frequencies (Winkowski and Kanold, 2013). JD was used to suppress non-reproducible activity. Fig. 6d, top shows the power-ratio score (left) and the topography of activity before denoising (center) and after denoising (right). Fig. 6d, bottom shows the time course of the activity of one neuron (arrow in top right) before denoising (left) and after denoising (right). The mean (blue) is similar before and after denoising, but the variability of this estimate (gray band) is greatly reduced. See Appendix 6 for more details on this and the other examples. These examples illustrate the flexibility of JD as a tool to clean and analyze multichannel electrophysiological data.

How does it work?

JD finds a set of weights to apply to sensors, electrodes, pixels, etc. that (a) suppresses the most prominent noise sources, and (b) preserves the activity of interest. This is similar to the principle of a beamformer. The weights are chosen such that the contribution of each noise source i is balanced out (the sum of the products of mixing weights v_{ij} and unmixing weights w_{jk} is zero, $\sum_j v_{ij} w_{jk} = 0$). The algorithm tries to find a set of weights such that this is satisfied for all noise sources, i while preserving the target source $i' : \sum_j v_{i'j} w_{jk} \neq 0$. JD can be understood as an efficient way to *search* within the *JK*-dimensional space of weights to find this solution.

As illustrated in Fig. 1(b), the key step of spatial whitening (decorrelation followed by normalization) removes all influence of variance, so that the data set has no preferred direction in *J*-dimensional space. The bias filter breaks the spherical symmetry, boosting the variance in the direction of the signal of interest, while shrinking variance in irrelevant noise directions. The final PCA aligns these directions with the component axes. The combination of spatial whitening and final PCA produces a linear transformation that increases the signal-to-noise ratio, where "signal" and "noise" are defined by the filtering operation. As a counterpart of optimizing the desired feature, another activity is minimized, and in this sense JD is a method to *denoise* the data.

Another way to conceptualize the effect of JD is to note that diagonalization of the data covariance matrix $\mathbf{C}_0$ defines a transform that allows the total power (variance) to be neatly "packaged" as a sum of powers (variances) of individual components, the cross correlation terms being zero. *Joint* diagonalization of $\mathbf{C}_0$ and $\mathbf{C}_1$ implies that the same packaging is valid for both the raw and the filtered data sets. Any difference in power between raw and "filtered" (for example a source active in one time interval but not the other) appears as a step in the power of a component, as in Fig. 6a where monkey ECoG activity is expressed as a sum of components that either turn on, or turn off, under anesthesia.

Each component is defined by a vector of weights (column of matrix $\mathbf{W}$, see Methods), and is associated with a *time series* (weighted sum of sensor signals). Cross-correlation between the component time series and the raw sensor waveforms yields another vector, of same size as the weights, that can be understood as a spatial pattern or *topography* (e.g. Fig. 3b). This pattern is an estimate of the amount of power accounted for by the component at each sensor. It is distinct from the pattern of weights, and usually more informative (Haufe et al., 2014).

Who invented it?

Fukunaga and Koontz suggested the present 2-step approach to joint diagonalization as a method to identify the difference in the spectrum of two signals. A similar generalized eigenvalue problem arose earlier already in the context of linear discrimination (Fisher, 1936; Rao, 1948). The concept of simultaneous diagonalization is well known in the context of commuting matrices going back to Frobenius in 1878 (Drazin, 1951). Simultaneous diagonalization of two covariance matrices, as discussed in the present paper, is the basis for the Common Spatial Pattern (CSP) method of Koles et al. (1990) that is popular in the Brain Computer Interface (BCI) literature (Blankertz et al., 2008; Dornhege et al., 2006; Lemm et al., 2011; Tangermann et al., 2011), and also appears repeatedly in the context of blind source separation and ICA (reviewed in Parra and Sajda, 2003).

The Denoising Source Separation method of Särelä and Valpola (2005), in its linear form, can be thought of as a generalization of CSP, and of a number of other source separation techniques that exploit temporal properties of the signals (Amari, 2000; Belouchrani et al., 1997; Blaschke et al., 2006; Cardoso, 2001; Molgedey and Schuster, 1994; Parra and Spence, 2000; Ziehe and Müller, 1998).

The contribution of the present paper is to emphasize the usefulness of the basic principle (diagonalization of raw and filtered covariance matrices) as a tool to perform a range of common tasks. The roots and relations between methods are further discussed in Appendix 3.

Overfitting and circularity

A basic weakness of JD, shared also by other techniques such as PCA, ICA, beam-forming, and clustering, is that the analysis is *data-dependent*: the matrix used to analyze the data depends on the data themselves. In the present case, JD selects a linear combination of sensors (i.e. one direction within the *J*-dimensional space) that maximizes a given optimality criterion. This is akin to data selection. The outcome of the analysis may then falsely appear to confirm the hypothesis that motivated the analysis, a problem known as circularity (Kriegeskorte et al., 2009). Over-fitting is most severe when the number of free parameters is large relative to the number of data that constrain them, magnifying random patterns and producing seemingly salient effects that are purely artifacts (as in Study A below). One must be alert

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

to this possibility and check whether effects observed are robust, for example using cross-validation or resampling techniques (Hyvarinen, 2012; Meinecke et al., 2002). As an example, the analysis of Figs. 3a–c was repeated 1000 times with surrogate data obtained by excising "trial" epochs at random positions within the MEG data. The 5–95% interval of the power-ratio statistic is plotted as a gray band in Fig. 3a. The power ratio values obtained for the real data are well outside of this range, giving us confidence that the pattern extracted by JD is real and not due to overfitting (which is manifest as the upward turn of the gray band near the left axis).

Other caveats and cautions

It is tempting to attribute JD components to individual neural sources, in the spirit of the blind source separation paradigm that motivates ICA. As noted earlier this is unlikely to be valid, if only because a small number of sensors cannot possibly resolve the many concurrent sources within the brain. In addition, the components obtained are mutually uncorrelated, whereas parts of the brain that work together are likely to have correlated activities. Rather, the best that we can say is that any subset of selected components defines a *subspace* of the data in which the activity of interest is concentrated.

Failure scenarios

The following examples are imaginary but based on real situations. The aim is to give hints as to what might go wrong. They are *not* a complete catalog. Appendix 7 contains more details including figures illustrating these effects.

Study A recorded cortical responses using a 440-channel MEG system. The data were low-pass filtered at 20 Hz, and organized into epochs. Unbeknownst to the experimenter, stimulation failed so there should have been *no* reproducible response. Nonetheless, when JD was applied to emphasize activity reproducible over epochs, a clear pattern emerged. What happened? The answer is: over-fitting. 440 free parameters were available to define each JD component, and the degrees of freedom available to constrain them were too few, in particular as lowpass filtering increases the serial correlation between samples. How to diagnose the problem? There are many techniques to test for overfitting. For example, repeat the analysis on a randomized version of the data (time markers are randomly shifted) so that reproducibility of a stimulus is not expected, and take the level of activity seen with such random data as an indication of chance performance. How to fix the problem? Apply PCA to reduce the dimensionality. Increase the number of trials. Consider removing lowpass filtering. And of course: check the stimulation.

Study B recorded responses to 100 repetitions of a stimulus. JD was applied in the hope of reinforcing the evoked response relative to strong 50 Hz power line noise. Unexpectedly the first few JD components contained *mainly 50 Hz and harmonics*. What happened? The experimenter made the mistake of presenting stimuli with inter-stimulus intervals that were all multiples of 1/50 Hz (20 ms). As a result, the 50 Hz activity was reproducible across trials, leading it to occupy the first JD components. How to fix? Make sure that stimulus presentation is incoherent with repeatable noise sources such as 50 Hz, heartbeat, and alpha activity. If data are already collected, use JD to isolate components dominated by 50 Hz and harmonics and project them out (as in example 1 above), prior to the main JD analysis. Or filter the data with a 20 ms boxcar window (to suppress 50 Hz and all harmonics), or a notch filter.

