# Detecting synchronization clusters in multivariate time series via coarse-graining of Markov chains

Carsten Allefeld[1,*] and Stephan Bialonski[2,3]

[1]*Department of Empirical and Analytical Psychophysics, Institute for Frontier Areas of Psychology and Mental Health, Wilhelmstrasse 3a, 79098 Freiburg, Germany*
[2]*Department of Epileptology, Neurophysics Group, University of Bonn, Sigmund-Freud-Strasse 25, 53105 Bonn, Germany*
[3]*Helmholtz Institute for Radiation and Nuclear Physics, University of Bonn, Nussallee 14-16, 53115 Bonn, Germany*

Synchronization cluster analysis is an approach to the detection of underlying structures in data sets of multivariate time series, starting from a matrix $R$ of bivariate synchronization indices. A previous method utilized the eigenvectors of $R$ for cluster identification, analogous to several recent attempts at group identification using eigenvectors of the correlation matrix. All of these approaches assumed a one-to-one correspondence of dominant eigenvectors and clusters, which has however been shown to be wrong in important cases. We clarify the usefulness of eigenvalue decomposition for synchronization cluster analysis by translating the problem into the language of stochastic processes, and derive an enhanced clustering method harnessing recent insights from the coarse-graining of finite-state Markov processes. We illustrate the operation of our method using a simulated system of coupled Lorenz oscillators, and we demonstrate its superior performance over the previous approach. Finally we investigate the question of robustness of the algorithm against small sample size, which is important with regard to field applications.

PACS number(s): 05.45.Tp, 05.45.Xt, 02.50.Ga, 05.10.−a

## I. INTRODUCTION

Studying the dynamics of complex systems is relevant in many scientific fields, from meteorology [1] and geophysics [2] to economics [3] and neuroscience [4,5]. In many cases, this complex dynamics is to be conceived as arising through the interaction of subsystems, and it can be observed in the form of multivariate time series where measurements in different channels are taken from the different parts of the system. The degree of interaction of two subsystems can then be quantified using bivariate measures of signal interdependence [6–8]. A wide variety of such measures has been proposed, from the classic linear correlation coefficient over frequency-domain variants such as magnitude squared coherence [9] to general entropy-based measures [10]. A more specific model of complex dynamics that has found a large number of applications is that of a set of self-sustained oscillators whose coupling leads to a synchronization of their rhythms [11,12]. Especially the discovery of the phenomenon of phase synchronization [13] led to the widespread use of synchronization indices in time series analysis [14–16].

However, by applying bivariate measures to multivariate data sets an $N$-dimensional time series is described by an $N \times N$ matrix of bivariate indices, which leads to a large amount of mostly redundant information. Especially if additional parameters come into play (nonstationarity of the dynamics, external control parameters, experimental conditions) the quantity of data can be overwhelming. Then it becomes necessary to reduce the complexity of the data set in such a way as to reveal the relevant underlying structures, that is, to use genuinely multivariate analysis methods that are able to detect patterns of multichannel interaction.

One way to do so is to trace the observed pairwise correspondences back to a smaller set of direct interactions using, e.g., partial coherence [6,17], an approach that has recently been extended to phase synchronization [18]. Another and complementary way to achieve such a reduction is cluster analysis, that is, a separation of the parts of the system into different groups, such that signal interdependencies within each group tend to be stronger than in between groups. This description of the multivariate structure in the form of clusters can eventually be enriched by the specification of a degree of participation of an element in its cluster. The straightforward way to obtain clusters by applying a threshold to the matrix entries has often been used [5,19,20], but it is very susceptible to random variation of the indices. As an alternative, several attempts have recently been made to identify clusters using eigenvectors of the correlation matrix [19,21,22], which were motivated by the application of random matrix theory to empirical correlation matrices [3,23].

In the context of phase synchronization analysis, a first approach to cluster analysis was based on the derivation of a specific model of the internal structure of a synchronization cluster [24,25]. The resulting method made the simplifying assumption of the presence of only one cluster in the given data set, and focused on quantifying the degree of involvement of single oscillators in the global dynamics. Going beyond that, the participation index method [26] defined a measure of oscillator involvement based on the eigenvalues and eigenvectors of the matrix of bivariate synchronization indices, and attributed oscillators to synchronization clusters based on this measure.

But despite the apparent usefulness of eigenvalue decomposition for the purposes of group identification, beyond some phenomenological evidence no good reason has been put forward why an eigenvector of a synchronization matrix should directly indicate the system elements taking part in a

_____
*allefeld@igpp.de

cluster. Moreover, in a recent survey of the performance of synchronization cluster analysis in simulation and field data [27] it has been shown that there are important special cases—clusters of similar strength that are slightly synchronized to each other—where the assumed one-to-one correspondence of eigenvectors and clusters is completely lost.

