A tutorial on generalized eigendecomposition for source separation in multichannel electrophysiology

Michael X Cohen[a]

[a]*Donders Centre for Medical Neuroscience, Radboud University Medical Center, the Netherlands,*

Abstract

The goal of this paper is to present a theoretical and practical introduction to generalized eigendecomposition (GED), which is a robust and flexible framework used for dimension reduction and source separation in multichannel signal processing. In cognitive electrophysiology, GED is used to create spatial filters that maximize a researcher-specified contrast, such as relative spectral power or experiment condition differences. GED is fast and easy to compute, performs well in simulated and real data, and is easily adaptable to a variety of specific research goals. This paper introduces GED in a way that ties together myriad individual publications and applications of GED in electrophysiology. Practical considerations and issues that often arise in applications are discussed.

Keywords: EEG, MEG, LFP, oscillations, source separation, GED, eigendecomposition, components analysis, covariance matrix

1. Background and motivation

1.1. Why multivariate analyses?

It is not contentious to write that electrophysiologists are uninterested in *electrodes*; instead, electrophysiologists are interested in using the numerical values that the electrodes produce to understand how the brain works. It is equally uncontentious to state that there is no one-to-one mapping of electrode to computational source. The neural computations underlying cognition are implemented by complex interactions across neural circuits comprising various types of cells, modulations by neurochemicals, etc.;

these circuit interactions produce electromagnetic fields that can be quantified using LFP, EEG, or MEG, but these electrical fields propagate to multiple electrodes simultaneously. Furthermore, each electrode simultaneously measures the electrical fields from multiple distinct neural circuits, plus artifacts from muscles and heart beats, and noise.

Thus, the manifest variables we are capable of measuring (voltage fluctuations in electrodes) reflect indirect mixtures of the latent constructs we seek to understand. The mixing of sources at the electrodes motivates *multivariate* analyses that identify patterns distributed across electrodes, as opposed to *univariate* analyses that consider each individual electrode to be a separate statistical unit (Figure 1).

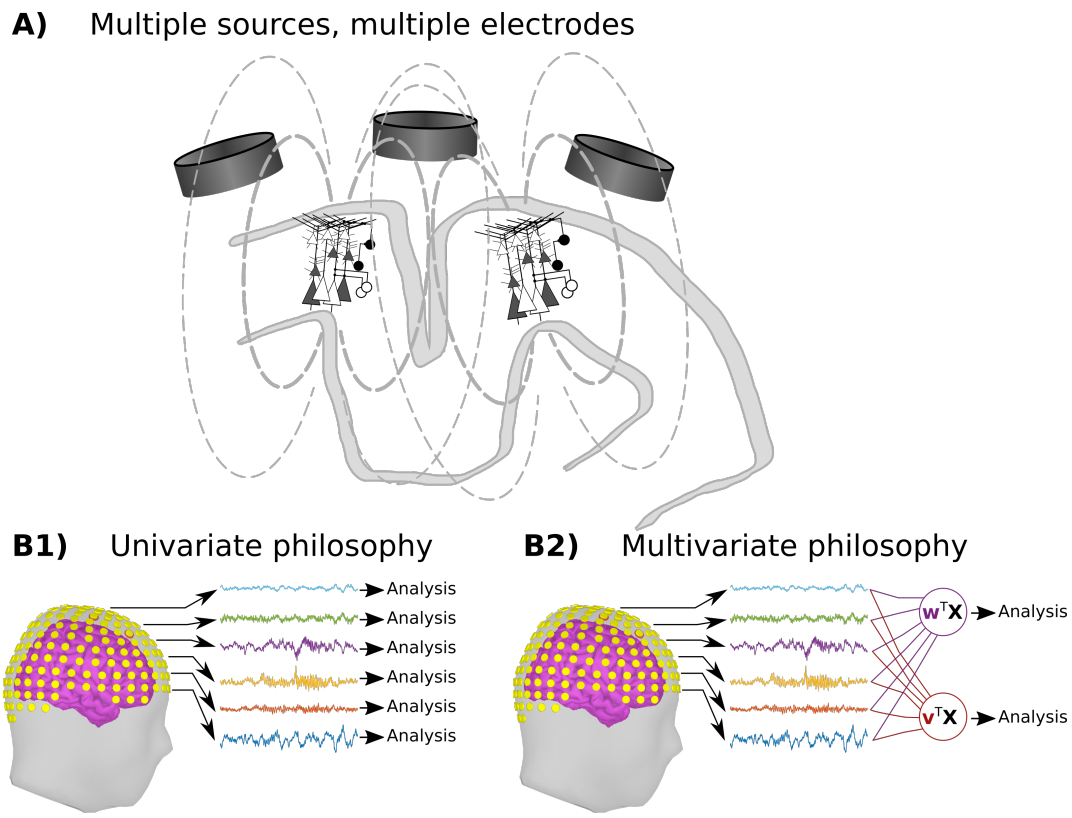

Figure 1: A) The source-separation problem. Electrophysiology involves measuring data from electrodes (black disks), which are manifest variables. But electrophysiologists are interested in understanding neural circuits, which are complex combinations of neurons and neurochemicals, and which generate electrical fields that propagate to the electrodes. The challenge is to leverage the spatial correlations across electrodes to separate activity from different neural circuits. The illustration here makes it seem like the problem is one of anatomical assignment, but because neural computations are spatially distributed, a single localized dipole is not physiologically plausible for all high-level sensory/cognitive/motor phenomena. B) The distinction between univariate (B1) and multivariate (B2) analyses lies in the conceptual and statistical use of the electrodes: The univariate philosophy is that each electrode is an independent measurement; the multivariate philosophy is that the signals of interest are embedded in patterns that span many electrodes, and thus isolating those signals requires appropriately designed spatial filters.

And yet, mass-univariate analyses have remained the most common analysis method for decades. The mass-univariate approach is nearly universally used because it has been so successful for much of the history of neuroscience. It also reflects the state of electrode technology and computational power that provided severe limits to the types of analyses that were feasible in past decades. However, these limitations no longer exist, and thus there is little justification for maintaining the position that analyses that dominated the literature 30 years ago are *prima facie* still optimal today.

Thus, the argument here is not that mass-univariate analyses are wrong or misleading; rather, the argument is that progress in neuroscience will be accelerated by shifting to ways of conceptualizing and analyzing data that are based on isolating and extracting information that is distributed across a collection of electrodes.

(Although the focus of this paper is on electrophysiology, the previous argument applies to any set of brain or behavioral manifest variables, including neurons, fMRI voxels, 2p imaging pixels, reaction time, questionnaire items, etc.)

To be sure, there are myriad modern data analysis methods that leverage increased understanding of neurophysiology and computational power. Here I will focus on one family of spatial multivariate analyses, which are used as dimension-reducing spatial filters. A spatial filter is a set of weights, such that the weighted combination of the manifest variables optimizes some criteria. That weighted combination results in a reduced number of data channels (often called "components") compared to the original dataset, and are designed such that each component isolates a pattern in the data that may be difficult to identify in any individual manifest variable.

The purpose of this tutorial paper is to explain one family of spatial filters that has been successful in electrophysiology data. This family of methods is based on a unifying mathematical framework: generalized eigendecomposition (GED). The GED seed can sprout many seemingly different multivariate applications, which makes it a powerful method to adapt to specific hypotheses, research goals, and dataset types. There are several other excellent general introductions to GED in the neuroscience literature (Parra et al., 2005; de Cheveigné and Parra, 2014; Haufe et al., 2014); the focus of the present paper is on practical issues that researchers might face when implementing GED.

1.2. A multitude of multivariate analyses

There are many multivariate analysis methods that have been introduced and validated on electrophysiology data. Below is a brief description of several commonly used methods. I focus on their limitations only to facilitate a contrast with the advantages of GED, without implying that these methods are flawed or should be avoided.

Principal components analysis (PCA) is a popular dimension-reduction method that finds a set of channel weights, such that the weighted combination of channels has maximal variance while forcing all components to be mutually orthogonal. PCA has three limitations with regards to source separation: It is descriptive as opposed to inferential, the components are constrained to be orthogonal, and maximizing variance does not necessarily maximize relevance (de Cheveigné and Parra, 2014).

Independent components analysis (ICA) is nearly ubiquitously used in M/EEG research to project out artifacts such as eye blinks and muscle activity. It is also used for data analysis (Debener et al., 2010), but less frequently than for data cleaning. ICA is generally used as a descriptive measure, although cross-validation methods exist for evaluating statistical significance (Hyvärinen, 2011). In our hands, ICA tends to have low accuracy at recovering ground-truth patterns in simulated data (Cohen, 2017a; Zuure and Cohen, 2021).

Decoding, a.k.a. multivariate pattern analysis, involves using machine-learning methods to classify experiment conditions (e.g., a particular motor response or visual stimulus) based on weighted combinations of brain signals (Cichy and Pantazis, 2017; King and Dehaene, 2014). Some linear classifiers (e.g., Fisher linear discriminant analysis) are built on GED, however, decoding methods generally threshold and binarize the data, thus discarding a considerable amount of rich and meaningful spectral and temporal variability.

Deep learning is a framework for mapping inputs to outputs, via myriad simple computational units, each of which implements a weighted sum of its inputs plus a nonlinearity (Schirrmeister et al., 2017). Deep learning has been transformative in many computational fields including computer vision and language translation. Outside of the visual system, deep learning applications in neuroscience have not (yet) made a major impact. Part of the difficulty of deep learning in neuroscience is that its representations are complex, nonlinear, and difficult to interpret. In other words, deep learning has major applications in society and engineering, but is (so far) of limited value for providing mechanistic insights (the same can be said of decoders more generally) (Ritchie et al., 2019). Furthermore, deep learning tends to underperform on problems that are simple or that have linear solutions.

There are several other multivariate components analyses that have unique applications but are less commonly used, such as factor analysis, Tucker decomposition, and nonnegative matrix factorization.

An exhaustive list and comparison of multivariate components analyses is beyond the scope of this paper.

1.3. Motivation for and advantages of GED

GED for source separation of multichannel data has several advantages.