Study C investigated a weak source activity time-locked to the stimulus. JD was applied to enhance it, but unexpectedly the best JD component was strongly affected by a noise source that did not seem particularly reproducible across trials. What happened? The target and noise happened to be *collinear* in the data, so that any transformation that selected one necessarily selected the other. How to fix? One way is to increase the number of sensors or electrodes so as to increase the dimensionality of the observations. Another is to try advanced techniques such as TSDSS (see Appendix 5).

Study D used EEG to probe stimulus-evoked activity. A slow drift was superimposed on the data producing relatively large DC offsets within some trials. To attenuate these offsets, the experimenter removed means from all trials. Unexpectedly, the first JD component appeared to be *superimposed on a ramp*. What happened? Removing the mean on each trial transformed the slow drift into a reproducible ramp pattern, that JD then enhanced, superimposing it on the genuine evoked response. How to fix? Do *not* remove the mean from each trial. It is not a good idea to remove trends trial by trial, be they constant, linear, or polynomial. Instead fit a polynomial to the data before cutting into trials, and subtract the fit. Another option is to use JD in a preprocessing stage to project out the slow drift, or else apply high-pass filtering prior to analysis.

Study E looked for 10 Hz oscillatory activity within an in-vitro preparation. Data were recorded with an electrode array, and JD was applied using a bandpass bias filter centered on 10 Hz. Bursts of 10 Hz oscillation were indeed found. However the experimenter also tried other bias filter frequencies, and found oscillations at those frequencies too, suggesting that something was wrong. What happened? Actually, the activity was not oscillatory but *propagatory*, consisting of bursts of activity that activated different electrodes in sequence. However, given the objective of emphasizing oscillatory activity, JD produced a *grid-shaped pattern of weights*, and the propagation of the bursts over this pattern produced the apparently oscillatory response. How to fix? There is no easy way to rule out this sort of artifact, but projecting the data back to sensor space should reveal the propagatory phenomenon. The experimenter must be attentive and question every effect found.

Study F searched for neural substrates differentially activated by two tasks. JD was applied to find the most discriminative linear combinations of channels. Unexpectedly, the first few components mainly contained small glitches or eye-related activity. What happened? JD is sensitive to any difference in variance. A glitch may be small, but if it only occurs in one interval and not the other it may take precedence over genuine activity. How to fix? One solution is to identify these artifactual components and project them out, prior to JD analysis. Another is to remove channels affected by glitches, or to apply temporal weighting to exclude the glitch intervals from the analysis. A third is to reduce the dimensionality of the data with PCA, so as to remove dimensions with low power, often dominated by glitches.

Study G applied JD to find activity time-locked to a repeated stimulus. Two highly repeatable components were indeed found, implying that response reflected at least two distinct neural sources, with distinct topographies and time courses. However, their shape was not consistent across subjects. The topographies did not fit the dipolar pattern expected of a single source, and their time-courses were also more complex than expected. What happened? JD recovers components that span the same subspace as the measurable stimulus-locked activity, but there is no guarantee that the components match neural sources, rather than being linear combinations of them. How to fix? Various techniques such as ICA, sparse component analysis, or canonical correlation analysis, may be useful to find meaningful directions within the selected subspace. These are beyond the scope of this paper.

A general tool for data analysis?

Many analysis techniques are available, often in multiple flavors, which is an obstacle when searching for a tool to perform a specific task. Trying out new tools is time consuming, and JD is no exception, but hopefully the investment is recouped over a range of tasks. JD can be used to enhance activity of interest, or to isolate unwanted activity and project it out of the data. It can be used repeatedly on the same data with different bias filters (de Cheveigné et al., 2012), to probe the data for different response characteristics, or in steps to isolate and

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

remove sources in succession. It is deterministic (whereas some ICA methods offer different solutions on different trials), it produces components in a well-defined order, and its computational cost is relatively low, so it can be applied to the large data sets typical of EEG or MEG. Finally, it is easy-to-understand, and gives insight into more sophisticated methods.

In summary

The JD algorithm addresses a variety of needs that arise in the analysis of multichannel electrophysiological data. Attractive features are (a) the algorithm is easy to understand, (b) processing is simple and efficient, (c) the method is flexible and can be reused for different tasks, and (d) the result is good.

Acknowledgments

Thanks to Kate Molloy, Nilli Lavie and Maria Chait for the MEG data in Figs. 5e–j. Thanks to Gareth Barnes for permission to use the MEG data from the study of Duncan et al. (2009), to Israel Nelken, Jan Schnupp and Andrew King for permission to use the optical imaging from Nelken et al. (2008), to Naotaka Fujii and Toru Yanagawa for making their monkey ECoG data publicly available (http://neurotycho.org), to Jonathan Ashmore and Siân Culley for the mouse cochlear hair cell two-photon imaging data, and to Patrick Kanold and Dan Winkowski for permission to use their mouse auditory cortex two-photon imaging data. Thanks to Maria Chait and Israel Nelken for comments on the manuscript. Thanks also to the many people who provided data sets on which the methods were developed, and to the authors of the JADE and FASTICA toolboxes for making their code available. Thanks to the reviewers of previous versions of this manuscript for their extensive comments.

Appendix 1. Precise description of JD

Given the matrix of observation signals $\mathbf{X}$, with dimensions $T \times J$, the first PCA matrix $\mathbf{P}$, with dimensions $J \times J$, is obtained by eigen-decomposition of the covariance matrix[1]:

$$\mathbf{C}_0 = \mathbf{X}^\top \mathbf{X}. \tag{2}$$

The eigen-decomposition of this matrix is given by:

$$\mathbf{C}_0 \mathbf{P} = \mathbf{P}\mathbf{D}, \tag{3}$$

where the columns of matrix $\mathbf{P}$ are the orthonormal eigenvectors and the diagonal matrix $\mathbf{D}$ holds the corresponding eigenvalues. Each eigenvalue represents the power (variance) of the data along the direction determined by the associated eigenvector. Setting $\mathbf{N} = \mathbf{D}^{-1/2}$, "sphered" signals are obtained by rotating and dividing each dimension by that scale:

$$\mathbf{Z} = \mathbf{X}\mathbf{P}\mathbf{N}. \tag{4}$$

This data matrix $\mathbf{Z}$ again has dimensions $T \times J$, but its covariance matrix is given by the identity matrix $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}$, i.e. the data are uncorrelated and have unit power (variance) in all dimensions. Next, we apply the bias filter $\mathbf{L}$ to $\mathbf{Z}$. By "bias filter" we mean here any linear transformation on the time domain:

$$\overline{\mathbf{Z}} = \mathbf{L}\mathbf{Z}, \tag{5}$$

where $\mathbf{L}$ is a matrix of dimensions T' by T. Importantly, this filter enhances the signal and suppresses noise. The covariance of the filtered data is:

$$\mathbf{C}_1 = \overline{\mathbf{Z}}^\top \overline{\mathbf{Z}}, \tag{6}$$

and its eigen-decomposition gives us the second rotation matrix $\mathbf{Q}$:

$$\mathbf{C}_1 \mathbf{Q} = \mathbf{Q}\mathbf{D}_{\overline{Z}}. \tag{7}$$

The rotation $\mathbf{Q}$ aligns the main axes of the bias-filtered data with the final components:

$$\overline{\mathbf{Y}} = \overline{\mathbf{Z}}\mathbf{Q}, \tag{8}$$

that are uncorrelated and ordered by decreasing variance. Once matrices $\mathbf{P}$, $\mathbf{N}$ and $\mathbf{Q}$ have been obtained, the same sequence of transformations can be applied also to the raw data without filtering:

$$\mathbf{Y} = \mathbf{X}\mathbf{P}\mathbf{N}\mathbf{Q}, \tag{9}$$

giving Eq. (1) in the main text with

$$\mathbf{W} = \mathbf{P}\mathbf{N}\mathbf{Q}. \tag{10}$$

We note that both the bias-filtered data $\overline{\mathbf{Y}} = \mathbf{L}\mathbf{X}\mathbf{W}$, and the unfiltered data $\mathbf{Y} = \mathbf{X}\mathbf{W}$ have now a diagonal covariance matrix, i.e. the time courses of these components (columns of $\mathbf{Y}$ and $\overline{\mathbf{Y}}$) are uncorrelated for both filtered and unfiltered data.

