Time-shift denoising source separation

Alain de Cheveigné a,b,*

a Laboratoire de Psychologie de la Perception, UMR 8581, CNRS and Université Paris Descartes, France
b Département d’Études Cognitives, École Normale Supérieure, France

ARTICLE INFO
orthogonal to the first, and so on. To obtain denoised data, a select subset of components \( y_k(t), j = 1, \ldots, K \) is projected back to sensor space:

\[
\tilde{x}_j(t) = \sum_k b_{jk} x_j(t)
\]

(2)

where \( \tilde{x}_j(t), j = 1, \ldots, J \) are the cleaned sensor signals and \( b_{jk} \) are projection coefficients. In this paper we focus on the case \( K = 1 \) where only the best component is retained. Additional components can be obtained by repeating the process after regressing out the best component ("deflation").

The required weight parameters \( a_{id} \) are obtained using the DSS algorithm (Särelä, 2004; Särelä and Valpola, 2005). In brief, DSS involves the following steps: (a) orthogonalization by PCA of the data matrix (here augmented with time shifts), (b) normalization of the principal components, (c) application of a contrast function to emphasize target components and/or deemphasize noise, and (d) a final PCA to rotate the data so that the first component has the highest possible value of the contrast function, the second the highest value within the subspace orthogonal to the first, etc. Details of the algorithm are given in the next section, and related work is reviewed in Section 4.

2. Methods

2.1. Signal model

Sensor signals \( \mathbf{X}(t) = [x_1(t), \ldots, x_J(t)]^T \) include interesting brain activity and uninteresting noise activity:

\[
\mathbf{X}(t) = \mathbf{X}_{\text{sp}}(t) + \mathbf{X}_{\text{sn}}(t).
\]

(3)

The first term results from the superposition within each sensor of sources of interest within the brain, \( \mathbf{X}_{\text{sp}}(t) = \mathbf{A}_{\text{sp}} \mathbf{S}_{\text{sp}}(t) \), and the second results from uninteresting sources within the environment, sensors, and subject’s body, \( \mathbf{X}_{\text{sn}}(t) = \mathbf{A}_{\text{sn}} \mathbf{S}_{\text{sn}}(t) \), where \( \mathbf{A}_{\text{sp}} \) and \( \mathbf{A}_{\text{sn}} \) are mixing matrices. Our aim is to attenuate \( \mathbf{X}_{\text{sn}}(t) \) so as to isolate \( \mathbf{X}_{\text{sp}}(t) \), thereby improving our observation of brain activity.

The sensor signals \( \mathbf{X}(t) \) span a space of dimensionality \( J \). Linear separation techniques such as ICA, beamforming or DSS assume that source and noise live in separate subspaces, the only difficulty being to find them. Here we are interested in the situation where target and interference components are partly colinear, and thus not separable using such techniques.

2.2. Contrast function

For clarity we will focus on the common situation where data are recorded in response to \( N \) presentations of the same stimulus, and we are interested in reproducible stimulus-evoked activity. The contrast function used in this paper is the average over trials:

\[
[X(t)] = \frac{1}{N} \sum_{n=1}^{N} \mathbf{X}^n(t)
\]

(4)

where \( \mathbf{X}^n(t) \) designates data for trial \( n \). Stimulus-evoked activity is reinforced by averaging, whereas stimulus-unrelated activity and noise are not, so the norm of \( [X(t)] \) is greater for evoked activity than for noise. Other contrast functions can be used but they are not considered in this paper.

2.3. Algorithm

Given a data matrix of size \( [J, T, N] \) (channel \( \times \) time \( \times \) trial), the following steps are performed:

1. Data are time-shifted over the range \( 0, \ldots, D - 1 \), the time-shifted data being treated like additional channels to form a matrix of size \( [JD, T, N] \).
2. Data are reorganized by concatenating trials along the time dimension to obtain a matrix of size \( [JD, TN] \). Each row corresponds to a particular channel/time shift, and each column to a particular time sample/trial.
3. PCA is performed, and the resulting component time series are normalized.
4. The data are reorganized again by trials to form a 3D matrix of size \( [JD, T, N] \) (component \( \times \) time \( \times \) trial). The contrast function defined in Eq. (4) is applied, resulting in a matrix of size \( [JD, T] \).
5. A second PCA is performed on this matrix.
6. The rotation matrix from the second PCA is applied to the whitened data from step 2 to obtain a matrix of components of size \( [JD, T, N] \) (component \( \times \) time \( \times \) trial). Each component is of size \( [T, N] \) (time \( \times \) trial).

