Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

KONSTANTIN FACKELDEY[1], ALEXANDER SIKORSKI[2],
MARCUS WEBER[3]

# Spectral Clustering for Non-reversible Markov Chains

[1]fackeldey@zib.de
[2]sikorski@zib.de
[3]weber@zib.de

ZIB Report 18-48 (August 2018)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

# Spectral Clustering for Non-reversible Markov Chains

Konstantin Fackeldey, Alexander Sikorski, Marcus Weber

**Abstract**

Spectral clustering methods are based on solving eigenvalue problems for the identification of clusters, e.g. the identification of metastable subsets of a Markov chain. Usually, real-valued eigenvectors are mandatory for this type of algorithms. The Perron Cluster Analysis (PCCA+) is a well-known spectral clustering method of Markov chains. It is applicable for reversible Markov chains, because reversibility implies a real-valued spectrum. We also extend this spectral clustering method to non-reversible Markov chains and give some illustrative examples. The main idea is to replace the eigenvalue problem by a real-valued Schur decomposition. By this extension non-reversible Markov chains can be analyzed. Furthermore, the chains do not need to have a positive stationary distribution. In addition to metastabilities, dominant cycles and sinks can also be identified. This novel method is called GenPCCA (i.e. Generalized PCCA), since it includes the case of non reversible processes. We also apply the method to real world eye tracking data.

## 1 Introduction

The analysis of Markov chains is used to figure out the transition behavior in many fields, ranging from the analysis of disease evolution in clinical data to molecular simulation. In the context of Google's page rank problem, for instance, it is the invariant measure of a Markov chain that provides a ranking of the relevant web pages [LM05, LM12]. These data can be interpreted as elements of a typically large space. More precisely, $\{X_i, i \in \mathbb{N}\}$ is a sequence of random variables with values $x_i$ in the finite state space $\Gamma = \{1, \dots, N\}$. A large number $N$ of states makes it difficult to reveal the general transition behavior of the chain. Clustering aims to reduce the number of states $n \ll N$ by still describing the underlying stochastic process correctly. Under the assumption that the stochastic process possesses metastable sets, which are subsets in the state space where the system or stochastic process spends a long time span before switching to a different metastability, the clustering is used to identify the rapidly mixing parts and to separate them from each other. This coarse graining of a Markov chain by partitioning the corresponding state spaces into subsets has been introduced in the context of economic systems by [SA61] and studied in [Cou75, HMN98].

Recent success with spectral clustering methods have been greatly celebrated [MLKCW03, Shu03, WL07, VLKG03, LBDD01, DHFS00, DW05]. Markov

State Models (e.g.[BPE14, SS12]) do not aim at a full eigen decomposition of the problem. In fact only the eigenvectors of the eigenvalues close to 1 are employed since these eigenvectors span an invariant subspace which can be used to analyze the metastable states of the process. Many methods in the field of Markov State Modeling assume that the eigenvalues and eigenvectors which are used for this analysis are real. This can be assured, e.g. if the Markov process or the Markov chain is reversible. This is the case if reversing the stochastic process leads to the same transition behavior. This property allows for a clustering in terms of metastable sets using the eigenvector data.

The authors of [FMSV07] proposed to replace the eigenvalue problem by a singular value decomposition if the process is non-reversible. This was further developed by [Tif11]. Moreover, in [Jac10] it is claimed that the singular vectors do not have the relevant sign structure to identify the metastable states, and thus it does not preserve the dynamical structure of the Markov chain. Nevertheless this method has been applied in [TBHY13] though in the context of identifying the collective variables.
In [Fil91] the eigenvalue bounds of the mixing rates for reversible Markov chains have been extended to non-reversible chains by reversing the non-reversible matrix. Based on this, clustering methods for non-reversible processes [RM07, HMS04], as well as also other approaches [Jac10, SS14][Ste94, Ch 1.7], have been developed.

In this article we introduce a novel clustering method (GenPCCA) aimed at grouping states of a Markov chain according to their transition behavior by replacing the eigenvalue decomposition with a Schur decomposition [FW17]. It turns out that this novel method offers a powerful analysis of the Markov chain which also includes the identification of coherent subsets (e.g. [FSM10]) and the freedom of regarding an arbitrary initial distribution of states. Moreover our method does not rely on the above mentioned invariant density of the Markov chain. Only the initial density is needed. Thus this novel method covers a broader class of applications by including non-reversible Markov chains. Since this method is a generalization of PCCA+ towards non-reversible processes it is named GenPCCA (Generalized-Perron Cluster Cluster Analysis).

## 2 Markov chains and clustering

A finite autonomous Markov chain is given by a sequence $X_0, X_1, \ldots$ of random variables $X_k, k = 0, 1, 2, \ldots$. Since the set of all states is finite, a transition probability matrix $(P_{ij})_{i,j=1,\ldots,N}$ can be given by

$$P_{ij} = \mathbb{P}(X_{k+1} = j; X_k = i) \quad i, j \in \{1, \ldots, N\}, k \in \mathbb{N},$$

where $\mathbb{P}$ denotes the conditional probabilities for reaching state $j$ in 1 step of the Markov chain, if the process has started in $i$. Obviously this matrix is non-negative and stochastic, i.e.

$$P_{ij} \geq 0 \quad \forall i, j \qquad \sum_{j=1}^{N} P_{ij} = 1 \qquad \forall i.$$

Let us furthermore denote the initial probabilities by $\eta_i = \mathbb{P}(X_0 = i)$, such that the vector $\eta = (\eta_1, \ldots, \eta_N)$ is the initial distribution.

The transition matrix $P(k)$ of the $k$th step then meets the semi group property given by

$$P(k) = (P(1))^k = P^k. \tag{1}$$

Equation (1) is also named "Markovianity"[1]. For the Markov chain represented by $P$, we aim at a *clustering*, i.e. to find a projection of a Markov chain from a high number of states $N$ to a small number of clusters $n$ where $n \ll N$.

The Markov chain represented by an $N \times N$-transition matrix $P$ is thus replaced by a $n \times n$-matrix $P_C$ providing the transition behavior between the $n$ clusters.