In this paper, we provide a better understanding of the role of eigenvectors in synchronization cluster analysis, and we present an improved method for detecting synchronization clusters, the eigenvector space approach. The organization of the paper is as follows: In Sec. II, we briefly recall the definition of the matrix of bivariate synchronization indices $R$ as the starting point of the analysis. We motivate its transformation into a stochastic matrix $P$ describing a Markov chain and detail the properties of that process. Utilizing recent results on the coarse-graining of finite-state Markov processes [28–30] we derive our method of synchronization cluster analysis, and we illustrate its operation using a system of coupled Lorenz oscillators. In Sec. III, we compare the performance of the eigenvector space method with that of the previous approach, the participation index method [26], and we investigate its behavior in the case of small sample size, which is important with regard to the application to empirical data.

## II. METHOD

### A. Measuring synchronization

Synchronization is a generally occurring phenomenon in the natural sciences, which is defined as the dynamical adjustment of the rhythms of different oscillators [11]. Because an oscillatory dynamics is described by a phase variable $\phi$, a measure of synchronization strength is based on the instantaneous phases $\phi_{im}$ of oscillators $i = 1, \dots, N$, where the index $m$ enumerates the values in a sample of size $n$. The nowadays commonly used bivariate index of phase synchronization strength [15,16,18,20,24,26,27] results from the application of the first empirical moment of a circular random variable [31] to the distribution of the phase difference of the two oscillators,

$$R_{ij} = \left| \frac{1}{n} \sum_{m=1}^{n} e^{i(\phi_{im}-\phi_{jm})} \right|. \tag{1}$$

The measure takes on values from the interval $[0,1]$, representing the continuum from no synchronization to perfect synchronization of oscillators $i$ and $j$; the matrix $R$ is symmetric, its diagonal being composed of ones. Special care must be taken in applying this definition to empirical data, because the interpretation of $R$ as a synchronization measure in the strict sense only holds if phase values were obtained from different self-sustained oscillators.

The determination of the phase values $\phi_{im}$ generally depends on the kind of system or data to be investigated. For the analysis of scalar real-valued time series $s_i(t)$ that are characterized by a pronounced dominant frequency, the standard approach utilizes the associated complex-valued ana-

lytic signal $z_i(t)$ [32], within which every harmonic component of $s_i(t)$ is extended to a complex harmonic. The analytic signal is commonly defined [13] as

$$z_i(t) = s_i(t) + iHs_i(t), \tag{2}$$

where $Hs_i$ denotes the Hilbert transform of the signal $s_i$,

$$Hs_i(t) = \frac{1}{\pi} \text{P.V.} \int_{-\infty}^{\infty} \frac{s_i(t')}{t-t'} dt', \tag{3}$$

and where P.V. denotes the Cauchy principal value of the integral. The instantaneous phase of the time series is then defined as

$$\phi_i(t) = \arg z_i(t). \tag{4}$$

Equivalently, the analytic signal can be obtained using a filter that removes negative frequency components,

$$z_i(t) = \mathcal{F}^{-1}\{\mathcal{F}[s_i(t)][1 + \text{sgn}(\omega)]\}, \tag{5}$$

where $\mathcal{F}$ denotes the Fourier transform into the domain of frequencies $\omega$ and sgn denotes the sign function [33]. This definition is more useful in practice because it can be straightforwardly applied to empirical time series, which are sampled at a finite number $n$ of discrete time points $t_m$, $s_{im} = s_i(t_m)$. If several time series (realizations of the same process) are available, the obtained phase values can be combined into a single multivariate sample of phases $\phi_{im}$, where the index $m = 1, \dots, n$ now enumerates the complete available phase data.

### B. Cluster analysis via Markov coarse-graining

In the participation index method [26], the use of eigenvectors of $R$ for synchronization cluster analysis was motivated by the investigation of the spectral properties of correlation matrices in random matrix theory [23]. Another context where eigenvalue decomposition turns up naturally is the computation of matrix powers, which becomes as simple as possible using the spectral representation of the matrix.

Powers of $R$ have a well-defined meaning in the special case of a binary-valued matrix ($R_{ij} \in \{0,1\}$), as it is obtained for instance by thresholding: The matrix entries of $R^a$ count the number of possible paths from one element to another within $a$ steps, i.e., they specify the degree to which these elements are connected via indirect links of synchrony. By analogy, we interpret $(R^a)_{ij}$ also in the general case as quantifying the degree of common entanglement of two elements $i$ and $j$ within the same web of synchronization relations. In the following we will call this quantity the $a$-step synchronization strength of two oscillators because it reduces to the original bivariate synchronization strength $R_{ij}$ in the case $a = 1$.

This synchronization strength over $a$ steps is relevant for synchronization cluster analysis, because in a system where synchronization clusters are present it is possible that the degree of direct bivariate synchrony of two elements is not very strong, but they are both entangled into the same web of links of synchrony. These indirect links, which make the two elements members of the same synchronization cluster, become visible in $R^a$.