First, it is based on specifying hypotheses. A cornerstone of experimental science is generating and testing null and alternative hypotheses. This manifests in statistical comparisons as a difference or a ratio between the means of two sample distributions. GED involves comparing two covariance matrices that are created from different features of the data — a "null" feature (later termed a "reference" feature) and an "alternative" feature (later termed a "signal" feature). As will be explained in a later section, the GED-defined spatial filters are hyperplanes that maximize the ratio of alternative:null (signal:reference). When the two covariance matrices — the two data feature distributions — are equivalent, the GED returns a ratio coefficient of 1, which is the expected null-hypothesis value.

Second, because of the inherent comparison of two covariance matrices, GED allows for inferential statistics to determine whether a component is significant. Methods for statistical inference are described in a later section.

Third, GED has few analysis parameters, which makes it easy to learn, apply, and adapt to new situations. Having few parameters also reduces researcher degrees of freedom and therefore increases reproducibility.

Fourth, and related to the previous point, there are no spatial or anatomical constraints. The order and relative locations of the physical data channels (electrodes, sensors, pixels, or voxels) is completely irrelevant to the analysis. This means that the spatial maps can be physiologically interpreted without concern for trivial biases imposed by an *a priori* anatomical model.

Fifth, GED allows for individual differences in topographies. For example, alpha-band activity might be maximal at electrodes Pz, POz, Oz, PO7, etc., in different individuals, leading to possible difficulties and subjectivity with electrode selection. A GED that maximizes alpha-band activity allows for idiosyncratic functional-anatomical distributions for different individuals, while ensuring that the components across

all individuals satisfy the same statistical criteria.

Sixth, GED is deterministic and non-iterative. This means that repeated decompositions of the same data give the same solution (this can be contrasted, for example, with ICA algorithms that are initialized with random weights). And it means that GED is fast. Indeed, the GED typically takes a modern computer a few milliseconds to compute; most of the total analysis time comes from data preparation such as temporal filtering.

Finally, GED has a long history of applications in statistics, machine learning, engineering, and signal processing (Parra et al., 2005). Although not always called "GED," generalized eigendecomposition provides the mathematical underpinning of many analysis methods, including linear discriminant analysis, common-spatial pattern (used in brain-computer interface algorithms), blind-source separation (Tomé, 2006; Blankertz et al., 2007; Parra and Sajda, 2003), and other methods discussed later. Although there are "tricks" for optimizing GED for specific applications, the general approach is well established in multiple areas of science.

1.4. GED in the wild

GED is widely used in neuroscience, although the terminology differs. In brain-computer interfaces, GED is called common spatial patterns (Blankertz et al., 2007) and is used to design spatial filters that facilitate neural control over a computer program. Nikulin et al. (2011) used GED to design a narrowband spatial filter, which they termed spatio-spectral decomposition. This method was extended to use broadband energy compared to narrow neighboring energy, over a successive range of center frequencies, by de Cheveigne, which they called spectral scanning (de Cheveigné and Arzounian, 2015) or, more generally, joint decorrelation (de Cheveigné and Parra, 2014). Dähne et al. (2014) used GED to design a spatial filter that maximizes a correlation between EEG and a behavioral measure like reaction time. Cohen (2017b) adapted GED to identify multivariate cross-frequency coupling. Several groups have used GED to optimize ERPs, which is particularly useful for single-trial analyses (Rivet and Souloumiac, 2013; Tanaka and Miyakoshi, 2019; Das et al., 2020). GED has also been used to obtain components that have maximal signal-to-noise characteristics in steady-state evoked potential studies (Dmochowski et al., 2015; Cohen and Gulbinaite, 2017).

There are many other applications of GED in electrophysiology research; the goal here is not to cite all of them, but instead to highlight that GED is a core part of the corpus of multivariate neuroscience analyses, even if its inclusion is not apparent from the titles of research papers.

1.5. Brief overview of GED

Before delving into the details, it will be useful to have the "bird's eye view" of GED (Figure 2). GED is a decomposition of two covariance matrices, here termed $\mathbf{S}$ and $\mathbf{R}$. These two covariance matrices come from different features of the data: an experiment condition and a control condition, a prestimulus period and a poststimulus period, or narrowband filtered and unfiltered data. The $\mathbf{S}$ matrix is the covariance of the "signal" data feature of interest, and the $\mathbf{R}$ matrix is the covariance of the "reference" data that provides a comparison.

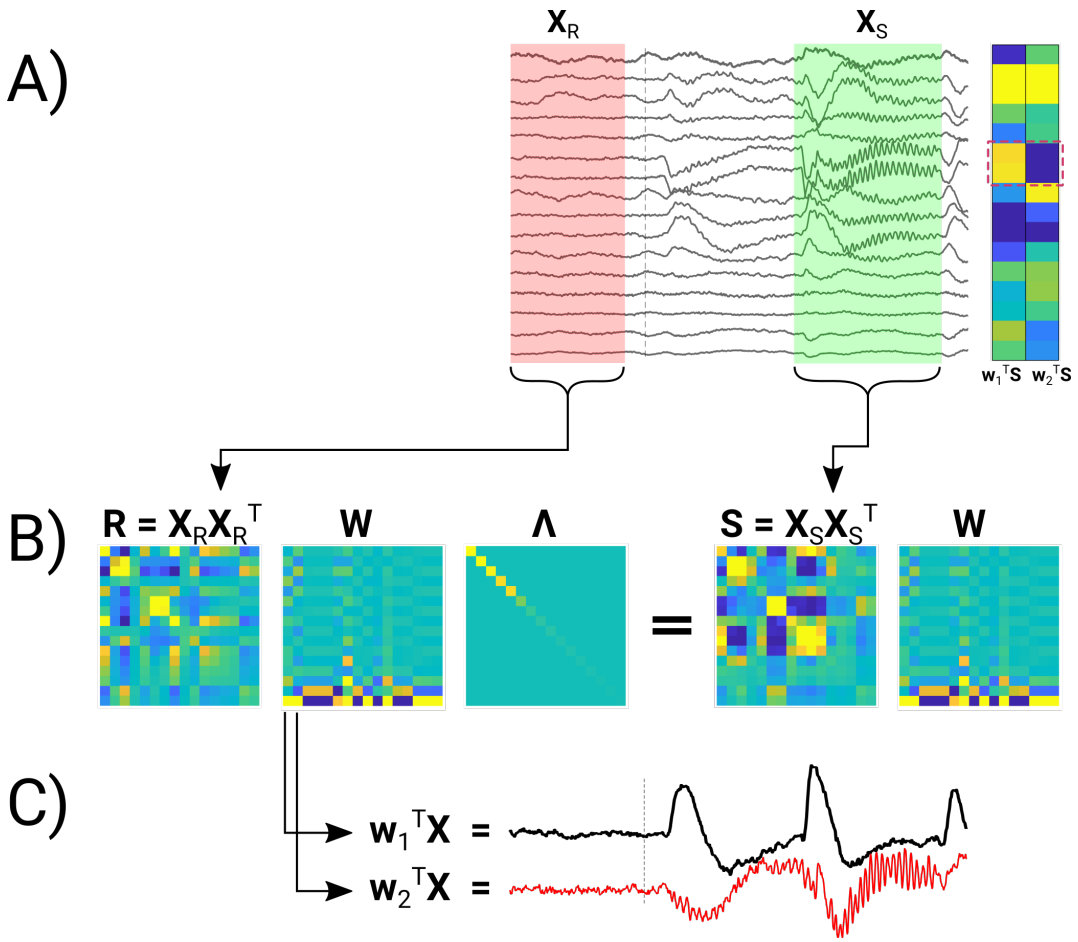

Figure 2: Graphical overview of GED. This example shows laminar recordings in mouse V1 after a visual stimulus onset. A) Two time windows are selected for the "reference" ($\mathbf{X_R}$) and "signal" ($\mathbf{X_S}$) features of the data. B) A generalized eigenvalue decomposition is performed on the two covariance matrices created from the two data features. The resulting eigenvectors (matrix $\mathbf{W}$) are the spatial filters, and their corresponding eigenvalues (diagonal elements of matrix $\mathbf{\Lambda}$) encode the ratio of $\mathbf{S}/\mathbf{R}$ along each direction $\mathbf{w_i}$. C) Each eigenvector times the data produces a component, with an accompanying time series and spatial map (spatial maps are visualized to the right of the time series in panel A). In this example, the first component captured a low-frequency response to the visual stimulus, while the second component isolated the gamma-band response. The red dashed box in the spatial maps indicates the approximate location of cortical layer 4.

The GED finds a weighting of the data channels that maximizes the ratio $\mathbf{S}/\mathbf{R}$. Data features that are common between $\mathbf{S}$ and $\mathbf{R}$ are ignored.

The GED returns a set of channel weights (eigenvectors) and accompanying SNR values (eigenvalues). The channel weights associated with the largest eigenvalue is the spatial filter, and the weighted sum of all channel time series is the component time series that maximizes the researcher-defined criteria

established by selecting data for the two covariance matrices. That component time series can be used in standard analyses such as ERP or time-frequency analysis, and its accompanying topography can be visualized for topographical or anatomical interpretation.

2. Mathematical and statistical aspects

2.1. The math of GED

GED is a covariance matrix-based decomposition. A covariance matrix is an $M \times M$ matrix in which each element contains the covariance between channel M_i and M_j. Thus, the entire matrix contains all linear pairwise interactions. A covariance is simply a non-normalized Pearson correlation coefficient, and thus a covariance matrix is a correlation matrix that retains the scale of the data (e.g., μV).

If the data are organized in a channels-by-time matrix $\mathbf{X}$, then the covariance matrix is given by

$$\mathbf{C} = \mathbf{X}\mathbf{X}^{\mathsf{T}}(n-1)^{-1} \tag{1}$$

In general, the size of the covariance matrix is features-by-features (that is, channels-by-channels), and so the multiplication would be expressed as $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ if the data were organized as time-by-channels. The division by $n-1$ is a normalization factor that prevents the covariance from increasing simply by increasing the number of observations (time points).