Appendix 2. The bias filter

We call bias filter any operation that can be performed by combining samples of a signal in time, the same operation being performed on all channels, and independently for each channel. With this definition, bias filtering is implemented by left-multiplying the data matrix with a matrix $\mathbf{L}$ as in Eq. (5).

Fig. 7 shows three examples of bias-filter matrix similar to those used in the examples. In Example 2 of the main text (stimulus-evoked response), the filtering operation consisted simply of averaging over trials. This is formalized as left-multiplication by a matrix $\mathbf{L}$ made by horizontal concatenation of n identity matrices of size $T' \times T'$ where T' is the length of an epoch and n is the number of trials, analogous to that shown in Fig. 7a. In the monkey ECoG example (effects of anesthesia), the filtering operation is formalized as a matrix $\mathbf{L}$ analogous to that shown in Fig. 7b, of size $T' \times T$ where T' is the length of the interval preceding the injection, and T that of the full data set (this is called "on/off-denoising" in Särelä and Valpola, 2005, or "maximum power-ratio" in Parra et al., 2005). In Example 4 of the main text (narrowband cortical activity), the narrowband filter centered on 10 Hz is formalized as left-multiplication by a matrix $\mathbf{L}$ of Toeplitz structure similar to that of Fig. 7c (referred to as "denoising based on frequency content" in Särelä and Valpola, 2005).

Other linear operations on the time/trial axis can be envisioned. Probably the earliest example is the blind source separation algorithm by Molgedey and Schuster (1994) which is recovered here if $\mathbf{L}$ implements a time delay. Another is slow-feature analysis in which $\mathbf{L}$ implements the temporal derivative and the goal is to find the components with the smallest derivatives (the "slow" components) (Blaschke et al., 2006; Wiskott and Sejnowski, 2002). In addition to linear filters, the DSS algorithm of Särelä and Valpola (2005) allows for non-linear filtering operations. However, the discussion in this paper is restricted to linear filtering only. Linear filters have a number of advantages over non-linear filtering: 1) they lead to the close-form solutions presented above, 2) the resultant algorithm can be implemented in a few lines of code using standard eigen-decomposition routines, 3) they allow us to make links to closely related classic signal analysis techniques, 4) they

[1] Note that we have not subtracted the mean so this is not strictly speaking a covariance matrix. But the subsequent discussion applies equally to the covariance matrix calculated after subtracting the mean.

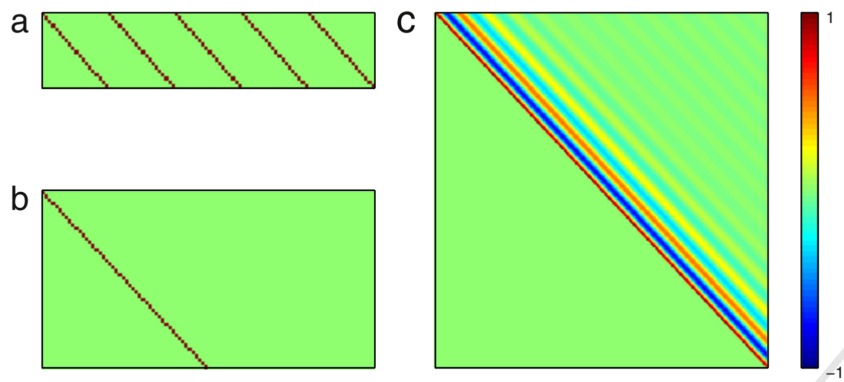


Fig. 7. Examples of matrix $\mathbf{L}$ corresponding to the three types of bias filter used in the examples. (a) Average over trials (here there are $n = 5$ trials). (b) Selection of a temporal interval (here the first half of the time axis is selected). (c) Bandpass filter (2nd-order resonator).

provide a simple geometric interpretation (Fig. 1), and finally 5) they allow us to prove optimality in terms of signal-to-noise ratio as we will demonstrate next.

Appendix 3. Roots of the approach, optimality

A similar two-step procedure for diagonalizing two covariance matrices was described by Fukunaga and Koontz in 1970 in the context of diagonalizing two correlation matrices $\mathbf{C}_0$ and $\mathbf{C}_1$ (Fukunaga, 1972, 1990; Fukunaga and Koontz, 1970). Their goal was to identify linear transformations that best distinguish between two signals characterized by their respective auto-correlation matrices. The same problem of finding the best linear subspace to distinguish between two classes was addressed by Fisher in 1936 (Fisher, 1936), and later extended by Rao to the multi-class problem in 1948 (Rao, 1948). Rao's approach is now known as Fisher Linear Discriminant (Duda et al., 2012). In their case $\mathbf{C}_0$ and $\mathbf{C}_1$ represent the between- and within-class covariances. Rao proposed that the eigenvectors of $\mathbf{C}_0^{-1}\mathbf{C}_1$ with the highest eigenvalues span a space that best separates these classes:

$$\mathbf{C}_0^{-1}\mathbf{C}_1\mathbf{W} = \mathbf{W}\mathbf{D}. \tag{11}$$

These directions maximize the differences between classes, let's call it “signal” variance, relative to the “noise” variance within each class. Quantitatively this is captured by the determinant ratio (see criterion (13) below). It is interesting to note that this $\mathbf{W}$ also diagonalizes both covariance matrices $\mathbf{C}_0$ and $\mathbf{C}_1$ individually (Fukunaga, 1972, 1990). While originally intended for within- and between-class covariance matrices, mathematically, the approach of Fukunaga–Koontz for diagonalizing two correlation matrices gives the same answer as the one-step solution using Eq. (11) (Fukunaga, 1972, 1990). What is perhaps even more intriguing is that the condition of simultaneous diagonalization, which is solved by this eigenvalue problem, reoccurs for a number of source separation problems. In source separation the first matrix often corresponds to the correlation matrix of the raw data $\mathbf{C}_0 = \mathbf{X}^\top\mathbf{X}$, as in the present case. The second matrix can take on different forms, depending on the assumptions made about the sources (non-Gaussianity, non-stationarity, non-whiteness) (Parra and Sajda, 2003). For the case of JD discussed here $\mathbf{C}_1$ corresponds to the covariance of the bias-filtered signal $\mathbf{C}_1 = \mathbf{X}^\top\mathbf{L}^\top\mathbf{L}\mathbf{X}$. The resulting $\mathbf{W}$ from Eq. (11) is identical to the solution of the two-step procedure (Eq. (10)), provided the arbitrary scaling of $\mathbf{W}$ is chosen to sphere $\mathbf{C}_0$ (see Fukunaga, 1972, 1990, Chapter 2, albeit in the context of classification and not source separation).

What is so special about the directions of the eigenvectors defined by these two symmetric matrices? As it turns out, these directions are optimal in a number of important ways, namely, the eigenvectors with the K largest eigenvalues ($K < J$) span the K-dimensional subspace with the maximum determinant-ratio as well as the maximum trace-ratio (Fukunaga, 1972, 1990, Chapter 10):

$$\mathbf{W} = \underset{\mathbf{W}\in\Re^{[J\times K]}}{arg\ max} \frac{|\mathbf{W}^\top\mathbf{C}_1\mathbf{W}|}{|\mathbf{W}^\top\mathbf{C}_0\mathbf{W}|} \tag{12}$$

$$= \underset{\mathbf{W}\in\Re^{[J\times K]}}{arg\ max}\, Tr\left\{\left(\mathbf{W}^\top\mathbf{C}_0\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{C}_1\mathbf{W}\right\}. \tag{13}$$

Importantly for the present case, from this follows that the top K eigenvectors maximize the summed power-ratio of the bias-filtered versus unfiltered component, if we add as a constraint that the components are uncorrelated in time:

$$\mathbf{W} = \underset{\mathbf{W}\in\Re^{[J\times K]},\text{s.c.}\mathbf{C}_y=\text{diag}}{arg\ max} \sum_{i=1}^{K}\frac{\sigma^2_{\bar{y}_i}}{\sigma^2_{y_i}}, \tag{14}$$

where $\sigma^2_{y_i}$ and $\sigma^2_{\bar{y}_i}$ are the power of the ith component for the raw and filtered versions of the data, i.e. the diagonal terms of the two covariance matrices. This finding is true for any $K < J$, in particular for $K = 1$, meaning that the first component has the largest possible power ratio (a criterion already proposed in Parra et al., 2005). The second component is uncorrelated from the first and, within that constraint, it captures again the largest power ratio, the third is uncorrelated from the first two and captures the next highest power ratio, and so on until finally the Jth component captures the smallest remaining power ratio. This means that the components extracted by JD are sorted by the power (variance) of the filtered signal relative to the raw data. Assuming that filtering enhances the signal of interest and attenuates uncorrelated noise, this implies that the eigenvectors capture uncorrelated components of the signal ordered by signal-to-noise ratio.

