PCCA+ on Markov Chains

October 15, 2014

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 [1] 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

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

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.

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 k is by means of a *characteristic vector* $\chi_i \in \{0,1\}^n$:

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

This *crisp clustering* approach, used by *Perron Cluster Analysis* (PCCA) of Deuflhard et al., 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 therefor developed a robust version, *Robust Perron Cluster Analyis* (PCCA+), by making use of a *fuzzy clustering* representing each cluster by an *almost characteristic vector* $\chi_k \in [0,1]^n$.

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

$$\sum_{i=1}^{n} \chi_i = 1 \tag{3.1}$$

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

3.2 Galerkin projection

3.2.1 Motivationn for the spectral decomposition

why using eigenvector data?

3.2.2 The coupling matrix

Given a set of *membership vectors* describing a clustering, we now aim at a Galerkin projection of P, to get a reduced version representing the dynamics on the clusters.

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

η vs π?

Thus we finally define the *coupling matrix* via

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

For a correct representation of the dynamics of the markov chain we require that discretization via χ and time propagation commute in the sense that

$$P\chi = \chi P_C. \tag{3.2}$$

This property ensures that the *coupling matrix* represents the right dynamics even, 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*

$$PX = X\Lambda \tag{3.3}$$

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

$$X^T D_\eta X = I. \tag{3.4}$$

Then the P_C is conjugate to Λ and discretization-propagation commutativity 3.2 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

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

□

3.2.3 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

3.3 PCCA+ and metastability

The goal is to linearly transform the matrix X spanning the invariant subspace to a "good", specified by some objective function, set of membership vectors.

3.3.1 Construction of Λ

3.3.2 Feasible Set

It is useful to think of the rows of X and χ as points in n-dimensional space.

The membership vector conditions, positivity and partition of unity, ensure that the rows of χ lie on the standard $(n-1)$-simplex.

Hence we look for a linear transformation A mapping the N points of each row of X onto that simplex.

$$restriction of feasible \tag{3.5}$$

3.3.3 Optimization

Since we have a lot of matrices satisfying the conditions, we now have the freedom to choose the clustering with the highest metastability.

METASTABILITIES:

During the last years several objectives developed.

Weber proposed to maximize the crispness objective

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

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

In 3.5.3 it is explained that I_1 is an upper bound for the metastability FOR REV. PROCESSES.

He also proposed the *metastability objective*:

$$I_2 := \operatorname{trace}(P_C)$$

Same trace because conjugate?

3.4 Nonreversible processes

3.4.1 Stochastic interpretation of the coupling matrix

3.4.2 Construction of Λ

3.4.3 Optimization

Röblitz reinterpretation of crispness obj.

new metastability objective

3.5 Summary

4 Eyetracking

5 Junk

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

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

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