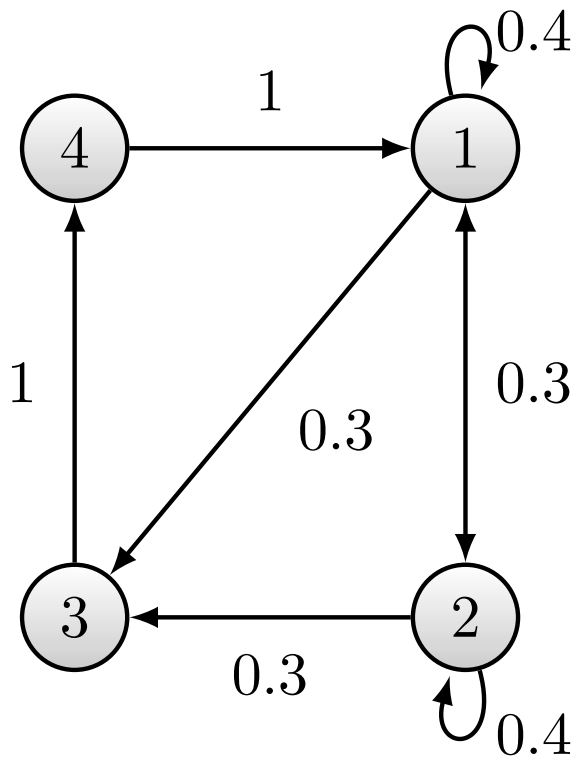
Markov State Space Aggregation via the Information Bottleneck Method

Bernhard C. Geiger
Joint Work with T. Petrov, G. Kubin, & H. Koeppl

Institute for Communications Engineering

Feb. 19, 2015

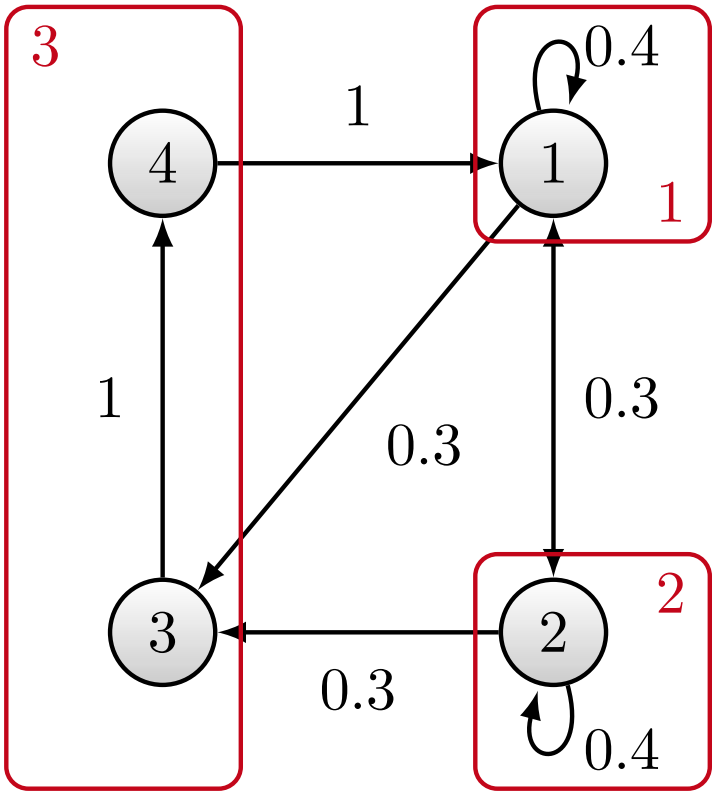
Problem Statement



$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$

- Stationary Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$
  - state space $\mathcal{X} = \{1, \ldots, N\}$,
  - transition matrix $\mathbf{P}$ ($\mathbf{X}$ is irreducible and aperiodic),
  - invariant distribution $\boldsymbol{\mu}$
- We want to *reduce the state space* to $\mathcal{Y} = \{1, \ldots, M\}$, $M < N$

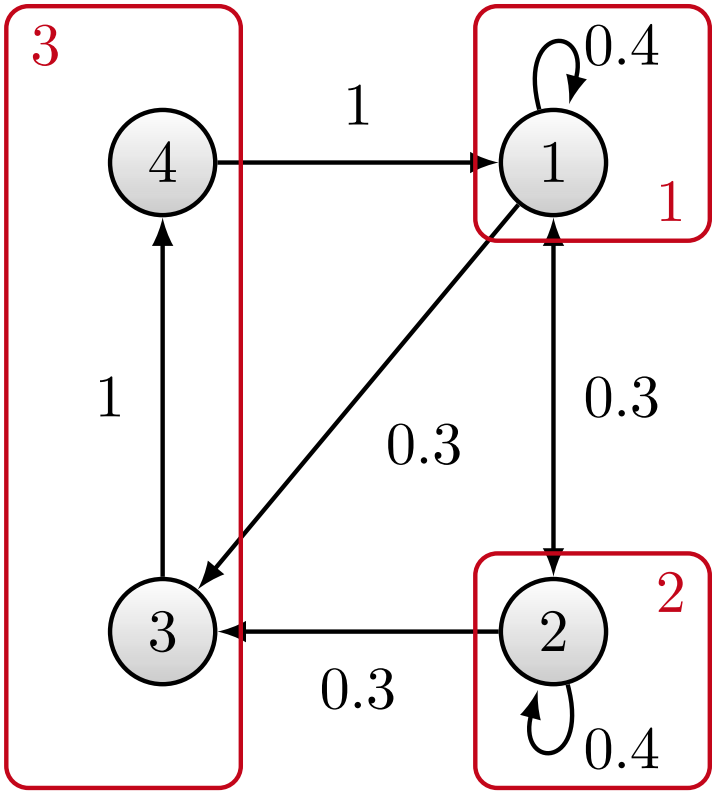
Problem Statement



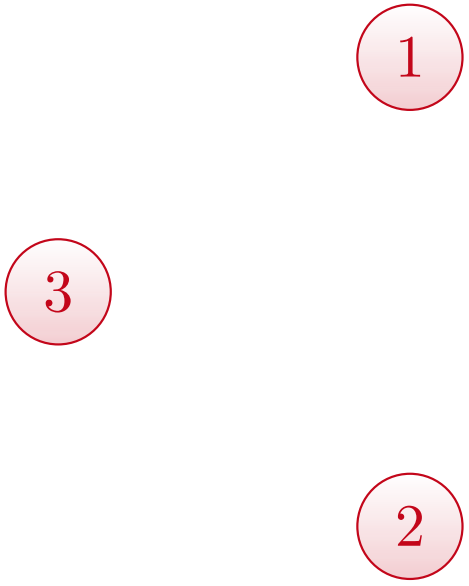
$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$

- Stationary Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$
  - state space $\mathcal{X} = \{1, \dots, N\}$,
  - transition matrix $\mathbf{P}$ ($\mathbf{X}$ is irreducible and aperiodic),
  - invariant distribution $\boldsymbol{\mu}$
- We want to *reduce the state space* to $\mathcal{Y} = \{1, \dots, M\}$, $M < N$, by a non-injective function g: $\mathcal{X} \to \mathcal{Y}$.