In fact, under the following set of assumptions the components can be shown to maximize signal to noise ratio. Assume that the observations represent the signal plus some additive uncorrelated noise, $\mathbf{X} = \mathbf{S} + \mathbf{N}$, so that the covariances are additive:

$$\mathbf{C}_0 = \mathbf{R}_X = \mathbf{S}^\top\mathbf{S} + \mathbf{N}^\top\mathbf{N} = \mathbf{R}_S + \mathbf{R}_N. \tag{15}$$

Assume in addition that filter $\mathbf{L}$ attenuates the noise with gain g_N and enhances the signal with gain g_S but leaves the correlation structure of each unchanged, $\mathbf{R}_{\bar{S}} = g_S^2\mathbf{R}_S$, $\mathbf{R}_{\bar{N}} = g_N^2\mathbf{R}_N$. Then:

$$\mathbf{C}_1 = \mathbf{R}_{\bar{X}} = \mathbf{S}^\top\mathbf{L}^\top\mathbf{L}\mathbf{S} + \mathbf{N}^\top\mathbf{L}^\top\mathbf{L}\mathbf{N} = g_S^2\mathbf{R}_S + g_N^2\mathbf{R}_N. \tag{16}$$

A matrix $\mathbf{W}$ that diagonalizes two symmetric matrices, say $\mathbf{R}_S$ and $\mathbf{R}_N$, also diagonalizes any linear combination of the two, in particular $\mathbf{R}_{\bar{X}}$ and

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

$\mathbf{R}_X$.[2] This means that solutions to the eigenvalue (Eq. (11)) with $\mathbf{C}_0 = \mathbf{R}_X$ and $\mathbf{C}_1 = \mathbf{R}_{\overline{X}}$ are also solutions to the same eigenvalue equation with $\mathbf{C}_0 = \mathbf{R}_N$ and $\mathbf{C}_0 = \mathbf{R}_S$. The order of eigenvalues is the same provided that $g_S > g_N$ (this can be shown using a similar argument as in Fukunaga, 1972, 1990, Chapter 2). Thus, the same projections of the data that maximize the power-ratio between filtered versus unfiltered signal – as in Eq. (13) – also maximize the power ratio between signal and noise. In short, JD maximizes SNR. The key assumption for this to hold is that the triplet (filter/signal/noise) satisfies conditions (15) and (16). Note that the filter does not have to be perfect at suppressing noise. Optimal SNR is achieved as long as the signal-gain is larger than the noise-gain. To our knowledge, this optimality had not been previously recognized.

Under which conditions is Eq. (16) satisfied? For the case that the bias filter implements trial averaging (Fig. 7a) Eq. (16) is satisfied if the reproducibility of the different signal components is the same, i.e. all signals of interest have the same level of variability across trials. For the case of a bias filter that defines the signal of interest by selecting a specific time-interval (Fig. 7b) all that is required is that noise components are (second-order) stationary across the different time intervals. Finally, for a shift-invariant temporal bias filter (Fig. 7c) this condition is satisfied if all signal components experience the same gain g_S and all noise components gain g_N. This does not necessarily require perfect separation in the frequency domain between the signal of interest and the noise — it suffices for the different signal components to have the same spectral content, and similarly for the noise to be spectrally the same across components.

Appendix 4. How to use JD repeatedly (deflation)

Removing components and projecting back into sensor space

In the main text we state that a subset of components was "projected out" of the data, or instead "projected back" into sensor space. What is meant is that the original data are replaced by a version that does not contain any activity correlated with the components that are removed.

This can also be understood in terms of subspaces of the vector space formed by all linear combinations of the J sensor signals. That space is of dimension at most J (it can be less if sensor signals are linearly dependent, in particular if $T < J$). The JD components form an orthogonal basis of that space. A subset of K components defines a *subspace*, orthogonal to the subspace spanned by the J–K remaining components. "Removing" the K components is the same as projecting the data on the orthogonal subspace.

A simple way to accomplish this is with the following operation:

$$\hat{\mathbf{X}} = \mathbf{XWEW}^{-1}, \tag{17}$$

where the diagonal matrix $\mathbf{E}$ has 0 for all the components that are to be removed and 1 for all that are to be preserved. In the case that dimensions have been omitted in the first PCA and $\mathbf{W}$ is rectangular the inverse here refers to the pseudo-inverse.

Deflation, dimensionality reduction

JD can be applied repeatedly to the same data set, projecting out selected components at each step (deflation). The rank of the data is reduced at each step. JD handles rank-deficient data in the initial PCA by removing eigenvectors with eigenvalues smaller than a threshold, so there is no problem with applying it repeatedly in this way.

The eigenvalue procedure of Eq. (11) is often considered inadequate in practice because it is very sensitive to estimation errors in the correlation matrices $\mathbf{C}_0$ and $\mathbf{C}_1$. Of particular concern is the inverse of $\mathbf{C}_0$, which may be dominated by very small noise contributions within the null space of the signal of interest. This problem may become more severe as the number of sensors increases and the activity of near-by sensors becomes strongly correlated. A simple and classic solution is to remove dimensions that carry little power in the data, i.e. remove all directions with a small eigenvalue of $\mathbf{C}_0$. In the two-step procedure of simultaneous diagonalization this is done by using only the eigenvectors with the largest eigenvalues of $\mathbf{C}_0$ in Eq. (4). If we keep $K < J$ dimensions, this means that $\mathbf{P}$ is of size $J \times K$, and that $\mathbf{Z}$ and $\mathbf{Y}$ are of size $T \times K$ and $\mathbf{W}$ of dimension $J \times K$, i.e. there are now only K components. This addresses the issue of the sensitivity of the null space of $\mathbf{C}_0$ to small amounts of noise. However this is assuming that activity of interest resides within the subspace spanned by the dimensions retained, which might not be the case if its variance is small.

Multiple-step JD

In several examples, JD was applied twice with different bias filters. At each step, JD optimizes the criterion at hand, and therefore one might expect that the outcome depends only on the second bias filter. What, then, is the advantage of the initial step? The first step allows the data to be projected to a smaller subspace, selected according to the first bias filter. The second step then finds an optimal solution according to the second filter within this subspace. The second JD operates in a smaller space and is less prone to over-fitting, and the solution thus favors the properties enforced by both filters.

Appendix 5. Relation to other methods

Extensions

Time shifts applied to the data allow JD solutions to implement multichannel finite impulse response (FIR) filters that can separate sources in the spatio-temporal domain. This is the idea behind the Time Shift DSS (TSDSS) method (de Cheveigné, 2010, see also Blankertz et al., 2008; Dornhege et al., 2006). A source that is not spatially separable from noise may nonetheless be resolved if the spectral characteristics (e.g. latency) of source and/or noise differ between sensors. In place of time shifts, other convolutional transforms can be used, for example a filter bank, and indeed the whole operation may be performed in the frequency domain. The one-channel case is that initially addressed by Fukunaga and Koontz (1970). *Cross-products* between channels allow JD to operate within the space of quadratic forms of the signals. This is the basis of the Quadratic Component Analysis (QCA) method (de Cheveigné, 2012) that finds components with power that obeys some criterion, for example repeatability across stimulus trials (such activity is referred to as *induced*, repeatable in power, as opposed to *evoked*, repeatable in both power and phase). In this paper we only considered linear bias filtering operations, for which a solution is found in a single step. The authors of Särelä and Valpola (2005) consider a wider range of operations for which the solution is found iteratively.

