An overview over PCCA+ and its application to eye tracking data

November 25, 2014

Contents

1 Introduction

In this thesis I will try to give an overview over the PCCA+ algorithm, as well as it's application.

I therefore will cover the basic PCCA+ setting, as primarily investigated by [5] in his dissertation.

I will furthermore include newest results on extending this to the setting of non-reversible chains, propose a new stochastic interpretation for fuzzy-set clustering and showcase an application to human eye-tracking data used for object recognition.

Mainly self-contained up to linear algebra knowledge.

2 Theoretical background

2.1 Introduction to Markov chains

Let S be a finite set. A Markov chain on S is a stochastic process, consisting of a sequence of random variables $X_i : \Omega \to S$, $i \in \mathbb{N}$ satisfying the Markov property:

$$P(X_{t+1} = x | X_1 = x_1, X_2 = X_2, ..., X_t = x_t) = P(X_{t+1} = x | X_t = x_t) \, \forall t \in \mathbb{N}.$$

It is common to interpret S as the state space of possible outcomes of measurements at the time t represented by X_t. The Markov property assures, that the transitions to the next timestep $t + 1$ only depend on the current state x_t. This means that the process at time t has no memory of its previous history $(x_1, ..., x_{t-1})$, thus this also sometimes called the memoryless property.

We will furthermore assume that S is finite and that the process is autonomous, i.e. not explicitly depending on the time:

$$P(X_{t+1} = x | X_t = y) = P(X_t = x | X_{t-1} = y) \forall t \in \mathbb{N}$$

This does not realy impose a restriction as any non-autonomous process can be turned into an autonomous one. By adding all possible times to the state space S taking the cartesian product $S' := \mathbb{N} \times S$ the explicit time-dependence

of the process on S can be implicitly subsumed by an autonomous process on S'.

For finite S we can, enumerating all states in S, encode the whole process in the transition matrix

$$P_{ij} := P\left(X_{t+1} = j | X_t = i\right)$$

A *stationary distribution* is a row vector π satisfying

$$\pi P = \pi$$

A markov chain is called *reversible* if there exists a stationary distribution π satisfying the detailed balance equation

$$\pi_i P_{ij} = \pi_j P_{ji} \tag{2.1}$$

which assures that the back and forth transitions between any two states π_i, π_j equalize.

Introducing the diagonal matrix

$$D_\pi := \mathrm{diag}\left(\pi\right)$$

this can also be written in matrix notation as

$$D_\pi P = P^T D_\pi.$$

2.2 Discretization of the state space

Although we only consider a discrete state space in this thesis, the results are extensible to continous state spaces as well.

The easiest way is using a set-based discretization, dividing the state space into a finite mesh of subsets.

For high dimensional state spaces, as for example met in molecular dynamics, this approach exhibits the curse of dimensionality, as the size of the mesh grows exponentially with the dimensions. As solution to this problem Weber developed a meshless version of PCCA+ using a global Galerkin discretization[5].

2.3 Clustering of the state space

As the goal of PCCA+ is to reduce the complexity of analysis of the markov chain by a dimension reduction we will now introduce the concept of clustering, that is subsuming different states of the state space to a smaller set of n clusters.

The simplest possibility is assigning each state $k = 1, ..., N$ to a cluster $i = 1, ..., n$, which can be encoded by means of the *characteristic vector* $\chi_i \in \{0, 1\}^N$:

$$\chi_{i,k} = \begin{cases} 1, & \text{if state } k \text{ belongs to cluster } i \\ 0, & \text{else} \end{cases}.$$

This *crisp clustering* approach, used by *Perron Cluster Analysis* (PCCA) of Deuflhard et al. [1], has the disadvantage of not beeing robust against small pertubations, as continious changes in P finally result in discontinous changes in the clustering.

Weber and Galliat, Deuflhard and Weber therefore developed a robust version, *Robust Perron Cluster Analyis* (PCCA+), by making use of a *fuzzy clustering* representing each cluster by an *almost characteristic vector*

$$\chi_i \in [0,1]^n . \tag{2.2}$$

A*lmost characteristic vectors* $\{\chi_i\}_{i=1}^n$ satisfying the partition of unity property

$$\sum_{i=1}^{n} \chi_i = 1 \tag{2.3}$$

are called *membership vectors* as they describe the relative membership of each state to each cluster.