Step 6 produces components that have two useful properties: (a) they are mutually orthogonal and (b) they are ordered by decreasing ratio of evoked power to total power. The first component is the best linear function of the data according to that criterion, and each subsequent component is the best within the subspace orthogonal to previous components. Eq. (1) defines a spatiotemporal filter: each component is a weighted sum of time-delayed channels. Intuitively it is easy to see why the algorithm maximizes evoked power: the first three steps create a data set that is “spherical” in the sense that all directions in multidimensional space have equal variance. Averaging over trial “tips the balance” by favoring certain directions that are selected by the second PCA. The evoked/raw power ratio is maximal along these directions.

The subspace spanned by the first few components contains the most reproducible brain activity. However, that subspace may also include time-shifted versions of brain activity, and thus all FIR-filtered versions of that activity, for filters of order up to \( D \). Components produced by step 6 may thus differ from real brain activity by the effect of an arbitrary FIR filter of order \( D \). This is not always a concern, as spectral filtering is commonly applied to improve quality of data. Nevertheless it may be desirable to minimize potential filtering effects and approximate unfiltered source activity. In that case an additional step may be performed:

7. Given a subset of components chosen for reproducibility, canonical correlation is used to find the linear combination most correlated to the raw data. Canonical correlation (Hotelling, 1936) is a technique that finds the linear combination of vectors within one set that best accounts for the variance in another set, and vice versa.

In the following we assume that only one component is retained (the most reproducible, or the most similar to unfiltered data within the chosen subspace). If stopped at step 6, the algorithm has one user-set parameter, \( D \), if continued to step 7 an additional parameter is the number of components within the chosen subset.

2.4. Implementation

The algorithm was implemented in Matlab. A few details are worth mentioning. The size of the data matrices grows proportionally with \( D \) and that of the covariance matrices with \( D^2 \), limiting the values of \( D \) that can be used. To save space, the time series may be divided into chunks and the covariance matrices calculated over these chunks and then summed. As in any calculation involving sums of squares, it is important to discount outliers. This is done by assigning them a zero weight in the calculations. PCA components with relative power below a threshold (e.g. \( 10^{-3} \)) are discarded in.
step 3 to save memory and computation time and avoid numerical problems.

2.5. Evaluation data

The evaluation is evaluated first with synthetic test data, and then with data from several biomagnetic sensor arrays: small animal MEG, spinal cord magnetometer, and human MEG. Small animal MEG data came from a system equipped with a 9-channel array of magnetometers in a 8 x 8 mm array at 3 mm from the external surface of the case placed over the animal’s head (Miyamoto et al., 2007, 2008). Three reference sensors sampled environmental noise, and an accelerometer sampled environmental vibration. Sensors and subject were placed within a magnetically shielded enclosure. Data were acquired at a sampling rate of 1000 Hz after high-pass filtering at 0.1 Hz and low-pass filtering at 200 Hz, and environmental and sensor noise attenuated using the TSPCA and SNS algorithms (de Cheveigné and Simon, 2007, 2008a). Data were recorded from guinea pig, gerbil and mouse in response to auditory stimuli presented diotically via flexible sound tubes ending at 1.5 cm from each ear and driven by Etymotics transducers placed outside the shielded chamber. Spinal cord data were recorded using a newly developed 120-channel vector magnetometer array placed within a dewar optimized for measurements from the spinal cord (Adachi et al., 2009). Data were filtered between 100 and 5000 Hz and recorded at a 10000 Hz sampling rate in response to 4000 repetitions of electric stimulation of a peripheral nerve. Human MEG data were acquired from a 160-channel system equipped with axial gradiometers (Kado et al., 1999). Data were sampled at 500 Hz after high-pass filtering at 0.1 Hz and low-pass filtering at 200 Hz, and environmental and sensor noise components were removed using the TSPCA and SNS algorithms. Data were recorded from human subjects in response to a set of auditory stimuli comprising clicks, short tones, noise bursts and tone ramps interleaved at random.

2.6. Evaluation statistics

The effect of denoising is quantified in terms of power (before vs. after denoising) and variance across stimulus repetitions (evoked/total power). The reliability of evoked responses is estimated using bootstrap resampling (Efron and Tibshirani, 1993). In brief: random sets of N trials are drawn with replacement, and the average over each set is calculated. This operation is repeated (typically 100 times) and the mean and standard deviation are calculated at each time sample within the epoch.