Moreover, with increasing power $a$ the patterns of synchrony within a cluster (the matrix columns) become more similar, approaching the form of one of the dominant eigenvectors. If there are different clusters in the system, a suitable $a$ can be chosen such that each cluster exhibits a different pattern, representing the web of synchronization relations it is comprised of. These patterns constitute *signatures* by which elements can be attributed to clusters. The cluster signatures are related to the dominant eigenvectors of $R$; by transition to larger $a$ they become even more dominant, leading to an effective simplification of the matrix.

For the identification of the members of a cluster only the patterns of synchrony are relevant, while the absolute size of elements of different columns diverges with $a$, so that some sort of normalization is called for. Different normalization schemes might be used for this purpose. However, using the $L^1$-norm the procedure can be simplified, because for the normalized version of the synchronization matrix, given by

$$P_{ij} = \frac{R_{ij}}{\sum_{i'} R_{i'j}}, \qquad (6)$$

it holds that powers of $P$ are automatically normalized, too. Moreover, the $L^1$-normalized matrix $P$ is a column-stochastic matrix, that is, it can be interpreted as the matrix of $i \leftarrow j$ transition probabilities describing a Markov chain, whose states correspond to the elements of the original system. Via this connection, the tools of stochastic theory and especially recent work on the coarse-graining of finite-state Markov processes [28–30] can be utilized for the purposes of synchronization cluster analysis.

The Markov process defined in this way possesses some specific properties [34]: It is aperiodic because of the nonzero diagonal entries of the matrix, and it is in general irreducible because the values of empirical $R_{ij}$ for $i \neq j$ will also almost never be exactly zero. For a finite-state process, these two properties amount to ergodicity, which implies that any distribution over states converges to a unique invariant distribution $p^{(0)}$, corresponding to the eigenvector of $P$ for the unique largest eigenvalue 1. This distribution can be computed from $R$ as

$$p_i^{(0)} = \frac{\sum_j R_{ij}}{\sum_{i'} \sum_j R_{i'j}}, \qquad (7)$$

where the vector components of $p^{(0)}$ are denoted by $p_i^{(0)}$. With the matrix $R$ also the stationary flow given by

$$P_{ij} p_j^{(0)} = \frac{R_{ij}}{\sum_{i'} \sum_{j'} R_{i'j'}} \qquad (8)$$

is symmetric, i.e., the process fulfills the condition of detailed balance $P_{ij} p_j^{(0)} = P_{ji} p_i^{(0)}$, which makes eigenvalues and eigenvectors of $P$ real-valued [28].

For the Markov process, the $a$-step synchronization strength considered above translates into transitions between states over a period of $\tau$ time steps. To compute the corresponding transition matrix $P^\tau$ the eigenvalue decomposition of $P$ is used. If $\lambda_k$ with $k=0,\ldots,N-1$ denote the eigenvalues of $P$, and the right and left eigenvectors $p_k$ and $A_k$ are scaled such that the orthonormality relation

$$A_k p_l = \delta_{kl} \qquad (9)$$

is fulfilled, the spectral representation of $P$ is given by

$$P = \sum_k \lambda_k p_k A_k, \qquad (10)$$

and consequently

$$P^\tau = \sum_k \lambda_k^\tau p_k A_k. \qquad (11)$$

We assume that eigenvalues are sorted such that $\lambda_0 = 1 > |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{N-1}|$. The scaling ambiguity left by the orthonormality relation is resolved by choosing

$$p_{ik} = p_i^{(0)} A_{ki} \qquad (12)$$

(where $p_{ik}$ and $A_{ki}$ denote the vector components of $p_k$ and $A_k$, respectively), which leads to the normalization equations

$$\sum_i \frac{p_{ik}^2}{p_i^{(0)}} = 1 \quad \text{and} \quad \sum_i p_i^{(0)} A_{ki}^2 = 1, \qquad (13)$$

with the special solutions $p_{i0} = p_i^{(0)}$ and $A_{0i} = 1$. Additionally, a generalized orthonormality relation

$$\sum_i \frac{p_{ik} p_{il}}{p_i^{(0)}} = \delta_{kl} \qquad (14)$$

follows for the right eigenvectors.

The convergence of every initial distribution to the stationary distribution $p^{(0)}$ corresponds to the fact that because of nonvanishing synchronies the whole system ultimately forms one single cluster. This perspective belongs to a time scale $\tau \to \infty$, at which all eigenvalues $\lambda_k^\tau$ go to 0 except for the largest one, $\lambda_0^\tau = 1$. In the other extreme of a time scale $\tau = 0$, $P^\tau$ becomes the identity matrix, all of its columns are different, and the system disintegrates into as many clusters as there are elements. For the purposes of cluster analysis, intermediate time scales are of interest on which many but not all of the eigenvalues are practically zero. If we want to identify $q$ clusters, we expect to find that many different cluster signatures, and that means we must consider $P^\tau$ at a time scale where eigenvalues $\lambda_k^\tau$ may be significantly different from zero only for the range $k=0,\ldots,q-1$.

