ARTICLE **Communicated by Tatyana Sharpee**

Orthogonal Connectivity Factorization: Interpretable Decomposition of Variability in Correlation Matrices

Aapo Hyvärinen
aapo.hyvarinen@helsinki.fi
Department of Computer Science and HIIT, University of Helsinki, Helsinki FI-00560, Finland

Jun-ichiro Hirayama
hirayama@atr.jp
Neural Information Analysis Laboratories, ATR Institute International, Kyoto 619-0288, Japan

Vesa Kiviniemi
vesa.kiviniemi@oulu.fi
Department of Diagnostic Radiology, Oulu University Hospital, Oulu FI-90220, Finland

Motoaki Kawanabe
kawanabe@atr.jp
Cognitive Mechanisms Laboratories, ATR Institute International, Kyoto 619-0289, Japan

In many multivariate time series, the correlation structure is nonstationary, that is, it changes over time. The correlation structure may also change as a function of other cofactors, for example, the identity of the subject in biomedical data. A fundamental approach for the analysis of such data is to estimate the correlation structure (connectivities) separately in short time windows or for different subjects and use existing machine learning methods, such as principal component analysis (PCA), to summarize or visualize the changes in connectivity. However, the visualization of such a straightforward PCA is problematic because the ensuing connectivity patterns are much more complex objects than, say, spatial patterns. Here, we develop a new framework for analyzing variability in connectivities using the PCA approach as the starting point. First, we show how to analyze and visualize the principal components of connectivity matrices by a tailor-made rank-two matrix approximation in which we use the outer product of two orthogonal vectors. This leads to a new kind of transformation of eigenvectors that is particularly suited for this purpose and often enables interpretation of the principal component as connectivity between two groups of variables. Second, we show how to incorporate

Neural Computation **28, 445–484** (2016)
doi:10.1162/NECO_a_00810

the orthogonality and the rank-two constraint in the estimation of PCA itself to improve the results. We further provide an interpretation of these methods in terms of estimation of a probabilistic generative model related to blind separation of dependent sources. Experiments on brain imaging data give very promising results.

1 Introduction

Estimation of nonstationary covariances, correlations, or other kinds of statistical connectivities is a topic of great interest in machine learning (Xuan & Murphy, 2007; Robinson & Hartemink, 2009; Kolar, Song, Ahmed, & Xing, 2010; Robinson & Priebe, 2012; Liu, Quinn, Gutmann, & Sugiyama, 2013) and neuroimaging (Kiviniemi et al., 2011; Allen et al., 2012; Leonardi et al., 2013; Monti et al., 2014). Such analysis would complement widely used analysis methods for stationary (nonchanging) connectivity. For example, spatial independent component analysis (ICA) is often used in fMRI (Kiviniemi, Kantola, Jauhiainen, Hyvärinen, & Tervonen, 2003; van de Ven, Formisano, Prvulovic, Roeder, & Linden, 2004; Beckmann, DeLuca, Devlin, & Smith, 2005) to find networks that have strong connectivities between the nodes (voxels). In addition to such ICA, it would be very useful to find networks that have strongly changing connectivities in terms of either internetwork or intranetwork connectivity, possibly ignoring networks where the connectivity is strong all the time and does not change.

It is important that in addition to estimation of nonstationary connectivities, we can summarize and visualize the nonstationarity of the system in an intuitively appealing way. A simple way of summarizing nonstationary behavior would be to compute some kind of connectivity statistics for different time segments and then perform PCA (Leonardi et al., 2013) or clustering (Allen et al., 2012) in the space of those connectivity statistics. The statistics of one segment could consist of all the elements of the covariance (or correlation) matrix computed inside that segment, for example. In a similar vein, network science researchers have developed methods for detection and tracking of communities in dynamically evolving networks (Greene, Doyle, & Cunningham, 2010; Tantipathananandh & Berger-Wolf, 2011).

The problem with applying existing unsupervised machine learning methods on the set of connectivity matrices is that they may not lead to very easily interpretable results due to the particularly high-dimensional and complex nature of the connectivity data. For example, the weight vectors of principal components, as well as cluster center points, are of the same form as the connectivity matrices. This means they are much more difficult to visualize than spatial patterns, which can be visualized like vectors in the

data space. Such a principal component (or center point) contains weights for all possible connections, and it is not clear how to summarize it verbally, for example, in terms of "connectivity between brain area A and brain area B." Imposing sparsity as in sparse PCA (Journée, Nesterov, Richtárik, & Sepulchre, 2010) may help but does not qualitatively change the problem. Alternatively, using the spatial 2D structure of image data as in 2D PCA (Yang, Zhang, Frangi, & Yang, 2004) can simplify the results in some cases, but does not seem to be very suitable for connectivity data.

Here we develop a framework for analyzing the variability of connectivities in terms of spatial patterns, that is, linear functions of the original data variables. In particular, we attempt to find pairs of components of the original data (i.e., pairs of spatial patterns) that have maximally variable connectivity, in the sense that the connectivity between the components is changing as strongly as possible. The connectivity between such a pair of spatial components then gives an approximation of a principal component of connectivity statistics. Such pairs are easier to interpret and visualize in most cases since we can simply visualize the two spatial patterns (e.g., brain areas).

The method takes a set of connectivity matrices as input and does not make any explicit assumptions on how those connectivities are obtained; in particular, they do not have to be from different time windows. They could come, for example, from different subjects or different experimental conditions.

Our method is based on developing a tailor-made low-rank approximation of connectivity matrices or their principal components. We approximate a principal component of connectivity by the outer product of two orthogonal vectors, motivated by the goal to model changes in connectivities between two different groups of variables. This is in stark contrast to conventional methods directly based on the eigenvalue decomposition (and closely related to a second-stage PCA) in which we take the outer product of an eigenvector with itself. We further develop a method that combines such a low-rank approximation with the computation of the connectivity-space PCA itself. The resulting method can also be interpreted as an estimation of the probabilistic generative model, related to blind separation of linearly correlated sources.

This article is structured as follows. A new theory of low-rank matrix approximation by two orthogonal vectors is motivated and presented in section 2. Objective functions and learning algorithms incorporating the orthogonal matrix approximation as a constraint in PCA are developed in section 3. A probabilistic generative model is proposed in section 4. Simulations are presented in section 5 and experiments on real brain imaging data in section 6. Connections to related methods are discussed in section 7, and section 8 concludes the article. Preliminary results were presented by Hyvärinen, Hirayama, and Kawanabe (2014).

2 Orthogonal Rank-Two Approximation of Connectivity Component Matrix

2.1 Motivation for New Matrix Approximation. Denote by $\mathbf{C}_\tau, \tau = 1, \ldots, k$ a number of connectivity matrices obtained by some well-known connectivity measures. For example, the $\mathbf{C}_\tau$ can be covariance matrices or correlation matrices. Typically the different matrices could be obtained from different time segments of a multivariate time series. However, for the methods developed in this article, it does not really matter what the difference between the connectivity matrices is; the index τ could just as well refer to different subjects, in which case we are analyzing interindividual differences or different experimental conditions. For simplicity of exposition, we usually consider the case of time segments, which leads to analysis of nonstationarity.

A basic approach for analyzing such matrices is to perform PCA on the vectorized forms of the matrices, collapsing each matrix into a vector by scanning it column by column. We consider the vectorized form of $\mathbf{C}_\tau$ an observation at time point τ and performing PCA in the usual way (this is called matrix PCA in the following). Thus, we obtain a number of principal component matrices that have the same form as $\mathbf{C}_\tau$. Let us denote a principal component thus obtained by $\mathbf{K}$ in the following.

A major problem is that such simple application of PCA leaves open the question of how to further analyze or visualize the obtained $\mathbf{K}$. In particular, visualization of such a component is often very difficult.

A low-rank approximation of $\mathbf{K}$ could be used to analyze its structure. A well-known approach is to use a rank-one approximation of $\mathbf{K}$ given by the dominant eigenvector. In such an approach, a single eigenvector is mainly able to represent a single clique (i.e., a group of closely connected variables) such that the correlations between those variables (within the clique) change together. Using many eigenvectors, we can represent many cliques, but interactions between the cliques (e.g., brain areas) are not explicitly represented.

Such a conventional low-rank analysis is useful in many cases, but here we wish to develop further analytical methods for several reasons:

1. The principles of rank-one approximation are already well known, as are the principles of successive (deflationary) rank-one approximations by the dominant eigenvectors, which are all accomplished simply by the eigenvalue decomposition of $\mathbf{K}$.
2. What would be very interesting in many applications (e.g., in brain imaging) would be to find two groups of variables such that the connectivity between the two groups is changing strongly. This needs more than conventional rank-one or rank-two approximations, as will be seen below.

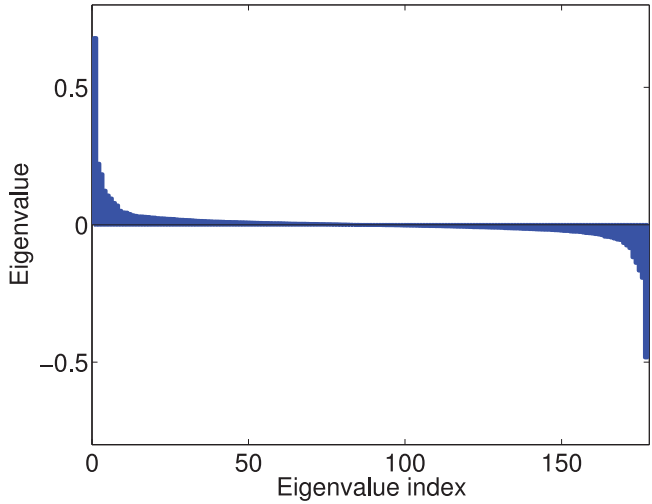

Figure 1: Illustration of a prototypical eigenspectrum of a matrix obtained by PCA of connectivities. This is from the first principal component of the functional connectivity MRI matrices analyzed in section 6. The spectrum is characterized by two dominant eigenvalues of opposite signs.

3. We have found empirically that the **K** in real data often cannot be meaningfully represented by a conventional low-rank approximation. An example of an eigenspectrum of **K** for a real data set is given in Figure 1, which shows an interesting pattern of two dominant eigenvalues with opposite signs. The intuitive and practical meaning of such an eigenstructure needs to be understood and properly interpreted. The goal of the developments below is to explain the meaning of such an eigenstructure. We will see that it is in fact closely connected to point 2 above: modeling changing connectivity between two groups of variables.

On the other hand, it is clear that we should try to improve the method by incorporating such low-rank constraints in the optimization problem itself. We do this in section 3.

2.2 Definition by Two Optimization Problems. Assume we are given a matrix **K** that has the same dimensions as a connectivity matrix obtained by a PCA of the connectivity matrices or some similar method. Next, we consider how to develop a low-rank approximation suitable for this purpose.

Denote by **w** and **v** two vectors in the data space that define brain areas or something similar. As an important departure from conventional analysis,

let us also assume that $\mathbf{w}$ and $\mathbf{v}$ are orthogonal. This is because we want to analyze connectivities between two different groups of variables (e.g., two different brain areas). In conventional rank one approximation, we would take the outer product of $\mathbf{w}$ with itself, and we would be analyzing connectivity inside a group of variables (those corresponding to nonzero entries in $\mathbf{w}$); a conventional rank-two approximation $\mathbf{w}\mathbf{w}^T + \mathbf{v}\mathbf{v}^T$ simply analyzes connections inside two groups separately from each other.

Thus, we use the outer product of $\mathbf{w}$ and $\mathbf{v}$ to model a pattern of connectivity between two regions. Due to the symmetry of the connectivity matrices, we further use a symmetrized version. This leads to an optimization problem in which we attempt to find $\mathbf{w}$ and $\mathbf{v}$ by minimizing

$$\min_{\mathbf{w}^T\mathbf{v}=0} \|\mathbf{K} - (\mathbf{v}\mathbf{w}^T + \mathbf{w}\mathbf{v}^T)\|^2, \tag{2.1}$$

where the norm is the Frobenius norm. This is a rather unconventional low-rank approximation since it uses the outer products of two orthogonal vectors. We call it the orthogonal rank-two approximation.[1]

The optimization problem can be related to a graph-theoretic problem where $\mathbf{K}$ is a zero-one adjacency matrix and $\mathbf{w}$, $\mathbf{v}$ are indicator functions of two sets of nodes. Orthogonality of $\mathbf{w}$, $\mathbf{v}$ then means that the two sets are disjoint. The optimization problem means that we approximate the graph by one whose connections consist exactly of all possible connections between the two sets. What we consider here is a (rather strongly) relaxed version of such a combinatorial problem, reminiscent of spectral clustering methods.