Problem Statement

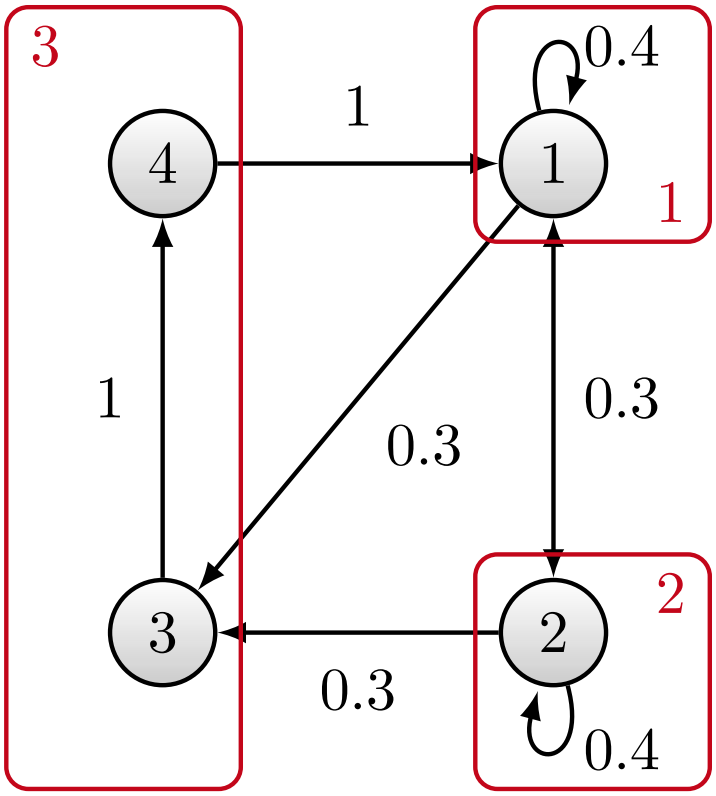


$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$

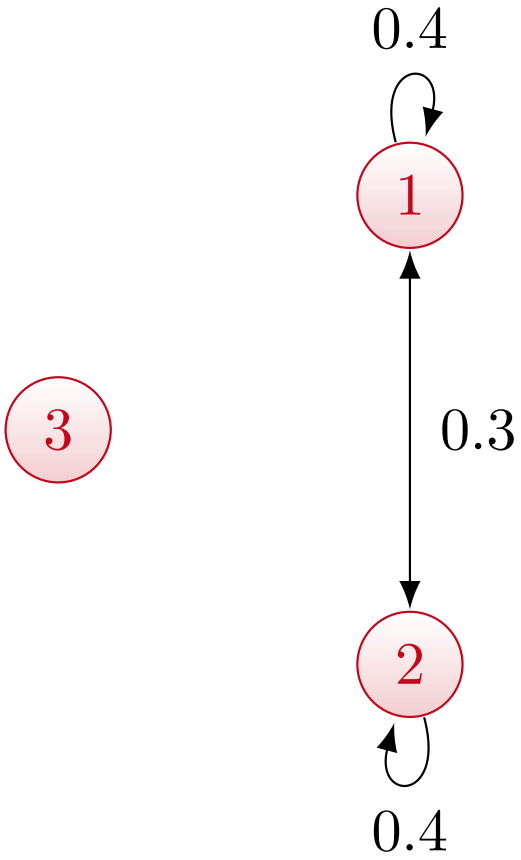


$$\mathbf{Y}'_g \sim \mathrm{Mar}(\mathbf{Q}, \boldsymbol{\nu}, \mathcal{Y})$$

Problem Statement

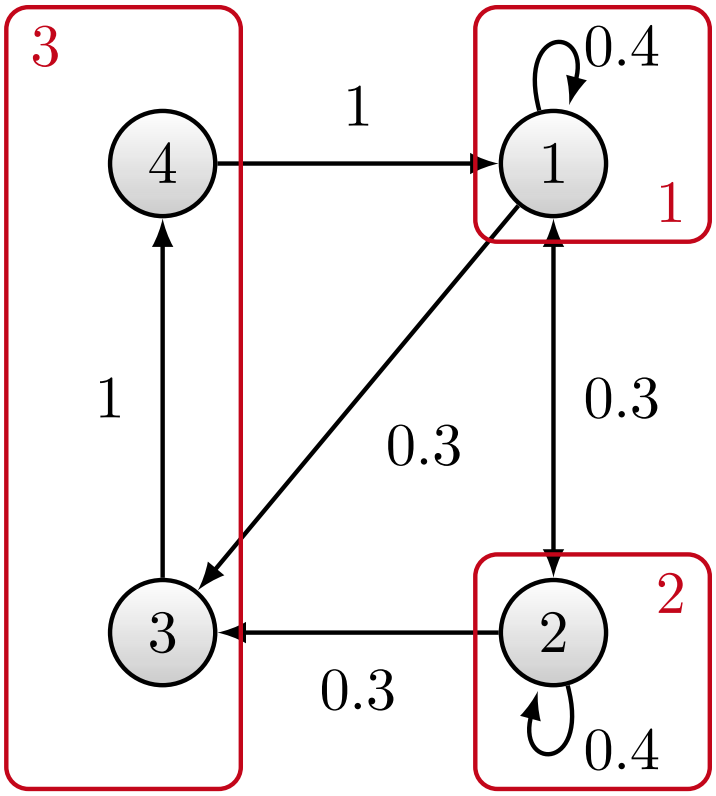


$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$

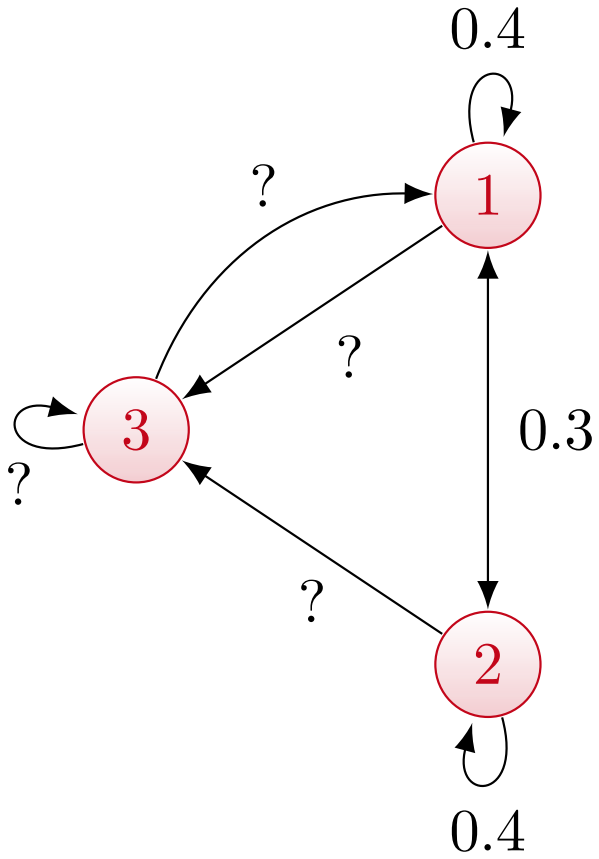


$$\mathbf{Y}'_g \sim \mathrm{Mar}(\mathbf{Q}, \boldsymbol{\nu}, \mathcal{Y})$$

Problem Statement

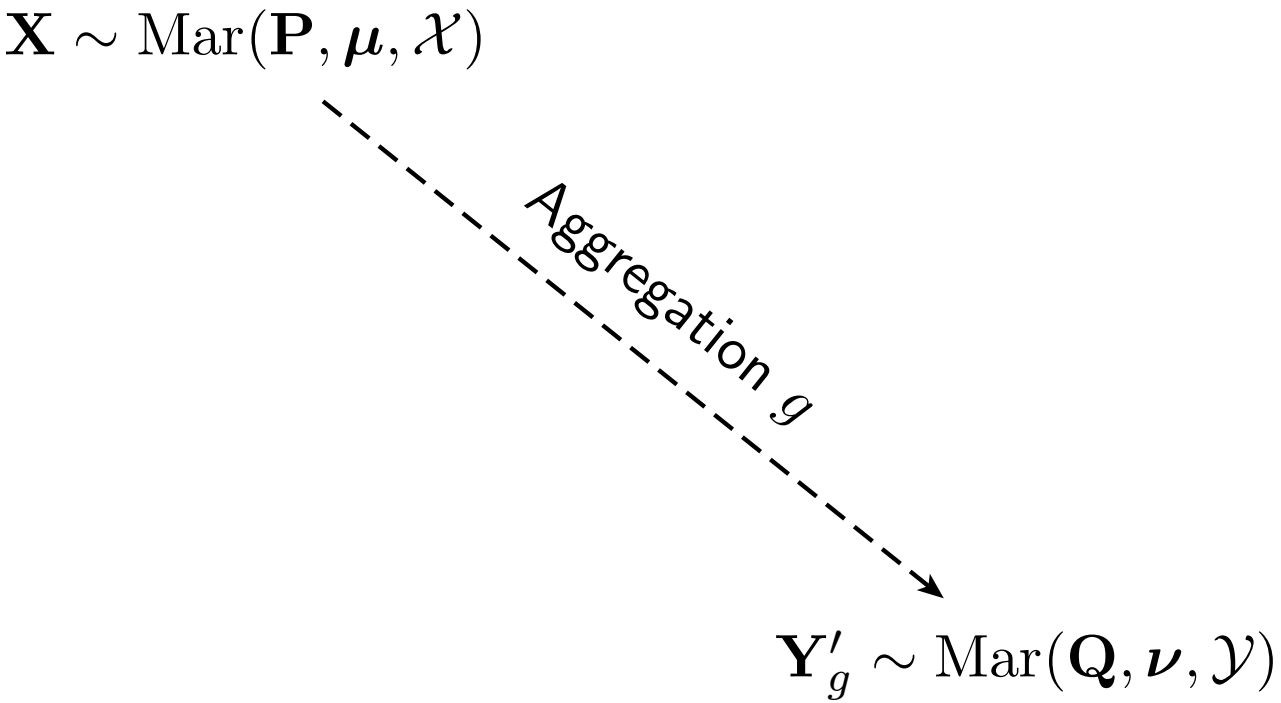


$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$

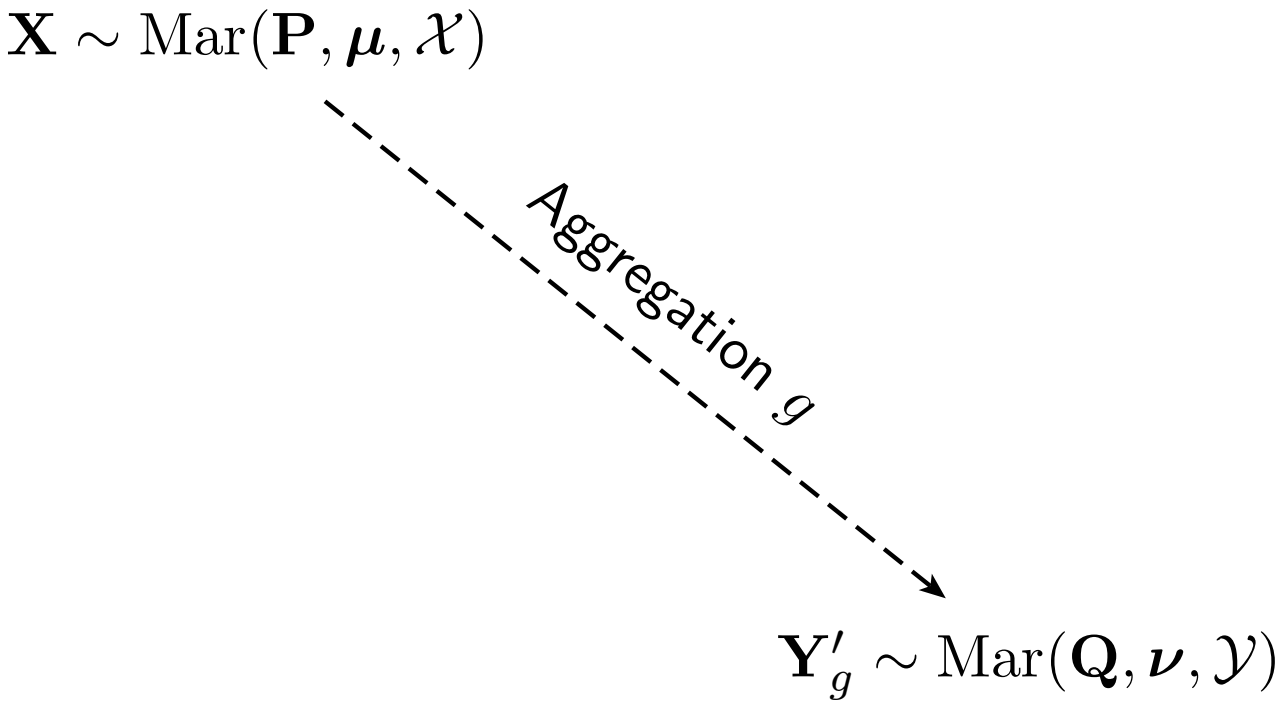


$$\mathbf{Y}'_g \sim \mathrm{Mar}(\mathbf{Q}, \boldsymbol{\nu}, \mathcal{Y})$$

Problem Statement (cont'd)