We will call the matrix $\chi := (\chi_i)_{i=1}^n \in \mathbb{R}^{N\times n}$ composed of the *membership vectors* a *clustering*, whereas in literature it is also referred to as *conformations*.

2.4 Galerkin projection of the transition matrix

2.4.1 The coupling matrix

To represent the dynamics on the reduced/clustered state space in the case of a crisp clustering χ, i.e. $\chi_i \in \{0,1\}$, Deuflhard et al. [1] introduced the *coupling matrix*

$$W_{ij} := \frac{\langle \chi_j, P\chi_i\rangle_\pi}{\langle \chi_i, \chi_i\rangle_\pi} = \frac{\chi_j^T D_\pi P \chi_i}{\chi_i^T D_\pi \chi_i}.$$

The entries W_{ij} can thus be interpreted as conditional transition probability from cluster i to cluster j, given the starting distribution π.

2.4.2 The propagator matrix

Unfortunately the projection via the *coupling matrix* does not commute with time propagation, which is desired for long term analysis of the markov chain which plays a role in (bio/med...).

Therefore Kube and Weber [3] proposed the *coarse propagator matrix*

$$P_C := \left(\chi^T D_\pi \chi\right)^{-1} \chi^T D_\pi P \chi,$$

which coincides with the *coupling matrix* W in the crisp clustering setting, but has the advantage of representing the right dynamics of the underlying markov chain on S, even on the coarser clustered state space, in the sense that discretization via χ and time propagation commute, i.e.

$$P\chi = \chi P_C. \tag{2.4}$$

This property ensures that the *coupling matrix* represents the right dynamics even for iterative application, i.e. $P^n\chi = \chi P_C^n$, which is not possible in general using a crisp clustering as in PCCA.

We now show this property is satisfied, assuming that χ is a linear combination of vectors spanning an P-invariant subspace satisfying an orthonormality condition, which will be satisfied by the PCCA+ approach.

Theorem 1. *Let $\chi = XA$, $X \in \mathbb{R}^{N\times n}$, $A \in \mathbb{R}^{n\times n}$ satisfying the subspace condition*

$$PX = X\Lambda \tag{2.5}$$

for some $\Lambda \in \mathbb{R}^{n\times n}$ and the orthonormality condition

$$X^T D_\pi X = I. \tag{2.6}$$

Then the P_C is conjugate to Λ and discretization-propagation commutativity 2.4 holds.

Proof. We calculate

$$\begin{aligned} P_C &= \left(\chi^T D_\pi \chi\right)^{-1} \chi^T D_\pi P \chi \\ &= \left(A^T X^T D_\pi X A\right)^{-1} A^T X^T D_\pi X \Lambda A \\ &= \left(A^T A\right)^{-1} \left(A^T \Lambda A\right) \\ &= A^{-1}\Lambda A, \end{aligned} \tag{2.7}$$

which implies

$$P\chi = PXA = X\Lambda A = XAA^{-1}\Lambda A = \chi P_C.$$

□

3 PCCA+

3.1 Motivation for the spectral decomposition

PCCA+ will construct the clusters, described by the *membership vectors*, as a linear combination of eigenvectors. By choosing the $n < N$ eigenvectors with the largest eigenvalues, one hopes to preserve the principal dynamics of P. The eigenvectors are good data for this goal, as an eigenvector with high eigenvalue ....

We now first present the PCCA+ algorithm in the case of reversible processes.

We first will construct the matrix X spanning the invariant subspace, then examine the possible linear transformations A mapping these to a set of membership vectors and finally propose an optimization problem to specify a "good" solution, representing the goal of metastability in the form of a objective function.

3.2 Reversible processes

3.2.1 Construction of X and Λ

To satisfy the orthonormality condition 2.6 we define the matrix $\tilde{P} := D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}}$, where D_π denotes the diagonal matrix with the stationary distribution π.