Some interpretations of the outer products in terms of nonstationarity in brain imaging are shown in Figure 2. In the most basic case, Figure 2a, we can think of $\mathbf{w}$ and $\mathbf{v}$ as zero-one indicator functions of two brain regions or other groups of variables. Then the idea is that it is the connectivity between those two areas that changes, possibly from zero to a positive value, as in this illustration. This is the simplest possible interpretation and could be considered approximately correct even if there are some weakly negative values in $\mathbf{w}$ and $\mathbf{v}$. If one of the vectors, say $\mathbf{v}$, has really significant negative values, what we are modeling is a more complicated connectivity pattern where the negative values correspond to a change of connectivity of the opposite sign. For example, this could mean a switching of connectivity. In some parts of the data, there is strong connectivity between the area defined by $\mathbf{w}$ and the area defined by the positive values of $\mathbf{v}$, while in other parts of the data, the strong connectivity is between the areas defined by $\mathbf{w}$ and

[1]The solution is clearly defined only up to a rescaling (and changing the signs) of the $\mathbf{v}$ and $\mathbf{w}$, since if we multiply $\mathbf{v}$ by a nonzero constant, dividing $\mathbf{w}$ with the same constant will lead to the same approximation error.

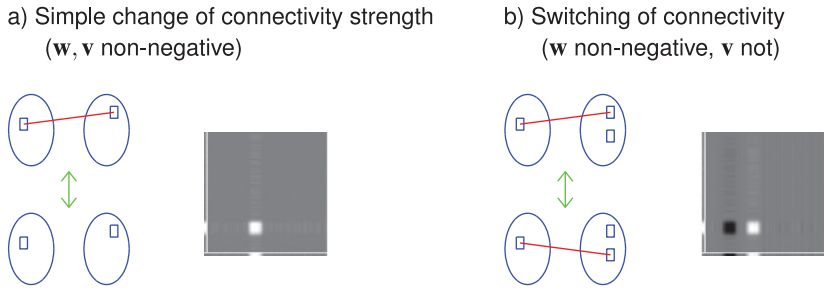

Figure 2: Illustration of different connectivity pattern changes modeled by the orthogonal rank-two framework in the case of brain imaging. In each panel, the left side gives an illustration of connectivities in terms of brain areas: each oval is the brain being analyzed, boxes are brain areas, and red lines are connectivities. The right side shows the matrix produced as an outer product of $\mathbf{w}$ and $\mathbf{v}$ (without symmetrization), with the vectors $\mathbf{w}$ and $\mathbf{v}$ given as the line plots at the left and lower edges; black is negative, white positive, gray zero. (a) If both $\mathbf{w}$ and $\mathbf{v}$ are nonnegative, we are modeling the change of connectivity between two groups of variables; in the simplest case, the connectivity could be zero for some τ and positive for others. (b) If one of the vectors, say $\mathbf{v}$, has negative values as well, we could be modeling a case where the connectivity from $\mathbf{w}$ switches between the two brain areas, defined by the positive and negative entries in $\mathbf{v}$, respectively.

the negative parts of $\mathbf{v}$. This simple case is illustrated in Figure 2b.[2] In the most general case, where $\mathbf{w}$ and $\mathbf{v}$ have both positive and negative values, the interpretation is more complex, and the most meaningful interpretation may be obtained by assuming that at least some of the negative values are small and insignificant.

The optimization problem in equation 2.1 is closely related to the following problem:

$$\max_{\|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \mathbf{w}^T\mathbf{K}\mathbf{v}. \tag{2.2}$$

In the following, we study these two simple optimization problems, and their connection.

[2]However, the exact interpretation depends on the baseline of connectivities, the interpretation above assuming that the connectivity principal component (dot product between $\mathbf{K}$ and $\mathbf{C}_\tau$) has the same sign (or is zero) for all τ. If the connectivity component fluctuates around zero, the connectivity would be switching in the sense that sometimes the connection between the area defined by $\mathbf{w}$ has connectivities of the same sign as the signs of entries in $\mathbf{v}$ and sometimes of the opposite sign.

2.3 Theory of Orthogonal Rank Two Approximation. To solve the optimization problem in equation 2.2, we have the following theorem. (While most proofs are given in the appendix, we prove this theorem here since the idea of the proof is important for the rest of the article.)

Theorem 1. *Assume that* $\boldsymbol{K}$ *is a symmetric (real-valued) matrix and the largest and smallest eigenvalues of* $\boldsymbol{K}$ *are distinct. (Here, "largest" and "smallest" mean according to ordinary sorting, not using absolute values.) Then a solution of the optimization problem, equation 2.2, is given by*

$$\boldsymbol{w} = \frac{1}{\sqrt{2}}(\boldsymbol{e}_{max} + \boldsymbol{e}_{min}), \qquad \boldsymbol{v} = \frac{1}{\sqrt{2}}(\boldsymbol{e}_{max} - \boldsymbol{e}_{min}) \tag{2.3}$$

where $\boldsymbol{e}_{max}$ *and* $\boldsymbol{e}_{min}$ *are the eigenvectors corresponding to the smallest and largest eigenvalues of* $\boldsymbol{K}$*. The whole set of solutions is given by* $\{(\boldsymbol{w}, \boldsymbol{v}), (-\boldsymbol{w}, -\boldsymbol{v}), (\boldsymbol{v}, \boldsymbol{w}), (-\boldsymbol{v}, -\boldsymbol{w})\}$*. Denoting by* λ_{max} *and* λ_{min} *the largest and smallest eigenvalues of* $\boldsymbol{K}$*, the value of the objective at optimum is equal to* $\frac{1}{2}(\lambda_{max} - \lambda_{min})$*.*

Proof. Make the change of variables

$$\mathbf{a} = \frac{1}{\sqrt{2}}(\mathbf{w} + \mathbf{v}), \qquad \mathbf{b} = \frac{1}{\sqrt{2}}(\mathbf{w} - \mathbf{v}), \tag{2.4}$$

which changes the optimization problem to

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^T\mathbf{b}=0} \frac{1}{2}[\mathbf{a}^T\mathbf{K}\mathbf{a} - \mathbf{b}^T\mathbf{K}\mathbf{b}], \tag{2.5}$$

where we have used the equality

$$\begin{aligned} 2\mathbf{w}^T\mathbf{K}\mathbf{v} &= \frac{1}{2}(\mathbf{w} + \mathbf{v})^T\mathbf{K}(\mathbf{w} + \mathbf{v}) - \frac{1}{2}(\mathbf{w} - \mathbf{v})^T\mathbf{K}(\mathbf{w} - \mathbf{v}) \\ &= \mathbf{a}^T\mathbf{K}\mathbf{a} - \mathbf{b}^T\mathbf{K}\mathbf{b}. \end{aligned} \tag{2.6}$$

To maximize this without the orthogonality constraint, it is clear that we need to choose $\mathbf{a}$ so that it is the eigenvector corresponding to the largest eigenvalue and $\mathbf{b}$ so that it is the eigenvector corresponding to the smallest eigenvalue. Furthermore, these two vectors are necessarily orthogonal, so they maximize the objective even with the orthogonality constraint. The inverse of the transformation in equation 2.4 is given by

$$\mathbf{w} = \frac{1}{\sqrt{2}}(\mathbf{a} + \mathbf{b}), \qquad \mathbf{v} = \frac{1}{\sqrt{2}}(\mathbf{a} - \mathbf{b}), \tag{2.7}$$

which gives equation 2.3. The value of objective function is then clearly given by half of the difference of eigenvalues. The whole set of solutions comes from considering the indeterminacy of the sign of normalized eigenvectors and the symmetry between **w** and **v**.

The optimization theorem applies for practically any connectivity structure since there is no assumption on the rank of the matrix **K** or the signs of its eigenvalues. However, the theorem is certainly more interesting in the case where the largest and smallest eigenvalues are very different (e.g., having different signs), so that the objective function obtains a relatively large value.

If the matrix has rank two, the whole structure of the matrix can be explained by two eigenvectors. We could analyze the variance explained by the orthogonal rank-two approximation, which should be large if the matrix is close to rank two. The following theorem provides the basis for such analysis and shows the connection between the two optimization problems, equations 2.1 and 2.2; they are essentially equivalent.

Theorem 2. *Consider the problem of approximating a symmetric matrix by a symmetrized outer product of two orthogonal vectors in equation 2.1 where the norm is the Frobenius norm. Make the assumptions of theorem 1. Then,*

1. *The set of optimizing* $\boldsymbol{w}$ *and* $\boldsymbol{v}$ *is the same, up to scaling constants, as the set of optimizing* $\boldsymbol{w}$ *and* $\boldsymbol{v}$ *in the optimization problem, equation 2.2, treated in theorem 1.*
2. *The value of the objective at optimum is equal to* $\|\boldsymbol{K}\|^2 - \frac{1}{2}(\lambda_{max} - \lambda_{min})^2$.
3. *The objective is zero iff* $\boldsymbol{K}$ *has rank two and* $\lambda_{min} = -\lambda_{max}$.

The proof is given in the appendix.

Note that in theorem 2, we cannot approximate every rank-two matrix exactly because we do not have separate scaling coefficients for the two eigenvectors. In a conventional rank-two approximation, we would have the two eigenvalues as such scaling coefficients, and we would be able to exactly approximate any rank-two matrix. Here, we obtain an exact approximation only if the two eigenvalues have opposite signs but equal absolute values. Thus, our approximation is exact precisely for matrices with the special plus-minus structure mentioned in section 2.1.

The important implication of the theory presented is that in our optimization problem, we do not use the eigenvectors themselves. This is in stark contrast to conventional low-rank approximations, which are based on using the outer products of the eigenvectors with themselves. Thus, in conventional methods, the eigenvectors themselves are supposed to give the interesting components in the data space. Here, we have shown that we need to transform the eigenvectors to have meaningful components in the low-rank approximation.

Note that our rank-two approximation is not positive semidefinite in general. As in ordinary PCA, we are mainly interested in modeling deviations from the mean, which precludes any constraints related to nonnegativity.

3 Constrained PCA Learning

The analysis in the preceding section gives new insight into the results obtained by applying PCA to the set of connectivity matrices $\mathbf{C}_\tau$, as performed, for example, by Leonardi et al. (2013). Next, we propose to directly integrate the orthogonal rank-two assumption into a PCA objective function. In particular, matrix PCA clearly suffers from the problem that if two eigenvalues of the covariance matrix of connectivities are equal, the corresponding eigenvectors are not well defined and are likely to mix together those two source pairs. Finding directly a low-rank solution to a PCA-like objective is likely to alleviate this problem, as will be confirmed by the simulations in section 5.

3.1 Definition of Constrained PCA Objective. Consider the original connectivity matrices by $\mathbf{C}_\tau$, $\tau = 1, \ldots, k$ as in the preceding section. In one time segment (or subject), the connectivity between the two areas can be intuitively defined as $\mathbf{w}^T\mathbf{C}_\tau\mathbf{v}$, which, in the case of covariance matrices, is actually the covariance of $\mathbf{w}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{x}$. We want to maximize the variance of this quantity in order to find components that explain as much of the nonstationary (or otherwise changing) connectivity structure as possible. We constrain $\mathbf{w}$ and $\mathbf{v}$ in the same way as in theorem 1: their norms are equal to unity, and they are orthogonal. Thus, we obtain the (preliminary form of the) optimization problem:

$$\max_{\|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \frac{1}{k}\sum_{\tau=1}^{k}(\mathbf{w}^T\mathbf{C}_\tau\mathbf{v})^2 - \left(\frac{1}{k}\sum_{\tau}\mathbf{w}^T\mathbf{C}_\tau\mathbf{v}\right)^2. \tag{3.1}$$

To simplify this, let us subtract the average connectivities to obtain the centered connectivity matrices,

$$\tilde{\mathbf{C}}_\tau = \mathbf{C}_\tau - \frac{1}{k}\sum_{i=1}^{k} C_i, \tag{3.2}$$

so that $\sum_\tau \tilde{\mathbf{C}}_\tau = \mathbf{0}$. Thus, we obtain the final optimization problem, which is equivalent to equation 3.1, as

$$\max_{\|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \sum_{\tau=1}^{k}(\mathbf{w}^T\tilde{\mathbf{C}}_\tau\mathbf{v})^2. \tag{3.3}$$

To understand the meaning of such constrained PCA, it is useful to consider the basic case of two connectivity matrices; $k = 2$. Then the subtraction of average connectivity means that $\tilde{\mathbf{C}}_1 = -\tilde{\mathbf{C}}_2 = \frac{1}{2}(\mathbf{C}_1 - \mathbf{C}_2)$. Constrained PCA is then simply based on analyzing the difference of the two connectivities, and equation 3.3 becomes the optimization problem in equation 2.2, the squaring being immaterial. In fact, the unconstrained principal component is simply obtained by $\mathbf{K} = \mathbf{C}_1 - \mathbf{C}_2$ up to a scaling. So the case of two connectivity matrices reduces to the theory in section 2, but for $k \geq 3$, we need to develop a new algorithm.