The data must be mean-centered before the multiplication. Mean-offsets in the data will cause the GED solutions to point in the direction of the offsets instead of the direction that maximizes the desired optimization criteria. Mean-centering means that the average value of each channel, over the time window used to compute the covariance matrix, is zero.

Variance normalization is not necessary if all channels are in the same scale (e.g,. microvolts). Multi-modal data might need within-modality normalization, which is discussed in section 6.1.

MATLAB code for computing the covariance matrix follows (variable `data` is channels-by-time).

```
data = data-mean(data,2); % mean-center
```

```
S = data*data' / (size(data,2)-1);
```

As introduced in the previous section, GED requires two covariance matrices: the $\mathbf{S}$ matrix computed from features of the data to highlight, and the $\mathbf{R}$ matrix computed from features of the data to use as reference ($\mathbf{S}$ for *signal* and $\mathbf{R}$ for *reference*). A simple example of how to create these two covariance matrices is that $\mathbf{S}$ is computed from a time window during stimulus presentation, and $\mathbf{R}$ is computed from a time window before stimulus presentation (the inter-trial interval). The resulting GED will therefore isolate the covariance patterns that maximally differentiate stimulus processing compared to the inter-trial-interval, while excluding any patterns in the data that are present before and during the stimulus period, such as ongoing spontaneous activity or slower neurocognitive processes that are unrelated to stimulus processing.

Once the two covariance matrices are formed, the goal is to find an M-element vector of weights (called vector $\mathbf{w}$; each element w_i is the weight for the i^{th} data channel), which acts as a *spatial filter* that reduces the dimensionality of the data from M channels to 1 component. The elements in $\mathbf{w}$ are constructed such that the linear weighted sum over all channels maximizes the ratio of "multivariate energy" in $\mathbf{S}$ to that in $\mathbf{R}$. Multivariate energy is expressed through the quadratic form $\mathbf{w}^\mathsf{T}\mathbf{S}\mathbf{w}$. The quadratic form is a single number that encodes the variance in matrix $\mathbf{S}$ along direction $\mathbf{w}$. Therefore, the goal of GED is to maximize the ratio of the quadratic forms of the two matrices.

$$\lambda = \frac{\mathbf{w}^\mathsf{T}\mathbf{S}\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{R}\mathbf{w}} \tag{2}$$

This expression is also known as the Rayleigh quotient. λ is the ratio of signal-to-reference energies along direction $\mathbf{w}$. Note that when the data covariance matrices are the same, $\lambda = 1$, which can be considered the null-hypothesis value ($\mathrm{H}_0 : \mathbf{S} = \mathbf{R}$).

The goal of GED is to find the vector $\mathbf{w}$ that maximizes λ. This is the objective function.

$$\arg\max_{\mathbf{w}} \frac{\mathbf{w}^\mathsf{T}\mathbf{S}\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{R}\mathbf{w}} \tag{3}$$

One can imagine a calculus-based solution to this optimization using Lagrange multipliers and the derivative with respect to $\mathbf{w}$ (Ghojogh et al., 2019). However, the linear algebra solution comes from considering a full set of M vectors, which means that equation 2 can be rewritten as follows.

$$\mathbf{\Lambda} = (\mathbf{W}^\mathsf{T}\mathbf{R}\mathbf{W})^{-1}\mathbf{W}^\mathsf{T}\mathbf{S}\mathbf{W} \tag{4}$$

where each column of $\mathbf{W}$ is a spatial filter, and each diagonal element of $\mathbf{\Lambda}$ is the multivariate ratio in the direction of the corresponding column in $\mathbf{W}$. Some algebraic manipulations brings us to the solution to the optimization problem:

$$\mathbf{\Lambda} = (\mathbf{W}^\mathsf{T}\mathbf{R}\mathbf{W})^{-1}\mathbf{W}^\mathsf{T}\mathbf{S}\mathbf{W} \tag{5}$$

$$\mathbf{\Lambda} = \mathbf{W}^{-1}\mathbf{R}^{-1}\mathbf{W}^{-\mathsf{T}}\mathbf{W}^\mathsf{T}\mathbf{S}\mathbf{W} \tag{6}$$

$$\mathbf{\Lambda} = \mathbf{W}^{-1}\mathbf{R}^{-1}\mathbf{S}\mathbf{W} \tag{7}$$

$$\mathbf{R}\mathbf{W}\mathbf{\Lambda} = \mathbf{S}\mathbf{W} \tag{8}$$

Equation 8 is known as a generalized eigendecomposition on matrices $\mathbf{S}$ and $\mathbf{R}$. This means that the set of weights that maximizes the multivariate signal-to-noise ratio — the spatial filter — is an eigenvector, and the value of that ratio is the corresponding eigenvalue. It is also useful to see that equation 8 is conceptually equivalent to a "regular" eigendecomposition on a matrix product.

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{S} \tag{9}$$

$$\mathbf{W}\mathbf{\Lambda} = \mathbf{B}\mathbf{W} \tag{10}$$

One of the limitations of PCA is that the eigenvectors matrix is orthogonal, meaning that all components must be orthogonal. That constraint comes from the eigendecomposition on a symmetric matrix, and all covariance matrices are symmetric. However, the product of two symmetric matrices is generally not symmetric, and so the eigenvectors of a GED may be correlated (though they must be linearly independent as long as the eigenvalues are distinct). Also note that matrix $\mathbf{R}$ is inverted "on paper"; in practice, MATLAB solves the GED problem without explicitly inverting matrices (Ghojogh et al., 2019), which means that GED can be performed on reduced-rank data.

The two matrices that GED returns contain pairs of eigenvectors and eigenvalues, with each eigenvector (spatial filter) having a corresponding eigenvalue (multivariate ratio). In this pairing, the eigenvector $\mathbf{w}$ points in a specific direction in the dataspace, but does not convey the importance of that direction (the $\mathbf{w}$'s are unit-length in the space of $\mathbf{R}$ because $\mathbf{W}^{-1}\mathbf{R}\mathbf{W} = \mathbf{I}$). In contrast, the corresponding eigenvalue λ encodes the importance of the direction, but is a scalar and therefore has no intrinsic direction. The implication of this is that the eigenvector associated with the largest eigenvalue is the spatial filter that maximizes the ratio $\mathbf{S}/\mathbf{R}$. The next-largest eigenvalue is paired with the eigenvector that maximizes that ratio while being linearly independent from the first direction. And so on for all other directions.

The interpretation of each eigenvalue λ as the "multivariate energy ratio" or "multivariate signal-to-noise ratio" comes from re-expressing the Rayleigh quotient with the data matrices instead of their covariance matrices.

$$\lambda = \frac{\mathbf{w}^\mathsf{T}\mathbf{X}_S\mathbf{X}_S^\mathbf{T}\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{X}_R\mathbf{X}_R^\mathsf{T}\mathbf{w}} \tag{11}$$

$$= \frac{(\mathbf{X}_S^\mathbf{T}\mathbf{w})^\mathbf{T}(\mathbf{X}_S^\mathbf{T}\mathbf{w})}{(\mathbf{X}_R^\mathbf{T}\mathbf{w})^\mathbf{T}(\mathbf{X}_R^\mathbf{T}\mathbf{w})} \tag{12}$$

$$= \frac{\|\mathbf{X}_S^\mathbf{T}\mathbf{w}\|^2}{\|\mathbf{X}_R^\mathbf{T}\mathbf{w}\|^2} \tag{13}$$

In other words, λ is the ratio of the magnitude of the "signal" data filtered through $\mathbf{w}$, to the magnitude of the "reference" data filtered through the same $\mathbf{w}$. If the two data matrices are thought of as point-clouds in the channel space, then $\mathbf{w}$ points along the direction where $\mathbf{X}_S$ is as far away from $\mathbf{X}_R$ as possible.

It is not reasonable to expect that all components are meaningful and can be interpreted. Indeed, if the two covariance matrices are similar to each other (e.g., created from two different experiment conditions that are matched on many perceptual and motor features), there may be only one significant component — or possibly no significant components. The GED simply returns all solutions without a p-value or confidence interval that would indicate interpretability. Inferential statistics can be applied to the eigenvalue spectrum, which is described in section 2.4.

GED is easy to implement in MATLAB. The returned solutions are not guaranteed to be sorted by eigenvalue magnitude, and thus it is convenient to sort the solution matrices.

```
[L,W] = eig(S,R);
[eigvals,sidx] = sort(diag(L),'descend');
eigvecs = W(:,sidx);
```

After sorting, the component time series and component map are created by multiplying the top eigenvector with, respectively, the multichannel data and the covariance matrix.

```
comp_ts  = eigvecs(:,1)' * data; % data are chansXtime
```

```
comp_map = eigvecs(:,1)' * S;
```

Trial-related data are often stored as 3D matrices (e.g., channels, time, trials), and so the code may need to be modified. The component map may be divisive-normalized by the quadratic form on $\mathbf{S}$ (Haufe et al., 2014), but this does not change the topography.

The computation of a component time series is an important divergence from typical machine-learning classification or discriminant analyses: The goal of traditional classification analyses is to use $\mathbf{w}$ to binarize the data and predict whether the data were drawn from "class A" or "class B". However, the application of GED described here results in a continuous time series that has considerably richer information than a binary class label.

2.2. Assumptions underlying GED

GED relies on several implicit and explicit assumptions.

1. Sources mix linearly in the physical data channels. This assumption is necessary because GED is a linear decomposition. For M/EEG/LFP, this assumption is readily feasible, because electrical fields propagate instantaneously (within measurement capabilities) and sum linearly (Nunez et al., 2006). For other measurement modalities like fMRI, multiunit activity, or 2-photon imaging, the linear-mixing assumption can be interpreted that spatially distributed manifestations of the neural computations are simultaneously active (within a reasonable tolerance given by the sampling rate) and sum linearly. This assumption appears to be met in fMRI data (van Dijk et al., 2020; Boynton et al., 2012).
2. The data are covariance-stationary. "Stationarity" means that the characteristics of the signal remain the same over time. Obviously the brain is a highly non-stationary complex system. However, the covariance matrices are computed from restricted time windows, typically on the order of a few hundred ms or a few seconds. Note that the assumption of transient stationarity in neural time series data has precedent in time-frequency analysis. Violations of the covariance stationarity assumption may decrease the separability of the covariance matrices, or may lead to multiple significant components from the same decomposition.
3. Covariance is a meaningful basis for source separation. Covariance is a second-order statistical mixed moment, and captures all of the pairwise linear interactions. This is notably distinct from,

for example, ICA, which begins by whitening the data and optimizing based on higher-order moments such as kurtosis or entropy. Validation studies in simulated data confirm the viability of this assumption.