A clustering can also be interpreted as a projection $G(P)$ of the matrix $P$ onto $n$ clusters. However, in order to guarantee that this projection is suitable (i.e. that the propagation commutes with the projection), it should also meet (1) in the sense:

$$(G(P))^k = G(P^k). \tag{2}$$

In general, the projection of a Markov chain is not Markovian and thus the stochastic process induced by the $n \times n$ transition matrix between the clusters is in general not a Markov process. The projected process is Markovian if the projection is based on an invariant subspace of the transition matrix of the high-dimensional process. In detail, Markovianity of the projection G(P) can be guaranteed, if the projection meets the

- *invariant subspace condition:* there exists a matrix $X \in \mathbb{R}^{N \times n}$ (for a suitable choice of $n$) which meets

  $$PX = X\Lambda \tag{3}$$

  for $\Lambda \in \mathbb{R}^{n \times n}$

- *orthogonaility relation*

  $$X^T D_\eta X = I_{n \times n}, \tag{4}$$

  where $D_\eta := \mathrm{diag}(\eta_1, ..., \eta_N)$ and $\Lambda \in \mathbb{R}^{n \times n}$, i.e. the $X$ are spanning an $n$ dimensional invariant subspace of $P$.

In this context $X$ is constructed to meet an orthogonality condition (4). $X$ are not necessarily eigenvectors of $P$. However, if the process is reversible, then there is a way to meet conditions (3) and (4) for eigenvectors $X$ of $P$ with the diagonal matrix $\Lambda$ of eigenvalues. If the process is non-reversible, then conditions (3) and (4) can be assured by using a suitable Schur decomposition of $P$ with a real Schur matrix $\Lambda$. It has been shown in [KW07] that conditions (3) and (4) of a projection $G$ are sufficient for Markovianity (2).

We remark that a singular value decomposition of $P$ does not meet (3) and consequently a Galerkin projection leads to a projection error [SS14, Chapter 5.2]. In the next section we show how the orthogonality relation and the invariant subspace condition are realized for reversible Markov chains.

[1] This is a consequence of the Chapman Kolmogorov equation.

## 2.1 Reversible Markov Chains

If we assume an irreducible and reversible Markov chain then it has a unique vector $\pi = (\pi_1, ...., \pi_N)^T$ such that

$$\pi^T P = \pi^T \text{ and } \sum_{i=1}^{N} \pi_i = 1,$$

where $\pi$ is an invariant distribution or equilibrium distribution. If one takes $\pi$ as the initial distribution, then the chain is stationary. We denote by $D_\pi \in \mathbb{R}^{N\times N}$ the matrix with the invariant distribution on its diagonal. For stationary Markov chains the detailed balance condition given by

$$\pi_i P_{ij} = \pi_j P_{ji} \tag{5}$$

is a necessary and sufficient characterization of reversibility in terms of transition probabilities and equilibrium distribution. In this special case, (3) is the eigenvalue equation of $P$, where $\Lambda$ is the diagonal matrix of the *real* eigenvalues near the Perron root $\lambda_1 = 1$ and $X$ are the corresponding *real* eigenvectors. Since the eigenvalues are real, they can be arranged in descending order, i.e.$1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$ . The orthogonality relation is only assured if the initial distribution equals the equilibrium distribution, i.e. $\eta = \pi$ in (4). The relation between eigenvalues of $P$ close to 1 and metastable sets of the Markov chain has been used by several authors in the past [Sch99, DHFS00, DW05, Web06, SS14].

The projection problem $(G(P))^k \neq G(P^k)$ has been discussed for the case of reversible Markov chains. In [Web06] this problem has been solved by looking at the Markov chain as a sequence of distributions instead as of a sequence of states [Web02, DW05, Web06]. The role of weighted inner products and the analysis of projections on the basis of invariant subspaces is a widely used tool in linear algebra and is also described in suitable text books. Schur decompositions have also been used previously in order to quantify projection errors or condition numbers of eigenvector problems [Ste94]. If $\eta \in \mathbb{R}^n$ is a probability distribution at a certain step of the chain, then $\hat{\eta} = P^T\eta$ denotes the probability distribution of states at the next step of the Markov chain. How does projection and propagation of distributions commute? This problem is solved by a subspace projection such that the projection error vanishes. The projection from $N$ states to $n$ clusters can be expressed by a membership matrix $C$. The non-negative entries $C_{ij}$ of this matrix denote how probable (or how intensive) it is that the state $i$ of the Markov chain belongs to the cluster $j$ of the projection. The row-sum of this matrix is 1. One part of solving the projection problem is: The membership matrix is constructed via PCCA+, i.e. $C = X\mathcal{A}$ is a linear combination of the leading eigenvectors of $P$, where $X \in \mathbb{R}^{N\times n}$ is the matrix of the $n$ leading eigenvectors and $\mathcal{A} \in \mathbb{R}^{n\times n}$ is the non-singular transformation matrix computed (as a solution of an optimization problem) by PCCA+ [Web06, DW05], which will be explained in Section 2.3. PCCA+ is a well-established method for clustering metastable, reversible Markov chains. This method uses the dominant eigenvalues of the corresponding transition matrix. These eigenvalues are real (because of the reversibility of $P$) and they are close to the Perron root $\lambda = 1$ . The Perron root is algebraically and geometrically simple if the matrix is irreducible and the Markov chain is aperiodic. If $\eta$ is an initial distribution of states, then $\eta_c = C^T\eta \in \mathbb{R}^n$ is its projection onto the

clusters, and thus the projection matrix is $\Pi^f = C^T = \mathcal{A}^T X^T$. In the reversible case, the matrix is projected via the following equation:

$$P_c = (C^T D_\pi C)^{-1}(C^T D_\pi P C). \tag{6}$$

Note that in the case of a reversible Markov chain, the left eigenvectors of $P$ are then given by

$$Y = D_\pi X,$$

such that the projection (6) clearly meets the orthogonality relation (4). Beyond that, orthonormality holds via $Y^T X = X^T D_\pi X = I$ , where $I \in \mathbb{R}^{n\times n}$ is the unit matrix. *We emphasize that, in this setting, the reversibility of the Markov chain implies the orthogonality relation.* By assuming that the starting distribution is a linear combination of the left eigenvectors of $P$ , i.e. $\eta = Y\alpha$ , where $\alpha \in \mathbb{R}^n$ is the vector of linear coefficients of this combination, the projection also meets the invariance condition (3). By the previous equations we get

$$\mathcal{A}^T\alpha = \mathcal{A}^T X^T D_\pi X\alpha = C^T Y\alpha = C^T\eta = \eta_c.$$