This is achieved by determining $\tau$ such that $|\lambda_q|^\tau \approx 0$. Using a parameter $\zeta \ll 1$ chosen to represent the quantity that is considered to be practically zero (e.g., $\zeta = 0.01$), from $|\lambda_q|^\tau = \zeta$ we calculate the appropriate time scale for a clustering into $q$ clusters as

$$\tau(q) = \frac{\ln \zeta}{\ln|\lambda_q|}. \quad (15)$$

The vanishing of the smaller eigenvalues at a given time scale describes the loss of internal differentiation of the clusters, the removal of the structural features encoded in the corresponding weaker eigenvectors. On the other hand, the differentiation of clusters from each other via the dominant eigenvectors will be clearer the larger the remaining eigenvalues are, especially $\lambda_{q-1}^{\tau}$. This provides a criterion for selecting the number of clusters $q$: the clustering will be better the larger $|\lambda_{q-1}|^{\tau(q)}$ is. Equivalently, we select $q$ based on the *timescale separation factor*

$$F(q) = \frac{\tau(q-1)}{\tau(q)} = \frac{\ln|\lambda_q|}{\ln|\lambda_{q-1}|}, \quad (16)$$

which is independent of the particular choice of $\zeta$, and invariant under rescaling of the time axis. This criterion gives a ranking of the different possible choices (from 1 to $N-1$). The fact that $\lambda_0 = 1$ and therefore $F(1) = \infty$ implies a limitation of this approach, since the choice $q = 1$ is always characterized as absolutely optimal. Therefore, the first meaningful—and usually best—choice is the second entry in the $q$ ranking list.

To determine which elements belong to the same cluster, we need a measure $d$ of the dissimilarity of cluster signatures, that is, of the column vectors of $P^{\tau}$. Since these vectors belong to the space of the right eigenvectors of $P$, the appropriate dissimilarity metric is based on the norm corresponding to the normalization equation for the right eigenvectors [Eq. (13), left-hand side],

$$\|p\| = \sum_i \frac{p_{ik}^2}{p_i^{(0)}}. \quad (17)$$

The resulting column vector dissimilarity

$$d^2(j,j') = \sum_i \frac{1}{p_i^{(0)}} |(P^{\tau})_{ij} - (P^{\tau})_{ij'}|^2 \quad (18)$$

has the convenient property that the dimensionality of the space within which the clustering must be performed can be reduced, because the expression obtained by inserting the spectral representation for the matrix entries of $P^{\tau}$,

$$(P^{\tau})_{ij} = \sum_k \lambda_k^{\tau} p_{ik} A_{kj}, \quad (19)$$

simplifies to

$$d^2(j,j') = \sum_k |\lambda_k|^{2\tau} (A_{kj} - A_{kj'})^2 \quad (20)$$

using the generalized orthonormality of the right eigenvectors, Eq. (14). Since for appropriately chosen $\tau = \tau(q)$ contributions for larger $k$ vanish starting from $k = q$, and because $A_{0i} = 1$ for all $i$, it is sufficient to let the sum run over the range $1, \ldots, q-1$. The dissimilarity $d$ can therefore be interpreted as the Euclidean distance within a $(q-1)$-dimensional left *eigenvector space*, where each element $j$ is associated with a position vector

$$\vec{o}(j) = (|\lambda_k|^{\tau} A_{kj}), \quad k = 1, \ldots, q-1. \quad (21)$$

To actually perform the clustering, we can in principle use any algorithm that is designed to minimize the sum of within-cluster variances. Our implementation derives from the observation that clusters in eigenvector space form a $q$-simplex [29,30]. A first rough estimate of the cluster locations can therefore be obtained by searching for the extreme points of the data cloud, employing a subalgorithm described in Ref. [30]: Determine the point farthest from the center of the cloud, then the point farthest from the first one; then iteratively the point farthest from the hyperplane spanned by all the previously identified points, until $q$ points are found. Using the result of this procedure as initialization, the standard $k$-means algorithm [35] that normally tends to get stuck in local minima converges in almost all cases quickly onto the correct solution.

In summary, the algorithmic steps [36] of the eigenvector space method introduced in this paper are: (1) Calculate the matrix of bivariate synchronization indices $R_{ij}$, Eq. (1). (2) Convert the synchronization matrix $R$ into a transition matrix $P$, Eq. (6). (3) Compute the eigenvalues $\lambda_k$ and left eigenvectors $A_k$ of $P$. (4) Select the number of clusters $q$, $q > 1$, with the largest time-scale separation factor $F(q)$, Eq. (16). (5) Determine the positions $\vec{o}(j)$, Eq. (21), in eigenvector space for $\tau = \tau(q)$, Eq. (15). (6) Search for $q$ extreme points of the data cloud. (7) Use these as initialization for $k$-means clustering.