3.2 Algorithm for Constrained PCA. Next, we consider the general optimization problem formulated in equation 3.3. To find a simple and efficient method for the optimization problem in equation 3.3, we formulate the following alternative problem, inspired by Journée et al. (2010),

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=\|\mathbf{r}\|=1} \sum_{\tau} r_\tau [\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b}], \tag{3.4}$$

with an auxiliary variable vector $\mathbf{r} = (r_1, \ldots, r_k)$, and $\mathbf{a}$, $\mathbf{b}$ defined in equation 2.4. This modified optimization problem is based on two ideas, already used in the proof of theorem 1. First, we have by simple linear algebra,

$$\mathbf{w}^T \mathbf{M} \mathbf{v} = \frac{1}{4}(\mathbf{w} + \mathbf{v})^T \mathbf{M}(\mathbf{w} + \mathbf{v}) - \frac{1}{4}(\mathbf{w} - \mathbf{v})^T \mathbf{M}(\mathbf{w} - \mathbf{v}) \tag{3.5}$$

for any symmetric matrix $\mathbf{M}$. Thus, by making a change of variables in equation 2.4, we obtain the bracketed expression above.

Second, the optimal r_τ are trivially obtained as

$$r_\tau = c[\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b}], \tag{3.6}$$

where the proportionality constant c is introduced due to the unit norm constraint but is immaterial. Plugging this into equation 3.4 gives something related to the sum of squares in equation 3.3, although the existence of c makes the situation a bit more complicated.

To formally show the equivalence of the two optimization problems, we have the following theorem:

Theorem 3. *The vectors* $\boldsymbol{w}$, $\boldsymbol{v}$ *are a solution of the optimization problem, equation 3.3, if and only if* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *as transformed in equation 2.4 are a solution (together with some* $\mathbf{r}$*) of the optimization problem, equation 3.4.*

The theorem is proven in the appendix.

The utility of this new formulation is that we can apply a simple alternating optimization method. The solution for $\mathbf{r}$ given $\mathbf{a}$, $\mathbf{b}$ was given in

equation 3.6, where the constant c can be set to any positive value without affecting the algorithm. On the other hand, given a running estimate $\mathbf{r}$ and considering it fixed, we have the optimization problem

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \mathbf{a}^T\mathbf{M}\mathbf{a} - \mathbf{b}^T\mathbf{M}\mathbf{b} \tag{3.7}$$

with

$$\mathbf{M} = \sum_\tau r_\tau \tilde{\mathbf{C}}_\tau, \tag{3.8}$$

which was already considered in the proof of theorem 1 and is solved by performing an eigenvalue decomposition of $\mathbf{M}$ and taking as $\mathbf{a}$ the eigenvector corresponding to the largest eigenvalue and as $\mathbf{b}$ the eigenvector corresponding to the smallest eigenvalue.

In fact, the eigenvectors $\mathbf{a}$ and $\mathbf{b}$ are necessarily orthogonal (unless in a very degenerate case, where all the eigenvalues are equal), and thus we obtain $\mathbf{a}^T\mathbf{b} = 0$ automatically. Here, we see another really useful property of this reformulation: the orthogonality constraint in equation 3.3 can be omitted since the optimum will necessarily produce orthogonal $\mathbf{a}$ and $\mathbf{b}$, which implies orthogonal $\mathbf{w}$ and $\mathbf{v}$.

Thus, we obtain a simple algorithm that contains no step-size parameters and in which every step is guaranteed, by construction, not to decrease the objective function. It is reasonable to start the optimization at a point given by the orthogonal rank-two approximation of matrix PCA, which means that maximization of this objective can be considered a fine-tuning of matrix PCA results.

3.3 Estimating Many Components. The methods so far presented in this section consider a single principal component pair of connectivities. Next, we consider estimation of many principal component pairs.

Consider first the case of constrained PCA. We extend it to many component pairs by well-known deflation (orthogonal projection) methods. However, here two distinct approaches seem possible:

1. A basic but very stringent condition would be to impose the orthogonality between all the vectors $\mathbf{w}_i$ and $\mathbf{v}_i$ estimated. However, this may be too strict, since we may want to have pairs in which one of the components is the same as or similar to a previously estimated one.
2. In the matrix space, we can restrict the different rank-two approximations $\mathbf{w}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{w}_i^T$ to be orthogonal in the sense that $\mathrm{tr}([\mathbf{w}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{w}_i^T][\mathbf{w}_j\mathbf{v}_j^T + \mathbf{v}_j\mathbf{w}_j^T]) = 0$ for $i \neq j$. This can be simply implemented by projecting the one-dimensional subspace spanned by $\mathbf{w}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{w}_i^T$ away from the set of connectivity matrices.

In the matrix-level PCA case in section 2, another solution would be to apply orthogonal rank-two approximation to principal component matrices obtained by conventional PCA of the connectivity matrices; in such conventional PCA, many components are defined in the usual way.

We propose here to use the orthogonality of rank-two approximations (case 2 above) in both the constrained and the matrix-level PCA cases. In practice, it does not seem to lead to too similar components, and thus the orthogonality does not seem to be too relaxed with this method. It is also more natural even in the case of matrix PCA, if the results of matrix PCA are to be reduced to the rank-two approximation afterward anyway.

3.4 Robust Variant. It is possible that the objective in equation 3.3 is too sensitive to connectivities in a single connectivity matrix, which might be due to outliers. To alleviate this problem, we can easily develop a more robust version, replacing squaring by an absolute value operator as

$$\max_{\|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \sum_{\tau=1}^{k} |\mathbf{w}^T \tilde{\mathbf{C}}_\tau \mathbf{v}|. \tag{3.9}$$

This objective can be optimized using a similar alternating variables method by defining the relaxed optimization problem as

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, r_\tau \in \{-1,+1\}} \sum_{\tau} r_\tau [\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b}]. \tag{3.10}$$

Given the r_τ, the optimal $\mathbf{a}$, $\mathbf{b}$ are obtained as in the basic constrained PCA algorithm. The only change is that here, the optimal r_τ are obtained as

$$r_\tau = \text{sign}(\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b}). \tag{3.11}$$

It is easy to show that this algorithm optimizes equation 3.9.

4 Probabilistic Generative Model

To get further insight into the methods in the preceding section, it is useful to formulate a generative probabilistic model such that the methods presented can be considered to estimate parameters in it. Again, we develop the model for the case where the connectivity matrices come from different time segments, but the model is applicable for the general case where the meaning of the index τ has an arbitrary meaning. The connectivity matrices are assumed to be covariance or correlation matrices here.

4.1 Basic Idea. To present the basic idea, we assume we have just two time segments. In both, the data follow a linear mixing model, with standardized, possibly gaussian components s_i

$$\mathbf{x} = \mathbf{H}\mathbf{s}, \tag{4.1}$$

where the number of s_i is equal to the number of the x_i and $\mathbf{H}$ is square and orthogonal. Assume the matrix $\mathbf{H}$ is the same for the entire data, but the statistics of the components change as follows. In the first segment, all the s_i are independent, but in the second segment, we have the perfect correlation $s_1 = s_2$, while all the other components are independent. Then we have

$$\mathbf{C}_1 = \mathbf{H}\mathbf{H}^T, \tag{4.2}$$

$$\mathbf{C}_2 = \mathbf{H}\mathbf{H}^T + \mathbf{h}_1\mathbf{h}_2^T + \mathbf{h}_2\mathbf{h}_1^T, \tag{4.3}$$

where $\mathbf{h}_i$ is the ith column of $\mathbf{H}$, and the matrix principal component, as shown in section 3.1, is given by

$$\mathbf{K} = \mathbf{C}_2 - \mathbf{C}_1 = \mathbf{h}_1\mathbf{h}_2^T + \mathbf{h}_2\mathbf{h}_1^T, \tag{4.4}$$

so by theorem 2, optimization of equation 2.2 will find $\mathbf{w} = \mathbf{h}_1$ and $\mathbf{v} = \mathbf{h}_2$. This shows that the methods considered in this article can be seen to analyze changes in correlation structure of components in a model related to blind source separation or independent component analysis (ICA) (Comon, 1994; Hyvärinen, Karhunen, & Oja, 2001). Various extensions relaxing the independent assumption have been developed before (see, in particular, Sasaki, Gutmann, Shouno, & Hyvärinen, 2014, for a related method considering linearly correlated components).

4.2 Definition of Probabilistic Model. Next, we define the generative model with more generality. To begin, assume the data are generated by a generative model as in equation 4.1, with gaussian and standardized s_i. In the following, we model the nonstationary covariance of two components, which can be assumed to be those with indices 1 and 2 without loss of generality. The nonstationary covariance of $\mathbf{h}_1$ and $\mathbf{h}_2$ is modeled by a latent variable $z(\tau)$. In particular, the ith, jth element of the precision matrix of $\mathbf{x}$ in time segment τ is assumed to be equal to $z(\tau)$,

$$\mathrm{cov}(\mathbf{s}(\tau))^{-1} = \mathbf{I} + z(\tau)[\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T], \tag{4.5}$$

where $\mathbf{e}_i$ is the ith canonical basis vector (all zeros but the ith entry equal one). We assume

$$\sum_{\tau} z(\tau) = 0, \tag{4.6}$$

which means that over the whole data set, the precision matrix of $\mathbf{s}$ is identity, as can be seen by summing equation 4.5 over τ. Further, we have to assume

$$|z(\tau)| < 1, \text{ for all } \tau, \tag{4.7}$$

which preserves the positive definiteness of the precision matrix by making it diagonally dominant.

We further assume that $\mathbf{H}$ is orthogonal. This assumption is made since we want the components to be mainly nonoverlapping, representing groups of variables that do not contain the same variables. (If the data were whitened, we would find the typical assumption of uncorrelatedness used in ICA.) This assumption may seem unrealistic, but we point out that it is made in this theoretical generative model only, and in the actual method explained in the preceding section, the constraint on orthogonality is essentially weaker, as explained in section 3.3.

To estimate the model, we have the following theorem, proven in the appendix:

Theorem 4. *Assume that the data follow the model specified by equations 4.5 to 4.7 and 4.1. Assume further that* $\boldsymbol{H}$ *is orthogonal, and that z follows a uniform prior in the infinitesimal interval* $[-\epsilon, \epsilon]$. *Then the Taylor approximation of the log likelihood of the model has the leading term (ignoring constant terms not depending on the parameters)*

$$log\ \tilde{L}(\boldsymbol{H}) = \frac{\epsilon^3}{6} \sum_{\tau} (\boldsymbol{h}_1^T \tilde{\boldsymbol{C}}_\tau \boldsymbol{h}_2)^2, \tag{4.8}$$

where $\tilde{\boldsymbol{C}}_\tau$ *is the covariance matrix in time segment* τ *with average covariances subtracted as in equation 3.2.*

Note that here, the first-order and second-order terms (in ϵ) disappear since they correspond to a gaussian likelihood of orthogonally transformed data, which is constant. Using such an infinitesimal ϵ is justified by the fact that we are considering an infinitesimal departure from stationarity, and in fact the term of $O(\epsilon^3)$ would be constant in the case of stationarity.

Also, since we used the orthogonality of $\mathbf{H}$ in computing its determinant (the normalization constant) in the proof, this likelihood is valid only for orthogonal $\mathbf{H}$. In fact, it explodes if $\mathbf{H}$ is allowed to go infinitely far from the origin.

What we see in this theorem is that the constrained PCA objective in equation 3.3 appears as a first-order approximation of the likelihood. The

first-order approximation means that the dependencies between the components are infinitesimal. This provides the link between our heuristically defined objective and estimation of the generative model.

Our model considers, for simplicity, a single component pair. More than one component pair can be estimated by well-known deflationary methods, discussed in section 3.3.

It should be noted that the probabilistic generative model as presented here is in fact a special case of generative models that can be estimated by the practical algorithms presented in earlier sections. In particular, we assumed here that the average connectivity is zero (see equation 4.6), which is not at all necessary in the practical algorithm. Likewise, we assumed that the whole matrix $\mathbf{H}$ is orthogonal, while in the practical algorithm, we use a more much more relaxed orthogonality constraint, as discussed in section 3.3. These extra constraints make the model much more similar to ICA and BSS methods and amenable to mathematical analysis. They should thus be seen as mathematical simplifications and not real restrictions on the data to be analyzed in practice.

4.3 More Realistic Model with Changing Variances. A problem in practice is that the simple model above ignores an essential aspect of many kinds of real data; the nonstationarity of the variances of observed variables. To make the model more realistic, we thus add separate terms modeling the changing variances (rescaling) of the data variables as

$$\mathbf{x} = \mathbf{D}_\tau \mathbf{H}\mathbf{s}, \tag{4.9}$$

where $\mathbf{D}_\tau$ is a random diagonal matrix modeling the changing variances of the data variables; its expectation is equal to identity. We do not make strong assumptions on $\mathbf{D}_\tau$, since in the developments below, we will see that it essentially disappears when we normalize the data covariance matrices by calculating the correlation coefficient matrices.