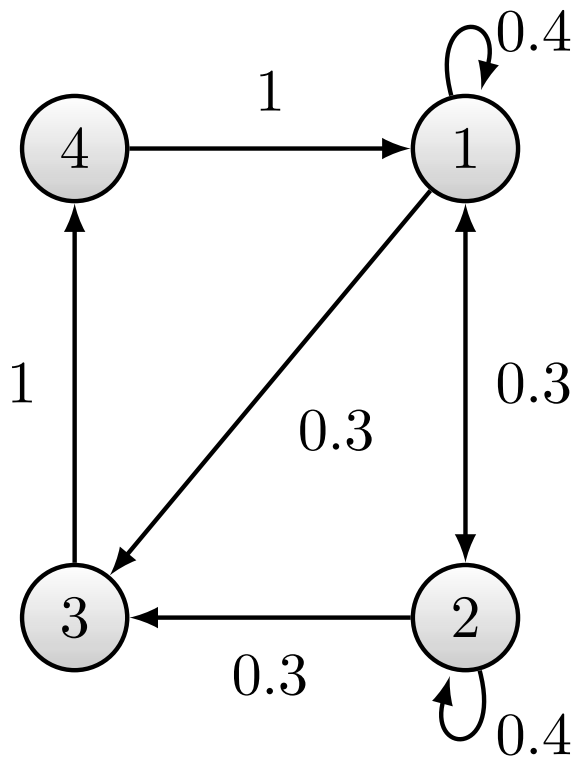
Problem Statement (cont'd)



Three Problems:

1. For given g, how to choose $\mathbf{Q}$?
2. For a given M, how to choose g?
3. (How to choose M?)

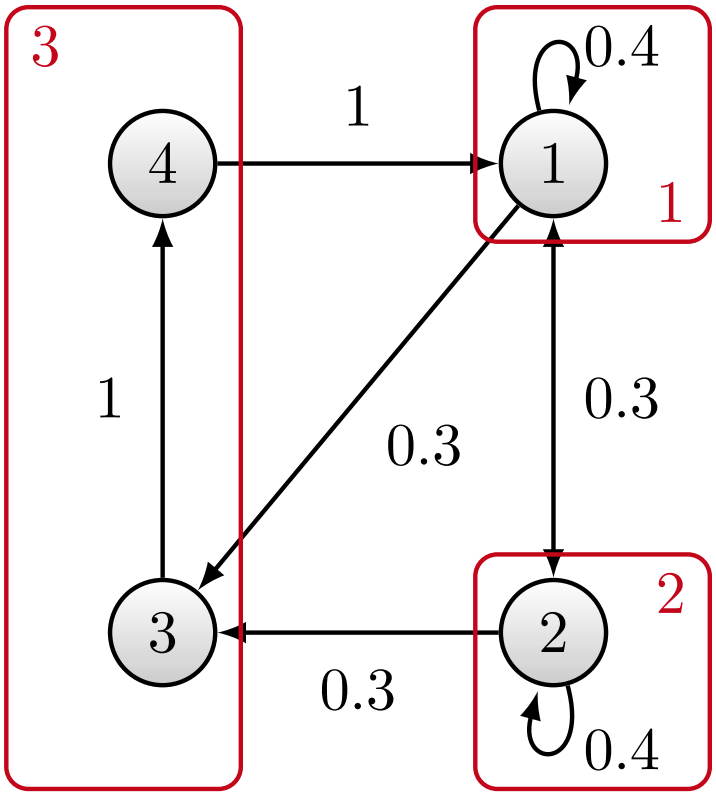
Problem 1: Choosing Q



$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$

- Let
  $X_m^n := (X_m, X_{m+1}, \dots, X_n)$
- $X_1^n = (2, 3, 4, 1, 2, 1, 3, 4, \dots)$

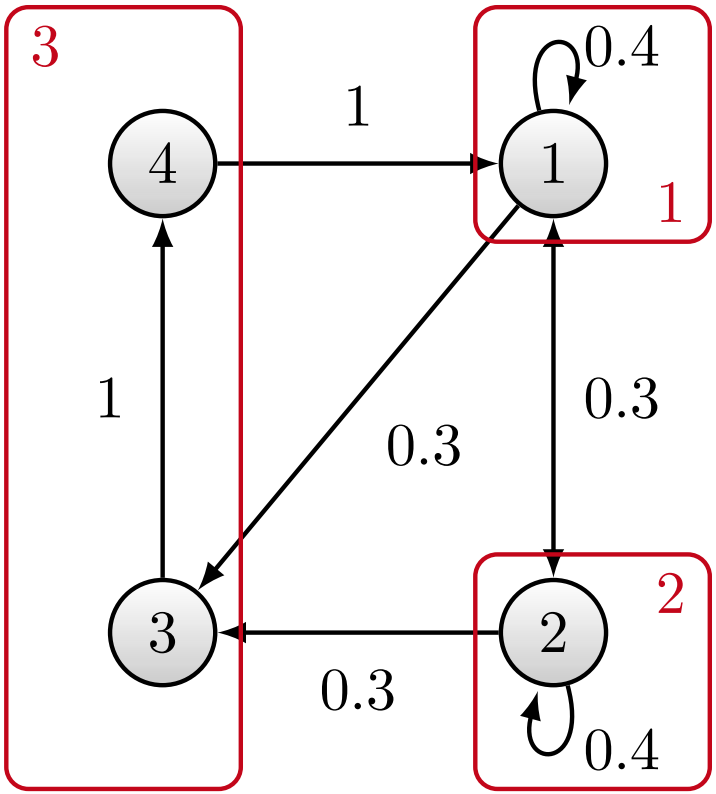
Problem 1: Choosing Q



$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$

- Let
  $X_m^n := (X_m, X_{m+1}, \dots, X_n)$
- $X_1^n = (2, 3, 4, 1, 2, 1, 3, 4, \dots)$
- $\forall k: \; Y_{g,k} := g(X_k)$
- $\mathbf{Y}_g$ is called *projection*
- $(Y_g)_1^n = (2, 3, 3, 1, 2, 1, 3, 3, \dots)$

Problem 1: Choosing Q

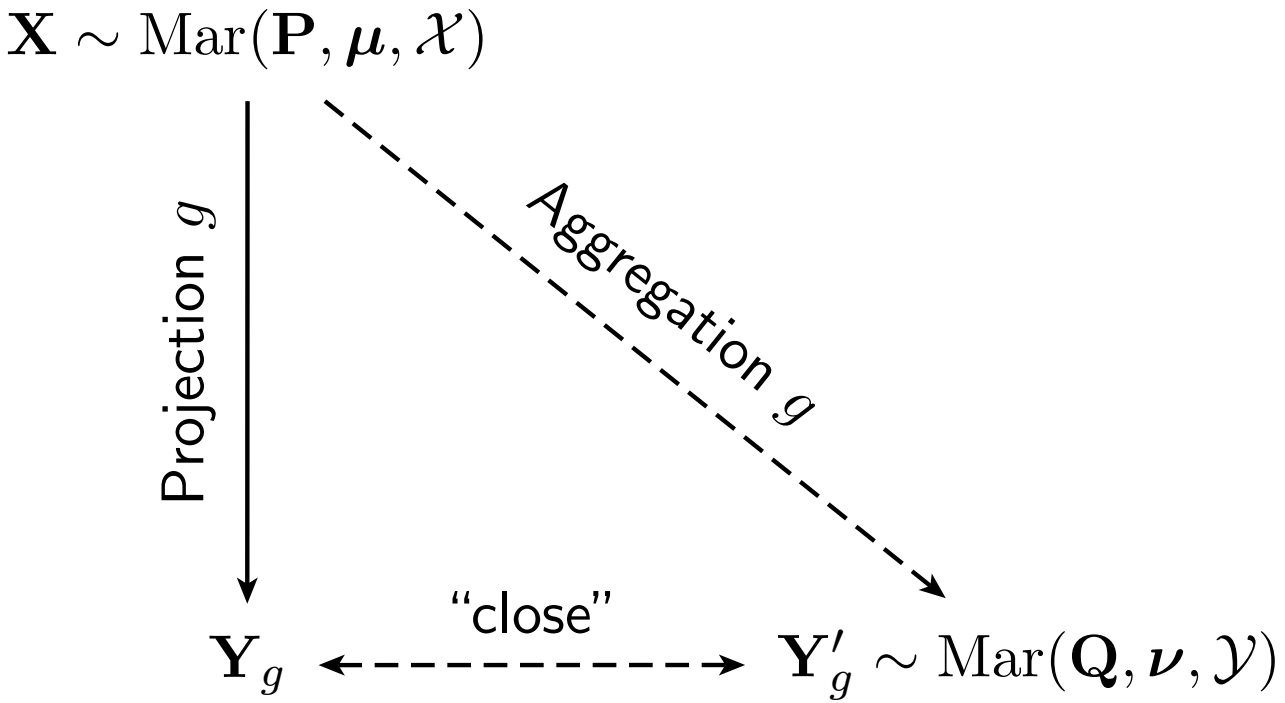


$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$

- Let $X_m^n := (X_m, X_{m+1}, \dots, X_n)$
- $X_1^n = (2, 3, 4, 1, 2, 1, 3, 4, \dots)$
- $\forall k: \; Y_{g,k} := g(X_k)$
- $\mathbf{Y}_g$ is called *projection*
- $(Y_g)_1^n = (2, 3, 3, 1, 2, 1, 3, 3, \dots)$

- $\mathbf{Y}_g$ is stationary but (in general) not Markov
- $\mathbf{Y}_g$ is Markov if $\mathbf{X}$ is *lumpable* w.r.t. g

Problem 1: Choosing $\mathbf{Q}$ (cont'd)



Problem 1: Choosing $\mathbf{Q}$ (cont'd)

Definition (Kullback-Leibler Divergence Rate [Gray, 1990])

The Kullback-Leibler divergence rate (KLDR) between two stationary processes $\mathbf{Z}$ and $\mathbf{W}$ on $\mathcal{Z}$ is

$$\bar{D}(\mathbf{Z}||\mathbf{W}) := \lim_{n\to\infty} \frac{1}{n} D(p_{Z_1^n}||p_{W_1^n})$$

provided the limit exists.