ICA

How does JD differ from other source separation techniques such as ICA? Conceptually, the difference is in the rule to calculate the matrix $\mathbf{C}_1$. If the data are indeed a mixture of J independent sources, then all choices of $\mathbf{C}_1$ should provide the identical answer (Parra and Sajda, 2003). In practice, however, the assumption that the data $\mathbf{X}$ is generated by J and only J independent sources is rarely correct, and as only a limited sample of data is observed the parameter estimates are imperfect. Thus, different techniques will provide different answers.

ICA is usually defined devoid of any temporal context, i.e. an ICA algorithm should give identical answers when applied to the same data but with samples that are scrambled in time. Thus, the algorithms

[2] The set of matrices that can be diagonalized by a single matrix forms a toral Lie algebra (Humphreys, 1972; Newman, 1967). The set of linear combinations of two covariance matrices is an example for such a toral Lie algebra.

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

must rely entirely on the non-Gaussian distribution of source signal samples. In contrast, second-order source separation algorithms, such as Belouchrani et al. (1997), Cardoso (2001), Molgedey and Schuster (1994), Parra and Sajda (2003), Wiskott and Sejnowski (2002), and Ziehe and Müller (1998) to list just a few, exploit the fact that sources have different temporal characteristics, for which the order of samples in time is essential. Temporal structure is what allows JD and other source separation methods to rely entirely on second order statistics. The field of blind source separation (BSS) methods, including ICA, has developed a wide range of sophisticated techniques (Cardoso, 2001; Choi et al., 2005; Cichocki, 2004; Parra and Sajda, 2003). Here we show that much can be achieved by one simple algorithm.

A decision tree

Which method to choose? The researcher setting out to analyze data is greeted by a daunting palette of methods. There is no overall "best": the choice of method depends on the nature of the data and the goals. These may become clear only during the analysis, so it is good to keep in mind a range of methods. JD constitutes a good starting point, because it is easy to understand and can address many tasks effectively. We offer here some hints as to how to orient oneself within the multitude of methods.

Average over trials?

Averaging, a standard tool to improve SNR, is applicable if a phenomenon repeats time-locked to an available reference (e.g. a trigger locked to stimulus or response). Downsides are that trial-specific patterns are lost, and the benefit increases only as $\sqrt{N}$ where N is the number of trials, i.e. it follows a law of diminishing returns.

Filter?

Filtering is another standard tool to improve SNR, useful when target and noise have different spectral properties. It involves convolution with an impulse response, and thus entails loss of temporal resolution and distortion of the waveforms (smoothing, ringing, etc.).

Select channels?

If SNR is good on one particular channel, that channel may be selected.

Average channels?

If SNR is good on a group of channels, those channels may be averaged. More generally, if the SNR map is known, it may be used to design a *matched spatial filter* where each channel is weighted by its SNR.

Common mode rejection?

If noise affects all channels equally, the average over channels may instead be subtracted from each channel. Alternatively, one may calculate the spatial *gradient*, or *Laplacian*. Such operations are routinely used in electrophysiology (e.g. "Current Source Density", CSD, or "re-referencing" in EEG).

Component analysis?

The previous are particular cases of a linear combination of channels. Given J channels, $J - 1$ parameters are available to fine-tune the noise rejection. Component analysis such as PCA, ICA, and JD can be understood as techniques to automatically find these parameters. In some cases it is possible to cancel the noise perfectly, for example if the noise is not of full rank (fewer noise sources than sensors). Granted that the solution found does not also cancel the target, the SNR improvement is infinite. JD and beamforming attempt to find such solutions, and blind separation techniques such as ICA may have a similar effect.

Do noise and target have the same correlation structure?

In this case component analysis is not useful, because any combination that cancels the noise also must cancel the target.

Are target-to-sensor mixing coefficients known?

Such is the case if the anatomical location of the source is known and a forward model is available. Beamforming (Hillebrand et al., 2005; Sekihara et al., 2006) can then be used to find a solution that minimizes the variance from other positions while preserving that of the source.

Does the target have a characteristic that can be enhanced by a bias filter?

Use JD to find components that best reflect the target, and project them back to get clean data.

Does the noise have a characteristic that can be enhanced by a bias filter?

Use JD to find components that best reflect the noise, and project them out to get clean data.

Are target and noise statistically independent?

Consider ICA. ICA methods (of which there are many) rely on some empirical measure of "independence". The sources must be at least one of: non-Gaussian, non-white, non-stationary (Cardoso, 2001; Parra and Sajda, 2003).

Does the instantaneous power of target and/or noise have a characteristic that can be enhanced by a bias filter?

Consider Quadratic Component Analysis (QCA) (de Cheveigné, 2012).

How to choose the bias filter?

The best bias filter depends on the task and the nature of the data. If target or noise is narrow-band, use a bandpass filter. If either is time-locked to a series of triggers, average over trigger-aligned epochs. If either is active within restricted time intervals, or its power is correlated with a known temporal masking function, then filter by weighting with that function.

What about PCA?

Principal Component Analysis transforms the data into components (PCs) that are mutually uncorrelated. Their variance equals that of the data, and most of it is packed into the first components, so that discarding the later components yields a low-dimensional approximation to the data. PCA is useful as a descriptive tool, to understand the correlation and variance structure of the data, and to reduce dimensionality before other forms of analysis (ICA, JD, etc.). It is usually less useful when applied directly to separate noise and target.

It is worth noting that certain of these approaches may be combined. For example JD can be combined with filtering and trial-averaging.

Appendix 6. Details of examples

This section provides additional details concerning the examples given in the main manuscript. The first five examples use the same MEG data set, the sixth uses a different MEG data set, and the last four examples involve data from other recording techniques (ECoG, intrinsic optical imaging, and 2-photon calcium imaging).

Power line noise

This example uses data from a published study that measured MEG responses of human subjects to visual stimulation (Duncan et al., 2009). During each 5 s trial, the subject fixated a cross during 2.5 s, followed by a grating within the lower right or left quadrant during 2.5 s. Stimuli were repeated for a total of 160 trials, of which a subset of 30 is used in the examples in this paper. Data were recorded with a 274-channel gradiometer MEG system (CTF) at a 600 Hz sampling rate. Further

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068

details can be found in the original study (Duncan et al., 2009). These data were also used for illustration in a recent study on induced responses (de Cheveigné, 2012).

JD was applied using a bias filter with peaks at 50 Hz and harmonics, and zeros elsewhere, implemented with a 1024-sample FFT. Each component produced by the JD analysis was examined to determine (a) that it was significantly dominated by 50 Hz and harmonics, and (b) that it did not contain appreciable stimulus-evoked activity. The first 20 components met these criteria and were projected out of the original data to obtain clean data. There is a tradeoff between the amount of remaining noise and the risk of projecting out brain activity collinear with the noise, but the choice in this case was not critical.

Stimulus-evoked activity

This example used the same data as the previous example after removal of 50 Hz components. JD was applied as described in the main text. A more detailed discussion of the use of JD to enhance stimulus-evoked activity is in de Cheveigné and Simon (2008a).

Cardiac artifacts

This example used the same MEG data as the first example after removal of 50 Hz components. An ECG signal was not available, and therefore a cardiac trigger signal had to be derived from the data. JD was used for this purpose, using a criterion that favors components with large kurtosis (i.e. localized large amplitude values interspersed with low amplitude values): matrix $\mathbf{C}_0$ was the covariance of the raw data, matrix $\mathbf{C}_1$ was the covariance of the signal weighted by a temporal mask function. This mask was calculated by taking the absolute value of the signal in each channel, and then averaging over channels. The mask emphasized intervals where the instantaneous amplitude is large, allowing JD to find components with locally large amplitudes, in this case cardiac components. Zero crossings of the first component were used as trigger points to define cardiac epochs.

On the basis of this cardiac trigger, JD was applied again, this time in the same way as for stimulus-evoked activity: matrix $\mathbf{C}_0$ was the covariance of the raw data, matrix $\mathbf{C}_1$ was the covariance of the data averaged over cardiac epochs. The plots in the main text are the result of this analysis.