3. Results

3.1. Simulated data

3.1.1. Simulation 1

This simulation illustrates the limits of static methods such as DSS. The artificial data had dimensions 10 channels x 1000 samples x 100 trials. The “signal” was a single period of a sinusoid as illustrated in Fig. 1 (top), identical in all channels and repeated over all trials, and the “noise” consisted of 9 (resp. 10) independent Gaussian white noise sources mixed within each of the 10 channels via a 9 x 10 (resp. 10 x 10) mixing matrix with random coefficients. The SNR is defined as the ratio of signal and noise power summed over channels, transformed to decibels (dB) according to the formula

\[ \text{SNR} = 10 \log_{10} \frac{P_s}{P_n} \]

Fig. 1 (bottom left) shows data for the “best” channel (the channel with the largest correlation with the clean signal) for SNR = -40 dB (RMS amplitude ratio 1/100, power ratio 1/10000). The blue line is the average over trials, and the gray band represents ±2 SD of a bootstrap resampling of that average. This plot is for 9 noise sources, the plot for 10 noise sources had a similar aspect (not shown). At this unfavorable SNR, the signal seems to be lost within the noise. The middle plot shows the first component of a DSS analysis (de Cheveigné and Simon, 2008b) for 9 noise sources. The recovered component is identical to the signal. In this case, signal and noise were separable within the 10-dimensional data space and DSS was successful in separating them. However, the right hand plot shows the same analysis for 10 noise sources, in which case the noise has the same dimensionality (10) as the data. Signal and noise are no longer linearly separable and the result of DSS is noisy (but nevertheless an improvement over no processing). This is the situation that we are interested in, where the dimensionality of the noise is too large to allow perfect linear separation from the signal.

3.1.2. Simulation 2

This simulation illustrates how TDDSS can be more effective than DSS. The signal was the same as in Simulation 1 (Fig. 1 top). The noise was produced from 5 independent noise sources mixed within each channel via a 5 x 10 random matrix, and the same 5 sources were shifted in time by 1 sample and added to each channel via another 5 x 10 random matrix. This noise has a similar covariance matrix as in the previous example, but there is now a serial correlation within and across channels.

Fig. 2 (top) shows the result of applying standard DSS at SNRs of -20, -40 and -60 dB. DSS is quite successful at -20 dB and partly successful at -40 dB (as with 10 independent sources, Fig. 1 bottom right), but it performs poorly at -60 dB. The lower plot shows the result of applying TDDSS (D = 1). Now the signal is almost perfectly recovered even at -60 dB (1/1000 RMS ratio, 10^-6 power...
Fig. 1. Effect of noise dimensionality on the performance of DSS. Artificial data involve 10 channels, each a mixture of one signal source (identical in all channels and all trials) and either 9 or 10 independent Gaussian noise sources mixed with random coefficients. Overall SNR is −40 dB (1/100 in amplitude). All plots are based on 100 trials, blue is the mean over trials, gray indicates ±2 standard deviations of a bootstrap resampling of the mean: (top) source waveform; (bottom left) waveform of the channel with best SNR; (middle) best DSS component for 9 noise sources; (right) best DSS component for 10 noise sources. DSS can extract the signal perfectly if target and source are linearly separable (middle), but not if they are not separable (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

tional denoising (see Section 2), the 9-channel data were subjected to three different processing schemes: PCA, DSS, and TSDSS with $D = 20$ as illustrated in Fig. 4 (top). Here PCA is used as a baseline to evaluate the improvement offered by DSS or TSDSS. Comparing the time course of the averages over trials (blue), the first principal component (left) appears to be noisier than the first DSS component (middle), itself noisier than the first TSDSS component (right), as confirmed by the bootstrap resampling (gray bands).

3.2.2. Spinal cord magnetometer

Data were recorded from a human subject in response to 4000 repetitions of an electrical stimulation of the median nerve of the left wrist. The 120-channel data were subjected to PCA, DSS, and TSDSS ($D = 4$) as illustrated in Fig. 4 (bottom). The first TSDSS component (right) is less noisy than the first DSS component (middle) itself less noisy than the first principal component (left).