Problem 1: Choosing $\mathbf{Q}$ (cont'd)

Definition (Kullback-Leibler Divergence Rate [Gray, 1990])

The Kullback-Leibler divergence rate (KLDR) between two stationary processes $\mathbf{Z}$ and $\mathbf{W}$ on $\mathcal{Z}$ is

$$\bar{D}(\mathbf{Z}||\mathbf{W}) := \lim_{n\to\infty} \frac{1}{n} D(p_{Z_1^n}||p_{W_1^n})$$

provided the limit exists.

Lemma (Markov Chains [Rached et al., 2004])

If $\mathbf{Z} \sim \mathrm{Mar}(\mathbf{P}^{(Z)}, \boldsymbol{\mu}^{(Z)}, \mathcal{Z})$ *and* $\mathbf{W} \sim \mathrm{Mar}(\mathbf{P}^{(W)}, \boldsymbol{\mu}^{(W)}, \mathcal{Z})$, *then*

$$\bar{D}(\mathbf{Z}||\mathbf{W}) = \sum_{i,j\in\mathcal{Z}} \mu_i^{(Z)} P_{ij}^{(Z)} \log \frac{P_{ij}^{(Z)}}{P_{ij}^{(W)}}$$

Problem 1: Choosing $\mathbf{Q}$ (cont'd)

Definition (Kullback-Leibler Divergence Rate [Gray, 1990])

The Kullback-Leibler divergence rate (KLDR) between two stationary processes $\mathbf{Z}$ and $\mathbf{W}$ on $\mathcal{Z}$ is

$$\bar{D}(\mathbf{Z}||\mathbf{W}) := \lim_{n\to\infty} \frac{1}{n} D(p_{Z_1^n}||p_{W_1^n})$$

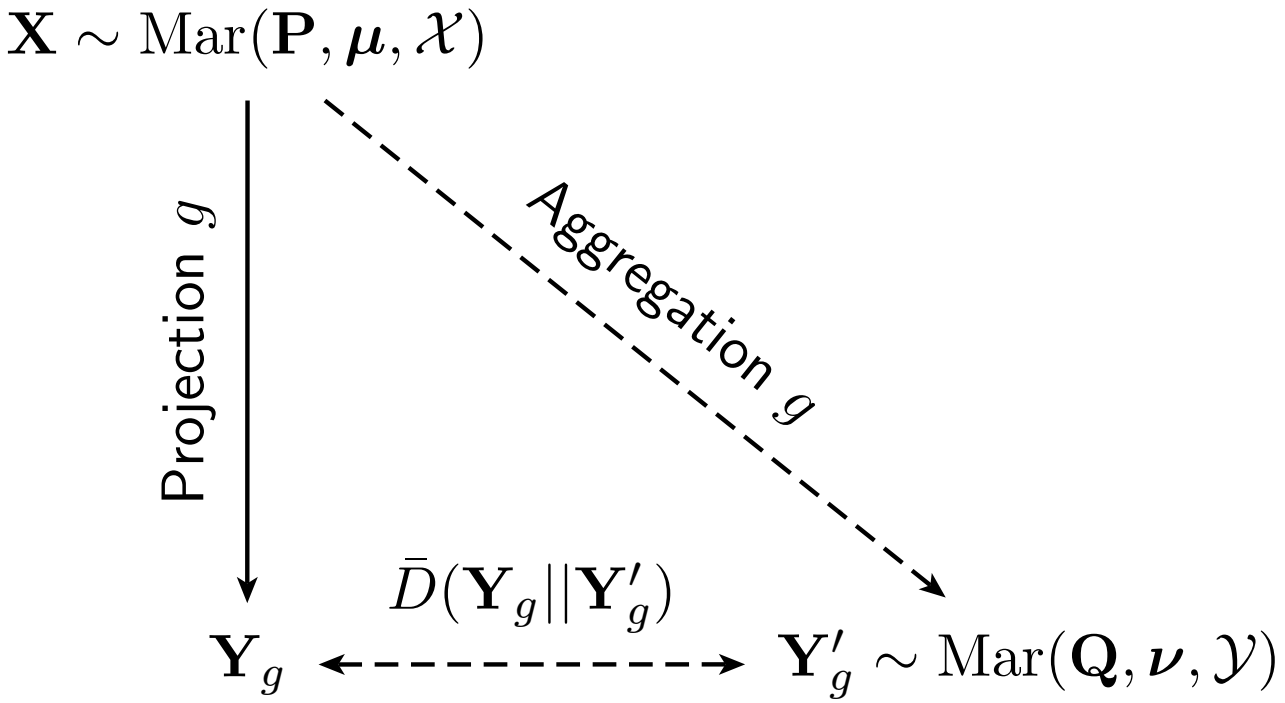
provided the limit exists.

Lemma (Process and Markov Chain [Gray, 1990, Cor. 7.4.1])

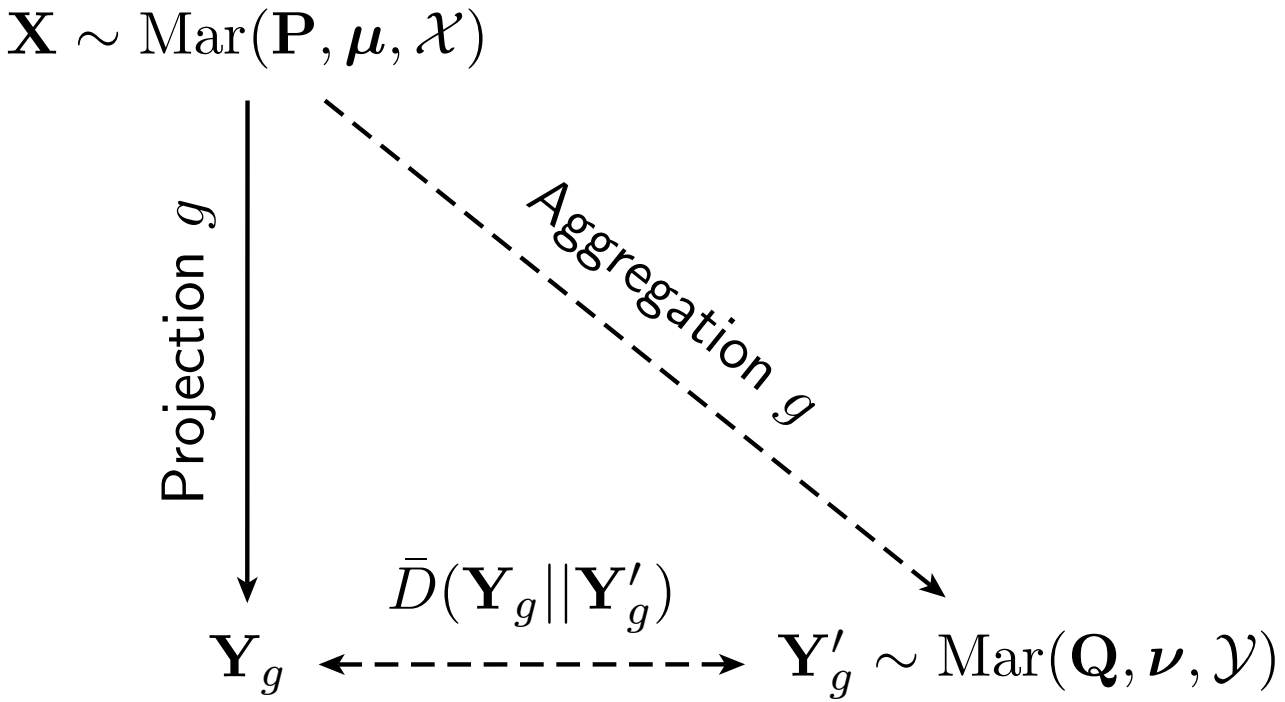
If $\mathbf{Z}$ is stationary and $\mathbf{W}$ a (dominating) Markov chain, then

$$\bar{D}(\mathbf{Z}||\mathbf{W}) = \mathbb{E}\left(D(p_{Z_2|Z_1}||p_{W_2|W_1})\right) + \lim_{n\to\infty} I(Z_{n+1}; Z_1^{n-1}|Z_n)$$

Problem 1: Choosing $\mathbf{Q}$ (cont'd)



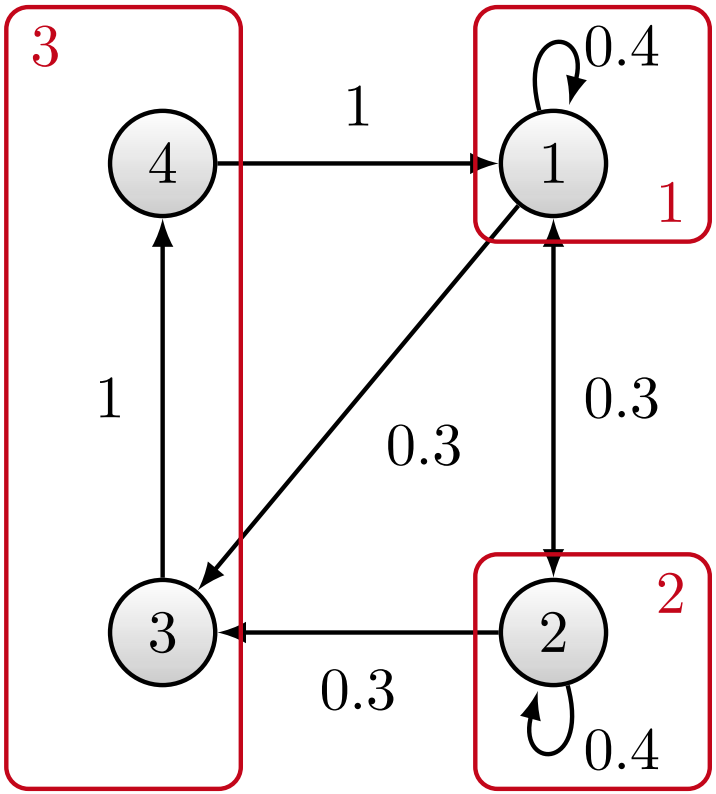
Problem 1: Choosing $\mathbf{Q}$ (cont'd)