4. The two covariance matrices are meaningfully separable. By carefully selecting which data features are used to create $\mathbf{S}$ and $\mathbf{R}$, the researcher assumes that the spatial filter that maximizes their ratio is meaningful, interpretable, and justified according to the hypotheses and goals of the analysis.

A final, implicit, assumption of GED is that the GED components "look reasonable" to our intuition. Solutions that don't "look nice" are likely to be rejected, or the analysis redone using different parameters or data features. This is actually not unique to GED — it is nearly ubiquitous in science that our expectations, intuitions, and experiences leads us to accept, reject, or modify experiments and data analyses. This is an important aspect of high-quality research, because artifacts, noise, and confounds can be detected by experienced scientists while being undetected by novices. But it is also important to be cognizant of the role — and potential bias — of researcher expectations on data analysis pipelines.

2.3. Understanding and avoiding overfitting

Overfitting is a term in statistics and machine learning that refers to models being optimized for a specific data feature, at the expense of generalizability. Overfitting is potentially dangerous because the model parameters can be driven by noise or other non-reproducible patterns in the data.

On the other hand, overfitting is a powerful and useful approach when used appropriately. GED is based on overfitting a spatial filter that maximizes the contrast $\mathbf{S}/\mathbf{R}$. Thus, using GED involves leveraging overfitting in a beneficial way without introducing systematic biases into the analyses that could confound the results. This requires some additional considerations, compared to blind decomposition methods such as ICA or PCA.

There are four solutions to avoid overfitting causing confounds in the analyses.

1. Apply statistical contrasts that are orthogonal to the maximization criteria. For example, use GED to create a spatial filter that maximizes gamma-band activity across all experiment conditions, and

then statistically compare gamma activity between conditions. In this case, a statistical test of the presence of gamma activity *per se* is biased by the creation of the spatial filter. On the other hand, a statistical test on condition differences in gamma is not biased by the filter construction.

2. Use cross-validation. Cross-validation is a commonly used method in machine learning to avoid over-fitting, and involves fitting the model (the spatial filter) using part of the data and applying the model to a small portion of the data that was not used to train the model. Typical train/test splits are 80/20% or 90/10%. One cross-validation fold would mean including only 10-20% of the experimental data, which is unlikely to provide a sufficient amount of data in typical electrophysiology experiments. An alternative is to use k-fold cross-validation, where the analysis is repeated, each time using a different 90/10 split of the data. Thus, after 10 splits, all of the data are included, without any of the test trials included in the GED construction.
3. Create the spatial filter based on independent data. This is similar to cross-validation and is a technique that is sometimes used in fMRI localizer tasks (for example, having a separate experiment acquisition to identify the fusiform face area). Thus, the GED spatial filter would be created from a different set of trials, possibly in a different experiment block.
4. Apply inferential statistics (via permutation-testing) to evaluate the probability that a component would arise given overfitting on data when the null hypothesis is true. The mechanism for this is described below.

2.4. Inferential statistics

As mentioned earlier, the eigenvectors simply point in a particular direction in the data space, while the eigenvalues encode the importance of that direction for separating $\mathbf{S}$ from $\mathbf{R}$. Thus, the goal of inferential statistics of GED is to determine the probability that the observed eigenvalues could have been obtained from overfitting data when the null hypothesis is true. The expected value of λ is 1 when the two covariance matrices are the same; but in real data, $\mathbf{S}$ and $\mathbf{R}$ can be expected to differ due to sampling variability and noise, even if they are drawn from the same population of data. Thus, the eigenvalues can be expected to be *distributed* around 1. Indeed, it is trivial that roughly 1/2 of the eigenvalues of a GED on null-hypothesis data will be larger than 1.

Inferential statistical evaluation of GED solutions involves creating an empirical null-hypothesis distribution of eigenvalues, and comparing the empirical eigenvalue relative to that distribution. One might

initially think of generating covariance matrices from random numbers. However, this approach would produce an inappropriately liberal statistical threshold, because real data have a considerable amount of spatiotemporal structure that must be incorporated into the null hypothesis distribution (Theiler et al., 1992). Thus, an appropriate approach is to randomize the mapping of data into covariance matrix, without generating fake data.

Consider an experiment with 100 trials, and the GED is based on the pre-trial to post-trial contrast. Each trial provides two covariance matrices, and thus there are 200 covariance matrices in total. Each covariance matrix is randomly assigned to be averaged into $\mathbf{S}$ or $\mathbf{R}$, and a GED is performed. From the resulting collection of M eigenvalues, the largest is selected. This is the largest eigenvalue that arose under the null hypothesis. This shuffling procedure is then repeated 1000 times, with each iteration having a new random assignment of data segment to covariance matrix. The resulting collection of 1000 pseudo-λ's (called ω below to distinguish from the real-data λ) is the distribution of the largest eigenvalues expected under the null hypothesis that $\mathbf{S} = \mathbf{R}$. The empirical λ (that is, the largest eigenvalue obtained from the GED without shuffling) can then be evaluated relative to this distribution as

$$\lambda_z = \frac{\lambda - \overline{\boldsymbol{\omega}}}{\sigma(\boldsymbol{\omega})} \tag{14}$$

where $\overline{\boldsymbol{\omega}}$ indicates the average of the 1000 largest eigenvalues from permutation shuffling, and $\sigma(\boldsymbol{\omega})$ is the standard deviation. A GED component can be considered statistically significant if λ_z exceeds some threshold, for example, $z > 1.63$ corresponding to $p < .05$. Note that these tests are one-tailed, as we are only interested in components that are *larger* than the null-hypothesis distribution.

Unfortunately, things are not always so simple. If $\mathbf{S}$ and $\mathbf{R}$ are not in the same scale, then the expected λ under the null hypothesis might be different from 1. This can occur, for example, if the $\mathbf{S}$ matrix comes from narrowband filtered data while the $\mathbf{R}$ matrix comes from broadband data. A narrowband signal will contain much less variance than the broadband signal from which it was extracted. Randomly shuffling data segments into $\mathbf{S}$ or $\mathbf{R}$ might produce an $\boldsymbol{\omega}$ distribution around 1, whereas the empirical largest λ might be .1 (indicating that the largest narrowband component has 10% of the energy of the broadband component). If overall differences in covariance matrix magnitude are expected, the matrices can be normalized. Normalization of the covariance matrices is discussed in section 3.3.

2.5. Sign uncertainty

Eigenvectors have a fundamental sign uncertainty. That is, because eigenvectors point along a 1D subspace (which is an infinitely long line that passes through the origin of the eigenspace), eigenvector $\mathbf{w}$ is the same as $-\mathbf{w}$. This can cause interpretational and averaging difficulties, because, for example, a P3 ERP component can manifest as a negative deflection.

Sign uncertainties do not affect spectral or time-frequency analyses, but they do affect time-domain (ERP) analyses and topographical maps.

There are two principled methods for fixing the sign of the eigenvector. One is to ensure that the electrode with the largest absolute value in the component map is positive. Thus, if the largest-magnitude electrode has a negative value, the entire eigenvector is multiplied by -1. (This method could be adapted to the ERP, for example ensuring that the ERP deflection at 300 ms is positive.) A second method is to compute a group-averaged ERP or topographical map, correlate each individual subject's data with the group average, and flip the eigenvector sign for any datasets that correlate negatively with the group average. This second method facilitates group-level data aggregation, but one must be cautious to avoid a biased statistical result if the group-level polarity is tested against zero.

3. Practical aspects

3.1. Preparing data for GED

GED doesn't "know" what is real brain signal and what is noise or artifact; it simply identifies the patterns in the data that maximally separate two covariance matrices. Therefore, the data should be properly cleaned prior to GED. This includes rejecting noisy trials, temporal filtering, ICA for projecting out non-brain sources, and excluding bad channels.

Channel interpolation is not necessary, because the interpolated channels are linear combinations of other channels, and thus do not contribute unique information to the decomposition. Removed channels can be interpolated in the component maps to facilitate averaging across subjects.

GED can still produce a good solution with reduced-rank data (discussed in section 3.4), so there is no concern about removing artifact ICs.

For EEG and LFP, the reference scheme does not impact the GED solution. Re-referencing is a linear operation, and GED is a linear decomposition. Of course, the channel weights will differ if a different reference is used. Referencing is a complicated issue in EEG research (Yao et al., 2019), and average reference is often recommended. The only real restriction on re-referencing with GED is that the same reference must be used to compute the GED and the components. For example, one should not construct the GED using earlobe reference and then apply the spatial filter to average referenced data.

It is also possible to clean the covariance matrices by excluding any segment covariances that are "far away" from their mean. This is discussed in section 3.3.

3.2. Selecting data features for S and R covariance matrices

Selecting two data features for the GED to separate is the single most important decision that the researcher makes during a GED-based data analysis. It is also the reason why GED is such a flexible and versatile analytic backbone (de Cheveigné and Parra, 2014; Parra et al., 2005). This means that GED forces researchers to think carefully and critically about their hypotheses and analyses, which is likely to have positive effects on the quality of the research.

The main hard constraint is that the data channels must be the same and in the same order; it is not possible to separate covariance matrices of different sizes, nor is a GED on covariance matrices with channels in different orders interpretable.

In general, the possibilities for selecting data can be categorized as optimizing one of the following.

1. **Condition differences**. The data for the $\mathbf{S}$ covariance matrix come from the condition of interest (e.g., trials with an informative attentional cue) and the data for the $\mathbf{R}$ matrix come from a control condition (e.g., trials with an uninformative attentional cue) (Zuure and Cohen, 2021). One must be mindful that experiment confounds might bias the GED result. For example, if condition $\mathbf{S}$ has faster reaction times than condition $\mathbf{R}$, then the GED result might reflect motor rather than attentional processing.