The justification for introducing the changing variances is the empirical observation that covariances can sometimes get unrealistically large values, possibly due to outliers or faulty sensors that lead to extremely high variances for some of the observed variables. In practice, most research uses the correlation coefficients instead of covariances, while the theory in the preceding section implies we should use covariances. As we see below, the model of changing variances removes this theoretical discrepancy, showing that correlation coefficients are actually the connectivity measure implied by this theory.

In this model, we need to introduce a further assumption on $\mathbf{h}_1$ and $\mathbf{h}_2$: they must be nonoverlapping in the sense that

$$\sum_i h_{1i}^2 h_{2i}^2 = 0. \tag{4.10}$$

This is necessary for the likelihood approximation because it means that the nonstationarity does not affect the diagonal of the inverse covariance matrix, and as a first-order approximation, it does not affect the covariances of the observed variables. (For more details, see the proof of theorem 5.) Such an assumption also makes intuitive sense, since it is essentially a stronger form of the orthogonality constraint, and in fact a form that more strictly formalizes the intuitive idea that $\mathbf{h}_1$ and $\mathbf{h}_2$ should model different sets of variables.[3]

To estimate the modified model, we have the following theorem (proven in the appendix), which again gives an approximate log likelihood is equal to our heuristic objective used in section 3.2:

Theorem 5. *Assume that the data follow the model specified by equations 4.5 to 4.7 and 4.9. Assume further that* $\boldsymbol{H}$ *is orthogonal and, in important contrast to theorem 4, that the first two columns are nonoverlapping in the sense of equation 4.10. Assume further that z follows a uniform prior in the infinitesimal interval* $[-\epsilon, \epsilon]$. *Then the Taylor approximation of the log likelihood of the model has the leading term (ignoring constant terms not depending on the parameters):*

$$\log \tilde{L}(\boldsymbol{H}) = \frac{\epsilon^3}{6} \sum_{\tau} (\boldsymbol{h}_1^T \tilde{\boldsymbol{C}}_\tau \boldsymbol{h}_2)^2, \tag{4.11}$$

where $\tilde{\boldsymbol{C}}_\tau$ *is the correlation coefficient matrix with average correlation coefficients removed.*

This approximation does not hold for covariances in this model, in contrast to theorem 4.

5 Simulations

We performed simulations to investigate how the methods presented are capable of estimating the components for data coming from our generative models in section 4.

We have three variants of orthogonal connectivity factorization:

- OCF1: The orthogonal rank-two decomposition of matrix PCA results (see section 2)
- OCF2: The constrained PCA algorithm or fine-tuning (see section 3)
- OCF3: The robust version of the constrained PCA algorithm (see section 3.4)

Furthermore, as baseline, we used

[3] This assumption of nonoverlappingness cannot be made on all the columns of $\mathbf{H}$ as it is full rank. Instead, it will only be made pairwise on the pairs correspond to correlated components.

EVD: Ordinary rank-two approximation using $\mathbf{e}_{\max}, \mathbf{e}_{\min}$ obtained by eigen-value decomposition of matrix PCA results

We applied these four methods on connectivity matrices given by correlation coefficients. In addition, we applied the basic method OCF1 on connectivities given by covariances, which is denoted by OCF1cov.

We simulated six kinds of data sets based on our generative models, giving rise to simulations 1 to 6:

Simulation 1: Basic case with one nonstationary component pair, where all OCF methods are supposed to work (in contrast to EVD)

Simulation 2: Introducing random scaling of observed variables as in section 4.3 to see what difference this makes for the different OCF methods

Simulation 3: Further introducing overlap between the spatial patterns, that is, violating equation 4.10 to see the significance of this constraint

Simulation 4: Going back to the basic case and introducing strong outliers to investigate the robustness of the different methods

Simulation 5: Like simulation 1, but now with two nonstationary components

Simulation 6: Making simulation 5 more difficult by equalizing the statistical properties (especially variances) of the two components

The dimension of the data was 12, which is rather small but enabled us to run the algorithm thousands of times in a reasonable time frame. The number of time points was 5000 which can be considered realistic in many brain imaging scenarios (for fMRI, assuming many subjects) and is not very different from what we have in the real-data experiments in the next section.

The changes in the correlations were created blockwise, so that in each block (segment) of 250 points (similar to a typical resting-state fMRI session), the correlation structure was randomly generated, with correlation coefficients between correlated source pairs drawn from a uniform distribution in $[-0.5, 0.5]$. In the algorithm, the size of each segment in which the connectivity was computed was not fixed to the right value but took different values: 50, 100, 250, 200, and 750 were used. For each setting, 1000 data sets (trials) were randomly generated, and the results shown are averaged over them. As in the theory in section 4, the matrix $\mathbf{H}$ was always constrained orthogonal.

In the first four simulations, there were 12 gaussian sources, all of them independent except for one pair. In the most basic one, simulation 1, there was no overlap between the spatial patterns between the sources, and there was no effect of variable scaling of the observed variables; that is, we used

the generating model in section 4.2 with constraint 4.10. In simulation 2, the observed variables had variable scaling in section 4.3. In simulation 3, the spatial patterns had an overlap of 25%. In simulation 4, two strong outliers (10 times the standard deviation of data) were introduced in the data.

Furthermore, we created two simulations (numbers 5 and 6) in which there were two pairs of connected sources (both without overlap and without variance scaling). In simulation 5, the connectivities between the pairs were created independently of each other, while in simulation 6, we created the two pairs so that the statistical properties (mean and variance) of the connectivities were the same for the two pairs.

The estimation error was computed as a simple mean-squared error and normalized so that it gives the proportion of error of the parameter vector norm. For visibility, the Briggs logarithms are plotted.

The results are in Figure 3. In simulation 1, we see that all proposed methods work quite well, while the baseline (EVD) is much worse. Unsurprisingly, the errors are minimized when the window size in the algorithm is equal to the window size in data generation. In simulation 2, we see that when the observed variables are randomly rescaled, using covariance does not work well at all, while correlations coefficients do work very well with all the methods. In simulation 3, we see that introducing overlap between the spatial patterns reduces the performance of all the methods rather uniformly but not dramatically. Introducing outliers in simulation 4 makes most methods fail, but the robust method (OCF3) still works. Interestingly, it is now important that the window size is not too large, presumably because with a large window, the proportion of connectivity matrices destroyed by outliers is larger. It should also be noted that in most simulations, the robust method (OCF3) has slightly larger errors, which indicates a typical trade-off with robust methods.

The basic case of two source pairs in simulation 5 is not very different from the case of a single source pair. However, we see some advantage for OCF2 (fine-tuning by constrained PCA), presumably because with more source pairs, matrix PCA with orthogonal rank-two approximation may mix them up if the eigenvalues are not clearly distinct. Such a phenomenon is simulated more explicitly in simulation 6, where the eigenvalues are set equal.

The case of two source pairs with identical connectivity statistics in simulation 6 is quite interesting, showing that OCF2 (fine-tuning by constrained PCA) results in a major improvement. This is understandable based on the theoretic argument that PCA cannot distinguish the eigenvectors in the case where the eigenvalues are equal. Actually, it is a bit surprising that OCF1 based on matrix PCA works at all in this case, which is presumably because the randomness in the data generation process created slightly unequal eigenvalues and also because any mixing of eigenvectors in the matrix PCA may be partly compensated by the EVD of the matrix principal component.

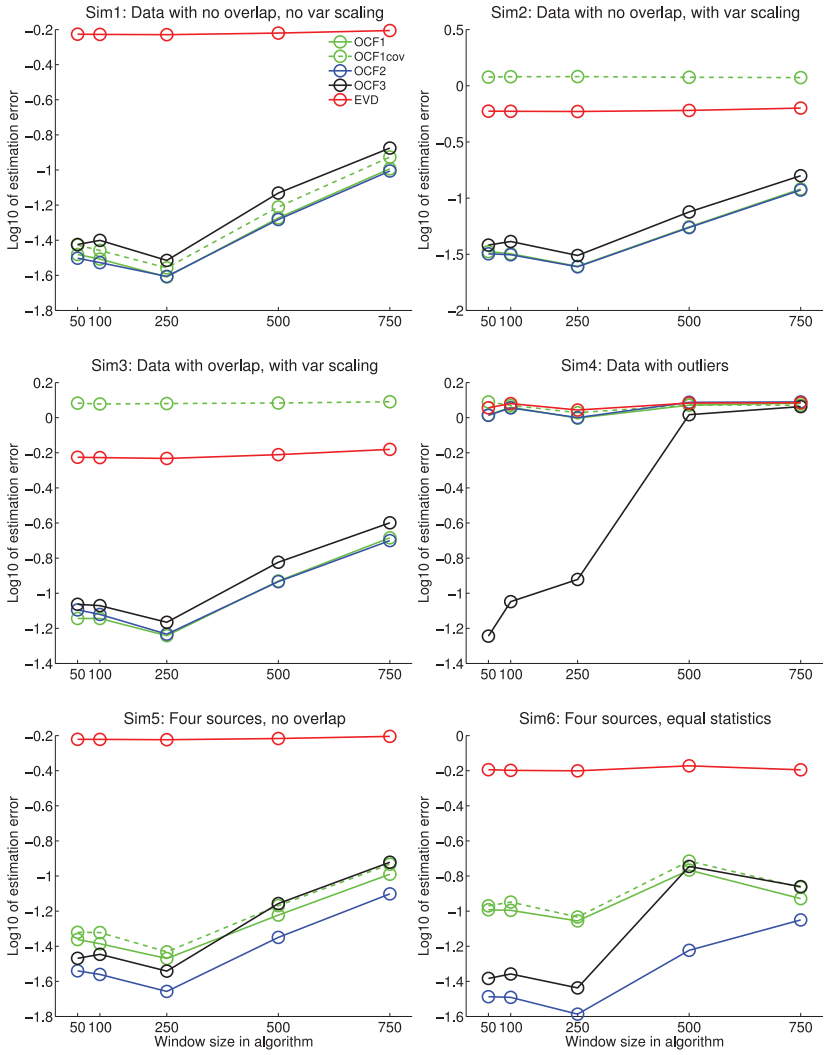

Figure 3: Main simulations with artificial data. Each panel is one simulation (type of data set), and each curve gives estimation errors for one variant of the algorithm. Vertical axis: Estimation errors, which were given a ceiling of one (zero in log scale) for visualization purposes. Horizontal axis: window size used in the algorithm (the window size in data generation was always the same). Standard errors of the mean are smaller than marker size and thus not plotted.

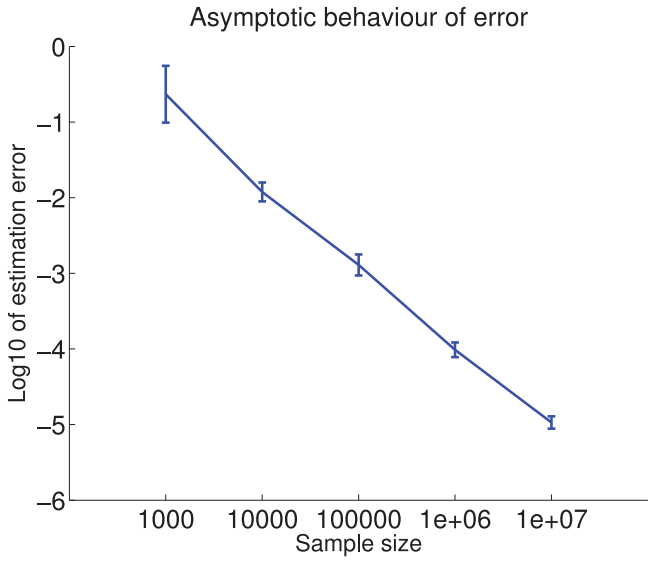

Figure 4: Further simulations with artificial data on consistency. Log-estimation error as a function of sample size when the sample size grows larger. Standard deviation given as error bar.

To recapitulate, OCF2 seems to be the best method in the comparison simulations, except that the robust variant OCF3 is better with outliers. OCF1 also performs quite well. We emphasize that OCF1 (matrix PCA with orthogonal rank-two approximation) is also a new method, so the fact that finetuning (OCF2) makes little difference in most simulations is not disappointing; it merely shows that our basic approach using the orthogonal rank-two approximation works well enough in itself and the more sophisticated developments may not always be needed. In contrast, the baseline (EVD) performed very badly.

Finally, we did a further simulation to investigate the consistency of the estimator. We took the basic setting in the first simulation and used OCF2, but varied the sample size to see if the estimation error seems to go to zero, which would indicate consistency of the estimator and identifiability of the model in this specific case. The result is in Figure 4. The error clearly seems to go to zero, which indicates consistency.

6 Experiments on Brain Imaging Data

To further validate the model, we performed experiments on two kinds of real brain imaging data, investigating both intersubject variability of connectivity and nonstationarity of connectivity.