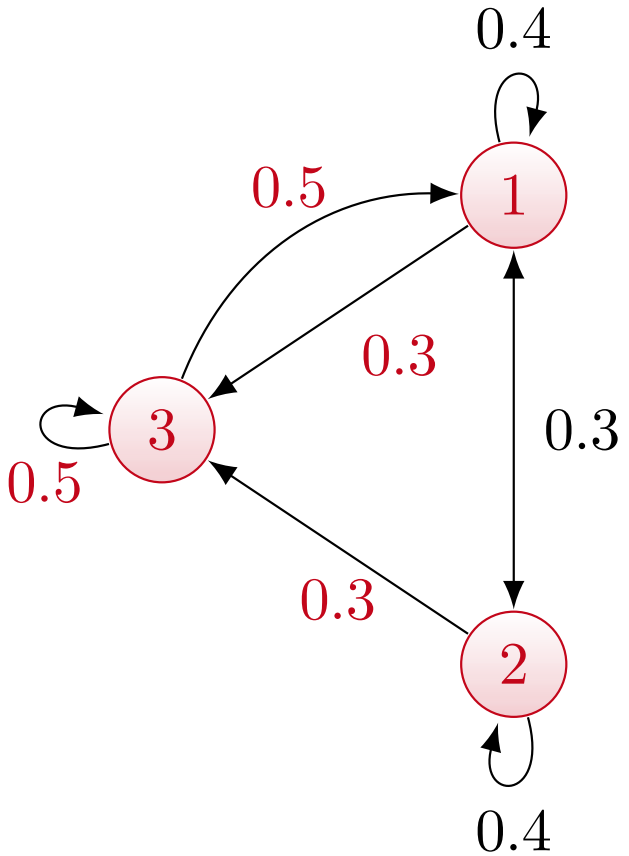
Lemma ([Gray, 1990, Cor. 7.4.2])

$Q_{ij} = p_{Y_{g,2}|Y_{g,1}}(j|i)$ *minimizes* $\bar{D}(\mathbf{Y}_g || \mathbf{Y}'_g)$

Problem 1: Choosing $\mathbf{Q}$ (cont'd)



$$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$$



$$\mathbf{Y}'_g \sim \mathrm{Mar}(\mathbf{Q}, \boldsymbol{\nu}, \mathcal{Y})$$

Problem 2: Choosing g

Related work using information-theoretic cost functions:

- Information-theoretic co-clustering [Dhillon et al., 2003]
- Information-theoretic pair-wise clustering of data given by pairwise similarities [Friedman and Goldberger, 2013, Tishby and Slonim, 2000]
- (Spectral) Aggregation of Markov chains [Deng et al., 2011, Goldberger et al., 2007, Vidyasagar, 2010] by maximizing *redundancy* $I(Y_{g,1};Y_{g,2})$.
- Aggregation of Markov chains by maximizing *lumpability* [Geiger et al., 2013]

Problem 2: Choosing g (cont'd)

New Problem Statement

Assume we have a Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$ and the cardinality M of $\mathcal{Y} = \{1, \dots, M\}$. Find

$$g^\bullet = \arg\min_g \bar{D}(\mathbf{Y}_g || \mathbf{Y}'_g).$$

Problem 2: Choosing g (cont'd)

New Problem Statement

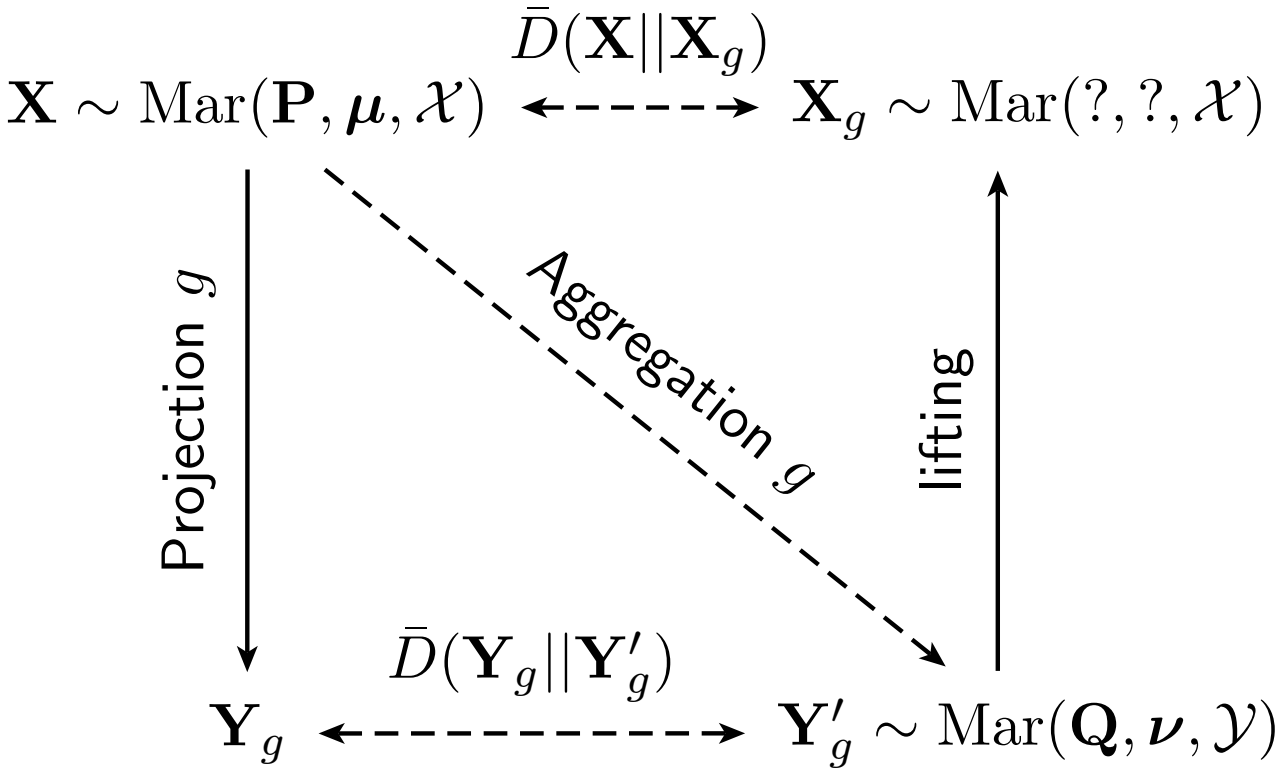
Assume we have a Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$ and the cardinality M of $\mathcal{Y} = \{1, \dots, M\}$. Find

$$g^\bullet = \arg\min_g \bar{D}(\mathbf{Y}_g || \mathbf{Y}'_g).$$

- KLDR between process and Markov chain difficult to compute (entropy rate of hidden Markov chain)
- KLDR between Markov chains easy to compute

$\Rightarrow$ "Lift" $\mathbf{Y}'_g$ back to a Markov chain $\mathbf{X}_g$ on $\mathcal{X}$

Lifting


$\bar{D}(\mathbf{X}||\mathbf{X}_g)$
$\mathbf{X} \sim \mathrm{Mar}(\mathbf{P}, \boldsymbol{\mu}, \mathcal{X})$
$\mathbf{X}_g \sim \mathrm{Mar}(?, ?, \mathcal{X})$
Projection g
Aggregation g
lifting
$\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$
$\mathbf{Y}_g$
$\mathbf{Y}'_g \sim \mathrm{Mar}(\mathbf{Q}, \boldsymbol{\nu}, \mathcal{Y})$

A little Lemma...

Lemma ([Geiger et al., 2013], based on, e.g., Gray, Pinsker,...)

Let $\mathbf{X}$ and $\mathbf{X}_g$ be Markov chains on the same state space, and let $\mathbf{Y}_g$ be the projection of $\mathbf{X}$ and $\mathbf{Y}'_g$ be the projection of $\mathbf{X}_g$. Let $\mathbf{Y}'_g$ be a Markov chain, i.e., let $\mathbf{X}_g$ be lumpable w.r.t. g. Then,

$$\bar{D}(\mathbf{X}||\mathbf{X}_g) \geq \bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g).$$

A little Lemma...

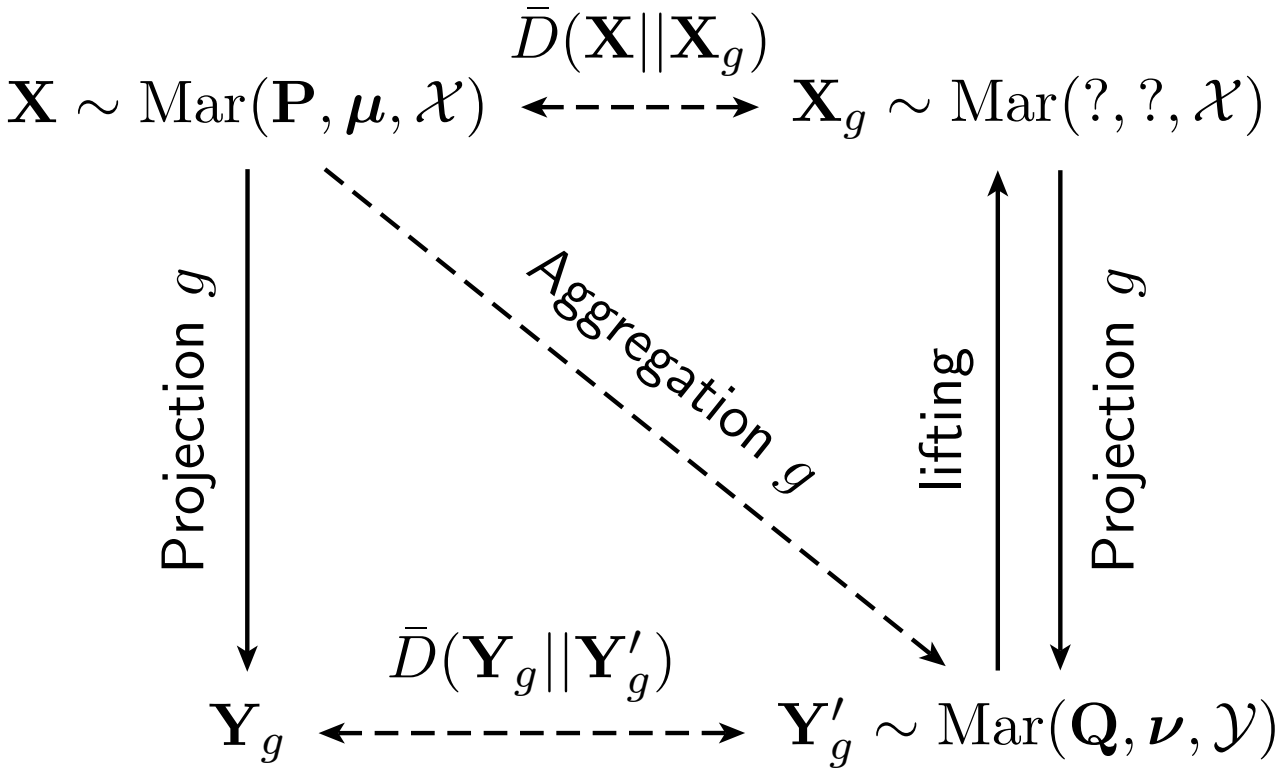
Lemma ([Geiger et al., 2013], based on, e.g., Gray, Pinsker,...)