As we assume a reversible process the detailed balance condition 2.1 holds, assuring that $\tilde{P}$ is symmetric:

$$\tilde{P}^T = D_\pi^{-\frac{1}{2}} P^T D_\pi^{\frac{1}{2}} = D_\pi^{-\frac{1}{2}} D_\pi P D_\pi^{-1} D_\eta^{\frac{1}{2}} = \tilde{P}$$

We therefore can diagonalize $\tilde{P}$ such that $\tilde{P}\tilde{X}' = \tilde{X}'\Lambda'$ with $\tilde{X}' \in \mathbb{R}^{N\times N}$ beeing orthonormal and $\Lambda' \in \mathbb{R}^{N\times N}$ beeing diagonal.

We then select the n largest eigenvalues $\Lambda \in \mathbb{R}^{n\times n}$ and the corresponding eigenvectors $\tilde{X} \in \mathbb{R}^{N\times n}$, and define $X := D_\pi^{-\frac{1}{2}} \tilde{X}$.

We then see that the conditions for 1 are satisfied:

$$\begin{aligned} \tilde{P}\tilde{X} &= \tilde{X}\Lambda \\ \Leftrightarrow D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}} D_\pi^{\frac{1}{2}} X &= D_\pi^{\frac{1}{2}} X \Lambda \\ \Leftrightarrow PX &= X\Lambda \end{aligned}$$

$$\begin{aligned} \tilde{X}^T\tilde{X} &= I \\ \Leftrightarrow X^T D_\pi^{\frac{1}{2}} D_\pi^{\frac{1}{2}} X &= I \\ \Leftrightarrow X^T D_\pi X &= I \end{aligned}$$

3.2.2 Feasible Set

Given a fixed eigenvector matrix X, we will now examine the set of feasible solutions $F_A \subset \mathbb{R}^{n\times n}$ for the transformation matrix A, defined by the leading to *membership vectors* $\chi = XA$.

Making use of the fact that $X_{i,1} = 1$, $i = 1, ..., N$ (reference?) one can reformulate the positivity 2.2 and partition of unity 2.3 conditions in terms of the matrices X and A, leading to the following constraints for A:

$$A_{1,j} \geq -\sum_{k=2}^{n} X_{ik} A_{kj},\ i = 1, ..., N,\ j = 1, ..., n \text{ (positivity)}, \tag{3.1}$$

$$A_{i,1} = \delta_{i,1} - \sum_{j=2}^{n} A_{ij},\ i = 1, ..., n \text{ (partition of unity)} \tag{3.2}$$

Since these constraints are linear in A the set F_A is a convex polytope, and it is not empty as the matrix $A^*_{ij} := \frac{\delta_{i,1}}{n}$ satisfies these conditions.

As Deuflhard and Weber [2] have shown, the set F_A is indeed uncountable. We therefore look for some criterion to choose a specific solution by means of choosing an objective function for an optimization problem. To motivate the specific choices we first try to gain some insight into the geometry of the clustering problem.

3.2.3 Geometric interpretation

If one considers the N rows of the matrix χ as points in the space $\mathbb{R}^n$, the positivity and 2.2 and partition of unity 2.3 conditions force these points to lie on the standard $(n-1)$-simplex Δ. Now, if $\chi = XA$ this means the that the matrix A maps the the N rows of the eigenvector matrix X onto that simplex.

Assuming that for every $j = 1, ..., N$ there exists an $i = 1, ..., n$ the convex hull of the rows of X already has the form of an $(n-1)$-simplex, we can now choose A uniquely (up to permutation) to map this exactly onto Δ, which among all the ways of mapping X into Δ gives us the highest distinguishability between the resulting clusters. It is easy to check that this assumption is equivalent to

$$\max_{i=1..N} \chi_{ij} = 1,\ j = 1, ..., n$$

and therefore also called the *maximality assumption*.

(Upper bound for metastability on W)

3.2.4 Maximal scaling condition

As in the general the *maximality assumption* is not met, it seems natural to turn it into an optimization problem. This has been done in [2, 5] by imposing maximization of the *maximal scaling condition*

$$I_1(A) := \sum_{j=1}^{n} \max_{i=1..N} \chi_{ij} \leq n_C.$$

Assuming that the *maximality assumption* is almost met, i.e. $\max_{i=1..N} \chi_{ij} \approx 1,\ j = 1, ..., n$, Weber [5] argues that the maximizing indices can be determined by the *index mapping algortithm*, turning this convex optimization problem into a linear one:

$$I_1(A) = \sum_{i,j=1}^{n} X_{ind(X)_j, i} A_{ij}$$

3.2.5 Maximal metastability condition

Another choice might be optimizing towards a maximal metastability, as done by Deuflhard and Weber [2, 5]

$$I_2(A) := \operatorname{trace}(W) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \frac{A(i,j)^2}{A(1,j)},$$

where they establish the latter equation making use of $\pi_i = A_{1,i}$ ([5], Lemma 3.6).