6.1 Multisubject Functional Connectivity MRI Data

6.1.1 Methods. We first applied the method on public domain functional connectomics data from the USC Multimodal Connectivity Database (http://umcd.humanconnectomeproject.org). The data consist of functional connectivity magnetic resonance imaging (fcMRI) resting-state data. We selected the 103 subjects measured at Oulu University Hospital, tagged "Oulu." The database provides one functional connectivity matrix per subject, consisting of correlations of activities between 177 regions of interest (ROI). Thus, we analyze the intersubject variability of connectivities in this experiment.

We used the same four different methods as in the simulations, that is, OCF1: Matrix PCA followed by orthogonal rank-two decomposition (see section 2); OCF2: constrained PCA (see section 3); and OCF3: robust variant of constrained PCA (see section 3.4). As a baseline for comparison, we also considered EVD, the original eigenvectors $\mathbf{e}_{\max}$, $\mathbf{e}_{\min}$ of the matrix principal components (which are related to the $\mathbf{w}$, $\mathbf{v}$ of the first method by a 90 degree rotation as in theorem 1).

The resulting 3D spatial patterns were plotted so that the center of each ROI was plotted as a dot. Very small values were plotted as gray dots, while clearly positive and negative values were plotted as red and blue dots, respectively, so that the diameter of the dot was proportional to the absolute value. To visualize the 3D structure, the dots were projected onto three planes: sagittal (brain viewed from the side), horizontal (brain viewed from above), and frontal (brain viewed from the front or, equivalently, from the back). The projections were transparent, using no occlusion information (i.e., the visualization used a "glass brain"). As a background, we plotted a 2D projection of the gray matter in the ICBM 152 atlas (version 2009a) template brain volume for normal population (Fonov et al., 2011). Due to an apparent scaling incompatibility between the two data sources, the coordinates of the ROI centers were increased by 7%, which seemed to provide a better fit.

Note that the signs of $\mathbf{w}$ and $\mathbf{v}$ are not determined by the estimation methods, similar to PCA or ICA. Purely for the purposes of visualization, we determine the signs so that the strongest coefficients are positive.

To further compare the methods quantitatively, we computed two statistics of the resulting representations: the sparsity of the spatial patterns, which was defined as a normalized version of the fourth (noncentered) moment, $(\sum_i w_i^4 + \sum_i v_i^4)/(\sum_i w_i^2 + \sum_i v_i^2)^2/n$, and a measure of overlap, which was a normalized version of the correlation of the squares, $(\sum_i w_i^2 v_i^2)/\sqrt{\sum_i (w_i^2)^2}\sqrt{\sum_i (v_i^2)^2}$. (In these measures, any mean was not subtracted.) We also computed the similarities of the resulting spatial patterns for the three proposed methods.

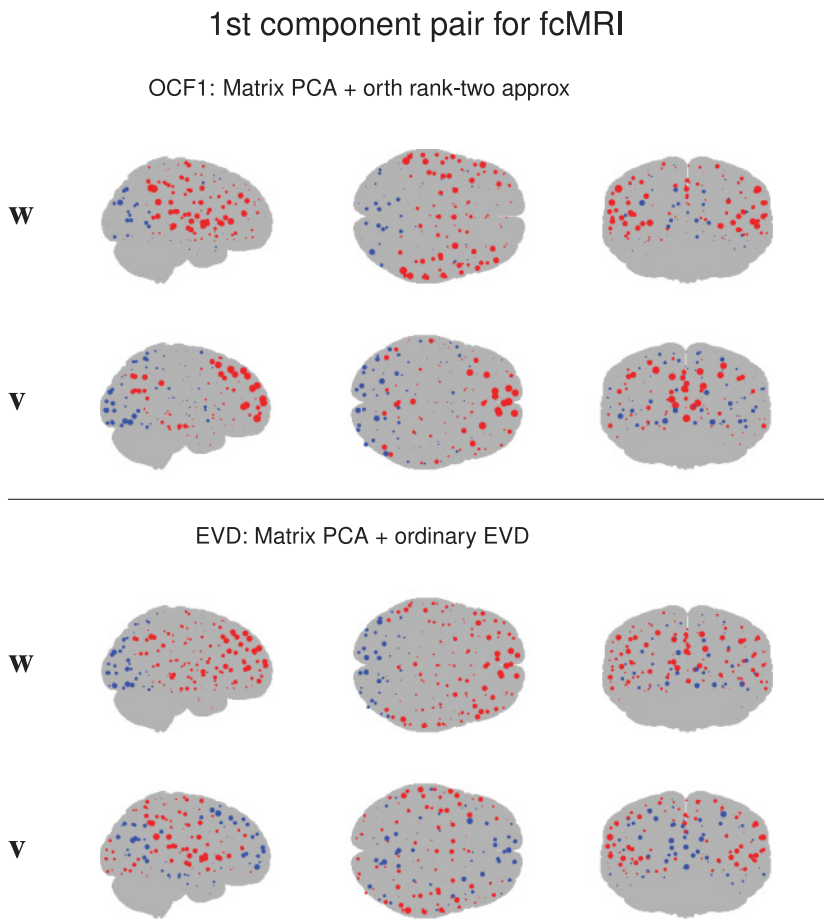

Figure 5: Results on multisubject fcMRI. The first principal component pair obtained by the two methods. (Top panel) Matrix principal component analyzed by orthogonal rank-two approximation. (Bottom panel) For comparison, matrix PC analyzed by ordinary EVD. In each panel, the upper and lower parts give the two spatial patterns **w**, **v** in the pair obtained, seen from three different angles. The brains are facing right.

6.1.2 Results and Discussion. The first two component pairs for OCF1, matrix PCA followed by the orthogonal rank-two decomposition, are shown in the upper panels of Figures 5 and 6.

The red (positive) areas of the first principal component pair (see Figure 5) reveal somewhat surprisingly two well-known resting state network configurations. The **v** resembles markedly the default mode network

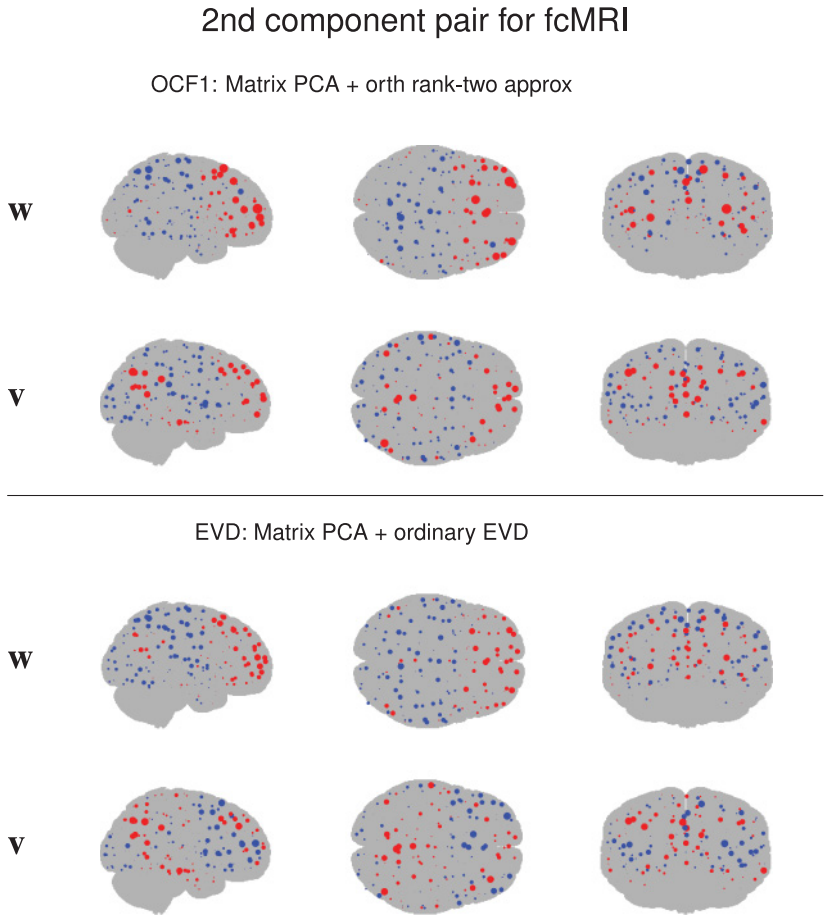

Figure 6: Results on multisubject fcMRI: the second principal component pair obtained by the two methods. For the legend, see Figure 5.

(DMN) with emphasis on the ventromedial prefrontal cortex, as well as some posterior cingulate (PCC) participation. The other spatial pattern, **w**, is formed from auditory, language areas, and sensorimotor cortices and can be regarded to form a task-positive network (TPN); however, it has to be noted that major medial cerebral arteries lie in these areas.

The second component pair (see Figure 6) again has DMN in **v**, with a bit more emphasis on the PCC and angular gyri and less on upper frontal areas. The **w** seems to consist of frontal areas of the central executive network (CEN), interestingly on both sides of the **v**'s DMN frontal regions.

These results are intriguing since the method finds well-known functional brain networks based on the interindividual variability. Previously, similar networks have been found by spatial ICA (Kiviniemi et al., 2003; van de Ven, Formisano, Prvulovic, Roeder, & Linden, 2004; Beckmann et al., 2005). In particular, the DMN seems to be most variable in its connectivity with TPN and CEN. These spatial patterns cover a substantial part of the frontoparietal functional networks. It is possible that this variability of connectivity is related to the intrasubject temporal nonstationarity of the DMN found in a previous ICA analysis (Kiviniemi et al., 2011). As the DMN is one of the most robust components in spatial ICA of resting-state activity in single subjects, the intersubject variability of DMN connectivity suggests that it performs different tasks reflecting the individual brain function of each subject.

For comparison, we also plot the eigenvectors $\mathbf{e}_{\min}$, $\mathbf{e}_{\max}$ of the matrix principal components corresponding to the smallest and largest eigenvalues in the lower panels of Figures 5 and 6. Clearly these spatial patterns are less localized and more overlapping, and thus it is difficult to associate them with individual brain areas.

Constrained PCA results (OCF2, OCF3) are not plotted here since they are visually indistinguishable from OCF1. This is at least partly attributable to the fact that OCF2 and OCF3 used OCF1 as the initial point of optimization iterations.

In addition to the spatial patterns, we can also compute the time courses of the components for each subject, which could give rise to further analyses. Here, we simply note that the average connectivity between the two areas in the first principal component were negative, while for the second principal component, they were positive (results not shown). These values are well in line with the interpretation given above.

The quantitative comparison of results for the four methods is in Figure 7. We see that all the three methods in this article improve the sparsity and reduce the overlap between the spatial patterns quite substantially compared to the baseline given by the eigenvectors of matrix PC (EVD). (Note that the reduced overlap has nothing to do with orthogonality; vector pairs are orthogonal for all the methods.) On the other hand, it is difficult to see any differences between the new proposed methods regarding these measures. Comparison of similarities of OCF1 and OCF2 components (pairs) in Figure 7c shows that while the two first component pairs are very similar, further components actually have some differences, and the fourth component pairs are completely different. In contrast, the robust version (OCF3) is very similar to OCF1 for all the five computed components (see Figure 7d).

Regarding the computation times in this analysis, we briefly mention that the total computation time was on the order of 1 minute for all the results in this section, most of which was spent in computation of basic constrained PCA (fine-tuning). Perhaps surprisingly, the robust version was much faster than the basic constrained PCA.

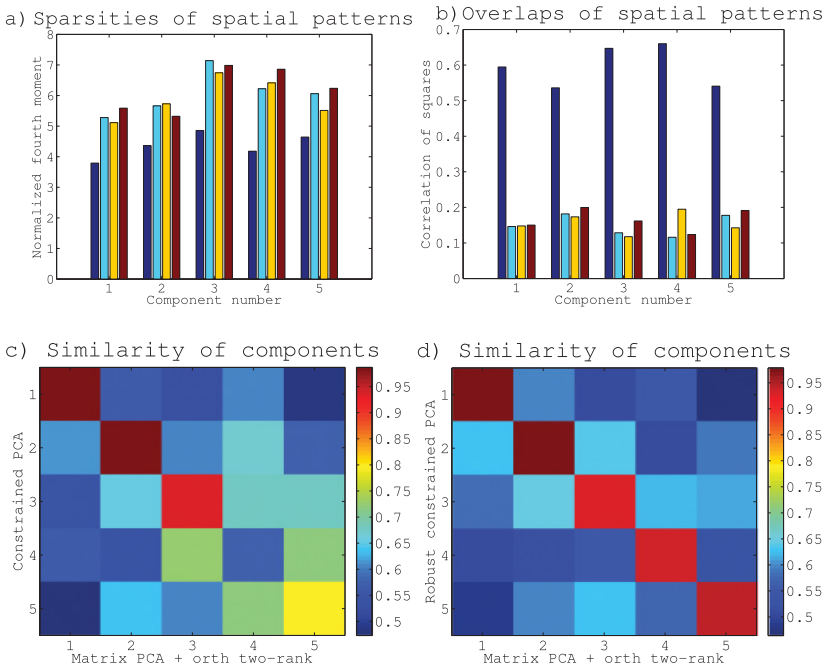

Figure 7: Quantitative comparison of methods based on fcMRI results. (a) The sparsities of fcMRI component spatial patterns for different methods. Methods are from left to right: EVD, the original eigenvectors of the matrix PCA (baseline); OCF1, matrix PCA with orthogonal rank-two decomposition; OCF2, constrained PCA; OCF3, constrained PCA with the robust version of objective. (b) The overlaps of fcMRI components for different methods; same as in panel a. (c) Similarities of estimated components for OCF1 and OCF2. The color codes the absolute values of correlations of the spatial patterns, taking the maximum over the two permutations of $\mathbf{v}$ and $\mathbf{w}$. (d) As in panel c, but using a robust version of constrained PCA (OCF3 instead of OCF2).