Let $\mathbf{X}$ and $\mathbf{X}_g$ be Markov chains on the same state space, and let $\mathbf{Y}_g$ be the projection of $\mathbf{X}$ and $\mathbf{Y}'_g$ be the projection of $\mathbf{X}_g$. Let $\mathbf{Y}'_g$ be a Markov chain, i.e., let $\mathbf{X}_g$ be lumpable w.r.t. g. Then,

$$\bar{D}(\mathbf{X}||\mathbf{X}_g) \geq \bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g).$$

- Solve problem sub-optimally by *minimizing an upper bound*
- Requires that $\mathbf{X}_g$ is *lumpable w.r.t.* g

A little Lemma... (cont'd)

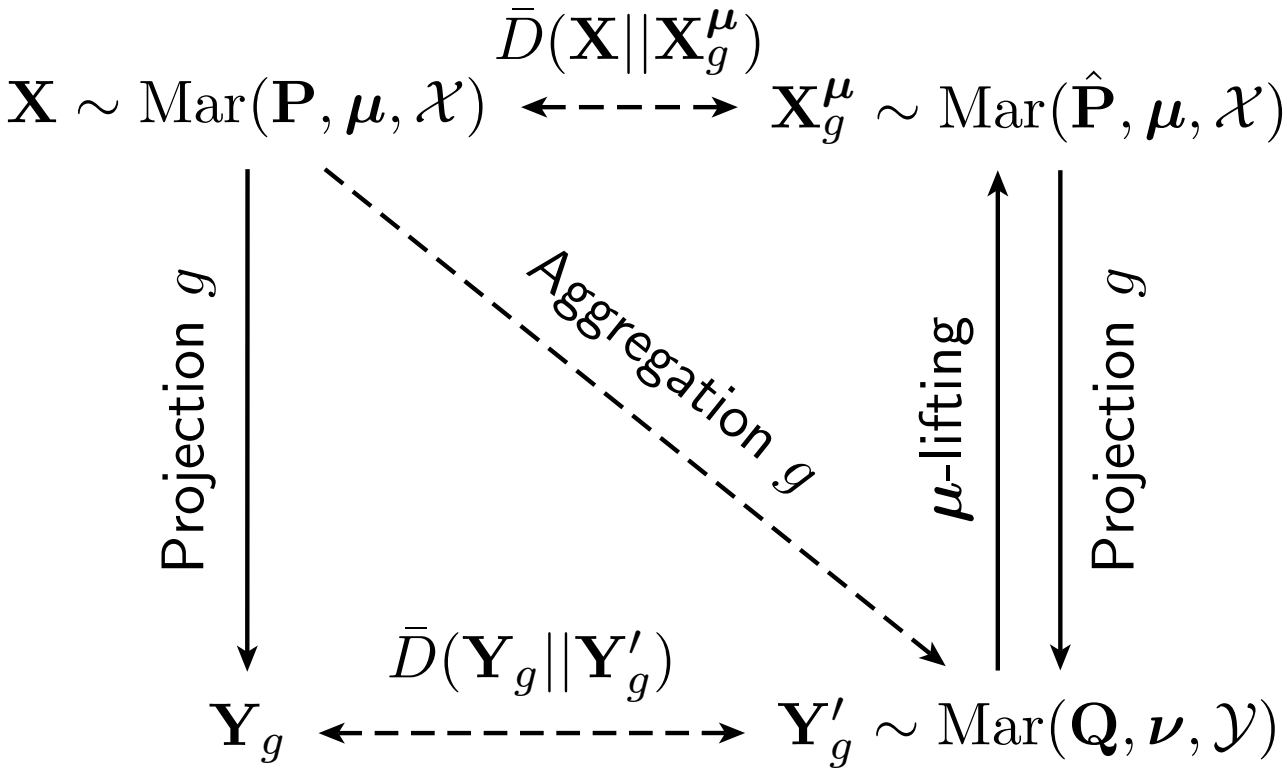


μ-Lifting

[Deng et al., 2011] proposed the following μ-lifting to get the transition matrix $\hat{\mathbf{P}}$ of $\mathbf{X}_g^{\boldsymbol{\mu}}$:

$$\hat{P}_{ij} := \frac{\mu_j}{\sum_{k \in g^{-1}(g(j))} \mu_k} Q_{g(i)g(j)}$$

μ-Lifting (cont'd)



μ-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)

μ-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$

μ-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$
- Stationary distribution of $\mathbf{X}_g^{\boldsymbol{\mu}}$ is $\boldsymbol{\mu}$

μ-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g || \mathbf{Y}'_g) \leq \bar{D}(\mathbf{X} || \mathbf{X}_g^{\boldsymbol{\mu}})$
- Stationary distribution of $\mathbf{X}_g^{\boldsymbol{\mu}}$ is $\boldsymbol{\mu}$
- $\bar{D}(\mathbf{X} || \mathbf{X}_g^{\boldsymbol{\mu}}) = I(X_1; X_2) - I(Y_{g,1}; Y_{g,2})$

μ-Lifting (cont'd)

Properties

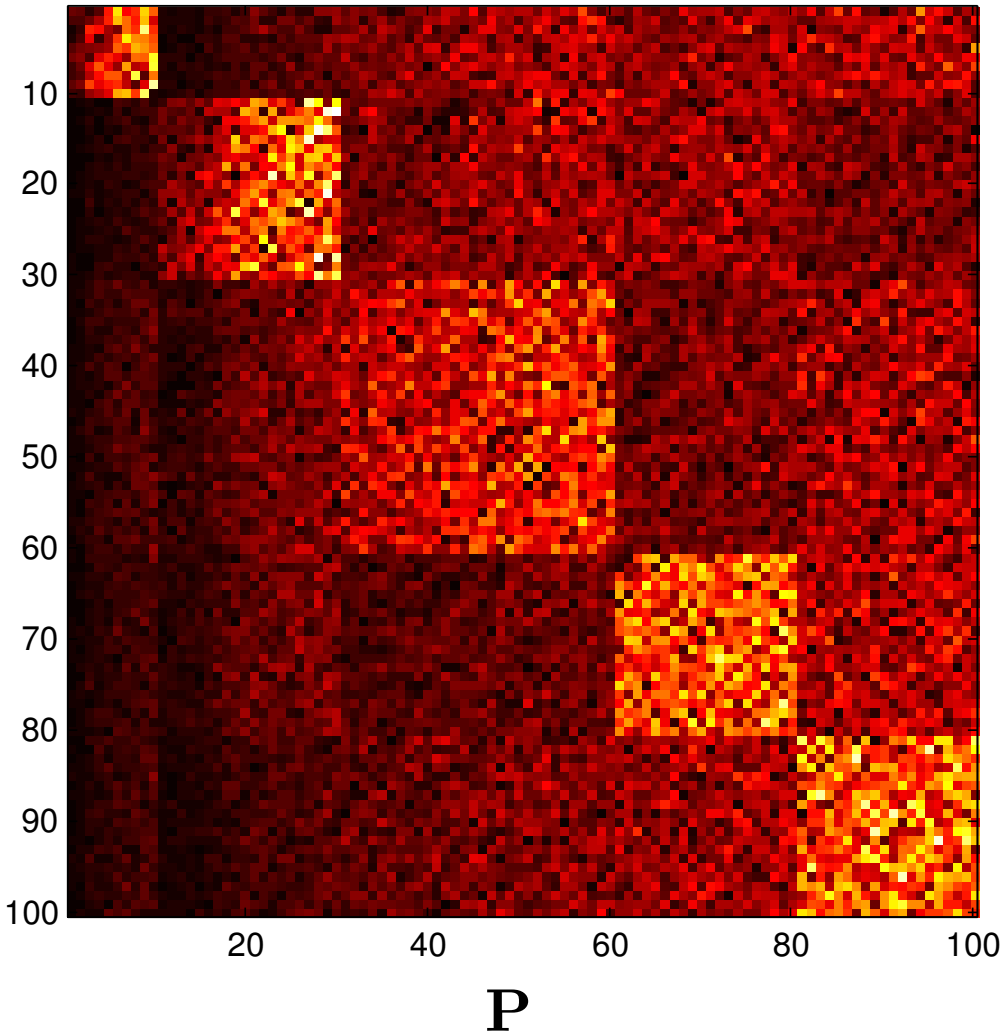
- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$
- Stationary distribution of $\mathbf{X}_g^{\boldsymbol{\mu}}$ is $\boldsymbol{\mu}$
- $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}}) = I(X_1;X_2) - I(Y_{g,1};Y_{g,2})$
- Bi-partition problem connected to spectral graph theory *under some conditions on eigenvalues*

μ-Lifting (cont'd)