2. **Task effects**. The data for the $\mathbf{S}$ covariance matrix comes from a within-trial time window (e.g., 0 to 800 ms post-trial-onset) and the data for the $\mathbf{R}$ matrix comes from a pre-trial baseline period (e.g., -500 to 0 ms) (Duprez et al., 2020). When data from all conditions are pooled together, this approach avoids any bias between different conditions.
3. **Spectral contrast**. The data for the $\mathbf{S}$ matrix are narrowband filtered in some range of interest (e.g., alpha, 10 Hz), and the data for the $\mathbf{R}$ matrix are broadband (de Cheveigné and Arzounian, 2015) or narrowband from neighboring frequencies (Nikulin et al., 2011).

3.3. Computing the covariance matrices

A covariance matrix requires at least two time points, though obviously longer time windows will lead to more stable estimates of the true covariance structure. The quality of the GED depends entirely on the quality of the covariance matrices, so it is important to make sure that the covariance matrices are made from clean data.

One way to increase the stability of a covariance matrix is to increase the number of time points in the data segment. However, the size of the time window presents a trade-off between cognitive specificity vs. covariance quality: Shorter time windows (e.g., 100 ms) better isolate phasic sensory/cognitive/motor events, but risk a noisier covariance matrix. On the other hand, longer time windows (e.g., 1000 ms) increase the quality of the covariance matrix but may span multiple distinct task events. Keep in mind, though, that the temporal resolution of the component time series is that of the data, and is not limited by the time windows used to create the covariance matrices (this is because the spatial filter is applied to the entire time series data, not only the data within the covariance time window). Thus, I tend to err on the side of longer time windows, typically 500-800 ms for a task-related design, or 2000 ms for spontaneous or resting-state data.

If the covariance matrix is computed from temporally narrowband filtered data, then the time windows should be at least one cycle, and preferably longer. For example, if the channel data are filtered at 4 Hz, then the time window to compute the covariance matrix should be at least 250 ms.

Ideally, the data for the two covariance matrices are of comparable quality. When possible, match the amount of data used, which means a comparable number of trials and/or a comparable number of time

points.

We have gotten better results by computing N covariance matrices of N data segments (for example, from N trials) and then averaging together, versus concatenating all segments into one wide data matrix (of size channels-by-timetrials) and then computing one covariance matrix. When working with continuous data (e.g., resting-state or very long experiment trials), the time series can be segmented into epochs of, say, two seconds, thus producing multiple small covariance matrices. This is analogous to using Welch's method for spectral analysis instead of computing one Fourier transform of a long time window.

Each individual data segment must be mean-centered before its covariance is computed. Illustrative MATLAB code shows how this can be implemented (`data` is a variable with dimensions: channel, time, trials).

```
covmat = zeros(nbchans);
for triali=1:ntrials
    seg = data(:,:,triali); % extract one trial
    seg = seg - mean(seg,2); % mean-center
    covmat = covmat + seg*seg'/(size(seg,2)-1);
end
covmat = covmat / triali;
```

Computing N covariance matrices also allows for an additional data cleaning step based on distance. In particular: The average covariance matrix $\overline{\mathbf{S}}$ is computed, and then the Euclidean distance (or any other distance metric) between each segment's covariance matrix $\mathbf{S}_n$ and $\overline{\mathbf{S}}$ is computed. Those N distances can be z-scored, and any covariance matrices with an excessive distance (e.g., $z > 2.31$ corresponding to $p < .01$) are excluded, and a new average $\overline{\mathbf{S}}$ is re-computed. The justification of this procedure is that large-distance covariance segments are multivariate outliers and may skew the results.

Because covariance matrices retain the scale of the data, normalization is not necessary. Even if $\mathbf{S}$ and $\mathbf{R}$ are in different scales, normalization may be unnecessary. This is because in most cases, the relative eigenvalues are important (e.g., for sorting), not the numerical values. Normalization is necessary only

when (1) permutation testing is used while $\mathbf{S}$ and $\mathbf{R}$ are in different scales, or (2) data are combined from different modalities that have very different numerical ranges (e.g., EEG and MEG).

It is possible to z-normalize each data channel, however, this will change the covariance matrix and thus can affect the GED solution. This is because z-normalizing each channel separately alters the magnitude of the between-channel covariances. In other words, channels with low variance are inflated whereas channels with high variance are deflated. Of course, this is the goal of z-normalizing, but some aspect of channel differences in variance are meaningful, for example alpha-band variance is higher at posterior-central channels compared to lateral temporal channels.

An alternative is to mean-center each channel separately, and then divide all channels by their pooled standard deviation. This approach preserves the relative covariance magnitudes within modality, while simultaneously ensuring that the total dataset has a pooled standard deviation of 1.

3.4. Regularization

Regularization involves adding some constant to the cost function of an optimization algorithm. Regularization has several benefits in machine learning, including "smoothing" the solution to reduce overfitting and increasing numerical stability, particularly for reduced-rank datasets.

There are several forms of regularization, including L1 (a.k.a. Lasso), L2 (ridge), Tikhonov, shrinkage, and others. Although there are mathematical differences between different regularization techniques, it is often the case that various methods produce comparable benefits (Wong et al., 2018, e.g.,).

Here I focus on shrinkage regularization, because it is simple and effective, and commonly used in GED applications (Lotte and Guan, 2010). Shrinkage regularization involves adding a small number to the diagonal of the $\mathbf{R}$ matrix (and, thus, $\widetilde{\mathbf{R}}$ replaces $\mathbf{R}$ in equation 8). That small number is some fraction of the average of $\mathbf{R}$'s eigenvalues.

$$\widetilde{\mathbf{R}} = \mathbf{R}(1-\gamma) + \gamma\alpha\mathbf{I}_M \tag{15}$$

$$\alpha = \sum_{i=1}^{M} \lambda_i \tag{16}$$

$\mathbf{I}_M$ is the $M \times M$ identity matrix, α is the average of all eigenvalues of $\mathbf{R}$, and γ is the regularization

amount. One should use as little regularization as possible but as much as necessary. A common value is $\gamma = .01$, which corresponds to 1% regularization.

Scaling down the $\mathbf{R}$ matrix by 1-γ ensures that the trace of $\widetilde{\mathbf{R}}$ and $\mathbf{R}$ are the same. This is useful because the trace is the sum of all eigenvalues, and thus the total "energy" of the eigenspectrum is preserved before and after regularization.

The MATLAB code is a direct implementation of equation 15.

```
gamma = .01;
Rr = R*(1-gamma) + gamma*mean(eig(R))*eye(length(R));
```

Our experience is that 1% shrinkage regularization noticeably qualitatively improves the solution for matrices that are reduced-rank, noisy, or are difficult to separate, while has little or no noticeable effect on the GED solution for clean and easily separable matrices. Note that when $\gamma = 1$, GED becomes a PCA. Thus, an interpretation of shrinkage regularization is that it pushes the GED slightly towards favoring high-variance solutions at the potential cost of reduced separability between $\mathbf{S}$ and $\mathbf{R}$.

Unpublished investigations in my lab (Gerova, 2021) comparing robust covariance estimators (standard covariance, 1% shrinkage, minimum covariance determinant, orthogonalized Gnanadesikan-Kettenring, and Olive-Hawkins) in simulated EEG data indicate that 1% shrinkage outperforms other methods, particularly for reduced-rank data contaminated by pink noise.

3.5. Which component to use

Theoretically, the component with the largest eigenvalue has the best separability. However, this should be visually confirmed before applying the spatial filter to data and interpreting the results, because the component that mathematically best separates two data features might not be a component of interest. For example, the top component in a GED that maximizes low-frequency activity may reflect eye-blink artifacts not entirely removed by ICA. In my lab, we typically produce a MATLAB figure for each dataset that shows the eigenspectrum, topoplots, and ERPs from the first 5 components (Figure

3). Our default choice is the largest component, but we sometimes select later components based on topography or ERP.

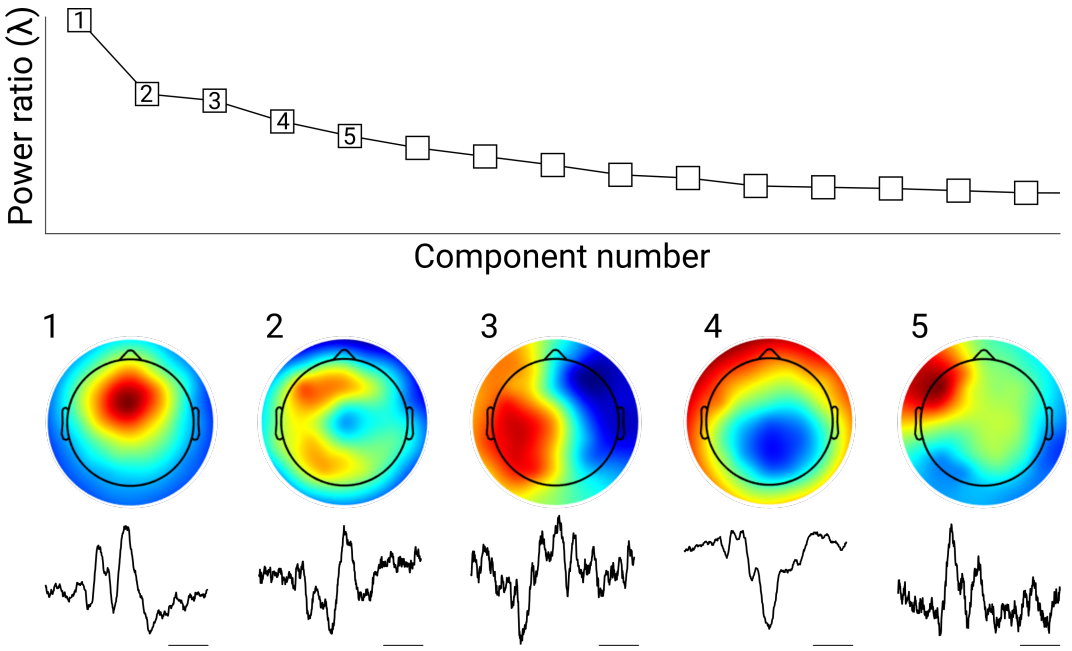

Figure 3: Picking a component based on visual inspection. A GED was performed on 64-channel EEG with the $\mathbf{S}$ matrix computed from 6-Hz filtered activity and the $\mathbf{R}$ matrix computed from broadband activity. Both matrices were created from data in a 100-800 ms window relative to trial onset. The top row shows the eigenspectum (also called a scree plot), with 15/64 components shown. The middle row shows the component maps, and the bottom row shows the ERP (the spatial filter was created from narrowband data but was applied back to the broadband signal). Horizontal bar indicates 300 ms. In this example, the first component isolated midfrontal theta. Component 4 appears to reflect parietal theta.