Narrow-band cortical activity

This example used the same MEG data as the first example after removal of 50 Hz components. JD was applied using as a bias filter a bandpass filter (second-order resonator). The analyses reported in the main text used filter center frequencies (10, 12,16, 30 Hz) chosen on the basis of a systematic scan of the data over a 1–100 Hz range (results not shown). The quality factor of the resonator filter (Q = 8 for 10, 12, 16 Hz, Q = 4 for 30 Hz) was chosen to roughly match the width of spectral peaks in the data, but its value did not appear to be critical. Power spectra in Fig. 4 were calculated with a 2.56 s Hanning window. It is worth noting that the scan failed to reveal one notable narrow-band component (stimulus-induced gamma oscillation near 50 Hz) that was found in the same data in other studies (de Cheveigné, 2012; Duncan et al., 2009). A likely reason for this failure is that the stimulus-induced gamma was collinear with lower-frequency activity, preventing it from emerging as a spatially distinct component in this study. The cited studies preprocessed the data with a high-pass filter, and this presumably allowed the oscillatory component to emerge.

Event-related desynchronization (ERD)

This example used the same MEG data as previous examples after removal of 50 Hz components. The analysis proceeded in two steps. In a first step, the data were normalized to give equal power to all channels, PCA was applied, and PCs with power greater than 0.1 were selected (n = 50). PCs with large power represent activity that is "shared" across sensors, and thus is likely to reflect a genuine cortical source. Conversely, PCs with small power are either specific to few sensors, or more widespread but with very low SNR on each sensor. The threshold chosen (0.1) was very conservative. Reducing dimensionality in this way (274 to 50) reduces the risk of over-fitting. The results shown in the main text were obtained by applying JD to the reduced data. The power spectrogram of Fig. 5d used a 640 ms window.

Two conditions, repeated trials

This example used a different set of MEG data derived from an unpublished study (Molloy et al., in preparation) Q10 that involved both visual (V) stimulation and combined auditory and visual (AV) stimulation. V and AV trials were randomly interleaved. Visual stimuli consisted of a small circle centered on the screen surrounded by letters, presented for a duration of 100 ms. Audio stimuli, when present, had the same onset and duration as the visual stimulus and consisted of a tone of one of four frequencies (0.5, 1, 2, 4 kHz) presented at a level of 10 dB SL. Subjects performed a search task on the visual stimulus and were not encouraged to attend to the auditory stimulus when it was present. The cortical response to this unattended sound was a focus of the study. Data were recorded from a 274-channel axial gradiometer system at a 600 Hz sampling rate. Analysis was performed on epochs of 1 s duration centered on the stimulus onset. The average of the data over the 500 ms pre-stimulus interval was subtracted prior to processing (baseline correction). JD analysis was carried out in two steps, as described in the main text. The first step found multiple highly-reproducible components, all of them with non-auditory topographies (only the first is shown in the paper). The second step, applied to the first 16 components from the first step, found two components with a clearly reproducible difference between trial-averaged responses. Both of these components had bilateral dipolar responses over the temporal region consistent with activity in the auditory cortex (only the first is shown in the paper).

Monkey ECoG

Data were taken from the NeuroTycho project web page (http://www.neurotycho.org/, data set "ECoG-100604"). Data were recorded from a 128-channel surface electrode array at a 1 kHz sampling rate over a 3200 s interval. Anesthetic (mixture of ketamine and medetomidine, Toru Yanagawa, personal communication) was injected half way through the interval. Before applying JD, the Sensor Noise Suppression (SNS) algorithm (de Cheveigné and Simon, 2008b) was used to remove electrode-specific activity, and the data were normalized to give equal power to each electrode, PCA was applied to the normalized data matrix, and a subset of 22 PCs with power greater than 0.5 was selected. These 22 PCs were then submitted to JD as described in the main text ($\mathbf{C}_0$ and $\mathbf{C}_1$ were covariance matrices of the full data and of the post-injection interval respectively).

Two photon imaging of a cochlear hair cell

Two-photon microscopy was used to image the calcium signals in a mouse cochlear inner hair cell, within a plane section at the base of the cell, using a fluorescent probe introduced through a recording patch pipette. The same pipette was used to depolarize the cell for 100 ms, opening channels in the cell membrane to increase the intracellular calcium. Images acquired at a 22 Hz rate were trimmed to a 105 × 90 pixel region containing one hair cell section (about 8 μm across). The stimulus was repeated 9 times (Culley and Ashmore, 2010, in Q11 preparation).

The data were treated as a time series with one channel per pixel ($J = 9450$). The mean of each channel signal was removed and the signals were scaled to equal variance and submitted to a PCA (using the Matlab function 'eigs' to speed the eigendecomposition of the 9450×9450 covariance matrix). PCs beyond the 40th were discarded, and JD was applied to the remaining PCs rather than to the original data. The analysis was performed in two stages, as described in the main text.

Intrinsic optical imaging of the auditory cortex of a ferret

Data were taken from a study that used intrinsic optical imaging to measure responses in auditory cortex of ferret to a pure tone with a frequency that was swept from 100 to 3200 Hz in 14 s (Nelken et al., 2008). Each sweep was repeated nine times. Images of size 76×63 were acquired at a rate of approximately 4.2 fps. Data were treated as a time series with one channel per pixel ($J = 4788$). The mean of each channel signal was removed and the signals were scaled to equal variance and submitted to a PCA. PCs beyond the 58th were discarded, and the JD analysis applied to these PCs rather than the original data. JD analysis was applied as described in the main text.

Two photon imaging of mouse auditory cortex

Two-photon calcium imaging was used to measure the response of neurons in the auditory cortex of mouse to acoustic stimulation (Winkowski and Kanold, 2013). Stimuli consisted of a series of 17 amplitude-modulated pure tones with carrier frequencies spaced at 0.25 octave intervals between 4 and 64 kHz. Tone duration was 1 s, sinusoidal modulation rate was 5 Hz, inter-onset interval was in the range of 6–7 s. Imaging frame rate was approximately 7 Hz, and 20 frames were acquired for each tone, with a 6-frame pre-onset interval. Responses to the 17 tones were concatenated, and the 113×128 pixel images were treated as a time series with one channel per pixel ($J = 14{,}464$). For each channel, the mean over the initial 6 frames of each trial was removed (baseline correction) and the signal was scaled to equal variance for all pixels and submitted to a PCA. PCs beyond the 100th were discarded, and JD analysis was applied to the remaining PCs rather than the raw data. The first 10 JD components were selected and projected back to pixel space to form "clean" data. The topographies in Fig. 6d (top middle and right) of the main paper were obtained by calculating the RMS of the data averaged over frames. The time courses (Fig. 6d (bottom)) were obtained by averaging all pixels within an 8×8 pixel patch centered on one neuron (arrow in Fig. 6d (top right)).

Appendix 7. Failure scenarios

The failure scenarios described in the main text are illustrated here in Fig. 8.

Study A was simulated using real data recorded from a 440-channel MEG system in the absence of a subject. The data were divided arbitrarily into 'epochs', and JD was applied to emphasize repeatable activity. The first component (Fig. 8a, left) indeed seems to be repeatable (the mean, blue, extends well beyond ± 2 standard deviations of a bootstrap resampling, gray), despite the absence of any genuine repeatable process. This is a spurious result of over-fitting. Applying PCA and truncating to 50 PCs before applying JD attenuate this effect (right).

Study B was simulated using a 'target' consisting of a cycle of a sinusoid repeated 100 times, superimposed on 'noise' recorded from a 160-channel MEG system in the absence of a subject, with an overall $\mathrm{SNR} = 0.01$. If the target is presented with an interstimulus interval multiple of 1/50 Hz, JD selects power-line activity present within the noise (left). If the interstimulus interval is incongruent with 1/50 Hz, JD selects the correct target activity (right).

Study C was simulated using the same target and noise as Study B, but an additional random Gaussian noise was added with the same source-to-sensor weights as the target (so that target and noise are colinear). In this situation, JD fails to resolve the target from this source of noise (left). Complementing the data with additional channels with a different target/noise ratio allows JD to extract the target (right).