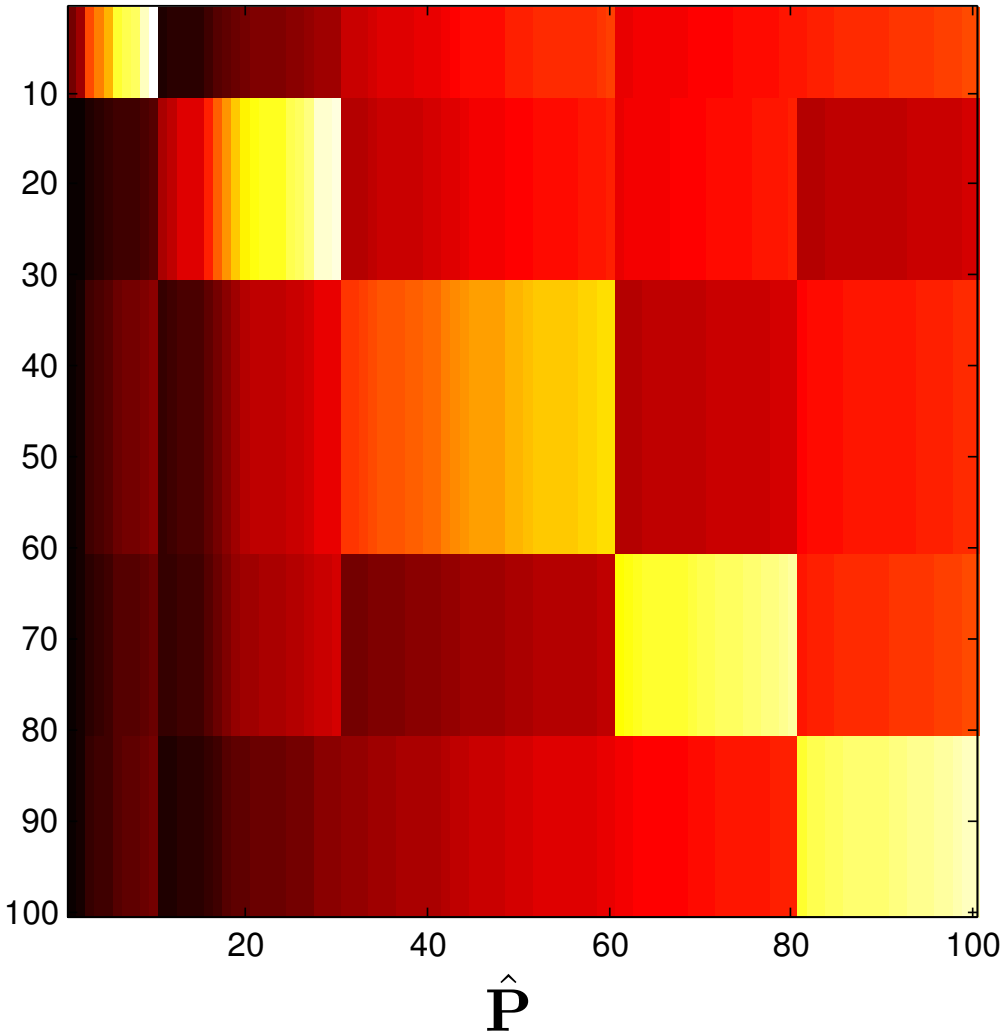
Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\boldsymbol{\mu}})$ (i.e., $\mathbf{X}_g^{\boldsymbol{\mu}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$
- Stationary distribution of $\mathbf{X}_g^{\boldsymbol{\mu}}$ is $\boldsymbol{\mu}$
- $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}}) = I(X_1;X_2) - I(Y_{g,1};Y_{g,2})$
- Bi-partition problem connected to spectral graph theory *under some conditions on eigenvalues*
- Structural properties of $\mathbf{P}$ not preserved in $\hat{\mathbf{P}}$ (e.g., impossible transitions become possible)

μ-Lifting (cont'd)



μ-Lifting (cont'd)

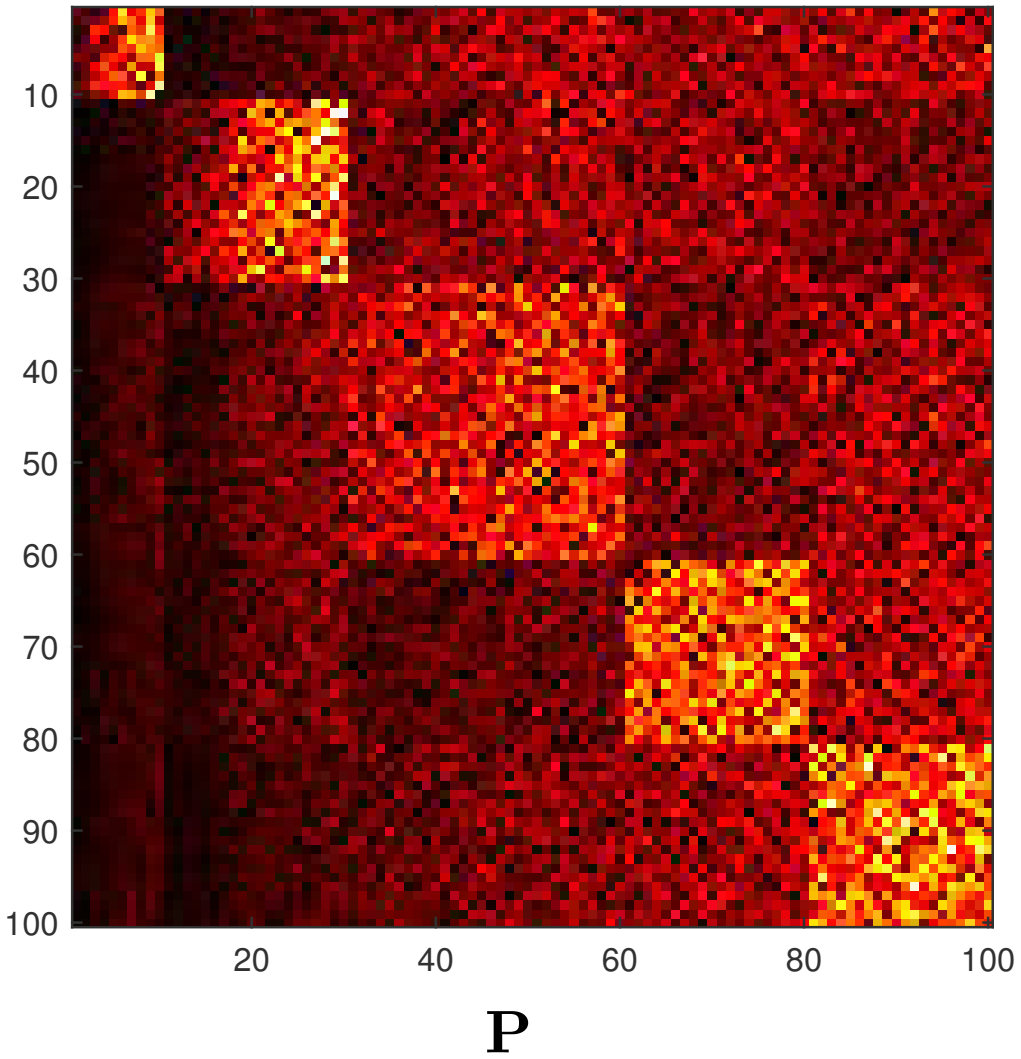


$\mathbf{P}$-Lifting

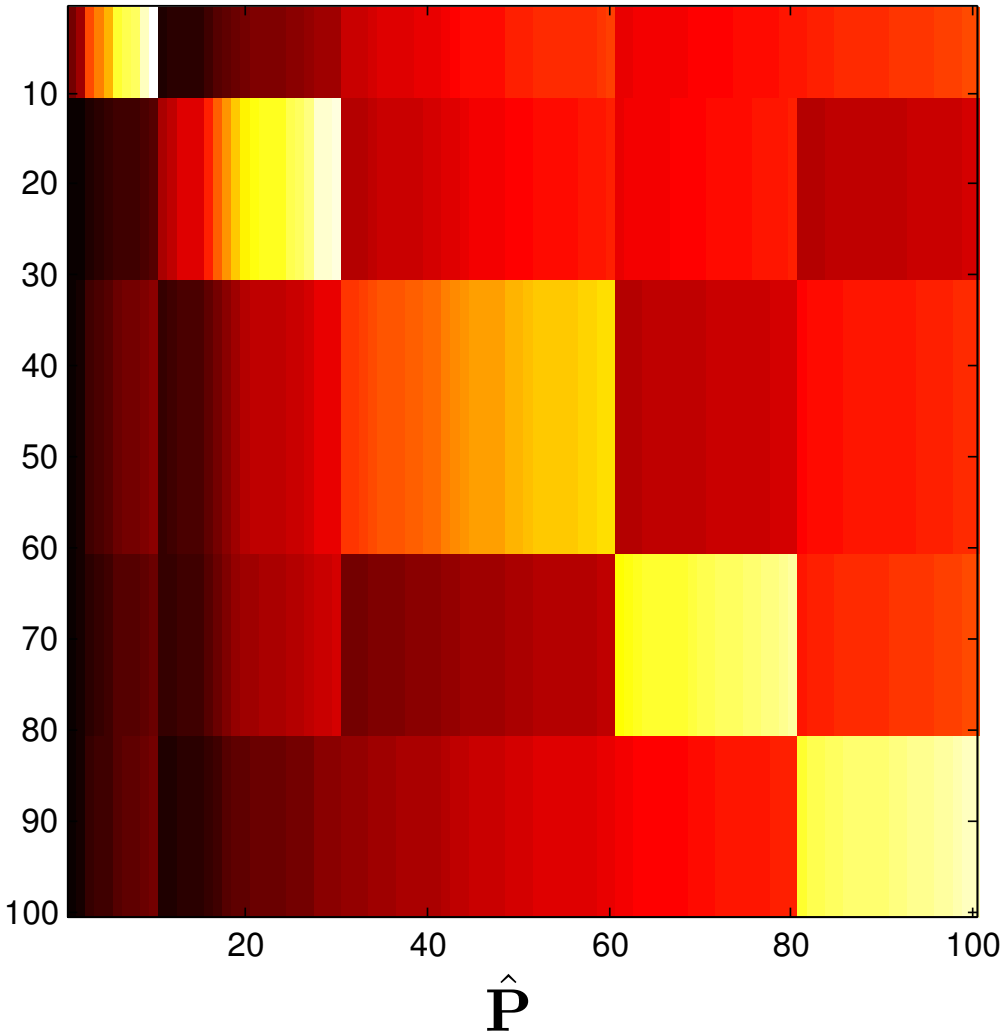
In [Geiger et al., 2013], we proposed an alternative lifting method, using the transition matrix instead of the stationary distribution to get the transition matrix $\tilde{\mathbf{P}}$ of $\mathbf{X}_g^{\mathbf{P}}$:

$$\tilde{P}_{ij} := \begin{cases} \frac{P_{ij}}{\sum_{k \in g^{-1}(g(j))} P_{ik}} Q_{g(i)g(j)}, & \text{if } \sum_{k \in g^{-1}(g(j))} P_{ik} > 0 \\ \frac{1}{\operatorname{card}(g^{-1}(g(j)))} Q_{g(i)g(j)}, & \text{if } \sum_{k \in g^{-1}(g(j))} P_{ik} = 0 \end{cases}$$