Because there is theoretical motivation and mathematical justification for selecting the component with the largest eigenvalue, it is not necessary to statistically evaluate the top component, unless there is a cause for concern about overfitting. For example, if a component is created based on maximizing alpha-band activity over all conditions, the top component can be computed and the inferential statistical analyses can be done on the differences in alpha power or phase dynamics across the experiment conditions, and not on the existence of the component *per se*. Analogously, in univariate analyses, one typically selects a data channel and then applies a statistical comparison across conditions, without necessarily testing whether that channel has a significant amount of activity per se.

3.6. Applying the spatial filter to the data

Each eigenvector (each column of the $\mathbf{W}$ matrix) is a set of weights for computing the weighted average of all data channel time series. This is the reason that the eigenvector is called a *spatial filter*. The vector-data multiplication reduces the dimensionality of the data from $M \times T$ to $1 \times T$ (for M channels and T time points).

The spatial filter is created from one feature of the data, but can be applied to other data. For example, a spatial filter created by contrasting 10 Hz activity ($\mathbf{S}$ matrix) to broadband activity ($\mathbf{R}$ matrix) can be applied to the broadband signal (this is done in the validation simulation, shown later). In this case, the component is designed to optimize alpha-band activity, but it is not constrained to pass through only alpha-band activity; activity in lower or higher frequencies may also be observed in the component time series. The interpretation of, for example, 40 Hz energy in the alpha-band filter would be that the higher-frequency activity, though spectrally separate from alpha, has a similar topography as alpha, and thus passes through the alpha-optimized spatial filter.

Applying the spatial filter to data beyond what was used to construct the $\mathbf{S}$ matrix also reduces the potential for bias or overfitting.

Finally, the filter must be applied to data with the same channels in the same order as was used to create the covariance matrices.

The component time series, computed as $\mathbf{w}^\mathsf{T}\mathbf{X}$, does not natively have the same units as the data $\mathbf{X}$ (e.g., μV or fT). This is because MATLAB scales the eigenvectors to be unit-norm in the space of $\mathbf{R}$, resulting from the constraint that $\mathbf{W}^\mathsf{T}\mathbf{R}\mathbf{W} = \mathbf{I}$ (the denominator of the Rayleight quotient). This means that the norm of the eigenvector is not necessarily 1, which means that $\|\mathbf{w}^\mathsf{T}\mathbf{X}\| \neq \|\mathbf{X}\|$.

The implication is that the component time series does not have the same units as the original data. For some applications, the units don't matter. Indeed, units do not affect qualitative or statistical comparisons across conditions or over time. Furthermore, spectral and time-frequency analyses usually involve normalization to decibel or percent change, meaning that the original units don't matter. Finally, measures of phase dynamics, synchronization, and correlation are unitless, and thus data scaling has no effect.

Nonetheless, it is sometimes useful to have the component time series in the scale as the original data. The solution is to unit-norm the eigenvector. The formula and MATLAB code are shown below.

$$\widetilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\| \tag{17}$$

```
w = evecs(:,1); % first eigenvector
w = w/norm(w); % scale to unit norm
```

Note that because the component reflects a weighted combination of data across all channels — and may include both positive and negative weights — the numerical values of the component time series are likely to be smaller than the numerical values at a given electrode. The numerical values at each electrode are the sum of many separate sources, while each GED component is one isolated source.

3.7. Complex GED solutions

The GED may return complex solutions, which means complex-valued eigenvalues and eigenvectors. When there are complex solutions to an eigenvalue problem, they appear in conjugate pairs. It is possible that the complex solutions are towards the bottom, and the largest components have zero-valued imaginary parts.

Although there is nothing mathematically wrong with complex solutions to a GED, it is usually a bad sign: It indicates that there was no good solution in the real domain, so the eigenvectors started rotating into the complex domain. This can arise in the presence of noisy covariance matrices, or when the $\mathbf{S}$ and $\mathbf{R}$ matrices were too close to each other.

When complex solutions are observed, one can consider using more data to create the covariance matrices (e.g., longer time windows or wider spectral bands) or redefining the optimization criteria (e.g., all conditions against the inter-trial-interval instead of one condition against another). Regularization nearly always leads to real-valued GED solutions, but this should be seen as treating the symptom rather than the cause of the problem.

3.8. Subject-specific or group-level data

If the same electrodes are placed in the same locations in different individuals, then the researcher has the choice to perform the GED separately on each individual, or to perform one GED on covariance matrices that are averaged across individuals. This is analogous to group-ICA, where data from all individuals are pooled to derive a single set of components that is based on data from all subjects.

In my mind, subject-specific GED is more sensible; indeed, one advantage of GED is its adaptability to individual topographical and anatomical variability. On the other hand, group-level GED will have higher signal-to-noise characteristics because of the increase in data, and will highlight the features that are consistent across individuals. Generalization across individuals is one of the goals of neuroscience, which clearly justifies group-level data processing methods. Parra et al. have presented a compelling argument and framework for identifying group-level sources (Parra et al., 2019).

Both individual and group-level GED have advantages and limitations; the decision of which level to implement GED can be made on a case-by-case basis.

3.9. Two-stage compression and separation

A two-stage GED involves (1) data compression via PCA and then (2) source separation via GED. This is useful when there are many data channels or for severely reduced-rank covariance matrices. The initial compression stage should be added to the analysis pipeline only when necessary, e.g., when GED returns unsatisfactory results while the data matrices are very large.

The goal of the first stage is to produce an $N \times T$ dataset, where N is the number of principal components with $N < M$. This is obtained as the eigenvectors matrix $\mathbf{V}$ times the data matrix (note that the eigenvectors $\mathbf{V}$ are from the PCA on the data, not from a GED).

$$\mathbf{Y} = \mathbf{V}_{1:N}^{\mathsf{T}}\mathbf{X} \tag{18}$$

The subscript $1{:}N$ indicates to use the first N columns in the eigenvectors matrix. The number of PCA components to retain (N) can be based on one of two factors. First, one can use the rank of the data matrix as the dimensionality. The advantage of this approach is that it prevents information loss

in the compression. In other words, the original and compressed data have the same information, but the post-PCA data matrix has fewer dimensions. This guarantees full-rank matrices in the GED and thus should improve numerical stability of the solution.

A second method is to select the number of compressed dimensions based on a statistical criterion. If the eigenvalues of the PCA are converted to percent total variance explained, then a variance threshold of, e.g., 0.1% can be applied. The compressed signal thus comprises only principal components that contribute more than 0.1% of the multivariate signal space. In this case, the data matrix used for GED contains less information than the original data. The motivation for this approach is that principal components that account for a tiny amount of variance might reflect noise or unrelated activity. On the other hand, the PCA is driven by total variance — and variance does not equal relevance (de Cheveigné and Parra, 2014). Thus, there is a risk that some of the information that was removed is important for separating $\mathbf{S}$ from $\mathbf{R}$.

GED then proceeds as described in the rest of this tutorial except using data matrix $\mathbf{Y}$ instead of $\mathbf{X}$.

It will be desirable to have the component maps in the original channel space (instead of the compressed PC space, which is unlikely to be anatomically interpretable). This is obtained by projecting the covariance matrix first through the GED eigenvector as described earlier, and then again through the PCA eigenvectors to "undo" the first-stage compression: $\mathbf{w}^\mathsf{T}\mathbf{S}\mathbf{V}_{1:N}$.

4. Validation in simulations

We and others have done many simulations over the years to demonstrate that GED is highly accurate at reconstructing simulated activity. Some of these simulations are published (Cohen, 2017a), many others were done during piloting, testing, explorations, and teaching. Simulations have the advantages of allowing full control over variables such as the amount and color of noise, the signal strength and characteristics, the number of trials and electrodes, and so on. On the other hand, simulations rarely capture the full complexity of EEG data and accompanying artifacts, and thus one should be not assume that GED (or any other method) always identifies The Truth in empirical data simply because it performed well in a simulation.

The goal of this section is to present a general framework for simulating data to evaluate GED. Researchers should modify this skeleton framework for their purposes, data characteristics, and methods.

Our data simulations involve simulating time series signals and noise at hundreds or thousands of dipole locations in the brain, passing those time series through an anatomical forward model (a.k.a. a leadfield matrix) to generate EEG activity at simulated electrodes, and then proceed to analyze the electrode-level activity. The results can be quantitatively compared to the signals that were generated in the dipoles. This pipeline is useful because it is fast and efficient, allows the researcher to control the strength, nature, and locations of the signal and noise, while also providing physiologically plausible EEG topographical characteristics.

On the other hand, we do not use biophysical models to simulate the data. Thus, we use simulated data only to evaluate the performance of GED or other analysis methods, not to make claims about neural computations. However, it is now feasible to combine biophysical computational models with anatomical forward models to generate more realistic EEG data (Næss et al., 2021).

The MATLAB code that accompanies this tutorial (available at github.com/mikexcohen/GEDtutorial) includes two simulations (no additional toolboxes are required). The first is very simple (1000 ms of data, of which 500 ms is a pure sine wave) and illustrates the high reconstruction accuracy of GED and the poor performance of PCA. This provides an introduction to implementing simulated data and performing GED, and is now shown here.