We can thus define $\Pi^b = D_\pi X\mathcal{A}^{-T}$ as the back projection, such that $\eta = \Pi^b\eta_c = D_\pi X\mathcal{A}^{-T}\eta_c$ and $\alpha = X^T\eta$. We are then in a position to prove the following:

**Lemma 2.1** *The propagation of the projected distributions commutes with the projection of the propagated distributions, i.e.*

$$\Pi^b(P_c^T)^k\Pi^f\eta = (P^T)^k\eta.$$

Proof: Let the number of steps be given by $k \in N$, then

$$\begin{aligned}
\Pi^b(P_c^T)^k\Pi^f\eta &= D_\pi X\mathcal{A}^{-T}[(C^T P^T D_\pi C)(C^T D_\pi C)^{-1}]^k\mathcal{A}^T X^T\eta \\
&= D_\pi X\mathcal{A}^{-T}[(\mathcal{A}^T X^T P^T D_\pi X A)(\mathcal{A}^T X^T D_\pi X A)^{-1}]^k\mathcal{A}^T X^T\eta \\
&= D_\pi X\mathcal{A}^{-T}[A^T\Lambda\mathcal{A}^{-T}]^k\mathcal{A}^T X^T\eta \\
&= D_\pi X\Lambda^k X^T\eta = D_\pi X\Lambda^k\alpha = (P^T)^k D_\pi X\alpha \\
&= (P^T)^k\eta,
\end{aligned}$$

where in $\Lambda \in \mathbb{R}^{n\times n}$ is the diagonal matrix of the the dominant real eigenvalues. □

Lemma 2.1 shows that for the distributions propagated via $P^T$, the projected distributions are propagated via $P_c^T$ (without error). Since the projection $\eta_c = C^T\eta$ is a non-negative vector with entry sum 1, it can be interpreted as a distribution on the $n$ clusters. Calculations [SA61, HMN98, Web11] also show that $P_c$ has the row-sum 1 and, thus, can be understood as the transition matrix of the corresponding projected Markov chain.

Summing up, Lemma 2.1 shows that the choice of a projection which meets the invariance condition (3) and orthogonality relation (4) leads to a commuting diagram. In this subspace projection, the initial distribution of the system is given by a linear combination of the left eigenvectors of $P$. The projected distribution $\eta_c = C^T\eta$ is propagated by a matrix $P_c^T$ , which can be computed according to (6). The diagram commutes if the membership matrix $C = X\mathcal{A}$ is a linear combination of the right eigenvectors of $P$.

## 2.2 Non-reversible Markov chains

In the foregoing section the orthogonality relation in the context of eigenvectors was realized by assuming that the underlying process is reversible. In fact Lemma 2.1 is only true if the underlying process is reversible. By resigning the reversibility of the underlying Markov chain, an interpretation of a transition matrix in terms of unconditional transition probabilities is not possible, since then the eigenvectors do not meet the invariance condition (3) and the subspace condition (4) in general.

Moreover for non-reversible processes, the spectrum of its corresponding transition matrix is in general not real but complex.

We thus take advantage of a Schur decomposition. Therefore let, $\widetilde{X}$ be $n$ Schur vectors of $\widetilde{P} = D_\eta^{0.5} P D_\eta^{-0.5}$. Then we have

$$\begin{aligned} & \widetilde{P}\widetilde{X} = \widetilde{X}\Lambda \\ \iff \quad & D_\eta^{0.5} P D_\eta^{-0.5}\widetilde{X} = \widetilde{X}\Lambda \\ \iff \quad & P D_\eta^{-0.5}\widetilde{X} = D^{-0.5}\widetilde{X}\Lambda \\ \iff \quad & PX = X\Lambda, \quad X = D_\eta^{-0.5}\widetilde{X}. \end{aligned}$$

We have thus shown that a Schur decomposition meets the invariant subspace condition (3) and the orthogonality condition (4). As a consequence, the projection

$$G(P) = (C^T D_\eta C)^{-1}(C^T D_\eta P C)$$

with Schur vectors $X$ meets (2). To show this, we have the following:

**Theorem 2.1** *Let $G(P) = (C^T D_\eta C)^{-1}(C^T D_\eta P C)$, where $X$ are the Schur vectors according to (7) and $C = X\mathcal{A}$ and $D_\eta$ are some initial distribution of the Markov chain, then*

$$(G(P))^k = G(P^k).$$

Proof:

$$\begin{aligned} G(P) &= (C^T D_\eta C)^{-1}(C^T D_\eta P C) \\ &= (\mathcal{A}^T X^T D_\eta X \mathcal{A})^{-1}(\mathcal{A}^T X^T D_\eta P X \mathcal{A}) \\ &= (\mathcal{A}^T X^T D_\eta X \mathcal{A})^{-1}(\mathcal{A}^T X^T D_\eta X \Lambda \mathcal{A}) \\ &= (\mathcal{A}^T \mathcal{A})^{-1}(\mathcal{A}^T \Lambda \mathcal{A}) \\ &= \mathcal{A}^{-1}\Lambda\mathcal{A}, \end{aligned}$$

such that $G(P)$ meets the desired criterion:

$$(G(P))^k = (\mathcal{A}^{-1}\Lambda\mathcal{A})^k = \mathcal{A}^{-1}\Lambda^k\mathcal{A} = G(P^k).$$

□

**Remark 2.1** *Note that in Theorem 2.1, the initial distribution $\eta$ does not have to be the stationary distribution. Theorem 2.1 may also be interpreted as commutativity between propagation in time ($k$ steps) and discretization $G$, which is a desired property for long term predictions.*

In the real Schur decomposition, the matrix $\Lambda$ is an upper triangle matrix with possibly $2\times2$-blocks on its diagonal. Moreover, the Schur vectors are orthogonal and define an insensitive invariant subspace (well conditioned) [KPC94][GvL85, Ch. 7]. We remark, that computing $n$ Schur values is challenging; efficient implementation by using the QR algorithm can be found in literature [GvL85, Vog04]. The remaining problem is that an arrangement of the Schur decomposition in descending order (of eigenvalues) is no longer possible. In [Bra02] it has been proposed to arrange the Schur-values according to an absolute distance to a given target value $z$. For the reversible case $z = 1$ must be chosen to guarantee that $P_C$ is close to the unit matrix allowing for a clustering into metastable states (i.e. the eigenvalues of $P_C$ correspond to these selected values).