3.2.3. Human MEG

Data were recorded from a human subject in response to 50 repetitions of a 1 kHz tone pip of duration 100 ms presented diotically at a comfortable level. These were interleaved with various other stimuli within a 5-min recording session. The original data comprised 160 channels, but for this simulation only 7 channels were retained (to obtain data with low spatial dimensionality to better illustrate the behavior of the TSDSS algorithm). After initial denoising to remove environmental and sensor noise (see Section
2), data were processed by the DSS algorithm (Fig. 5 left), and by the TSDSS algorithm with $D = 30$ (Fig. 5 right). The blue line indicates the average over trials, and the gray band ±2 SD of a bootstrap resampling. The TSDSS algorithm appears to yield a cleaner result, with a smoother average and narrower bootstrap interval. One might wonder whether the benefit of TSDSS is greater than that obtained by simply smoothing the data with kernel of size $D = 30$. To answer this question, data were smoothed with such a kernel before applying DSS (middle). The result is less reproducible than with TSDSS, as reflected by the width of the bootstrap band (gray). With a full data set (150 channels rather than 7) the benefit of TSDSS was less impressive (not shown), presumably because that

![Figure 3](image1.png)

**Fig. 3.** The target is the same as in Figs. 1 and 2. The noise consists of 50 sinusoidal sources added with random amplitude and phase, different for each of the 100 trials: (top) DSS algorithm and (bottom) TSDSS algorithm implemented with $D = 10$. The performance of DSS degrades as SNR decreases (top row), whereas that of TSDSS remains good for SNRs as low as $-60$ dB ($1/1000$ in amplitude, 1 PPM in power).

![Figure 4](image2.png)

**Fig. 4.** Comparison of TSDSS and DSS on real data. (Top) Data from a 9-channel magnetometer for small animals. The fields originate from the auditory cortex of a guinea pig in response to 6400 repetitions of a 2 kHz tone pip of duration 30 ms: (left) first PCA component; (middle) first DSS component; (right) first TSDSS component ($D = 20$). (Bottom) Data from a 120-channel magnetometer for spinal cord evoked field measurements. The fields originate from the spinal cord of a human subject in response to 4000 repetitions of an electrical stimulation: (left) first PCA component; (middle) first DSS component; (right) first TSDSS component ($D = 4$).
number of channels was sufficient to represent the noise without delays.

4. Discussion

TSDSS performs by introducing time shifts that allow the algorithm to synthesize an optimal multichannel FIR spatiotemporal filter, rather than just a spatial filter. The benefit is clear on the artificial data, on real data it is more limited but nevertheless appreciable.

4.1. How it works

The principle of DSS is well explained by Särrelä (2004); Särrelä and Valpola (2005), and has been illustrated in the context of an evoked-response paradigm by de Cheveigné and Simon (2008b). In brief, the initial PCA and normalization steps of the DSS algorithm produce a data set that is “spherical” in the sense that it extends equally in every direction of the multidimensional signal space, and is thus free to rotate. The contrast function emphasizes certain directions of interest (here: directions of greatest repeatability), and the second PCA aligns them with the coordinate axes. In this way DSS finds the optimal linear combinations for the chosen criterion. The effectiveness of DSS depends on the dimensionality of the data, itself determined by the number of sensors. Intuitively, one might expect time shifts to be beneficial simply by increasing the dimensionality. Furthermore, the ability to synthesize multichannel FIR filters might give leverage in at least two ways. The first is by designing a spectral filter that emphasizes spectral features of the target relative to the noise (as in a Wiener filter), regardless of spatial structure. The second is by giving the algorithm more flexibility to suppress interfering sources on the basis of their spatial structure. As an example, suppose two noise sources such that the target is not collinear with either of them, but nevertheless falls within the hyperplane that they span. If either noise source were alone the target would be separable, but if both are active at the same time it is not. However, if the two noise sources have different spectral properties, TSDSS may be able to find a solution involving two spatial filters, each optimized for a different spectral region.

4.2. Caveats and cautions

The algorithm extracts weak sources by cancelling strong interference components. This delicate “balancing act” might go wrong in a variety of ways, a remark that was made previously for DSS (de Cheveigné and Simon, 2008b), but that applies even more to TSDSS because of the larger number of parameters, and because TSDSS may be applied at lower SNRs. It is essential to understand what the algorithm is doing, and to corroborate its results by cross-validation.

The method involves a large number of free parameters ($J, D, K'$) are the number of sensors, shifts, and retained components, respectively), and thus it is susceptible to over-fitting: even within a random data set there may exist linear combinations that appear “repeatable” to some degree. The investigator should worry this possibility by applying cross-validation tests. For example, the denoising stage may be included within the resampling loop of a bootstrap procedure (Efron and Tibshirani, 1993). Concretely: sets of trials are drawn with replacement, the denoising algorithm is applied to each set, and statistics (mean and standard deviation) are calculated over the population of denoised sets.