The second simulation shows a better use-case: isolating an alpha-band component embedded in noise during simulated resting-state data. Figure 4A shows the projection of the "alpha dipole" activation onto the scalp. A GED was computed on narrowband-filtered signals compared to broadband. Figure 4B shows the scree plot, which clearly indicates one component that dominates the analysis. For reference, a traditional univariate analysis was applied (Welch's method on each electrode). At this SNR level, the true alpha dynamics are barely visible in the scalp and power spectrum. At higher signal gain values, the electrode data more accurately recovered the ground truth activity (not shown here, but easy to demonstrate with a minor adjustment to the code).

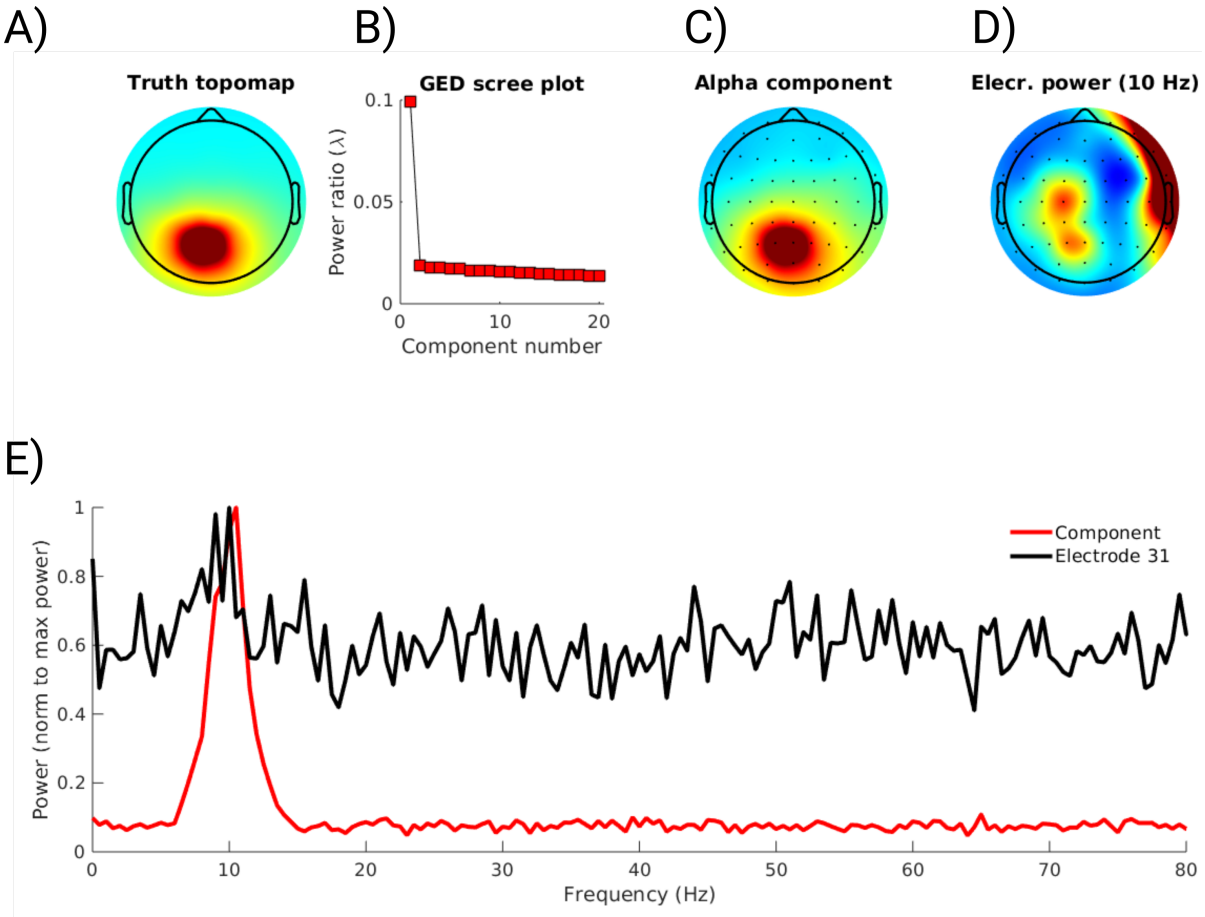

Figure 4: EEG simulation demonstrates the advantage of GED over electrode-level analyses. Data were generated in 2004 brain dipoles and projected to 64 scalp electrodes (black dots). One dipole in parietal cortex was simulated as a non-stationary alpha oscillation with .5 gain; all other dipoles contained Gaussian random noise with a $\sigma = 1$ standard deviation. The dipole projection is visualized in panel A. 100 segments of 2 seconds each were generated to simulate resting-state EEG data. A GED was run on 10 Hz filtered signal ($\mathbf{S}$) vs. broadband ($\mathbf{R}$). The eigenspectrum is shown in panel B. Note that all components are much smaller than 1, because there is more energy in the entire spectrum compared to only 10 Hz. The top component had a topography (panel C) that closely matched the ground truth. Power spectra were computed at each of 64 electrodes, and the 10 Hz power is shown in panel D. Panel E shows the power spectra from electrode 31 (closest to the ground truth dipole maximal projection) and the 10 Hz component.

5. Section 5. Interpretation

5.1. Are components dipoles?

GED components are a linear weighted combination of data channels, where the weights are defined by maximizing distance between two covariance matrices. There are no constraints or assumptions about

spatial smoothness or anatomical localization. This means that GED components do not necessarily correspond to a single dipole. And there is no philosophical need for them to be dipolar. This contrasts with a common view of ICA: Indeed, independent components are often evaluated in terms of their "dipolarity" (that is, their fit to a single dipole projection) (Delorme et al., 2012). Although dipolarity may indicate physiological plausibility for sensory or motor processes that are likely to be driven by a spatially restricted population of neurons, neuroscience is increasingly moving towards the idea that many cognitive processes are implemented by spatially distributed networks (Bassett and Sporns, 2017). This suggests that the fit to a single dipole is not necessarily an important consideration when evaluating a spatial filter.

We have observed that GED components are sometimes dipole-like and other times spatially distributed (e.g., figure 3). In simulations, non-dipolar topographies can arise from noisy data, although the component time series accurately reflects the simulated ground truth. Therefore, the component topography is one indicator of the spatial filter quality, but it should not be the sole basis for selecting or rejecting a GED component.

It is possible to project the component map onto the brain given a suitable forward model (Cohen and Gulbinaite, 2017; Hild and Nagarajan, 2009), but this is simply a useful visualization, not a guarantee of an anatomical origin.

The recommendation here is to interpret the GED components in a conservative manner corresponding to how the components were created, which is maximizing a statistical contrast in multivariate data. The physiological interpretation is that each GED component captures a functionally cohesive and temporally coherent network, which could be spatially restricted or spatially distributed.

5.2. Multiple significant components

Although the component with the largest eigenvalue is theoretically the most relevant, it is possible that multiple linearly separable dimensions separate $\mathbf{S}$ from $\mathbf{R}$. As a simple cognitive example, there are multiple perceptual and motor operations that separate task performance from resting state. This was also shown in figure 1. When appropriate, permutation testing can be used to determine the number of statistically significant components from a GED (Zuure et al., 2020).

The exact interpretation of multiple significant components is not entirely clear. The simple interpretation is that they reflect distinct brain features that separate the two covariance matrices. In other words, that there is a multidimensional subspace that separates $\mathbf{S}$ and $\mathbf{R}$, and the number of significant components is the dimensionality of that subspace.

However, it is difficult to rule out alternative hypotheses, for example of a 1-dimensional nonlinear component that is captured by 2-3 linear components (eigendecomposition is a linear decomposition). Distinguishing linear vs. nonlinear dimensions in more than 3 Euclidean dimensions (that is, when the data cannot be visualized in its entirety) is nearly impossible with noisy data. Linear decompositions tend to be more robust and numerically stable, and provide for visually interpretable topographies, and are therefore advantageous. Of course, the linearity of GED only concerns the way the components are created (scalar multiplication and sum); data analyses on those components are not constrained to be linear.

5.3. Interpreting eigenvectors vs. component maps

The anatomical interpretability of the component maps — and the lack of direct anatomical interpretability of the eigenvectors — was elegantly explained by Haufe et al (Haufe et al., 2014). Briefly, eigenvectors contain a mixture of boosting signal and suppressing noise, as well as counteracting "smearing" from volume conduction. A conceptualization is that the eigenvectors represent how to contort the channel data from “outside” to see the underlying source, whereas the component maps represent the underlying source projecting outwards onto the electrodes. This is why the component map is obtained as the covariance matrix "filtered through" (multiplied by) the eigenvector.

That said, the eigenvectors contain rich and high-spatial-frequency information, and have been used to identify empirical frequency band boundaries based on sudden changes in spatial correlation patterns of eigenvectors across neighboring frequencies (Cohen, 2021).

5.4. There is no escaping the ill-posed problem

Finally, it is important to keep in mind that all source separation problems — be they statistical or anatomical — are fundamentally *ill-posed*.

Ill-posed means that there are more unknown factors than observable measurements. Just because GED gives a deterministic solution does not mean that it gives the *correct* solution, or even the only solution. There is an infinite number of possible brain generator configurations that could produce an observed EEG signal; our objective is to select the configuration that best satisfies a set of assumptions (outlined in section 2.2).

Ultimately, the ill-posed problem means that our interpretation of GED (or any other multivariate method) should be appropriately conservative: We cannot know if we are uncovering true sources in the brain, but we can be confident that the components reflect patterns in the data that boost an hypothesis-relevant signal while also minimizing noise and individual variability that contaminate traditional univariate data selection methods.

6. Future directions and conclusions

6.1. Multimodal datasets

GED does not "know" or "care" about the origin or type of data. It is possible to include multiple sources of data into the same data matrix, for example, EEG+MEG (Zuure et al., 2020), LFP+single units (Cohen et al., 2021), EEG+FMRI, or any other combinations of brain, behavior, and body read-outs. Any continuous signals that are relevant for an hypothesis can be included in the covariance matrices. Temporal up/downsampling may be necessarily to have data values at each time point.

Multimodal signals, however, do need to be suitably normalized. For example, EEG data in microvolts and MEG data in Tesla differ in raw numerical values by roughly 14 orders of magnitude. See section 3.3 for discussion of covariance matrix normalizations.