Study D was simulated using the same target and noise as Study B, but a slow ramp (linearly increasing voltage) was added to the data before dividing into epochs. Subtracting the mean from each epoch causes JD to incorrectly select the ramp as the most repeatable component (left). If this (harmful) step is omitted, JD correctly finds the target (right).

Study E was simulated by creating a 'target' consisting of a pulse with an increasing delay across an array of 50 sensors (i.e. 'propagating' across the sensors, left, top). JD applied with a bandpass bias filter centered at 10 Hz resulted in a series of weights with alternating positive and negative values (left, bottom). The resulting component waveform seems oscillatory (right), despite the absence of any oscillatory process within the original data.

Study F was simulated by creating a 'target' consisting of a burst of random-phase sinusoidal activity occurring within the initial part of each epoch. JD was applied using covariance matrices calculated from the initial and final parts of the epoch, the expected outcome being to extract the target. Instead, the first component was a glitch that occurred by chance in the first part of one trial (left). If such glitches are masked (by applying zero weight to high-amplitude portions in the covariance calculation), JD correctly finds the target (right).

Study G was simulating by creating two targets, consisting of 1 or 2 cycles of a sinusoid (Fig. 8g, left). These were repeated on every trial, and added to the same noise as Study B with $\mathrm{SNR} = 0.01$. The two targets had distinct mixing matrices. JD was applied to find components that optimize the signal-to-noise ratio on the basis of repeatability over trials. Two components are indeed found to have high scores (Fig. 8g, right). They span the same subspace as that spanned by the targets, but neither component matches a target.

Appendix 8. Practical considerations

Implementation

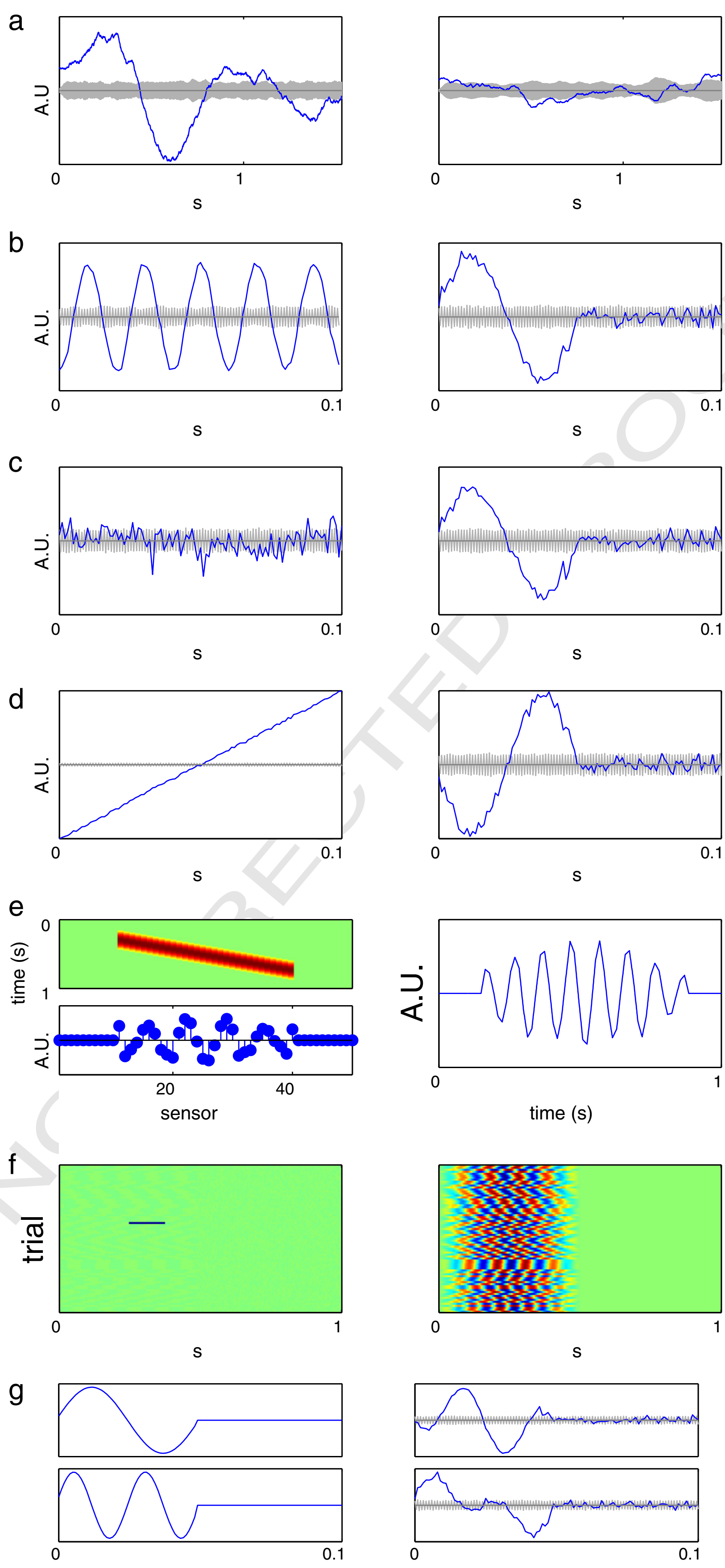
The basic algorithm can be implemented in a few lines of Matlab. Supposing that data for two conditions to be contrasted are in matrices 'x1' and 'x2' (time $\times$ channels), the solution that maximizes activity in 'x1' relative to 'x2' is found by:

```
c0 = x1' * x1 + x2' * x2; c1 = x1' * x1;
[V, D] = eig(c0, c1); V = real(V); D = real(D);
[~, idx] = sort(diag(D), 'descend'); V = V(:, idx);
z1 = x1 * V; z2 = x2 * V;
```

where the 'z1' and 'z2' are the matrices of JD components, the first column of which has the highest possible power ratio of the first condition relative to the second, and the last column the smallest. Implementations are also available in the NoiseTools toolbox (http://audition/ens/fr/adc/NoiseTools/) and the DSS toolbox (http://www.cis.hut.fi/projects/dss/package/). Asymptotic space requirements are dominated by the need to store covariance matrices, which is $O(J^2)$. The covariance matrices may be calculated chunk-by-chunk, so the full data set does not need to fit in memory. Asymptotic runtime requirements are dominated by the cost of eigenvalue decomposition which is $O(J^3)$ where J is the number of channels. The dependency on number of samples T is linear.

Preprocessing

Prior to PCA it may be useful to apply the SNS algorithm (de Cheveigné and Simon, 2008b) to remove channel-specific activity, defined as variance uncorrelated with any other channel. Channel-specific activity may reflect sensor noise (EEG, MEG), or brain activity proximal to the sensor or electrode (LFP, ECoG). By definition,


a
A.U
0
1
s
b
A.U.
0.1
c
d
e
time (s)
sensor
20
40
f
trial
g

channel-specific activity does not benefit from (or contribute to) component analysis, and it is best studied on a per-channel basis.

If some proportion of the noise variance can be suppressed before applying JD, for example by preprocessing the data with a filter that attenuates spectral components remote from the activity of interest, degrees of freedom that would have been used to remove that variance become available to suppress other noise sources. For example if the brain activity of interest is well below 50 Hz, convolving the data with a square window of size 1/50 Hz (with zeros at 50 Hz and all harmonics) will obviate the need to project out spatial components dominated by line power. For similar reasons it may be useful to remove slow trends by fitting a polynomial to the raw data and subtracting the fit. It is important that such a fit be calculated on the full data before dividing into epochs. Polynomial trends usually should *not* be removed on a trial-by-trial basis (see Failure Scenario D).

In general, second order-statistics are very sensitive to outliers. Even a single large outlier can end up dominating the largest eigenvectors of $\mathbf{C}_0$ and $\mathbf{C}_1$. This is one reason why blind source separation techniques are often used for artifact detection and subtraction. However, when we are really interested in the components of neural signals, sensitivity to noise and outliers is not desired. Data should be screened for outliers prior to calculation, and also possibly at intermediate stages because new outliers may become apparent after strong components have been removed.