### C. Illustration of the eigenvector space method

In order to illustrate the operation of the method, we apply it to multivariate time series data obtained from simulated nonlinear oscillators, coupled in such a way as to be able to observe synchronization clusters of different size as well as unsynchronized elements. The system consists of $N = 9$ Lorenz oscillators that are coupled diffusively via their $z$-components,

$$\dot{x}_j = 10(y_j - x_j),$$

$$\dot{y}_j = 28x_j - y_j - x_j z_j,$$

$$\dot{z}_j = -\frac{8}{3} z_j + x_j y_j + \epsilon_{ij}(z_i - z_j). \quad (22)$$

The coupling coefficients $\epsilon_{ij}$ were chosen from $\{0,1\}$ to implement the coupling configuration depicted in Fig. 1(a), such that oscillators number 2–4 are unidirectionally driven by number 1, oscillators number 7 and 8 driven by number 9, and numbers 5 and 6 are uncoupled. These differential equations were numerically integrated using a step size of $\Delta t = 0.01$, starting from randomly chosen initial conditions. After discarding an initial transient of $10^4$ data points, further $4 \times 10^4$ samples entered data processing. Instantaneous phases $\phi_{jm}$ of oscillators $j$ at time points $t_m = (m-1)\Delta t$ were determined from the $z$-components using the analytic signal approach after removal of the temporal mean, and bivariate synchronization strengths $R_{ij}$ were computed.
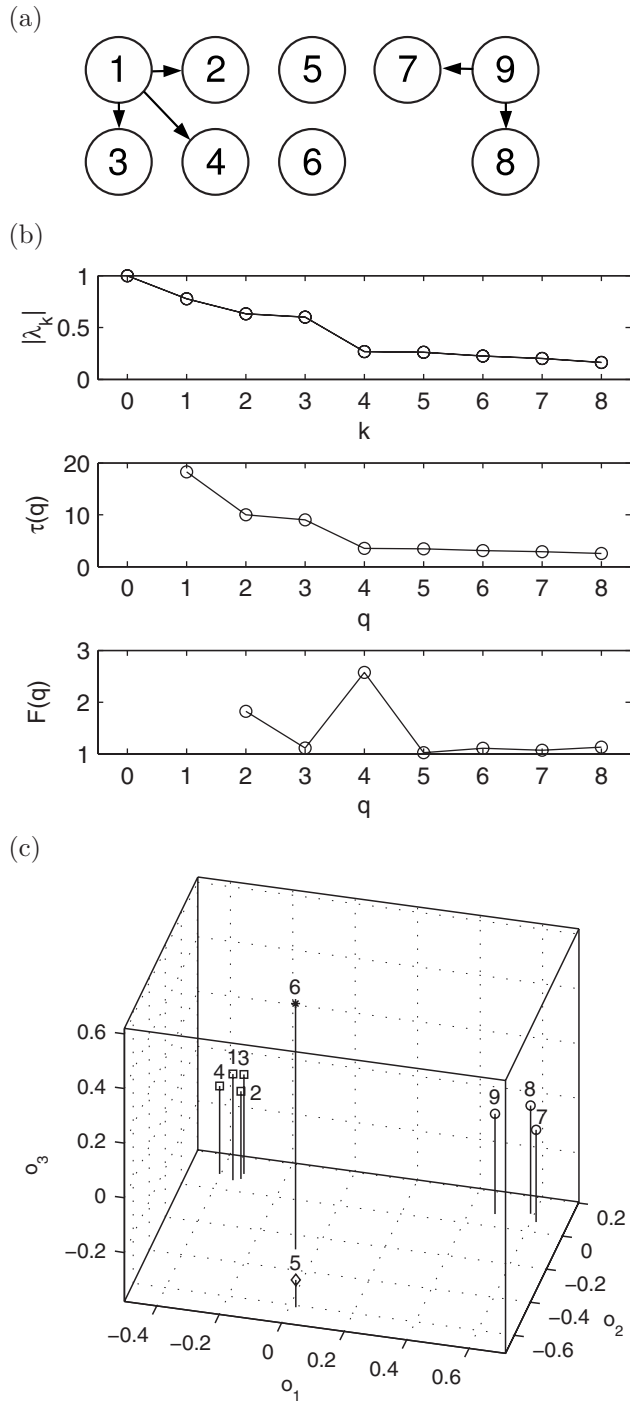
FIG. 1. Application of the eigenvector space method to a system of nine partially coupled Lorenz oscillators. (a) Coupling configuration: The left-hand group of four oscillators is driven by number 1, the right-hand group of three is driven by number 9, the remaining two are uncoupled. (b) Eigenvalues $\lambda_k$, time scales $\tau(q)$, and time-scale separation factors $F(q)$. The maximal separation factor $F(4)$ indicates the presence of four clusters. (c) Positions attributed to oscillators in three-dimensional eigenvector space $(o_1, o_2, o_3)$. The clustering by the $k$-means algorithm results in a cluster composed of oscillators number 1–4 ($\square$), two single-element clusters consisting in oscillators number 5 ($\diamond$) and number 6 ($*$), respectively, and a cluster composed of oscillators number 7–9 ($\bigcirc$).