6.2. Better neurobiological and computational interpretations

As a statistical method to facilitate practical data analysis, GED (and many other multivariate techniques) is unambiguously useful. Its mathematical foundations are clear, and uncountable simulation and empirical studies demonstrate its flexibility and power.

However, as neuroscientists, we are not interested in data methods *per se*; we are interested in using the results of data analyses to understand brain function and its role in behavior and disease. Therefore, it would be ideal if GED components could be interpreted in a more physiological manner. It is currently premature to make claims about neural circuit configurations purely on the basis of GED topographies or eigenspectra, but there are two paths to being able to link GED to neural circuit configurations.

The first avenue is empirical multimodal studies. The idea would be to use combined EEG+fMRI, or combined EEG/LFP+spikes. One could then empirically evaluate the sensitivity of GED to distinct spatial or neural configurations.

A second promising avenue for linking GED to neural circuits comes from developments in computational modeling that are increasingly biophysically and morphologically accurate (e.g., Næss et al., 2021). Such generative forward models allow researchers to simulate data at the neural level (including specifying different types of neurons across different cortical layers, density of inter-neuronal connectivity, etc.) and then produce scalp-level EEG data. One could then apply GED to the biophysically simulated EEG data after manipulating neurophysiologically meaningful circuit and connectivity parameters.

6.3. Conclusions

The goal of this paper was to present an approachable and generic tutorial on the use of GED for multivariate source separation. There are several other lucid introductions to the theory and advantages of GED (Parra et al., 2005; de Cheveigné and Parra, 2014; Haufe et al., 2014; Tomé, 2006; Blankertz et al., 2007); the contribution of this paper is to focus on intuition and practical aspects that researchers might encounter while implementing GED. Although the discussion centered on electrophysiology, these methods are generic and can be applied to any type of multichannel time series data, including fMRI or calcium imaging.

I believe that the traditional mass-univariate approach to neuroscience is reaching its limits, and that progress in neuroscience will require multivariate source separation methods. GED is certainly not the only useful multivariate method, nor is it suitable for all situations and datasets. But it has many statistical and practical advantages, and its flexibility and high SNR should make it a tool in the toolbox of every neuroscientist collecting multichannel time series data.

Acknowledgements

Many colleagues and students have helped shape my understanding and practical experience of GED. In addition to the scholars listed in the references section, following is an alphabetized list of people I have personally interacted with who indirectly contributed to this manuscript: Adam Dede, JJ Morrow, Joan Dupre, Lucas Parra, Marrit Zuure, Mihaela Gerova, Rasa Gulbinaite, Vignesh Muralidharan.

References

Bassett, D.S., Sporns, O., 2017. Network neuroscience. Nat. Neurosci. 20, 353–364. doi:10.1038/nn.4502, arXiv:28230844.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Muller, K.R., 2007. Optimizing spatial filters for robust eeg single-trial analysis. IEEE Signal processing magazine 25, 41–56.

Boynton, G.M., Engel, S.A., Heeger, D.J., 2012. Linear Systems Analysis of the fMRI Signal. Neuroimage 62, 975. doi:10.1016/j.neuroimage.2012.01.082.

de Cheveigné, A., Arzounian, D., 2015. Scanning for oscillations. Journal of neural engineering 12, 066020.

de Cheveigné, A., Parra, L.C., 2014. Joint decorrelation, a versatile tool for multichannel data analysis. Neuroimage 98, 487–505.

Cichy, R.M., Pantazis, D., 2017. Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. Neuroimage 158, 441–454. doi:10.1016/j.neuroimage.2017.07.023, arXiv:28716718.

Cohen, M.X., 2017a. Comparison of linear spatial filters for identifying oscillatory activity in multichannel data. Journal of neuroscience methods 278, 1–12.

Cohen, M.X., 2017b. Multivariate cross-frequency coupling via generalized eigendecomposition. ELife 6, e21792.

Cohen, M.X., 2021. A data-driven method to identify frequency boundaries in multichannel electrophysiology data. Journal of Neuroscience Methods 347, 108949.

Cohen, M.X., Englitz, B., França, A.S.C., 2021. Large- and multi-scale networks in the rodent brain during novelty exploration. eNeuro doi:10.1523/ENEURO.0494-20.2021.

Cohen, M.X., Gulbinaite, R., 2017. Rhythmic entrainment source separation: Optimizing analyses of neural responses to rhythmic sensory stimulation. Neuroimage 147, 43–56. doi:10.1016/j.neuroimage.2016.11.036, arXiv:27916666.

Dähne, S., Meinecke, F.C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.R., Nikulin, V.V., 2014. Spoc: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. NeuroImage 86, 111–122.

Das, N., Vanthornhout, J., Francart, T., Bertrand, A., 2020. Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research. Neuroimage 204, 116211. doi:10.1016/j.neuroimage.2019.116211.

Debener, S., Thorne, J., Schneider, T.R., Viola, F.C., 2010. Using ICA for the Analysis of Multi-Channel EEG Data, in: Simultaneous EEG and fMRI. Oxford University Press, Oxford, England, UK, pp. 121–135. doi:10.1093/acprof:oso/9780195372731.003.0008.

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S., 2012. Independent EEG Sources Are Dipolar. PLoS One 7, e30135. doi:10.1371/journal.pone.0030135.

van Dijk, J.A., Fracasso, A., Petridou, N., Dumoulin, S.O., 2020. Linear systems analysis for laminar fMRI: Evaluating BOLD amplitude scaling for luminance contrast manipulations. Sci. Rep. 10, 1–15. doi:10.1038/s41598-020-62165-x.

Dmochowski, J.P., Greaves, A.S., Norcia, A.M., 2015. Maximally reliable spatial filtering of steady state visual evoked potentials. Neuroimage 109, 63. doi:10.1016/j.neuroimage.2014.12.078.

Duprez, J., Gulbinaite, R., Cohen, M.X., 2020. Midfrontal theta phase coordinates behaviorally relevant brain computations during cognitive control. NeuroImage 207, 116340.

Gerova, M., 2021. The effect of robust covariance estimators on the performance of generalized eigenvalue decomposition (GED) in simulated settings. Technical Report. Donders Center for Medical Neuroscience, Radboud University Medical Centre.

Ghojogh, B., Karray, F., Crowley, M., 2019. Eigenvalue and Generalized Eigenvalue Problems: Tutorial. arXiv URL: `https://arxiv.org/abs/1903.11240v1`, `arXiv:1903.11240`.

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, 96–110.

Hild, K.E., Nagarajan, S.S., 2009. Source localization of EEG/MEG data by correlating columns of ICA and lead field matrices. IEEE Trans. Biomed. Eng. 56, 2619–2626. doi:10.1109/TBME.2009.2028615, arXiv:19695993.

Hyvärinen, A., 2011. Testing the ica mixing matrix based on inter-subject or inter-session consistency. NeuroImage 58, 122–136.

King, J.R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends in cognitive sciences 18, 203. doi:10.1016/j.tics.2014.01.002.

Lotte, F., Guan, C., 2010. Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. IEEE Transactions on biomedical Engineering 58, 355–362.

Næss, S., Halnes, G., Hagen, E., Hagler, Jr., D.J., Dale, A.M., Einevoll, G.T., Ness, T.V., 2021. Biophysically detailed forward modeling of the neural origin of EEG and MEG signals. Neuroimage 225, 117467. doi:10.1016/j.neuroimage.2020.117467, arXiv:33075556.

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, 1528–1535.

Nunez, P.L., Nunez, E.P.B.E.P.L., Srinivasan, R., Press, O.U., Srinivasan, A.P.C.S.R., 2006. Electric Fields of the Brain. Oxford University Press, Oxford, England, UK.

Parra, L., Sajda, P., 2003. Blind source separation via generalized eigenvalue decomposition. The Journal of Machine Learning Research 4, 1261–1269.

Parra, L.C., Haufe, S., Dmochowski, J.P., 2019. Correlated components analysis - extracting reliable dimensions in multivariate data. arXiv:1801.08881.

Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. Neuroimage 28, 326–341. doi:10.1016/j.neuroimage.2005.05.032, arXiv:16084117.

Ritchie, J.B., Kaplan, D.M., Klein, C., 2019. Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. British J. Philos. Sci. 70, 581–607. doi:10.1093/bjps/axx023, arXiv:31086423.

Rivet, B., Souloumiac, A., 2013. Optimal linear spatial filters for event-related potentials based on a spatio-temporal model: Asymptotical performance analysis. Signal Processing 93, 387–398.

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. 38, 5391. doi:10.1002/hbm.23730.

Tanaka, H., Miyakoshi, M., 2019. Cross-correlation task-related component analysis (xTRCA) for enhancing evoked and induced responses of event-related potentials. Neuroimage 197, 177–190. doi:10.1016/j.neuroimage.2019.04.049, arXiv:31034968.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Doyne Farmer, J., 1992. Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77–94. doi:10.1016/0167-2789(92)90102-S.

Tomé, A.M., 2006. The generalized eigendecomposition approach to the blind source separation problem. Digital Signal Processing 16, 288–302.

Wong, D.D.E., Fuglsang, S.A., Hjortkjær, J., Ceolini, E., Slaney, M., de Cheveigné, A., 2018. A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding. Front. Neurosci. 12. doi:10.3389/fnins.2018.00531.

Yao, D., Qin, Y., Hu, S., Dong, L., Vega, M.L.B., Sosa, P.A.V., 2019. Which Reference Should We Use for EEG and ERP practice? Brain Topogr. 32, 530–549. doi:10.1007/s10548-019-00707-x.

Zuure, M.B., Cohen, M.X., 2021. Narrowband multivariate source separation for semi-blind discovery of experiment contrasts. Journal of Neuroscience Methods 350, 109063.

Zuure, M.B., Hinkley, L.B., Tiesinga, P.H.E., Nagarajan, S.S., Cohen, M.X., 2020. Multiple Midfrontal Thetas Revealed by Source Separation of Simultaneous MEG and EEG. J. Neurosci. 40, 7702–7713. doi:10.1523/JNEUROSCI.0321-20.2020.