6.2 Nonstationarity in Magnetoencephalographic Data

6.2.1 Methods. Second, we investigated nonstationary connectivity in magnetoencephalographic (MEG) data from Ramkumar, Parkkonen, Hari, and Hyvärinen (2012). The data comprised a single 12 minute session recorded on a Elekta Neuromag 306-channel neuromagnetometer. The subject received alternating stimulation of visual, auditory, and tactile modalities, interspersed with rest periods. As basic preprocessing, we applied the signal space separation method (Taulu, Kajola, & Simola, 2004) to attenuate artifacts and motion correction and downsampled the data to 150 Hz.

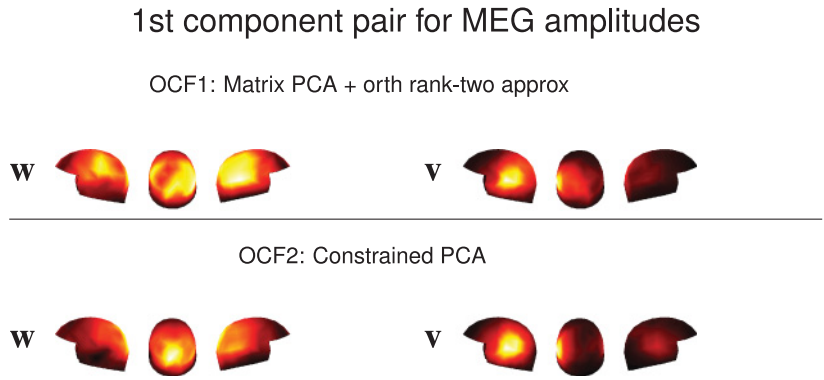

Figure 8: Results on the nonstationarity of MEG amplitudes: the first principal component pair obtained by the two methods. (Top panel) Matrix principal component analyzed by orthogonal rank-two approximation. (Bottom panel) Constrained PCA results. In each panel, the left and right parts give the two spatial patterns in the pair obtained.

We first Morlet-filtered the data using a single complex-valued Morlet filter with a center frequency of 10 Hz to extract alpha-range activity. We next performed ICA on the Morlet-filtered MEG data to separate sources of rhythmic activity. Finally, we computed the energies (envelopes) of the sources and used these as input to the proposed method.

Correlation coefficient matrices between the energies of independent components were computed in nonoverlapping time windows of a length of 5 seconds. Two variants of the method, OCF1 or matrix PCA followed by rank-two approximation (see section 2), and OCF2 or constrained PCA (see section 3) were applied on the data. Note that $\mathbf{w}$ and $\mathbf{v}$ could be interchanged, and we have here manually switched them in some plots to make the results of the two methods as similar as possible for ease of visualization.

6.2.2 Results and Discussion. We show the first component pair $(\mathbf{w}, \mathbf{v})$ for the two methods in Figure 8 and the second component pair in Figure 9. The spatial patterns are visualized on the measurement helmet by adding together the spatial patterns (squares of columns of mixing matrix) of the underlying independent components that contribute to each connectivity component. Thus, we obtain a rough plot of the spatial extents of the components.

In general, the results using constrained PCA look cleaner and more plausible. The first component pair (see Figure 8, lower panel) shows that the strongest nonstationarity is in the connectivity between an occipito-parietal area and the left temporal area. The second component pair (see Figure 9, lower panel) shows that another strongly nonstationary

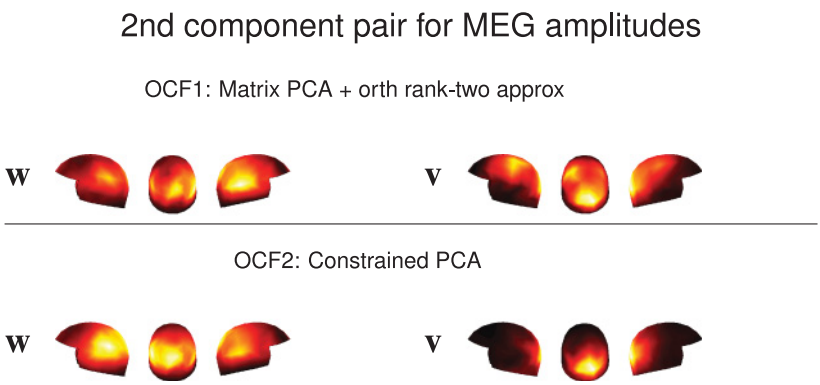

Figure 9: Results on the nonstationarity of MEG amplitudes: the second principal component pair obtained by the two methods. See the Figure 8 legend.

connectivity is between inferior occipital areas and slightly superior occipito-parietal areas.

7 Discussion

Next, we discuss the relation of our methods to well-known methods.

7.1 PARAFAC as Constrained PCA. The method has interesting relations to parallel factor analysis (PARAFAC) (Harshman, 1970), also called CANDECOMP, one of the simplest tensor decompositions (the collection of covariances can be considered a tensor). Let us first consider basic PARAFAC for the centered connectivity matrices $\tilde{\mathbf{C}}_\tau$. Considering just a single component for simplicity, we would be minimizing

$$\min_{\mathbf{w},\mathbf{v},\mathbf{r}} \sum_\tau \|\tilde{\mathbf{C}}_\tau - r_\tau \mathbf{w}\mathbf{v}^T\|^2, \tag{7.1}$$

where the idea is to approximate each of the matrices $\tilde{\mathbf{C}}_\tau$ by the same outer product $\mathbf{w}\mathbf{v}^T$ but allowing for a rescaling by the scalar r_τ. The optimal vectors would clearly be such that $\mathbf{w} = \mathbf{v}$ due to symmetry of the $\tilde{\mathbf{C}}_\tau$. Ignoring irrelevant constants, we can write the objective as

$$\min_{\mathbf{w},\mathbf{r}} \sum_\tau r_\tau^2 \|\mathbf{w}\|^4 - 2r_\tau \mathbf{w}^T \tilde{\mathbf{C}}_\tau \mathbf{w}. \tag{7.2}$$

Due to the scaling by r_τ, we can further assume without loss of generality that $\|\mathbf{w}\| = 1$, and we can solve the optimal $\mathbf{r}$ as

$$r_\tau = \mathbf{w}^T \tilde{\mathbf{C}}_\tau \mathbf{w}. \tag{7.3}$$

Plugging this into the objective, we obtain the final equivalent objective for PARAFAC as

$$\max_{\|\mathbf{w}\|=1} \sum_\tau (\mathbf{w}^T \tilde{\mathbf{C}}_\tau \mathbf{w})^2, \tag{7.4}$$

which basically means that PARAFAC is trying to find a single spatial pattern $\mathbf{w}$ such that the variability inside that spatial pattern is maximized. Superficially, this objective looks algebraically quite similar to our constrained PCA objective, but with the crucial difference that the vectors multiplying $\tilde{\mathbf{C}}_\tau$ are here equal, while in our constrained PCA, they are two different vectors constrained orthogonal.

7.2 Connection to BSS methods. Next, we point out that such PARAFAC can be seen as a blind source separation method based on the nonstationarity of variances (Matsuoka, Ohya, & Kawamoto, 1995; Pham & Cardoso, 2001; see also De Lathauwer (2006) for related developments). Consider data generated by a linear mixing $\mathbf{x} = \mathbf{Hs}$, where the variances $\sigma_{i,\tau}$ of the s_i change independently of each other. The covariance at a time segment is then equal to $\mathbf{H}\mathrm{diag}(\sigma_{i,\tau}^2)\mathbf{H}^T$, and the centered covariance is obtained by replacing the σ's by the centered versions. Assume the data have been whitened so $\mathbf{H}$ is orthogonal, and define $\mathbf{q} = \mathbf{H}^T\mathbf{w}$. Then the objective in equation 7.4 equals, in the limit of infinite number of time windows,

$$\begin{aligned}\sum_\tau (\mathbf{w}^T \tilde{\mathbf{C}}_\tau \mathbf{w})^2 &\to E_\tau\{(\mathbf{w}^T\mathbf{H}[\mathrm{diag}(\sigma_{i,\tau}^2) - \mathrm{diag}(E_\tau\{\sigma_{i,\tau}^2\})]\mathbf{H}^T\mathbf{w})^2\} \\ &= E_\tau\left\{\left(\sum_i q_i^2[\sigma_{i,\tau}^2 - E\{\sigma_{i,\tau}^2\}]\right)^2\right\} = \sum_i q_i^4 \mathrm{var}(\sigma_{i,\tau}^2).\end{aligned} \tag{7.5}$$

If we constrain $\|\mathbf{w}\| = \|\mathbf{q}\| = 1$, this reduces to a well-known optimization problem usually formulated in terms of kurtosis in ICA theory (Comon, 1994; Delfosse & Loubaton, 1995) and is well known to be solved by taking $\mathbf{w}$ to be equal to one of the columns of $\mathbf{H}$, thus finding one of the original sources s_i. Thus, we see how PARAFAC of the covariance matrices can separate sources in the same way as methods by Matsuoka et al. (1995) and Pham and Cardoso (2001).

7.3 Our Method as Orthogonally Constrained PARAFAC. We can further show that our constrained PCA method can be obtained from a PARAFAC with special constraints, followed by a transformation. Let us consider a PARAFAC with two components (here, each component

corresponding to one outer product) where we constrain the r_τ to have opposite signs and the same absolute value for the two components, and the two vectors $\mathbf{a}$ and $\mathbf{b}$ to be orthogonal:

$$\min_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^T\mathbf{b}=0} \sum_\tau \|\tilde{\mathbf{C}}_\tau - r_\tau(\mathbf{a}\mathbf{a}^T - \mathbf{b}\mathbf{b}^T)\|^2. \tag{7.6}$$

In other words, we approximate the matrices $\tilde{\mathbf{C}}_\tau$ by the difference of the two outer products, so that the scaling r_τ is applied on that difference instead of having a separate scaling variable for each outer product.

Here, we assume that the PARAFAC uses the outer product $\mathbf{a}\mathbf{a}^T$ instead of using two different vectors, say, $\mathbf{a}\mathbf{c}^T$. This is not really a constraint since it is clear that the optimal approximation will use the same vector due to the symmetry of $\tilde{\mathbf{C}}_\tau$. Expanding the squared norm, we obtain

$$\begin{aligned}\min_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^T\mathbf{b}=0} &\sum_\tau -2r_\tau(\mathbf{a}^T\tilde{\mathbf{C}}_\tau\mathbf{a}^T - \mathbf{b}^T\tilde{\mathbf{C}}_\tau\mathbf{b}) \\ &+ r_\tau^2(\|\mathbf{a}\|^4 + \|\mathbf{b}\|^4 - 2(\mathbf{a}^T\mathbf{b})^2) + \text{const.},\end{aligned} \tag{7.7}$$

from which we get $r_\tau = \mathbf{a}^T\tilde{\mathbf{C}}_\tau\mathbf{a}^T - \mathbf{b}^T\tilde{\mathbf{C}}_\tau\mathbf{b}$, and an equivalent optimization problem is

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^T\mathbf{b}=0} \sum_\tau (\mathbf{a}^T\tilde{\mathbf{C}}_\tau\mathbf{a} - \mathbf{b}^T\tilde{\mathbf{C}}_\tau\mathbf{b})^2. \tag{7.8}$$

This is the same as our reformulated constrained PCA in equation 3.3 with optimal r_τ plugged in, except that we have the constraint of orthogonality. However, it was shown in the proof of theorem 1 that the optimal $\mathbf{a}$ and $\mathbf{b}$ in the optimization of equation 3.4 are always orthogonal, so the same $\mathbf{a}$ and $\mathbf{b}$ maximize this objective as well.