The outcomes of the eigenvector space method applied to the resultant matrix of bivariate indices are presented in Fig. 1. Figure 1(b) shows the spectrum of eigenvalues $\lambda_k$ of the transition matrix $P$ with the corresponding time scales $\tau(q)$ and time-scale separation factors $F(q)$. A gap in the eigenvalue spectrum between indices $k=3$ and 4 translates into a maximum time-scale separation factor for $q=4$, which recommends a search for four clusters in the eigenvector space for time scale $\tau=3.5$. This three-dimensional space is depicted in Fig. 1(c), where the expected grouping into four clusters can be clearly recognized in the arrangement of elements $j$ with positions $\vec{o}(j)$. These four clusters that correspond to the two groups of driven oscillators and the two uncoupled oscillators (each of which forms a single-element cluster) are easily identified by the $k$-means algorithm. The results shown here were obtained using $\zeta=0.01$; alternative choices of 0.1 and 0.001 yielded the same clustering.

## III. PERFORMANCE

To assess the performance of the eigenvector space method introduced in this paper, we compare it with the previous approach to synchronization cluster analysis based on spectral decomposition. For reference, we briefly recollect the important details.

The participation index method [26] is based on the eigenvalue decomposition of the symmetric synchronization matrix $R$ itself, into eigenvalues $\eta_k$ and $L^2$-normalized eigenvectors $v_k$. Each of the eigenvectors that belong to an eigenvalue $\eta_k > 1$ is identified with a cluster, and a system element $j$ is attributed to that cluster $k$ in which it participates most strongly, as determined via the participation index

$$\Pi_{jk} = \eta_k v_{jk}^2, \qquad (23)$$

where $v_{jk}$ are the eigenvector components of $v_k$. The method performs quite well in many configurations, but it encounters problems when confronted with clusters of similar strength that are slightly synchronized to each other, which was demonstrated in Ref. [27] using a simulation.

Here we employ a refined version of that simulation to compare the two methods. We consider a system of $N=32$ oscillators forming two clusters, and check whether the methods are able to detect this structure from the bivariate synchronization matrix $R$ for different degrees of intercluster synchrony $\rho_{int}$. The cluster sizes are controlled via a parameter $r$, such that the first cluster comprises elements $j = 1, \ldots, r$, the second $(r+1), \ldots, N$.

To be able to time-efficiently perform a large number of simulation runs and to have precise control over the structure of the generated synchronization matrices, we do not implement the system via a set of differential equations. Instead, our model is parametrized in terms of the population value of the bivariate synchronization index, Eq. (1),

$$\rho_{ij} = |\langle \exp[i(\phi_i - \phi_j)]\rangle| \qquad (24)$$

(where $\langle \cdots \rangle$ denotes the expectation value), which is the first theoretical moment of the circular phase difference distribution [31]. For $i, j$ within the same cluster $\rho_{ij}$ is fixed at a

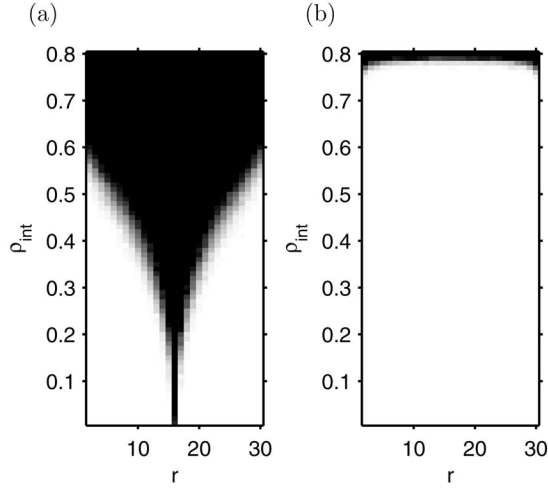(a)                                    (b)



FIG. 2. Comparative performance of the participation index (a) and the eigenvector space method (b). The methods are tested on a system of $N=32$ elements, divided into two clusters containing $r$ and $N-r$ elements, respectively. The intercluster synchronization strength $\rho_{\text{int}}$ is varied from 0 up to the value of intracluster synchronization 0.8. Synchronization matrices are generated based on samples of size $n=200$. The quantity shown is the relative frequency (over 100 trials) with which the respective algorithm failed to recover exactly the given two-cluster structure; it is visualized in gray scale, covering the range from 0 (white) to 1 (black). Comparison shows that in a large area along $r=16$ where the participation index method fails, the eigenvector space method introduced in this paper performs perfectly.

value of $\rho_1=\rho_2=0.8$. For intercluster synchronization relations it is set to $\rho_{\text{int}}$, which is varied from 0 up to 0.8 such that the two-cluster structure almost vanishes.

