PCCA+ on Markov Chains

November 18, 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 [6] 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 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

$$
\begin{aligned}
\pi_i P_{ij} &= \pi_j P_{ji} \qquad (2.1)\\
D_\pi P &= P^T D_\pi
\end{aligned}
$$

D_pi?

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

2.1 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[6].

2.2 Clustering and Metastability

Abstract explanation of clustering as reduction of state space / coupling matrix

Commutativity

Introduction of the concept of metastability

3 Spectral Clustering

3.1 Membership vectors

We now look for a way to represent the clustering into n clusters.

One possibility of assigning states to a cluster i is by means of a *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*

$$\chi_i \in [0,1]^n . \qquad (3.1)$$

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

$$\sum_{i=1}^{n} \chi_i = 1 \tag{3.2}$$

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 of characteristic vector a *clustering.*

3.2 Galerkin projection

3.2.1 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

3.2.2 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.

Unfortunately the stochastic interpretation of W fails for a fuzzy clustering, not least due to the fact that it's entries may become negative.

3.2.3 The propagator matrix

Assuming a *starting distribution* $\eta \in [0, 1]^N$, we define the diagonal matrix $D_\eta := diag\,(\eta_1, ..., \eta_N) \in [0, 1]^{NxN}$.

Kube and Weber [3, 4] proposed the *coarse propagator matrix*

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

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.

$$P\chi = \chi P_C. \tag{3.3}$$

This property ensures that the *coupling matrix* represents the right dynamics even for iterative application, as we can see by following both outer paths of the following diagram.

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*

$$PX = X\Lambda \tag{3.4}$$

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

$$X^T D_\eta X = I. \tag{3.5}$$

Then the P_C is conjugate to Λ and discretization-propagation commutativity 3.3 holds.

Proof. We calculate

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

which implies

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

□

3.3 PCCA+ and metastability

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.

3.3.1 Construction of X and Λ

To satisfy the orthonormality condition 3.5 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 and eigenvectors $\tilde{X} \in \mathbb{R}^{n\times n}$, $\Lambda \in \mathbb{R}^{n\times n}$ and define $X := D_\pi^{-\frac{1}{2}}\tilde{X}$.

We then see that the conditions for 1 are satisfied:

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

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

3.3.2 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 leading to the *membership vectors* χ.

Making use of the fact that $X_{i,1} = 1$, $i = 1, ..., N$ (reference?) one can reformulate the positivity 3.1 and partition of unity conditions 3.2 in terms of the matrices X and A:

$$\begin{aligned}
A_{1,j} &\geq -\sum_{k=2}^{n} X_{ik}A_{kj},\; i = 1, ..., N,\; j = 1, ..., n \text{ (positivity)}, & (3.6) \\
A_{i,1} &= \delta_{i,1} - \sum_{j=2}^{n} A_{ij},\; i = 1, ..., n \text{ (partition of unity)} & (3.7)
\end{aligned}$$

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

3.3.3 Optimization

As Deuflhard and Weber [2] have shown, the set F_A is indeed uncountable, one can now specify an objective function leading to a optimization problem to find a specific solution $A \in F_A$. I will now give an overview over the objectives, which were developed during the last years.

In [6] Weber proposed the maximization of the following two objective functions

$$I_1 := \sum_{j=1}^{n} \max_{l=1..N} \chi_{l,j} \leq n_C$$

$$I_2 := \operatorname{trace}(W).$$

I_1 can be interpr

assuring a high correspondance of each state to some cluster, thus leading to a almost crisp clustering.

However, as these in [5] it is argued that this metastability is

3.4 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.

3.4.1 Stochastic interpretation of the coupling matrix

3.4.2 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 3.3.1 $\tilde{P} := D_\eta^{\frac{1}{2}} P D_\eta^{-\frac{1}{2}}$ 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 in3.3.1).

Note that we made no use of η beeing a stationary distribution so far.

3.4.3 Optimization

Röblitz reinterpretation of crispness obj.

new metastability objective

3.5 Summary

4 Application to eyetracking data

This algorithm was applied to eyetracking data.

4.1 The experiment

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:

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

We compute the stationary distrubution 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 3.4.2 and pass is to PCCA+, which in return gives us the clustering χ.

We then use the stochastic interpretation 3.4.1 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: $\left(\chi_i^T D_\eta\right)(P\chi_j)$.

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 $\chi_j = XA_{:,j}$:

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

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] Fischer Schütte Deuflhard, Huisinga. Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Algebra Appl 313:39-59*, 2000.

[2] Weber Deuflhard. Robust perron cluster analysis in conformation dynamics. *Linear Algebra Appl 398*, 2005.

[3] Weber Kube. Coarse grained molecular kinetics. *ZIB-Report 06-35, Zuse Institute Berlin*, 2006.

[4] Weber Kube. A coarse graining method for the identifiation of transition rates between molecular cconformations. *J Chem Phys 126(2)*, 2007.

[5] Weber Röblitz. Fuzzy spectral clucluster by pcca+: application to markov state models and data classification. *Avd Data Anal Classif (2013) 7:147-179*, 2013.

[6] Marcus Weber. *Meshless methods in Conformation Dynamics*. PhD thesis, 2006.