It is customary to remove the mean prior to calculation of a covariance matrix or PCA, but this is not necessary, or desirable if a deviation of the mean from zero is meaningful. For example if the mean was set to zero over a pre-stimulus interval (baseline correction), removing the global mean would undo that correction.

References

Amari, S., 2000. Estimating functions of independent component analysis for temporally correlated signals. Neural Comput. 12, 2083–2107.

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

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 (1), 41–56.

Blaschke, T., Berkes, P., Wiskott, L., 2006. What is the relation between slow feature analysis and independent component analysis? Neural Comput. 18 (10), 2495–2508.

Buzsáki, G., 2006. Rhythms of the Brain. Oxford Univ. Press,.

Cardoso, J.F., 2001. The three easy routes to independent component analysis; contrasts and geometry. Proc. Int. Conf. on Independent Component Analysis and Blind, Source Separation (ICA01), pp. 1–6.

Choi, S., Cichocki, A., Park, H.-M., Lee, S.-Y., 2005. Blind source separation and independent component analysis: a review. Neural Inf. Process. Lett. Rev. 6, 1–57.

Cichocki, A., 2004. Blind signal processing methods for analyzing multichannel brain signals. Int. J. Bioelectromagn. 6 (1).

Culley, S., Ashmore, J., 2010. Identification of simultaneous calcium entry sites in cochlear inner hair cells of the adult mouse. Proc Physiol Soc, 19, p. PC37.

de Cheveigné, A., 2010. Time-shift denoising source separation. J. Neurosci. Methods 189 (1), 113–120 (May).

de Cheveigné, A., 2012. Quadratic component analysis. NeuroImage 59 (4), 3838–3844 (Feb.).

de Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift PCA. J. Neurosci. Methods 165 (2), 297–305 (Sep.).

de Cheveigné, A., Simon, J.Z., 2008a. Denoising based on spatial filtering. J. Neurosci. Methods 171 (2), 331–339 (Jun.).

de Cheveigné, A., Simon, J.Z., 2008b. Sensor noise suppression. J. Neurosci. Methods 168 (1), 195–202 (Feb.).

de Cheveigné, A., Edeline, J.-M., Gaucher, Q., Gourévitch, B., 2012. Component analysis reveals sharp tuning of the local field potential in the guinea pig auditory cortex. J. Neurophysiol. 109, 261–272.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Müller, K.-R., 2006. Combined optimization of spatial and temporal filters for improving brain–computer interfacing. IEEE Trans. Biomed. Eng. 53 (11), 2274–2281.

Drazin, M., 1951. Some generalizations of matrix commutativity. Proc. Lond. Math. Soc. 3 (1), 222–231.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons,.

Duncan, K.K., Hadjipapas, A., Li, S., Kourtzi, Z., Bagshaw, A., Barnes, G., 2009. Identifying spatially overlapping local cortical networks with MEG. Hum. Brain Mapp. 31 (7), 1003–1016 (Dec.).

Efron, B., Tibshirani, R., 1993. Introduction to the bootstrap. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Ann. Hum. Genet. 7 (2), 179–188.

Fukunaga, K., 1972, 1990. Introduction to Statistical Pattern Recognition. Academic Press,.

Fukunaga, K., Koontz, W.L.G., 1970. Application of the Karhunen–Loeve expansion to feature selection and ordering. IEEE Trans. Comput. C-19 (4), 311–318.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87 (C), 96–110 (Feb.).

Hillebrand, A., Singh, K.D., Holliday, I.E., Furlong, P.L., Barnes, G.R., 2005. A new approach to neuroimaging with magnetoencephalography. Hum. Brain Mapp. 25 (2), 199–211.

Humphreys, J.E., 1972. Introduction to Lie Algebras and Representation Theory, vol. 1980. Springer, New York,.

Hyvarinen, A., 2012. Independent component analysis: recent advances. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 371 (1984), 20110534 (Dec.).

Hyvärinen, J., Karhunen, J., Oja, E., 2009. Independent Component Analysis. Wiley Interscience,.

Koles, Z.J., Lazar, M.S., Zhou, S.Z., 1990. Spatial patterns underlying population differences in the background EEG. Brain Topogr. 2 (4), 275–284.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540 (Apr.).

Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56 (2), 387–399 (May).

Meinecke, F., Ziehe, A., Kawanabe, M., Muller, K., 2002. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. IEEE Trans. Biomed. Eng. 49 (12), 1514–1525.

Molgedey, L., Schuster, H.G., 1994. Separation of a mixture of independent signals using time delayed correlations. Phys. Rev. Lett. 72 (23), 3634–3637.

Nelken, I., Bizley, J.K., Nodal, F.R., Ahmed, B., King, A.J., Schnupp, J.W.H., 2008. Responses of auditory cortex to complex stimuli: functional organization revealed using intrinsic optical signals. J. Neurophysiol. 99 (4), 1928–1941 (Jan.).

Newman, M., 1967. Two classical theorems on commuting matrices. J. Res. Natl. Bur. Stand. B Math. Math. Phys. 71 B (2, 3).

Nikulin, V.V., Nolte, G., Curio, G., 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. NeuroImage 55 (4), 1528–1535 (Apr.).

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

Parra, L., Spence, C., 2000. Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Process. 8 (3), 320–327.

Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. NeuroImage 28 (2), 326–341.

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

Rao, C., 1948. The utilization of multiple measurements in problems of biological classification. J. R. Stat. Soc. Ser. B Methodol. 10 (2), 159–203.

Särelä, J., Valpola, H., 2005. Denoising source separation. J. Mach. Learn. Res. 6, 233–272.

Sekihara, K., Hild, K., Nagarajan, S., 2006. A novel adaptive beamformer for MEG source reconstruction effective when large background brain activities exist. IEEE Trans. Biomed. Eng. 53, 1755–1764.

Stevenson, I.H., Kording, K.P., 2012. How advances in neural recording affect data analysis. Nat. Neurosci. 14 (2), 139–142.

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Müller-Putz, G.R., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., Blankertz, B., 2011. Review of the BCI competition IV. Front. Neurosci. 6 (55), 1–31 (Dec.).

Wang, Y., Berg, P., Scherg, M., 1999. Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study. Clin. Neurophysiol. 110 (4), 604–614 (Mar.).

Winkowski, D.E., Kanold, P.O., 2013. Laminar transformation of frequency organization in auditory cortex. J. Neurosci. 33 (4), 1498–1508.

Wiskott, L., Sejnowski, T., 2002. Slow feature analysis. Neural Comput. 14, 715–770.

Yeung, N., Bogacz, R., Holroyd, C.B., Cohen, J.D., 2004. Detection of synchronized oscillations in the electroencephalogram: an evaluation of methods. Psychophysiology 41 (6), 822–832 (Nov.).

Ziehe, A., Müller, K.-R., 1998. TDSEP—an efficient algorithm for blind separation using time structure. Proc Int Conf on Artificial Neural Networks ICANN, 98, pp. 675–680.

Fig. 8. Failure scenarios. (a) Study A: JD applied to a 440-channel data set produces a spurious component (left) due to over-fitting. Dimensionality reduction to 50 dimensions attenuates this spurious response (right). (b) Study B: Due to an unfortunate choice of trial-to-trial interval size, JD selects power line components (left) instead of the expected response (right) that would otherwise be found. (c) Study C: JD fails to extract a target component because it is collinear with a large noise source (left). Complementing the data with additional channels so that target and noise are no longer collinear allows JD to extract the target (right). (d) Study D: In the presence of a slow drift, removal of the mean from each trial causes JD to select a spurious "ramp-shaped" component (left). Omitting that processing step allows JD to find the correct target (right). (e) Study E: In the presence of activity that propagates across the sensor array (left, top), JD with a bias filter centered on 10 Hz produces a spurious oscillatory component (right). The weights associated with this component consist of alternating positive and negative values (left, bottom). (f) Study F: In the presence of a glitch, JD fails to reveal underlying induced activity (left). Assigning zero weight to the glitch allows JD to find the expected activity (right). (g) Study G: Applied to data containing two distinct targets, JD finds two components that do not match them (although they span the same subspace).

Please cite this article as: de Cheveigné, A., Parra, L.C., Joint decorrelation, a versatile tool for multichannel data analysis, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.068