For the non-reversible case, however, we can apply another method by arranging the Schur-values according to a distance from the unit circle. In this case $P_C$ has eigenvalues close to the unit circle and is thus similar to a permutation matrix, which can be seen as a clustering of states in the sense of coherent sets [FPG14]. This feature of GenPCCA is shown in the section of illustrative examples below.

## 2.3 GenPCCA

So far we have not yet explained how the matrix $\mathcal{A}$ is obtained. In the framework of GenPCCA, this step is identical to PCCA+ [Web06, DW05]. The problem of finding the matrix $\mathcal{A}$ can be converted to an optimization problem. More precisely, GenPCCA finds a transformation matrix $\mathcal{A}$ mapping the column vectors of Schur vectors $X$, spanning the invariant subspace, to the basis $C = X\mathcal{A}$ used for the projection $G(P)$. The aim of this algorithm is to find an optimal $n \times n$-basis transformation matrix $\mathcal{A}$. In the following we will assume that the dimension $n$ of the projected process is given. However, the problem of choosing an appropriate value of $n$ is not trivial. There exist many suitable algorithms which try to solve this problem from different perspectives. Methods which search for a spectral gap are very helpful if $P$ has a real-valued spectrum. Conditions for the existence of a spectral gap with only a few discrete large eigenvalues have been given by Huisinga [Hui01]. Methods which compute the crispness of the resulting clustering are also suitable for the Schur decomposition [WRS06]. In principle, large numbers $N$ make this identification problem more complicated. The matrix $X$ of the invariant subspace is needed as input. The output of GenPCCA is the above mentioned matrix of membership vectors $C$. The column vectors of both matrices, $X$ and $C$, span the same subspace. Thus, GenPCCA provides an invariant subspace projection of $P$, such that the subspace spanning vectors $C$ have an interpretation in terms of membership vectors. To do so, the matrix $C$ has to meet the following properties explaining the simplex structure of $C$:

- $\sum_{J=1}^{n_C} C_{iJ} = 1$ (partition of unity)
- $C_{iJ} \geq 0 \qquad \forall i \in \{1, ..., n\}$ (positivity)
- $C = X\mathcal{A}$, $\mathcal{A}$ non-singular, (invariance).

These conditions imply the feasible set of transformations $\mathcal{A}$. The selection of $\mathcal{A}$ is realized by a convex maximization problem [DW05]. In PCCA+ the function

$f(\mathcal{A}) = \mathrm{trace}(\mathrm{diag}(\mathcal{A}(1,:))^{-1}\mathcal{A}^T\mathcal{A})$ is maximized [Web06] aiming at increasing the angle between the column vectors of $C$. Here, $\mathrm{diag}(\mathcal{A}(1,:))$ is the diagonal matrix constructed from the first row of $\mathcal{A}$. The procedure of GenPCCA is depicted in Algorithm 1.

**Algorithm 1** GenPCCA applied to a transition matrix $P$

1. Given an initial positive distribution $\eta$, compute the diagonal matrix $D_\eta^{0.5}$.

2. Compute a real Schur decomposition of $\widetilde{P} = D_\eta^{0.5} P D_\eta^{-0.5}$. From a spectral gap analysis, determine the number $n$ of clusters.

3. With the aid of the method SRSchur [Bra02], sort the Schur matrix in such a way that the $n$ Schur values close to the Perron root 1 are in the top left part of the Schur matrix. Extract the leading Schur vectors $\widetilde{X}$ from the Schur matrix.

4. Compute $X = D_\eta^{-0.5}\widetilde{X}$.

5. Apply PCCA+ to the matrix $X$ to yield $C = X\mathcal{A}$.

# 3 Examples

In this section we investigate two types of examples. These examples will show that GenPCCA is indeed a powerful generalization of PCCA+. Instead of computing a projected Markov chain of a reversible metastable process, it can be used to rigorously analyze non-reversible chains or in order to find transient states which have a common target set of states. Recently the GenPCCA has been applied to biomolecules with an electric field [RWF+18].

## 3.1 Example: Illustrative Metastability

In the first example, we analyze the following transition network in Figure 1. The corresponding transition matrix has one real eigenvalue i.e. $\lambda_1 = 1$ and eight complex eigenvalues. Out of these, the two eigenvalues with the highest real absolute value are $\lambda_{2,3} = 0.9483 \pm 0.0279i$. These values are close to $\lambda_1 = 1$ and indicate in total three metastabilities. Analyzing this network via PCCA+ is impossible. If we make it reversible before applying PCCA+, we spoil the directed structure of the network (Fig. 2).

In contrast to that, GenPCCA can directly be applied to the Schur vectors of the system. We assume an equal initial distribution. The matrix$\Lambda$ is then given by

$$\Lambda = \begin{pmatrix} 1.0000 & 0 & 0 \\ 0 & 0.9483 & 0.0279 \\ 0 & -0.0279 & 0.9483 \end{pmatrix},$$

which corresponds to the eigenvalue analysis. After taking a proper linear combination of the leading Schur vectors, the result of GenPCCA clearly shows the different grades of membership that reflect the directed structure of the network (Fig. 3).

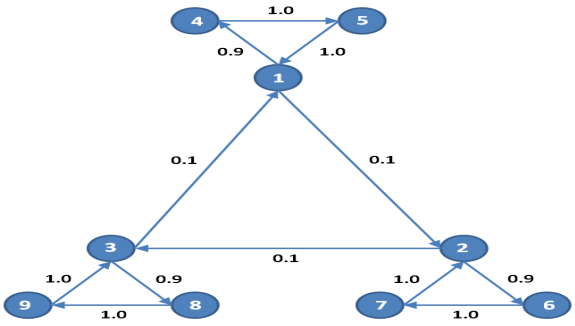



Figure 1: Transition network with three metastable states, but with a directed, non-reversible transition pattern.