3.2.6 Crispness objective

Röblitz [4] argues that the use of W has no stochastic interpretation in the fuzzy setting. Optimization of the trace of P_C makes no sense as it is similar to $\Lambda 2.7$ and therefore independent of A. Defining

$$\mathcal{S} = ...$$

she therefore suggests maximization of

$$I_3 := \text{trace}\,(\mathcal{S}) = \sum_{i,j=1}^{n} \frac{(A_{ij})^2}{A_{1,j}}$$

which is similar to I_2 ....

This condition minimizes the off-diagonal entries of $\mathcal{S}$ which means that the corresponding clustering χ is as crisp as possible.

3.2.7 Unconstrained Optimization

Due to the high number of inequality constraints 3.1 solving these linear or convex problems may still be very time consuming. Therefore Deuflhard and Weber [2, 5] have shown how to turn this constrained into an unconstrained optimization problem.

Assuming that the objective function is convex over F_A, which is the case for the here given objective functions, is attains its maximum at one of the vertices $v\,(F_A)$.

This allows the determination of the $2n$ active constraints

$$\begin{aligned} A_{i,1} &= \delta_{i,1} - \sum_{j=2}^{n} A_{ij},\ i = 1, ..., n \\ A_{1,j} &= -\min_{l=1,...,N} \sum_{i=2}^{n} X_{li} A_{ij},\ j = 1, ..., n \end{aligned}$$

which define a subspace $F_A^{'} \subset F_A$ still containing all vertices $v\,(F_A) \subset F_A^{'}$ and thus containing the maximizin matrix A. These equations furthermore reduce the search space to $(n-1)^2$ unknowns.

Side condition...

3.2.8 Index mapping algorithm

3.2.9 The PCCA+ Algorithm

3.2.10 Determination of the number of clusters n

So far we have imposed a desired number of clusters n.

3.3 Nonreversible processes

The whole algorithm is also applicable to nonreversible processes, with some smaller adjustments concerning the construction of the invariant subspace and the stochastic interpretation, as well as a new optimization goal.

3.3.1 Stochastic interpretation of the coupling matrix

3.3.2 Construction of X and Λ

When the underlying stochastic process is not reversible the matrix P is no more real diagonalizable. We therefore make use of the *real Schur decomposition*, decomposing a matrix $A = QTQ^{-1}$ into a orthonormal matrix Q, called the *Schur vectors,* and an upper quasi-triangular (1-by-1 and 2-by-2 blocks on its diagonal) matrix T, called the *Schur form.* The columns of Q are called the Schur vectors of P. The eigenvalues of P appear on the diagonal of T, where complex conjugate eigenvalues correspond to the 2-by-2 blocks.

To compute an orthonormal basis for an invariant subspace belonging to n eigenvalues one can reorder the diagonal blocks of T such that the upper left $n \times n$ block contains these n eigenvalues. Then the first n columns of the updated transformation matrix Q form a basis for the desired subspace [**?**].

So we define, analogously to 3.2.1 $\tilde{P} := D_\eta^{\frac{1}{2}} P D_\eta^{-\frac{1}{2}}$ and compute the *real Schur decomposition* of $\tilde{P}$. We then select the $n \times n$ blocks corresponding to the n eigenvalues with the highest absolute value by the reordering procedure. Let us denote the resulting *Schur vectors* by $\tilde{X}$ and the *Schur form* by Λ.

Now $X := D_\pi^{-\frac{1}{2}} \tilde{X}$ and Λ satisfy the conditions for 1 (same calculations as in3.2.1).

3.3.3 Optimization

Not only metastable, but also cyclic dynamics.

3.4 Summary

4 Application to eyetracking data

This algorithm was applied to eyetracking data.

4.1 The experiment

Universität Potsdam

We measure the fixations $f_i \in \mathbb{R}^2$

4.2 Implementation

As state space we define the coordinates of each fixation $S := \{s_i\}$, $s_i = f_i$, though if one has too many fixations we recommend using for example a k-means clustering of the fixations.