7.4 Checking and Comparing Model Assumptions. Thus, we see that the constrained PCA method can be seen to be a constrained form of PARAFAC. This might seem to imply that there is little reason to apply constrained PCA since PARAFAC is a more general method. Such reasoning is invalid, because in the PARAFAC framework, it is the $\mathbf{a}$ and $\mathbf{b}$ that are considered to define the components (e.g., in terms of spatial patterns). In our framework, we transform them into $\mathbf{w}$, $\mathbf{v}$, by taking sums and differences, in order to get the spatial patterns. Thus, the components obtained are completely different. Furthermore, one of the main motivations for the constrained PCA method is that by inputting the constraints in the optimization problem itself, we can get better estimates of the relevant parameters.

Basically, which method should be used depends on whether the assumptions in our model are correct. In particular, we assume that the main

nonstationarity is in correlations between the components, while the BSS framework, which is equivalent to constrained PARAFAC, assumes that it is the variances of the components that are nonstationary. For example, in raw EEG/MEG data (as opposed to energies of sources, which we used in our experiments in section 6.2), the nonstationarity of variances is a well-known phenomenon (Hyvärinen, Ramkumar, Parkkonen, & Hari, 2010), and it would probably be a bad idea to apply our method on such data. The same applies for natural image sequences (Hyvärinen, Hurri, & Hoyer, 2009). The situation is very different in fMRI. The nonstationarity of variances does not seem to be a phenomenon of much interest in the neuroimaging community in contrast to nonstationarity in connectivity, a topic of great current interest (Leonardi et al., 2013). Thus, it may be reasonable to assume that the nonstationarity of connectivity is a more prominent phenomenon.

If the assumptions in our model are not fulfilled, it is possible that our algorithm is in fact doing something like blind source separation based on nonstationarity, where the sources will not be given by the $\mathbf{w}, \mathbf{v}$ but by $\mathbf{a}, \mathbf{b}$ (at least assuming that the data are white or close to white). Such a possibility could in some cases be checked by investigating which pair gives a more plausible estimate of spatial patterns. For example, if we assume that the spatial patterns are sparse, the sparsities of these solutions could be compared, which is what we did for the fMRI data in section 6.1.

7.5 Relationship to Common Spatial Patterns. The method has some similarity to common spatial patterns (CSP). Typically in CSP, we assume just two connectivity matrices and analyze their difference. As noted in section 3.1, in the case of just two time segments, PCA in the connectivity matrix space in fact results in the analysis of the difference of the two connectivity matrices in our framework as well. Extensions to many connectivity matrices exist but are not too widely used (Dornhege, Blankertz, Curio, & Müller, 2008).

So the crucial difference is in how the difference of connectivity matrices is analyzed. In CSP, we would use well-known low-rank approximations—vectors equal to the eigenvectors. This in in contrast to our method, where we transform the eigenvectors by taking sums and differences. On the other hand, CSP has an interesting affinity to our method in the sense that vectors corresponding to both the dominant positive and negative eigenvalues are used.

Another important difference is that CSP is usually formulated in a supervised case, where the segmentation of the data corresponds to some externally defined classes. However, nothing would prevent CSP from being applied, for example, to the first half and the second half of resting-state data. Further related work in classification was published by Tomioka and Müller (2010), where a penalty for the rank of a matrix classifier leads to a low-rank classifier.

We can easily define a supervised variant of our method, simply by defininining the segments in which connectivity is computed to correspond to segments predefined by an experimental protocol or other supervisory labels. In such a case, we could even use the components for classification. However, we leave a detailed investigation of such supervised methods for future research and concentrate on purely unsupervised learning here.

8 Conclusion

We proposed a new method, orthogonal connectivity factorization, for analyzing the dynamics of connectivity patterns or, in general, the differences of a set of connectivity matrices. The main novelty here is to find components that are linear in the original data. This is in contrast to conventional methods that find linear components in the connectivity matrix space. The main goal is to analyze the variability (nonstationarity) of connectivities in a way that is intuitively comprehensible and easy to visualize. Furthermore, our method considerably reduces the number of free parameters to estimate and is thus likely to lead to better results when data are scarce.

The method finds two linear, orthogonal components of the data such that the connectivity between them is maximally nonstationary. The orthogonality constraint is the main mathematical novelty and sets the method apart from related methods, such as common spatial patterns (CSP), or blind source separation using nonstationarity (Matsuoka et al., 1995). Intuitively, the method is related to a constrained form of PCA, but according to the simulations, it seems to have better identifiability, similarly to ICA or PARAFAC. The orthogonality constraint improved the fMRI analysis results substantially.

We presented three variants of the method: OCF1 to OCF3. The first is based on further analyzing the PCA of connectivity matrices by a dedicated, orthogonal rank-two approximation, and the two others are based on formulating a constrained PCA objective function and optimizing it by a tailor-made algorithm. Constrained PCA produced better results on some simulated data, especially when there are component pairs with similar statistics, which leads to the well-known identifiability problem of PCA. Constrained PCA seemed to produce better results even with some real data, although only on one of the data sets, and the validation of the results is purely visual since ground truth was not known. However, the basic idea of performing the novel orthogonal rank-two approximation on results of matrix PCA is surprisingly efficient as well.

While the method was mainly motivated by brain imaging data in this article, it seems to have extremely wide applicability, and we did not make any assumptions that would somehow be particular to brain imaging data. For example, it is easy to imagine applications on word co-occurrence matrices, correlations in financial time series, and regional feature covariances in images.

A Matlab package implementing the method is publicly available on the Internet (http://www.cs.helsinki.fi/u/ahyvarin/code/ocf/).

Appendix: Proofs of Theorems

A.1 Proof of Theorem 2. We can equivalently formulate the problem as

$$\min_{\alpha, \|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \left\| \mathbf{K} - \frac{\alpha}{2}(\mathbf{v}\mathbf{w}^T + \mathbf{w}\mathbf{v}^T) \right\|^2, \tag{A.1}$$

where α is a scalar and the norm is the Frobenius norm. Then we have

$$\left\| \mathbf{K} - \frac{\alpha}{2}(\mathbf{v}\mathbf{w}^T + \mathbf{w}\mathbf{v}^T) \right\|^2 = \|\mathbf{K}\|^2 - 2\alpha \mathbf{v}^T\mathbf{K}\mathbf{w} + \frac{\alpha^2}{2} \tag{A.2}$$

under the unit norm and orthogonality constraints. Maximizing this with respect to α gives $\alpha = 2\mathbf{v}^T\mathbf{K}\mathbf{w}$, and plugging this alpha into the objective function transforms the optimization problem into

$$\min_{\|\mathbf{w}\|=\|\mathbf{v}\|=1, \mathbf{w}^T\mathbf{v}=0} \|\mathbf{K}\|^2 - 2(\mathbf{w}^T\mathbf{K}\mathbf{v})^2, \tag{A.3}$$

which is equivalent to equation 2.2 since we can always switch the sign of, say, $\mathbf{w}$ and thus have to actually maximize the absolute value of the bilinear form in equation 2.2. Using theorem 1, we further see that the value of the objective is as announced in the theorem.

The squared Frobenius norm of a symmetric matrix equals, by diagonalization, the sum of the squared eigenvalues. Denote by c the sum of the squares of the eigenvalues of $\mathbf{K}$, excluding the smallest and the largest. We therefore have

$$\|\mathbf{K}\|^2 = \lambda_{\max}^2 + \lambda_{\min}^2 + c. \tag{A.4}$$

Thus, the optimal value of the objective function can be manipulated as

$$\begin{aligned} \|\mathbf{K}\|^2 - 2(\mathbf{w}^T\mathbf{K}\mathbf{v})^2 &= (\lambda_{\max}^2 + \lambda_{\min}^2 + c) - \frac{1}{2}(\lambda_{\max} - \lambda_{\min})^2 - \\ &= \frac{1}{2}(\lambda_{\max} + \lambda_{\min})^2 + c. \end{aligned} \tag{A.5}$$

This is clearly always nonnegative, and it is zero if and only if $c = 0$, which means that the matrix has rank two, and $\lambda_{\min} = -\lambda_{\max}$.

A.2 Proof of Theorem 3. Given $\mathbf{a}$, $\mathbf{b}$, the optimal r_τ is as in equation 3.6, with

$$c = \frac{1}{\sqrt{\sum_\tau [\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b}]^2}}. \tag{A.6}$$

Plugging this into the objective, equation 3.4, we obtain the optimization problem,

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \sqrt{\sum_\tau (\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b})^2}. \tag{A.7}$$

Maximizing this objective is equivalent to maximizing

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \sum_\tau (\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b})^2, \tag{A.8}$$

since an increasing transformation of the objective does not change the optimizing point.

Now assume $\mathbf{a}$, $\mathbf{b}$, $\mathbf{r}$ maximize the transformed objective, equation 3.4. Then, by the logic just given, the $\mathbf{a}$, $\mathbf{b}$ maximize equation A.8. On the other hand, for any given $\mathbf{r}$, the optimization of equation 3.4 is equivalent to the problem

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \mathbf{a}^T \mathbf{M} \mathbf{a} - \mathbf{b}^T \mathbf{M} \mathbf{b} \tag{A.9}$$

with

$$\mathbf{M} = \sum_\tau r_\tau \tilde{\mathbf{C}}_\tau. \tag{A.10}$$

Since this is of the same form as the problem in the proof of theorem 1, the $\mathbf{a}$ and $\mathbf{b}$ are orthogonal, as shown in the proof of theorem 1. Thus, the optimization of equation 3.4 gives the same $\mathbf{a}$, $\mathbf{b}$ as the same optimization of equation A.8, with an added orthogonality constraint:

$$\max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^T\mathbf{b}=0} \sum_\tau (\mathbf{a}^T \tilde{\mathbf{C}}_\tau \mathbf{a} - \mathbf{b}^T \tilde{\mathbf{C}}_\tau \mathbf{b})^2. \tag{A.11}$$

In this optimization, we can transform the variables as in equation 2.7, and we obtain the original constrained PCA objective, equation 3.3. Thus, an optimum of equation 3.4 is an optimum of equation 3.3.

Conversely, assume that $\mathbf{w}$, $\mathbf{v}$ optimize the original objective, equation 3.3. Transforming as in equation 2.4, the ensuing $\mathbf{a}$, $\mathbf{b}$ equivalently solve

the problem in equation A.11. As shown in the preceding paragraph, the constraint of orthogonality on $\mathbf{a}, \mathbf{b}$ does not change the optimum; that is, the $\mathbf{a}, \mathbf{b}$ solve equation A.8 as well. Thus, see that the $\mathbf{a}, \mathbf{b}$ also optimize equation 3.4, together with the $\mathbf{r}$ given in equation 3.6.

A.3 Proof of Theorems 4 and 5. We prove here theorem 5. The proof of theorem 4 is obtained as a simplification of the proof below by replacing $\mathbf{D}_\tau^{-1}$ by identity and any correlation coefficient matrices by covariance matrices.

First, we have for the precision matrix of observed data

$$\begin{aligned}\operatorname{cov}(\mathbf{x}(\tau))^{-1} &= [\mathbf{D}_\tau \mathbf{H} \operatorname{cov}(\mathbf{s}(\tau)) \mathbf{H}^T \mathbf{D}_\tau]^{-1} = \mathbf{D}_\tau^{-1} \mathbf{H} \operatorname{cov}(\mathbf{s}(\tau))^{-1} \mathbf{H}^T \mathbf{D}_\tau^{-1} \\ &= \mathbf{D}_\tau^{-1} [I + z(\tau)(\mathbf{h}_1 \mathbf{h}_2^T + \mathbf{h}_2 \mathbf{h}_1^T)] \mathbf{D}_\tau^{-1}. \end{aligned} \tag{A.12}$$