### 3.2 Example: GenPCCA and Symmetrized PCCA+

In order to illustrate which kind of results are to be expected by GenPCCA, we construct a random matrix. First, three 10-by-10 random matrices A1, A2, and A3 are constructed using the Matlab-routine [MAT10] "RAND(10,10)". Another 10-by-10 zero matrix Z is constructed such that "C=[[Z, A1,Z];[Z,Z,A2];[A3, Z,Z]];". After adding a random matrix with entries between 0 and 0.1 to C, the rows of this matrix are rescaled such that the resulting matrix $P$ is stochastic. In Fig. 4 this transition matrix is depicted with its clearly visible block structure.

According to theory, any positive initial distribution $\eta > 0$ is possible. We will chose a random initial distribution. The rescaled matrix used for a Schur decomposition is given by $\widetilde{P} = D_\eta^{0.5} P D_\eta^{-0.5}$. This matrix has a partial real Schur decomposition of the form $\widetilde{P}\widetilde{X} = \widetilde{X}\Lambda$ with a non-diagonal matrix $\Lambda$. In our realization,

$$\Lambda = \begin{pmatrix} 1.0000 & -0.0267 & -0.0928 \\ 0 & -0.3884 & -0.6836 \\ 0 & 0.6426 & -0.3884 \end{pmatrix}.$$

Besides the diagonal element "1" (Perron root) there is a 2-by-2-block on the diagonal of $\Lambda$, which belongs to a complex eigenvalue pair $-0.3884 \pm 0.6628i$ near the unit circle. The absolute values of these three eigenvalues are well separated from the other absolute values. The matrix of rescaled Schur vectors used for GenPCCA is constructed by $X = D_\eta^{-0.5}\widetilde{X}$. Note that the first column vector of $X$ is constant, i.e. each element is "1". This is a necessary condition for the GenPCCA algorithm. Using this matrix for GenPCCA provides a 30-by-3-membership matrix $C = X\mathcal{A}$. The three columns of this matrix are plotted in Figure 5. They correspond to the three different clusters of states which have a

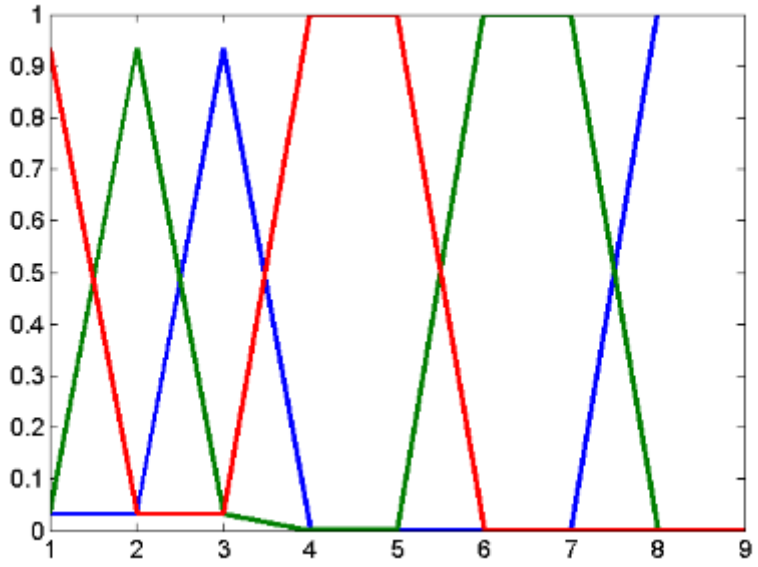



Figure 2: Applying PCCA+ to the network in Fig. 1 after making it reversible. The figure shows the membership $\chi$ of the 9 states to the three clusters (colored curves). One can see that state 3 belongs to the "green" and "red" cluster with the same grade of membership, although, the directed graph has only direct transitions from the "blue" to the "red" cluster.

similar transition pattern. This transition pattern is revealed by computing

$$P_C = \begin{pmatrix} 0.0417 & \mathbf{0.8543} & 0.1041 \\ 0.0993 & 0.0591 & \mathbf{0.8416} \\ \mathbf{0.8416} & 0.0762 & 0.0823 \end{pmatrix},$$

where the highest entries are marked. This matrix can be interpreted as the transition matrix between the three clusters of states. Note that this matrix is not diagonal dominant.

### 3.3 Example: Eye Tracking Data

In this example the GenPCCA algorithm is applied to experimental eye tracking data obtained by the Department of Psychology of the University of Potsdam with the goal to detect objects as metastable clusters using just the dynamics of the human eye, i.e. without any data of the image itself, thus providing a way of interpreting humans' object recognition expressed through eye movements (see also [KLK15]).

A group of test persons was presented with different pictures for about ten seconds, during which an eye tracker measured their fixations $f_i \in \mathbb{R}^2$ and their respective durations, $t_i \in \mathbb{R}$. For subsequent analysis it is necessary to group different areas of the image into areas of interest (AOE), which correspond to subjectively identified objects in the corresponding picture. To apply GenPCCA

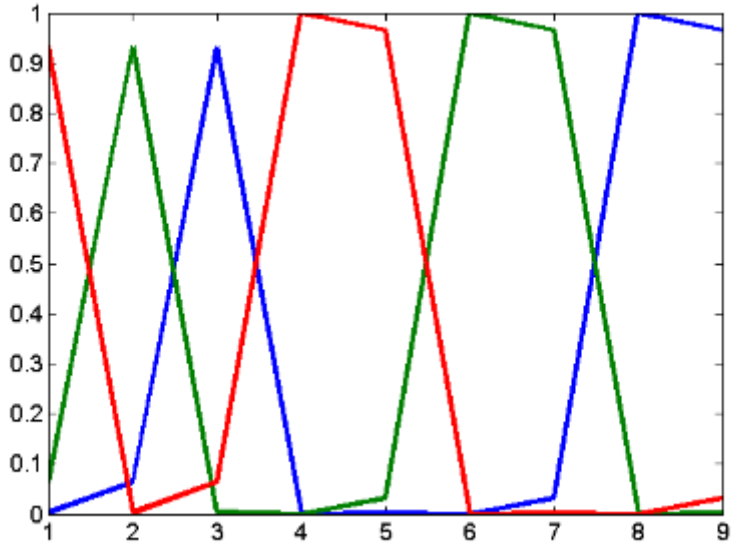



Figure 3: Applying GenPCCA directly to the network in Fig. 1. The figure shows the membership $\chi$ of the 9 states to the three cluster. One can see that, e.g., state 3 belongs to the "green" and "red" cluster with a different grade of membership, as we expect it from the directed graph.