To be able to properly account for the effect of random variations of $R_{ij}$ around $\rho_{ij}$ due to finite sample size $n$ we generated samples of phase values, using an extension of the single-cluster model introduced in Ref. [24]: The common behavior of oscillators within each of the two clusters is described by cluster phases $\Phi_1$ and $\Phi_2$. The phase differences between the members of each cluster and the respective cluster phase

$$\Delta\phi_j = \begin{cases} \phi_j - \Phi_1 & \text{for } j = 1, \ldots, r, \\ \phi_j - \Phi_2 & \text{for } j = (r+1), \ldots, N, \end{cases} \quad (25)$$

as well as the phase difference of the two cluster phases,

$$\Delta\Phi = \Phi_2 - \Phi_1, \quad (26)$$

are assumed to be mutually independent random variables, distributed according to wrapped normal distributions [31] with circular moments $\rho_{1C}$, $\rho_{2C}$, and $\rho_{CC}$, respectively. Since the summation of independent circular random variables results in the multiplication of their first moments [24], for the relation of model parameters and distribution moments holds,
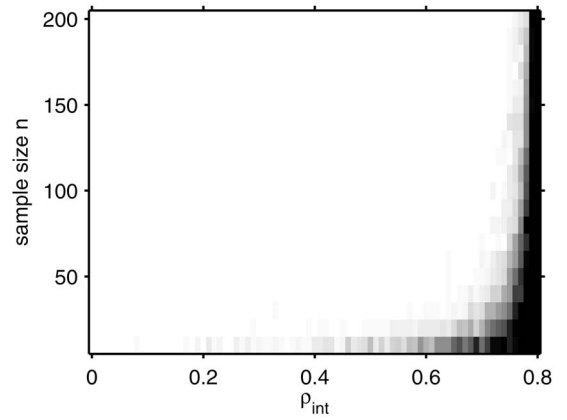
$$\rho_1 = \rho_{1C}^2,$$



FIG. 3. Performance of the eigenvector space method depending on the sample size $n$, investigated using the two-cluster system of Fig. 2 for different values of the intercluster synchronization strength $\rho_{\text{int}}$. The quantity shown here is the proportion of values of the parameter $r$ (controlling cluster sizes) at which the two-cluster structure failed to be recovered, visualized on a gray scale from 0 (white) to 1 (black). The plot shows that the ability of the eigenvector space method to correctly identify clusters up to high values of $\rho_{\text{int}}$ breaks down only for very small sample sizes $n$.

$$\rho_2 = \rho_{2C}^2,$$

$$\rho_{\text{int}} = \rho_{1C}\rho_{CC}\rho_{2C}. \quad (27)$$

For the performance comparison of the two methods, $n=200$ realizations of this model of the multivariate distribution of phases $\phi_j$ were generated for each setting of the parameters, and synchronization indices $R_{ij}$ were calculated via Eq. (1).

The clustering results are presented in Fig. 2 for the participation index and the eigenvector space method (using $\zeta=0.01$). The quantity shown is the relative frequency (over 100 instances of the matrix $R$) of the failure to identify correctly the two clusters built into the model. Figure 2(a) shows that the participation index method fails systematically within a region located symmetrically around $r=N/2=16$ (clusters of equal size). The region becomes wider for increasing $\rho_{\text{int}}$ but is already present for very small values of the intercluster synchronization strength. In contrast, the eigenvector space approach (b) is able to perfectly reconstruct the two clusters for all values of $r$ up to very strong intercluster synchronization. It fails to correctly recover the structure underlying the simulation data only in that region where intercluster synchronization indices attain values comparable to those within clusters, i.e., only where there are no longer two different clusters actually present. These results demonstrate that the eigenvector space method is a clear improvement over the previous approach.

For real-world applications, a time series analysis method must be able to work with a limited amount of data. In the case of synchronization cluster analysis, a small sample size attenuates the observed contrast between synchronization relations of different strength, making it harder to discern clusters. Using the two-cluster model described above, in a fur-

ther simulation we investigated the effect of the sample size $n$ on the performance of the eigenvector space method. Parameters $\rho_{int}$ and $r$ were varied as before, and synchronization matrices were generated for $n = 10, \ldots, 200$ (in steps of 10). For each value of $\rho_{int}$ and $n$, as a test quantity the proportion of $r$ values for which the algorithm did not correctly identify the two clusters was calculated. The result shown in Fig. 3 demonstrates that the performance of the eigenvector space method degrades only very slowly with decreasing sample size. The method seems to be able to provide a meaningful clustering down to a data volume of about $n = 30$ independent samples, making it quite robust against small sample size.