We then compute the membership of each fixation f_i to each state s_j based on the gaussian of the distance:

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

Then summing over all measured transitions $f_a \to f_b$ and normalizing to row sum 1 we compute the resulting transition matrix P:

$$P_{ij} = \frac{\sum_{f_a \to f_b} M_{ai} M_{bj}}{\sum_{k,l} \sum_{f_a \to f_b} M_{ak} M_{bl}}$$

We compute the stationary distrubution (why) as:

$$this.pi = sum(this.C, 2)/sum(sum(this.C));$$

Once we have constructed P this way we now compute the invariant eigenspace using the weighted Schur decomposition as in 3.3.2 and pass is to PCCA+, which in return gives us the clustering χ.

We then use the stochastic interpretation 3.3.1 to calculate a a transition matrix on our reduced state space.

Then we compare the two testgroups by first reordering the clusters to maximize the correlation between same-numbered clusters.

4.3 Results

5 Junk

5.0.1 Stochastic interpretation

By multiplying the conditional transitions of P with the actual starting distribution we get the unconditional transitions matrix $D_\eta P$, which we can interpret as amount of actual transitions between the states. Now multiplying with a membership vector χ_i^T from the left gives the numbers of transitions starting in χ_i and finally multiplying with χ_j from the right measures the amount of these transitions landing in χ_j, i.e. $\chi_i^T D_\eta P \chi_j$.

gives the unconditional number of transitions from cluster i to cluster j. Note that this interpretation is associative in the sense that we come to the same results by for example first interpreting $\chi_i^T D_\eta$ as actual amount of states in the cluster χ_i and measure the overlap with where the transitions to χ_j, come from: $\left(\chi_i^T D_\eta\right)\left(P\chi_j\right)$.

We can vectorize this equation to compute the unconditional transitions by means of the *membership matrix* $\chi = (\chi_1, ..., \chi_n)$*:*

$$\chi^T D_\eta P \chi$$

We then turn this unconditional transitions into conditinal by dividing by the number of states belonging to χ_i and have to make up the fact that we do not count all transitions... Baustelle

5.1 Junk2

The key idea of extracting metastable behaviour via PCCA+ is transforming the eigenvectors X of P to membership vectors representing the clusters of S. Taking linear combinations A of n eigenvectors $X = (x_1, ..., x_n)$ with eigenvalues $\lambda_1, ..., \lambda_n \approx 1$ guarantees "metastable" (in some way) properties of the cluster $\chi_j = XA_{:,j}$:

$$\begin{aligned}
\langle P\chi_j, \chi_j \rangle &= \left\langle \sum_k PX_{ik}A_{kj}, \sum_k X_{ik}A_{kj} \right\rangle \\
&= \left\langle \sum_k \lambda_k X_{ik}A_{kj}, \sum_k X_{ik}A_{kj} \right\rangle \\
&= \sum_{i,k} \lambda_k \left(X_{ik}A_{kj}\right)^2 \\
&\geq \min_k \lambda_k \sum \left(X_{ik}A_{kj}\right)_{i,k}^2 \\
&= \min_k \lambda_k \langle \chi_j, \chi_j \rangle \\
\Rightarrow \frac{\langle P\chi_j, \chi_j \rangle}{\langle \chi_j, \chi_j \rangle} &\approx 1
\end{aligned}$$

Also these so called perron vectors only exist when the matrix is diagonalizable (the process is reversible), this approach can be extended to non-reversible processes by using a Schur decomposition instead of a diagonalization. Reordering the Schur decomposition and selecting only the highest eigenvalue eigenvectors then leads to a similar P-invariant subspace.

References

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

[2] P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005.

[3] S. Kube and M. Weber. A coarse graining method for the identifiation of transition rates between molecular conformations. *Journal of Chemical Physics*, 126(2), 2007.

[4] S. Röblitz and M. Weber. Fuzzy spectral clustering by pcca+: application to markov state models and data classification. *Advances in Data Analysis and Classification*, 7:147–179, 2013.

[5] M. Weber. *Meshless methods in Conformation Dynamics*. PhD thesis, Freie Universität Berlin, department of mathematics, 2006.