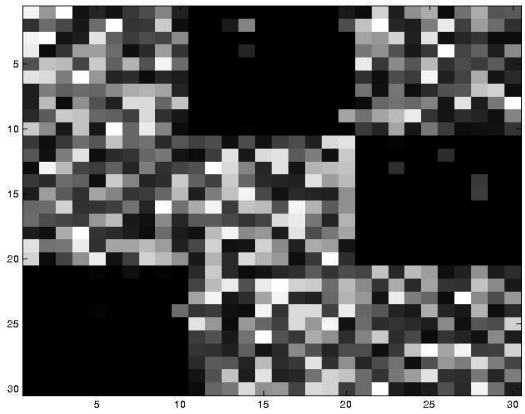



Figure 4: A realization of a 30-by-30-transition matrix. Gray scale of the entries from white to black.

we need to turn this spatial time series into a Markov chain. We model each fixation as a random choice on a spatial grid weighted by a Gaussian of the distance to the grid points, and then we construct a Markov chain by counting the induced transitions on the grid points. Assuming that humans, when

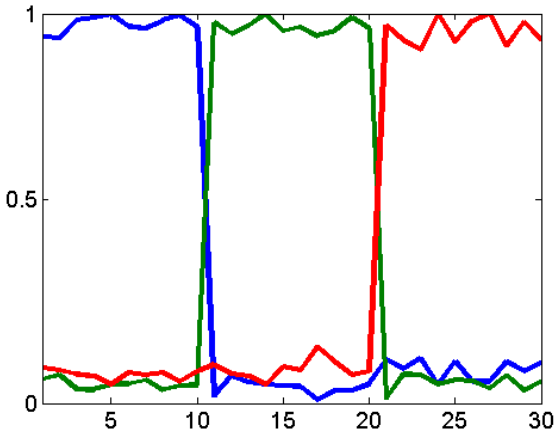



Figure 5: GenPCCA on the basis of Schur vectors can reveal the block-structure of the random matrix in Fig. 4. The three columns of the matrix $C$ are indicating the states which are assigned to each of the clusters. Note, that these are not metastable cluster of states. The states assigned to one of the clusters simply have a common transition pattern.

looking at the pictures, do not jump randomly between all recognized objects but remain for some fixations inside one AOE, this behavior should recur as high metastability of a clustering corresponding to the AOEs. As state space we choose a spatial grid $S := \{s_i\}$, where a natural choice is using all fixation coordinates as a grid, i.e. $s_i = f_i$.

Introducing a parameter $\sigma$, we assign a membership of each fixation to each grid point weighted by a Gaussian of the distance between them, i.e. for each fixation $f_i$ and each state $s_j$:

$$M_{ij} := \frac{e^{-\frac{|f_i - s_j|^2}{2\sigma^2}}}{\sum_j e^{-\frac{|f_i - s_j|^2}{2\sigma^2}}} \tag{7}$$

The mass matrix $M$ assures that nearby fixations "overlap", adopting the metric information contained in the fixation data to the Markov process. Thus, the parameter $\sigma$, scaling the distance between points, can be interpreted as a spatial coupling constant.

We then choose a fixed time step $\Delta\tau$ as grid size for the time discretization, along which we count the transitions between the states weighted with the corresponding fixation transitions, and row-normalize it to generate a transition matrix. In detail, for the transitions from state $i$ to $j$, we have

$$P_{ij} \quad = \quad \frac{\sum_{s=0} M_{f_s,i} M_{f_{s+1},j}}{\sum_{s=0} M_{f_s,i}},$$

where $f_s$ denotes the current fixation at time $s\Delta\tau$. Obviously, this Markov chain is in general not reversible, i.e. (5) does not hold. If and only if the matrix

$S = diag(\pi)P$ is symmetric, then the detailed balance condition (5) holds. As a measure of non-reversibility, we therefore compute $\delta = \|S - S^T\|/\|S\|$ as being the deviation from reversibility.

We then apply GenPCCA (with the objective function from [RW13], 4.3.1) to obtain a fuzzy clustering. Afterwards each state $s_i$ gets assigned the cluster $c_i$ with the maximal share,

$$c_i = \text{argmax}_j \chi_{ij}. \tag{8}$$

Note, that by choosing this discretization of the fuzzy clustering, some clusters may never be assigned when being dominated by other clusters on every grid point. The desired number of clusters, $n$, was chosen near the number of objects recognized by the experimenter. This, of course, is a subjective choice, but the number of clusters in general depends on the desired resolution of the clustering and thus on the future application. The time step size $\Delta\tau$ should be chosen as large as possible without skipping too many transitions. If it is chosen too large some fixations will be skipped resulting in loss of information and thus leading to a worse clustering. If, on the other hand, it is chosen too small we count one fixation as multiple self-transitions, thus weakening the effect of the real transitions and favoring the spatial over the dynamic information.

The parameter $\sigma$ introduces the spatial information and can thus be considered as a weight between dynamic and spatial clustering and is therefore inherently necessary. While small $\sigma$ values favor the dynamic information, this can lead to scattered clusters neglecting the spatial component. Large $\sigma$ values will lead to a more regular and convex clustering by enforcing a stronger spatial coupling between nearby fixations. Each of the following clusterings was computed based on about 2000 fixations, giving rise to the clustering problem on $N \approx 2000$ states. Each fixation is marked as a dot, with the color representing the corresponding cluster.