## IV. CONCLUSION

We introduced a method for the identification of clusters of synchronized oscillators from multivariate time series. By translating the matrix of bivariate synchronization indices $R$ into a stochastic matrix $P$ describing a finite-state Markov process, we were able to utilize recent work on the coarse-graining of Markov chains via the eigenvalue decomposition of $P$. Our method estimates the number of clusters present in the data based on the spectrum of eigenvalues, and it represents the synchronization relations of oscillators by assigning to them positions in a low-dimensional space of eigenvectors, thereby facilitating the identification of synchronization clusters. We showed that our approach does not suffer from the systematic errors made by a previous approach to synchronization cluster analysis based on eigenvalue decomposition, the participation index method. Finally, we demonstrated that the eigenvector space method is able to correctly identify clusters even given only a small amount of data. This robustness against small sample size makes it a promising candidate for field applications, where data availability is often an issue. Concluding we want to remark that though the method was described and assessed in this paper within the context of phase synchronization analysis, our approach might also give useful results when applied to other bivariate measures of signal interdependence.

[1] M. S. Santhanam and P. K. Patra, Phys. Rev. E **64**, 016102 (2001).

[2] N. Marwan, M. Thiel, and N. R. Nowaczyk, Nonlinear Processes Geophys. **9**, 325 (2002).

[3] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[4] E. Pereda, R. Quian Quiroga, and J. Bhattacharya, Prog. Neurobiol. (Oxford) **77**, 1 (2005).

[5] C. Zhou, L. Zemanová, G. Zamora, C. C. Hilgetag, and J. Kurths, Phys. Rev. Lett. **97**, 238103 (2006).

[6] D. Brillinger, *Time Series: Data Analysis and Theory* (Holden-Day, New York, 1981).

[7] M. B. Priestley, *Nonlinear and Non-Stationary Time Series Analysis* (Academic, New York, 1988).

[8] J. S. Bendat and A. G. Piersol, *Random Data Analysis and Measurement Procedure* (Wiley, New York, 2000).

[9] S. M. Kay, *Modern Spectral Estimation* (Prentice-Hall, New York, 1988).

[10] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[11] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths, *Synchronization. A Universal Concept in Nonlinear Sciences* (Wiley, New York, 2000).

[12] S. Boccaletti, J. Kurths, G. Osipov, D. L. Valladares, and C. S. Zhou, Phys. Rep. **366**, 1 (2002).

[13] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, Phys. Rev. Lett. **76**, 1804 (1996).

[14] P. Tass, M. G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, and H.-J. Freund, Phys. Rev. Lett. **81**, 3291 (1998).

[15] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, Hum. Brain Mapp. **8**, 194 (1999).

[16] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, Physica D **144**, 358 (2000).

[17] C. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series* (Princeton University, Princeton, NJ, 1964).

[18] B. Schelter, M. Winterhalder, R. Dahlhaus, J. Kurths, and J. Timmer, Phys. Rev. Lett. **96**, 208103 (2006).

[19] D.-H. Kim and H. Jeong, Phys. Rev. E **72**, 046133 (2005).

[20] E. Rodriguez, N. George, J.-P. Lachaux, J. Martinerie, B. Renault, and F. J. Varela, Nature **397**, 430 (1999).

[21] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).

[22] A. Utsugi, K. Ino, and M. Oshikawa, Phys. Rev. E **70**, 026110 (2004).

[23] M. Müller, G. Baier, A. Galka, U. Stephani, and H. Muhle, Phys. Rev. E **71**, 046116 (2005).

[24] C. Allefeld and J. Kurths, Int. J. Bifurcation Chaos **14**, 417 (2004).

[25] C. Allefeld, Int. J. Bifurcation Chaos **16**, 3705 (2006).

[26] C. Allefeld, M. Müller, and J. Kurths, Int. J. Bifurcation Chaos **17**, 3493 (2007).

[27] S. Bialonski and K. Lehnertz, Phys. Rev. E **74**, 051909 (2006).

[28] G. Froyland, Physica D **200**, 205 (2005).

[29] B. Gaveau and L. S. Schulman, Bull. Sci. Math. **129**, 631 (2005).

[30] P. Deuflhard and M. Weber, Linear Algebra Appl. **398**, 161 (2005).

[31] K. V. Mardia and P. E. Jupp, *Directional Statistics* (Wiley, New York, 2000).

[32] D. Gabor, J. Inst. Electr. Eng., Part 3 **93**, 429 (1946).

[33] R. Carmona, W.-L. Hwang, and B. Torrésani, *Practical Time-Frequency Analysis* (Academic, New York, 1998).

[34] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability* (Springer, New York, 1993).

[35] J. B. MacQueen, in *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol 1*, edited by L. M. Le Cam and J. Neyman (University of California Press, Berkeley, 1967), pp. 281.

[36] An implementation of the algorithm for MATLAB and OCTAVE is available from the corresponding author.