Another concern is spectral distortion. Suppose that the algorithm succeeds in suppressing all noise components. The SNR is now infinite, but the recovered data may nevertheless differ from true brain activity, as TSDSS components are weighted sums of sensor signals and their time-shifted versions. This can obviously affect the latency (within a range of size $D$), but also the shape of the response. To illustrate this point, Fig. 6 shows the result of a PCA on the time-shifted versions of the signal depicted in Fig. 1 (top left) for $D = 5$. The first principal component (PC) is a slightly-smoothed version of the signal, but none of the subsequent PCs resembles the signal. The subspace of time-shifted signals includes such distorted waveforms, and TSDSS components are potentially susceptible to similar distortion. One can argue that TSDSS components are optimized for repeatability, and thus any filtering will emphasize the “true” evoked response. Nevertheless the optional step 7 of the algorithm offers a means to minimize this distortion by rotating the subspace spanned by the selected components to maximize similarity to the undelayed data (at the expense of repeatability).

It is important to stress a potential pitfall of using TSDSS or other data-dependent algorithms such as DSS, ICA or beamforming. If the algorithm is applied separately to each condition of an experiment, differences in data-dependent processing might masquerade as effects. One way to avoid this is to interleave conditions to ensure that the same processing is applied to all conditions.

4.3. Relation with previous methods

The DSS method was proposed by Valpola and Pajunen (2000)
al., 2008; Koles et al., 1990) that have been widely used for classification in the brain computer interface (BCI) literature. These methods have earlier roots (Fukunaga, 1990; Parra et al., 2005). In de Cheveigné and Simon (2008b) we applied DSS (without time shifts) to denoising of stimulus-evoked data.

The idea of augmenting multichannel data with time shifts has been proposed in several contexts. The initial two steps of TSDSS correspond to the spatiotemporal whitening proposed by Beucker and Schlitt (1997). Time delays are involved in various other analysis techniques (Friman et al., 2002; Gruber et al., 2006; Sander et al., 2002; Woon and Lowe, 2004; Ziehe et al., 2000). TSDSS exploits the serial correlation structure of the data, that can also be captured in the frequency domain (Anemüller et al., 2003) or by multichannel linear prediction (e.g. Xie et al., 1992). Dornhege et al. (2006) have proposed to include time shifts in CSP to derive classification features in a brain computer interface application, a proposition closely related to ours. An appeal of DSS is that it is comparably easy to understand, an advantage for non-specialized users.

4.4. Summary

TSDSS is effective in extracting extremely weak signals from noise. It capitalizes on the dimensionality of large multisensor arrays such as used in MEG, EEG or multielectrode electrophysiology, and on the spatiotemporal disparities between target and noise and/or between different noise components. In this respect, it pushes the limit of what can be done to retrieve signals at extremely unfavorable SNRs using purely linear techniques. Applied to data from human MEG, small animal MEG, and spinal cord biomagnetometry, TSDSS offered an improvement over DSS in the quality and reliability of extracted components as assessed by bootstrap resampling. An similar benefit was observed in simulations, for which the method succeeded over a wide range of conditions for which the best alternative algorithm that we know (DSS) failed. This suggests that there may be applications for which TSDSS is the only effective solution to extract exploitable data.

Acknowledgments

Thanks to Yoshiaki Adachi for providing the spinal cord data, and to Maria Chait for recording the human MEG data. Jonathan Simon, Israel Nelken and one anonymous reviewer provided useful comments on previous versions of this paper. This work was supported in part by BBSRC grant BB/H006958/1, by the Neuroinformatique 2009 program of CNRS, and by a collaboration grant awarded by NTT. The small animal MEG system was provided by Kanazawa Institute of Technology.

References


Sekihara K, Nagarajan S, Poeppel D, Miyashita Y. Reconstructing spatio-temporal
activities of neural sources from magnetoencephalographic data using a vector
Sekihara K, Hild KE, Nagarajan SS. A novel adaptive beamformer for MEG source
reconstruction effective when large background brain activities exist. IEEE Trans
Taulu S, Simola J, Kajola M. Applications of the signal space separation method. IEEE
Valpola H, Pajunen P. Fast algorithms for Bayesian independent component analysis.
In: Proceedings of the second international workshop on independent compo-
Vrba J, Robinson SE, McCubbin J. How many channels are needed for MEG?
Neurology, Neurophysiology and Neuroscience 2004; 99:1–5. Available at:
http://www.neurojournal.com/article/view/313. Date accessed: 12 November
2009.
Woon WL, Lowe D. Can we learn anything from single-channel unaveraged MEG
Xie F, Le Cerf P, Van Compernolle D. Linear prediction analysis of multichannel
and signal processing, 1992; 277–82.
Ziehe A, Müller K-R, Nolte G, Mackert BM, Curio G. Artifact reduction in magnetoneu-
rography based on time-delayed second-order correlations. IEEE Trans Biomed