**The Sistine Madonna**

Figure 6(a) depicts Raphael's "Sistine Madonna" oil painting. In (b) we chose a small $\sigma$ value to emphasize the dynamic share of the clustering. As a consequence Saint Sixtus and Saint Barbara are clustered into one cluster, indicating back and forth movements of the observer, although they are separated by the Madonna. This non-convex clustering is a feature specific to the analysis of the dynamics which can not be reached by purely spatial clustering, e.g. $k$-means. Additionally, PCCA+ is not useful in this situation because there exist eigenvalues of $P$ which are complex valued. The matrix $P$ is not reversible with $\delta = 0.7$ which indicates that the deviation from reversibility is about 7%. With GenPCCA the clustering has been successful. The Madonna with her child and the Papal tiara are separated. A higher $\sigma$ value and the higher cluster number in (c) allow a separation between the left and right saints. Further refinement (d) leads to a seperation of Madonnas feet, her skirt and the left curtain, as well as an artifact cluster consisting of a single point in the bottom right. In this example the non-reversibility of $P$ was about 10%. Leading eigenvalues were clearly complex-valued thus "classical" spectral clustering methods based on eigenvalues are not useful. GenPCCA provides a meaningful clustering. How-

(a) Sistine Madonna

(b) $n = 5$, $\Delta\tau = 50$, $\sigma = 40$

(c) $n = 6$, $\Delta\tau = 50$, $\sigma = 50$

(d) $n = 10$, $\Delta\tau = 50$, $\sigma = 50$

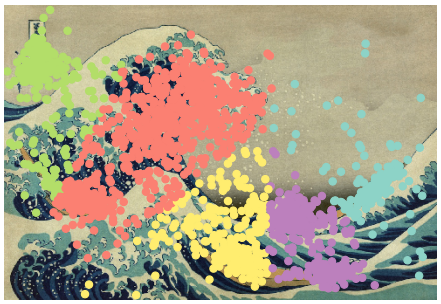
Figure 6: The Sistine Madonna (1512) by Raphael [wik18]

ever, even with Schur and GenPCCA we were not able to separate the Madonna from her child.

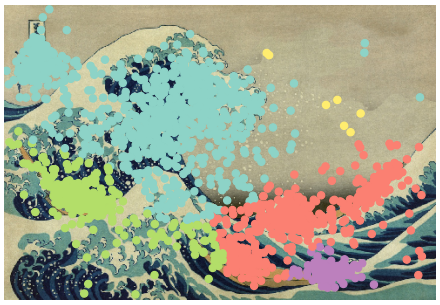
### The Great Wave of Kanagawa

In Figure 7 we compare different clustering approaches at the hands of Hokusai's "The Great Wave of Kanagawa". Part (b) depicts the k-Means clustering applied to the fixations. The purely spatial approach leads to a rather artificial seeming linear edge between the green and yellow cluster. In (c) we can see the result of applying PCCA+ to the symmetrized transition matrix leading to more natural divisions but also a single cluster for the outliers in the top right. Finally (d) shows the application of GenPCCA to the original transitions. Whilst the front boats assignment is split between the left wave and rear boat clusters in the symmetrized version, the non-symmetric approach assigns its own clusters to each boat. Note furthermore the theoretical advantage that GenPCCA provides: by circumventing the symmetrization of the transitions, the resulting clustering still admits the coarse-grained dynamics of the underlying non-reversible Markov

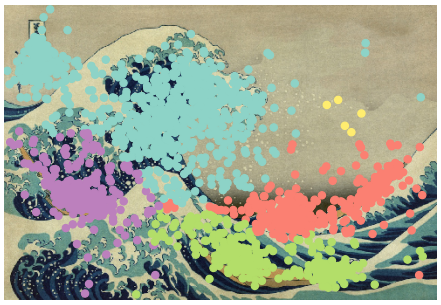
(a) The Great Wave of Kanagawa



(b) k-Means



(c) PCCA+ with symmetrized data



(d) GenPCCA

Figure 7: The Great Wave off Kanagawa (1831) by Katsushika Hokusai[wik18]. We used $k = n = 5$, $\Delta\tau = 80$, $\sigma = 40$ for the corresponding clusterings.

Chain in the form of its projected transition matrix (2), allowing for further analysis of the non-reversibility.

# 4 Conclusions

The Galerkin projection of a Markov operator $P$ onto a coarse grained matrix $P_C$ has to be chosen with care, since in general the Markovianity, which is necessary to map the correct dynamics, is not preserved, which is necessary to map the correct dynamics. We showed, that each projection, which meets the invariant subspace condition (3) and the orthogonality relation (4) preserves Markovianity. The novel method GenPCCA is also capable of treating non-reversible Markov chains by using a Schur decomposition.

**Acknowledgment.** The authors thank Reinhold Kliegl and Jochen Laubrock from the Division of Cognitive Psychology at the University of Potsdam for the eye tracking data. The work of Marcus Weber has been done for the DFG Collaborative Research Center 765. Konstantin Fackeldey would like to thank the Matheon. This is a post-peer-review, pre-copyedit version of an article published in Computational and Applied Mathematics, Springer.

# References


[BPE14] Gregory R. Bowman, Vijay S. Pande, and Frank Noé (Eds.). An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Springer, 2014.

[Bra02] J.H. Brandts. Matlab code for sorted real schur forms. Num Lin Alg App, 9(3):249–261, 2002.

[Cou75] P. J. Courtois. Error analysis in nearly-completely decomposable stochastic systems. Econometrica, 43(4):691–709, 1975.

[DHFS00] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schuette. Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. Linear Algebra and its Applications, 315(1-3):39 – 59, 2000.

[DW05] P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. Linear Algebra and its Applications, 398(0):161 – 184, 2005. Special Issue on Matrices and Mathematical Biology.

[Fil91] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. Ann. Appl. Probab., 1(1):62–87, 1991.

[FMSV07] D. Fritzsche, V. Mehrmann, D.B. Szyld, and E. Virnik. An SVD approach to identifying metastable states of Markov chains. ETNA, 29:46–69, 2007.

[FPG14] G. Froyland and K. Padberg-Gehle. Almost-invariant and finite-time coherent sets: directionality, duration, and diffusion. In Gary Froyland Wael Bahsoun, Chris Bose, editor, Ergodic Theory, Open Dynamics, and Coherent Structures, volume 70 of Proceedings in Mathematics and Statistics, pages 171–216, 2014.

[FSM10] Gary Froyland, Naratip Santitissadeekorn, and Adam Monahan. Transport in time-dependent dynamical systems: Finite-time coherent sets. Chaos: An Interdisciplinary Journal of Nonlinear Science, 20(4):043116, 2010.