Denoting the observed covariance matrix in time segment τ by $\hat{\mathbf{C}}_\tau$, we can formulate the gaussian likelihood of the model defined by equations 4.9 and 4.5 as

$$\begin{aligned}\log L(\mathbf{H}, z(\tau), \mathbf{D}_\tau) = &- \sum_\tau \frac{1}{2} \operatorname{tr}(\hat{\mathbf{C}}_\tau^T \mathbf{D}_\tau^{-1} (I + z(\tau)[\mathbf{h}_1 \mathbf{h}_2^T + \mathbf{h}_2 \mathbf{h}_1^T]) \mathbf{D}_\tau^{-1}) \\ &+ \frac{1}{2} \log |\det(I + z(\tau)[\mathbf{h}_1 \mathbf{h}_2^T + \mathbf{h}_2 \mathbf{h}_1^T])| \\ &+ \log |\det \mathbf{D}_\tau^{-1}| + \log p(z(1), \ldots, z(k)), \end{aligned} \tag{A.13}$$

where $p(z(1), \ldots, z(k))$ is the (prior) pdf of the $z(\tau)$, and the summation is taken over all the terms. Using logic similar to the proof of theorem 1 and the orthogonality of $\mathbf{H}$, we see that the eigenvalues of the first matrix whose determinant is being computed above are equal to $1 - z(\tau)$ and $1 + z(\tau)$, and the rest are ones. Furthermore, we can arrange the product in the first trace term as according to the well-known commutativity of the trace. Thus, we have

$$\begin{aligned}\log L(\mathbf{H}, z(\tau), \mathbf{D}_\tau) = &- \sum_\tau z(\tau) \mathbf{h}_1^T \mathbf{D}_\tau^{-1} \hat{\mathbf{C}}_\tau \mathbf{D}_\tau^{-1} \mathbf{h}_2 + \frac{1}{2} \log |1 - z(\tau)^2| \\ &+ \log |\det \mathbf{D}_\tau^{-1}| + \frac{1}{2} \operatorname{tr}(\mathbf{D}_\tau^{-1} \hat{\mathbf{C}}_\tau \mathbf{D}_\tau^{-1}) \\ &+ \log p(z(1), \ldots, z(k)) + \text{const.}, \end{aligned} \tag{A.14}$$

where the constant term does not depend on z or $\mathbf{H}$. Now, in theorem 5, we approximate $\mathbf{D}_\tau$ by the standard deviations of the data variables (i.e., its sample version). This approximation is exact in the limit of infinitesimal $z(\tau)$. Together with the nonoverlapping property, the approximation leads in fact to an error of $O(z^2)$ if we neglect finite-sample errors because we

have

$$\begin{aligned}\operatorname{cov}(\mathbf{x}(\tau)) = I - z(\tau)(\mathbf{h}_1\mathbf{h}_2^T + \mathbf{h}_2\mathbf{h}_1^T) - z(\tau)^2(\mathbf{h}_1\mathbf{h}_1^T \\ + \mathbf{h}_2\mathbf{h}_2^T) + O(z^3),\end{aligned} \tag{A.15}$$

as can be verified by simply multiplying this by $\operatorname{cov}(\mathbf{x}(\tau))^{-1}$. Now, due to the nonoverlapping property, $\mathbf{h}_1\mathbf{h}_2^T$ has zero diagonal, which means that the variances are changed only by $O(z^2)$. Thus,

$$\operatorname{cov}(\mathbf{x}(\tau)) = \mathbf{D}_\tau[I - z(\tau)(\mathbf{h}_1\mathbf{h}_2^T + \mathbf{h}_2\mathbf{h}_1^T)]\mathbf{D}_\tau^{-1} + O(z^2). \tag{A.16}$$

With this approximation, $\mathbf{D}_\tau^{-1}\hat{\mathbf{C}}_\tau\mathbf{D}_\tau^{-1}$ is nothing else than the sample correlation coefficient matrix $\bar{\mathbf{C}}_\tau$ in time segment τ, and we have the approximation of the likelihood

$$\begin{aligned}\log L(\mathbf{H}, z(\tau)) = -\sum_\tau z(\tau)\mathbf{h}_1^T\bar{\mathbf{C}}_\tau\mathbf{h}_2 + \frac{1}{2}\sum_\tau \log|1 - z(\tau)^2| \\ + \log p(z(1), \ldots, z(k)) + \text{const.} + O(z^3),\end{aligned} \tag{A.17}$$

where the constant term does not depend on z or $\mathbf{H}$.

Next, we implement the zero-mean constraint of z in equation 4.6. We use the simple identity $E\{(x - Ex)(y - Ey)\} = E\{x(y - Ey)\} = E\{(x - Ex)y\}$ applicable for any two random variables, which in this case, applied on z and $\mathbf{h}_1^T\bar{\mathbf{C}}_\tau\mathbf{h}_2$ and averaging taking place over τ, implies

$$\begin{aligned}\sum_\tau z(\tau)\mathbf{h}_1^T\bar{\mathbf{C}}_\tau\mathbf{h}_2 &= \sum_\tau z(\tau)(\mathbf{h}_1^T\bar{\mathbf{C}}_\tau\mathbf{h}_2 - \frac{1}{T}\sum_\tau \mathbf{h}_1^T\bar{\mathbf{C}}_\tau\mathbf{h}_2) \\ &= \sum_\tau z(\tau)\mathbf{h}_1^T[\bar{\mathbf{C}}_\tau - \frac{1}{k}\sum_\tau \bar{\mathbf{C}}_\tau]\mathbf{h}_2,\end{aligned} \tag{A.18}$$

and thus we define the correlation coefficient matrices with the average correlations removed as

$$\tilde{\mathbf{C}}_\tau = \bar{\mathbf{C}}_\tau - \frac{1}{k}\sum_\tau \bar{\mathbf{C}}_\tau. \tag{A.19}$$

We can write the approximate likelihood as

$$\begin{aligned}\log L(\mathbf{H}, z(\tau)) = -\sum_\tau z(\tau)\mathbf{h}_1^T\tilde{\mathbf{C}}_\tau\mathbf{h}_2 + \frac{1}{2}\sum_\tau \log|1 - z(\tau)^2| \\ + \log p(z(1), \ldots, z(k)) + \text{const.} + O(z^3),\end{aligned} \tag{A.20}$$

which holds even for z without the zero-mean constraint. Thus, we see that we can equivalently analyze the correlation coefficients whose global means have been removed, which makes intuitive sense, and was done above.

As in the theorem, we define

$$p(z) = \frac{1}{2\epsilon} I_{[-\epsilon,\epsilon]}. \tag{A.21}$$

To integrate out $z(\tau)$, let us first make a Taylor expansion for each of them separately. Denoting $c_\tau = \mathbf{h}_1^T \tilde{\mathbf{C}}_\tau \mathbf{h}_2$, simple Taylor approximations give for a single time point (where indices t are dropped for simplicity):

$$\begin{aligned} L(\mathbf{H}, z) &= \exp(\log L(\mathbf{H}, z)) = \exp(-zc)\sqrt{1 - z^2} + O(z^3) \\ &= [1 - zc + \frac{1}{2}z^2c^2 + o(z^2)]\left[1 - \frac{1}{2}z^2 + o(z^2)\right] + O(z^3) \\ &= 1 - zc + \frac{1}{2}(c^2 - 1)z^2 + o(z^2). \end{aligned} \tag{A.22}$$

We need to integrate this for all τ, over $[-\epsilon, \epsilon]^T$ with the constraint of zero mean in equation 4.6:

$$\begin{aligned} &\frac{1}{(2\epsilon^2)} \int_S \prod_\tau [1 - z(\tau)c_\tau + \frac{1}{2}(c_\tau^2 - 1)z(\tau)^2 \\ &\quad + o(z(\tau)^2)] dz(1), dz(2), \ldots, dz(k), \end{aligned} \tag{A.23}$$

where $S = \{x \in [-\epsilon, \epsilon]^T, \sum_\tau x_\tau = 0\}$. When the product in the integrand is written out, it will have a constant term equal to one, not depending on the parameters (i.e., c); the first-order terms $-z(\tau)c_\tau$, which integrate to zero by symmetry; and second-order terms, which are the relevant ones. They are of two kinds, $\frac{1}{2}(c_\tau^2 - 1)z(\tau)^2$ and $z(\tau)c_\tau z(\tau')c_{\tau'}$. The terms $\frac{1}{2}z(\tau)^2$ can be ignored since they do not depend on c_τ. The remaining terms can be arranged as $\frac{1}{2}(\sum_\tau c_\tau z(\tau))^2$. Since the c_τ have zero mean, this term does not depend on the mean of the $z(\tau)$, and it in fact is equal to $\frac{1}{2}\sum_\tau (c_\tau z(\tau))^2$. Thus, we are left with an integral in which all variables are separable, and we simply need to compute (ignoring lower-order terms)

$$\frac{1}{(2\epsilon^2)} \int_{-\epsilon}^{\epsilon} \int_{-\epsilon}^{\epsilon} \cdots \int_{-\epsilon}^{\epsilon} \frac{1}{2} \sum_\tau (c_\tau z(\tau))^2 dz(1), dz(2), \ldots, dz(k) = \frac{\epsilon^3}{6} \sum_\tau c_\tau^2, \tag{A.24}$$

which gives the approximation in the theorem.

Acknowledgments

We are grateful to Okito Yamashita for information on fcMRI databases. A.H. was supported by Academy of Finland, Centre-of-Excellence in Inverse Problems Research and Centre-of-Excellence Algorithmic Data Analysis. J.H. and M.K. were supported by the Japanese Ministry of Internal Affairs and Communication, project "Novel and Innovative R&D Making Use of Brain Structures." J.H. was further partially supported by JSPS KAKENHI grant 25730155.

References

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, *24*, 663–676.

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.*, *360*(1457), 1001–1013.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing*, *36*, 287–314.

De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM J. on Matrix Analysis and Applications*, *28*(3), 642–666.

Delfosse, N., & Loubaton, P. (1995). Adaptive blind separation of independent sources: A deflation approach. *Signal Processing*, *45*, 59–83.

Dornhege, G., Blankertz, B., Curio, G., & Müller, K.-R. (2008). Boosting bit-rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, *51*(6), 993–1002.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327.

Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *Proceedings of the Int. Conf. on Advances in Social Networks Analysis and Mining* (pp. 176–183). Piscataway, NJ: IEEE Press.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, *16*(1), 1–84.

Hyvärinen, A., Hirayama, J., & Kawanabe, M. (2014). Dynamic connectivity factorization: Interpretable decompositions of non-stationarity. In *Proc. Int. Workshop on Pattern Recognition in Neuroimaging* (pp. 1–4). Piscataway, NJ: IEEE Press.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics*. New York: Springer-Verlag.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Hoboken, NJ: Wiley Interscience.

Hyvärinen, A., Ramkumar, P., Parkkonen, L., & Hari, R. (2010). Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *NeuroImage*, *49*(1), 257–271.

Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Machine Learning Research*, *11*, 517–553.

Kiviniemi, V., Kantola, J. H., Jauhiainen, J., Hyvärinen, A., & Tervonen, O. (2003). Independent component analysis of nondeterministic fMRI signal sources. *NeuroImage*, *19*(2), 253–260.

Kiviniemi, V., Vire, T., Remes, J., Abou-Elseoud, A., Starck, T., Tervonen, O., & Nikkinen, J. (2011). A sliding time-window ICA reveals spatial variability of default mode network in time. *Brain Connectivity*, *1*, 339–347.

Kolar, M., Song, L., Ahmed, A., & Xing, E. P. (2010). Estimating time-varying networks. *Annals of Applied Statistics*, *4*(1), 94–123.

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.-M., Schluep ... Van De Ville, D. (2013). Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage*, *83*, 937–950.

Liu, S., Quinn, J. A., Gutmann, M. U., & Sugiyama, M. (2013). Direct learning of sparse changes in Markov networks by density ratio estimation. In *Machine Learning and Knowledge Discovery in Databases* (pp. 596–611). New York: Springer.

Matsuoka, K., Ohya, M., & Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, *8*(3), 411–419.

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., & Montana, G. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, *103*, 427–443.

Pham, D.-T., & Cardoso, J.-F. (2001). Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, *49*(9), 1837–1848.

Ramkumar, P., Parkkonen, L., Hari, R., & Hyvärinen, A. (2012). Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. *Human Brain Mapping*, *33*(7), 1648–1662.

Robinson, J. W., & Hartemink, A. J. (2009). Non-stationary dynamic Bayesian networks. In D. Köller, D. Schuurmans, & L. Bottou (Eds.), *Advances in neural information processing systems*, 21 (pp. 1369–1376). Cambridge, MA: MIT Press.

Robinson, L. F., & Priebe, C. E. (2012). *Detecting time-dependent structure in network data via a new class of latent process models*. Arxiv preprint arXiv:1212.3587.

Sasaki, H., Gutmann, M. U., Shouno, H., & Hyvärinen, A. (2014). Estimating dependency structures for non-gaussian components with linear and energy correlations. In *Proc. Int. Conf. on Artificial Intelligence and Statistics* (vol. *33*, pp. 868–876). JMLR.

Tantipathananandh, C., & Berger-Wolf, T. Y. (2011). Finding communities in dynamic social networks. In *Int. Conf. on Data Mining* (pp. 1236–1241). Piscataway, NJ: IEEE Press.

Taulu, S., Kajola, M., & Simola, J. (2004). Suppression of interference and artifacts by the signal space separation method. *Brain Topography*, *16*, 269–275.

Tomioka, R., & Müller, K.-R. (2010). A regularized discriminative framework for EEG analysis with application to brain–computer interface. *NeuroImage*, *49*(1), 415–432.

van de Ven, V. G., Formisano, E., Prvulovic, D., Roeder, C. H., & Linden, D. E. (2004). Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Human Brain Mapping*, *22*(3), 165–178.

Xuan, X., & Murphy, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 1055–1062). JMLR.

Yang, J., Zhang, D., Frangi, A. F., & Yang, J.-Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 131–137.

Received July 28, 2015; accepted October 21, 2015.