[FW17] K. Fackeldey and M. Weber. Genpcca – markov state models for non-equilibrium steady states. In Big data clustering: Data preprocessing, variable selection, and dimension reduction, pages 70–80, 2017.

[GvL85] G. Golub and C.F. van Loan. Matrix Computation. John Hopkins University Press, 1985.

[HMN98] D. J. Hartfiel, Carl D. Meyer, and An N. On the structure of stochastic matrices with a subdominant eigenvalue near 1. Linear Algebra Appl, 272:272–193, 1998.

[HMS04] W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability in markovian and molecular systems. Ann. Appl. Probab., 14(1):419–458, 2004.

[Hui01] W. Huisinga. Metastability of Markovian systems. PhD thesis, Freie Universität Berlin, 2001.

[Jac10] M.N. Jacobi. A robust spectral method for finding lumpings and meta stable states of non-reversible Markov chains. ETNA, 37:296–306, 2010.

[KLK15] R. Kliegl, J. Laubrock, and A. Köstler. Augenblicke bei der Bildbetrachtung. Eine kognitionswissenschaftliche Spekulation zum Links und Rechts im Bild. In Verena M. Lepper, Peter Deuflhard, and Christoph Markschies, editors, Räume - Bilder - Kulturen, pages 77–90. de Gruyter, Berlin/Boston, 2015.

[KPC94] M. M. Konstantinov, P. H. Petkov, and N. D. Christov. Nonlocal perturbation analysis of the schur system of a matrix. SIAM Journal on Matrix Analysis and Applications, 15(2):383–392, 1994.

[KW07] Susanna Kube and Marcus Weber. A coarse graining method for the identification of transition rates between molecular conformations. Journal of Chemical Physics, 126(2), 2007.

[LBDD01] C. Li, G. Biswas, M. Dale, and P. Dale. Building models of ecological dynamics using hmm based temporal data clustering - a preliminary study. In F. Hoffmann et al, editor, Advances in Intelligent Data Analysis, volume 2189 of Lecture Notes in Computer Science, pages 53–62. Springer Berlin Heidelberg, 2001.

[LM05] Amy N. Langville and Carl D. Meyer. A survey of eigenvector methods for web information retrieval. SIAM Review, 47(1):135–161, 2005.

[LM12] Amy N. Langville and Carl D. Meyer. Google's PageRank and Beyond The Science of Search Engine Rankings. Princeton University Press, Princeton, 2012.

[MAT10] MATLAB. version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts, 2010.

[MLKCW03] C. S. Möller-Levet, F. Klawonn, K.-H. Cho, and O. Wolkenhauer. Fuzzy clustering of short time-series and unevenly distributed sampling points. In LNCS, Proceedings of the IDA2003, pages 28–30. Springer Verlag, 2003.

[RM07] T. Runolfsson and Yong Ma. Model reduction of nonreversible markov chains. In Decision and Control, 2007 46th IEEE Conference on, pages 3739–3744, 2007.

[RW13] Susanna Röblitz and Marcus Weber. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. Advances in Data Analysis and Classification, 7(2):147–179, Jun 2013.

[RWF+18] B. Reuter, M. Weber, K. Fackeldey, S. R´oblitz, and M. Garcia. A generalized markov state modeling method for non-equilibrium biomolecular dynamics – exemplified on peptide conformational dynamics driven by an oscillating electric field. Journal of Chemical Theory and Computation, (accepted for publication(DOI: 10.1021/acs.jctc.8b00079), 2018.

[SA61] H.A. Simon and A. Ando. Aggregation of variables in dynamic systems. Econometrica, 29:111–138, 1961.

[Sch99] Ch. Schütte. Conformational Dynamics: Modelling, Theory, Algorithm and Application to Biomolecules. Habilitation thesis, Freie Universität Berlin, 1999.

[Shu03] R. H. Shumway. Time-frequency clustering and discriminant analysis. Stat Prob Let, 63(3):307–314, 2003.

[SS12] C. Schütte and M. Sarich. Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches. AMS, Courant Lecture Notes Volume: 24n, 2012.

[SS14] M. Sarich and C. Schütte. Utilizing hitting times for nding metastable sets in non-reversible markov chains. Technical Report 14-32, ZIB, Takustr.7, 14195 Berlin, 2014.

[Ste94] W.J. Stewart. Introduction to the Numerical Solution of Markov Chains. Princeton University Press, 1994.

[TBHY13] J. D. Tjakra, J. Bao, N. Hudon, and R. Yang. Analysis of collective dynamics of particulate systems modeled by markov chains. Powder Technology, 235(0):228 – 237, 2013.

[Tif11] R. Tifenbach. On an SVD-based Algorithm for Identifying Metastable States of Markov Chains. ETNA, 38:17–33, 2011.

[VLKG03] M. Vlachos, J. Lin, E. Keogh, and D. Gunopulos. A wavelet-based anytime algorithm for k-means clustering of time series. In In Proc. Workshop on Clustering High Dimensionality Data and Its Applications, pages 23–30, 2003.

[Vog04] W. Vogt. Zur numerik großdimensionaler eigenwertprobleme, 2004. Preprint 22-4.

[Web02] M Weber. Clustering by using a simplex structure, 2002.

[Web06] M. Weber. Meshless Methods in Conformation Dynamics. PhD thesis, Freie Universität Berlin, 2006.

[Web11] M. Weber. A subspace approach to molecular markov state models via a new infinitesimal generator, 2011. habilitation thesis.

[wik18] From Wikimedia Commons, the free media repository, files: RAFAEL–Madonna Sixtina (Gemäldegalerie Alter Meister, Dresden, 1513-14. Òleo sobre lienzo, 265 x 196 cm).jpg; Tsunami by hokusai 19th century.jpg, 2018.

[WL07] T. Warren Liao. A clustering procedure for exploratory mining of vector time series. Pattern Recogn., 40(9):2550–2562, 2007.

[WRS06] M. Weber, W. Rungsarityotin, and A. Schliep. An Indicator for the Number of Clusters Using a Linear Map to Simplex Structure. In From Data and Information Analysis to Knowledge Engineering, 29th Annual Conference of the German Classification Society 2005, March 9-11, Studies in Classification, Data Analysis, and Knowledge, pages 103–110. Springer, Heidelberg, 2006.