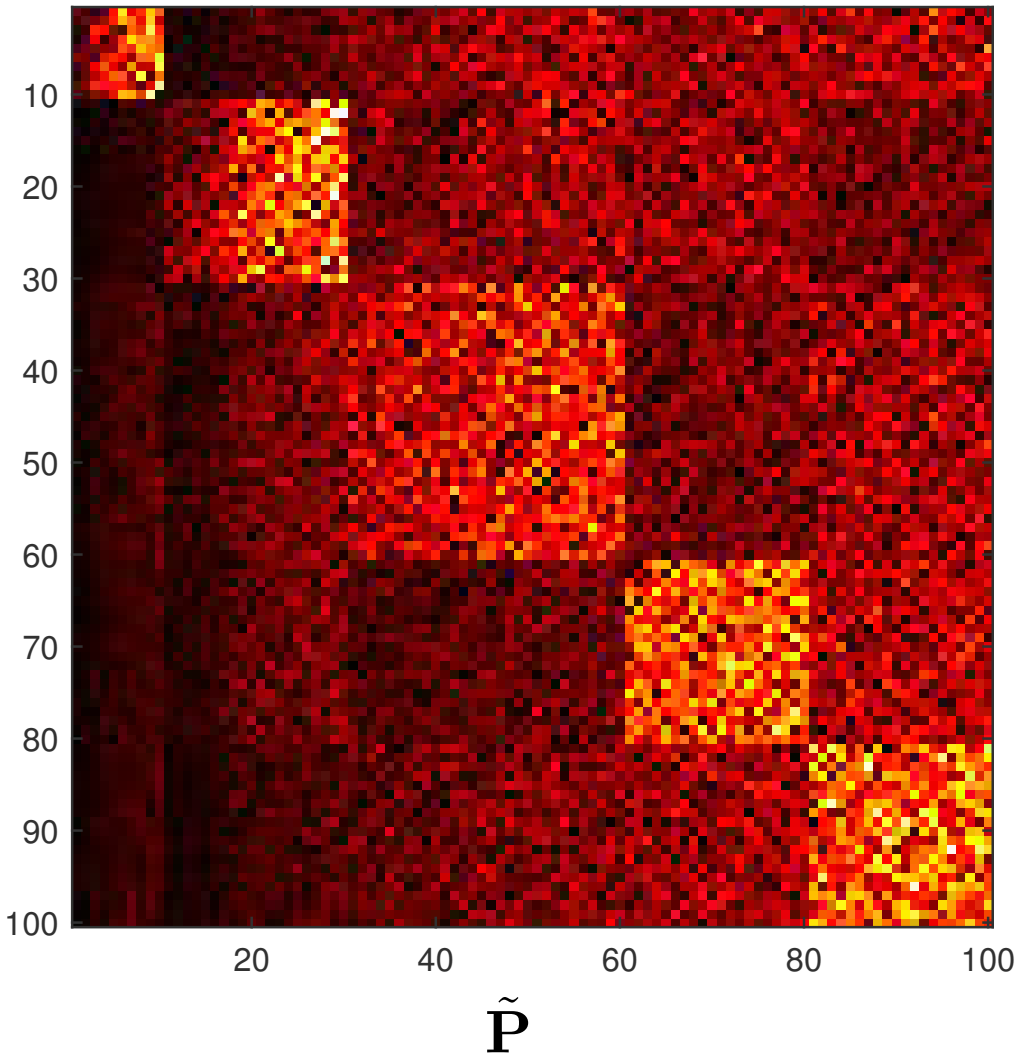
$\mathbf{P}$-Lifting (cont'd)



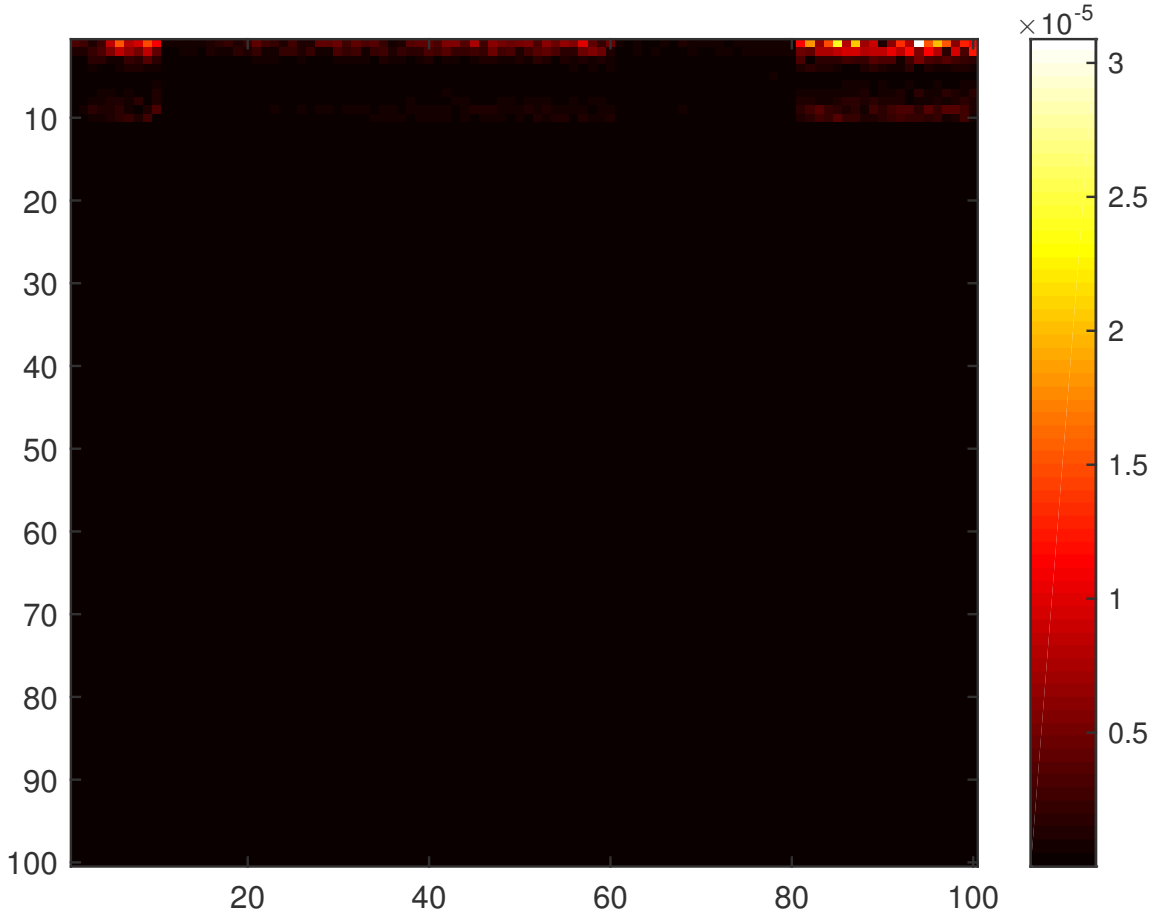
$\mathbf{P}$-Lifting (cont'd)



$\mathbf{P}$-Lifting (cont'd)



$\mathbf{P}$-Lifting (cont'd)



$$|\mathbf{P} - \tilde{\mathbf{P}}|$$

Lumpability

Lemma ([Geiger and Temmel, 2012])

Let $\mathbf{X}$ and be a Markov chain and let $\mathbf{Y}_g$ be the projection of $\mathbf{X}$. $\mathbf{Y}_g$ is a Markov chain, i.e., $\mathbf{X}$ is lumpable w.r.t. g iff

$$H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1) = 0.$$

Lumpability

Lemma ([Geiger and Temmel, 2012])

Let $\mathbf{X}$ and be a Markov chain and let $\mathbf{Y}_g$ be the projection of $\mathbf{X}$. $\mathbf{Y}_g$ is a Markov chain, i.e., $\mathbf{X}$ is lumpable w.r.t. g iff

$$H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1) = 0.$$

This lemma is closely related to first-order bounds ($n = 1$) on the entropy rate of a projection of a Markov chain:

Lemma ([Cover and Thomas, 2006, Birch, 1962])

Let $\mathbf{X}$ and be a Markov chain and let $\mathbf{Y}_g$ be the projection of $\mathbf{X}$. The entropy rate $\bar{H}(\mathbf{Y}_g)$ of $\mathbf{Y}_g$ satisfies, for all $n \geq 1$,

$$H(Y_{g,n+1}|(Y_g)_2^n, X_1) \leq \bar{H}(\mathbf{Y}_g) \leq H(Y_{g,n+1}|(Y_g)_1^n)$$

$\mathbf{P}$-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\mathbf{P}})$ (i.e., $\mathbf{X}_g^{\mathbf{P}}$ is strongly lumpable w.r.t. g)

$\mathbf{P}$-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\mathbf{P}})$ (i.e., $\mathbf{X}_g^{\mathbf{P}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}}) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$

$\mathbf{P}$-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}_g^{\mathbf{P}})$ (i.e., $\mathbf{X}_g^{\mathbf{P}}$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}}) \leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$
- (in some sense) best upper bound on $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$

$\mathbf{P}$-Lifting (cont'd)

Properties

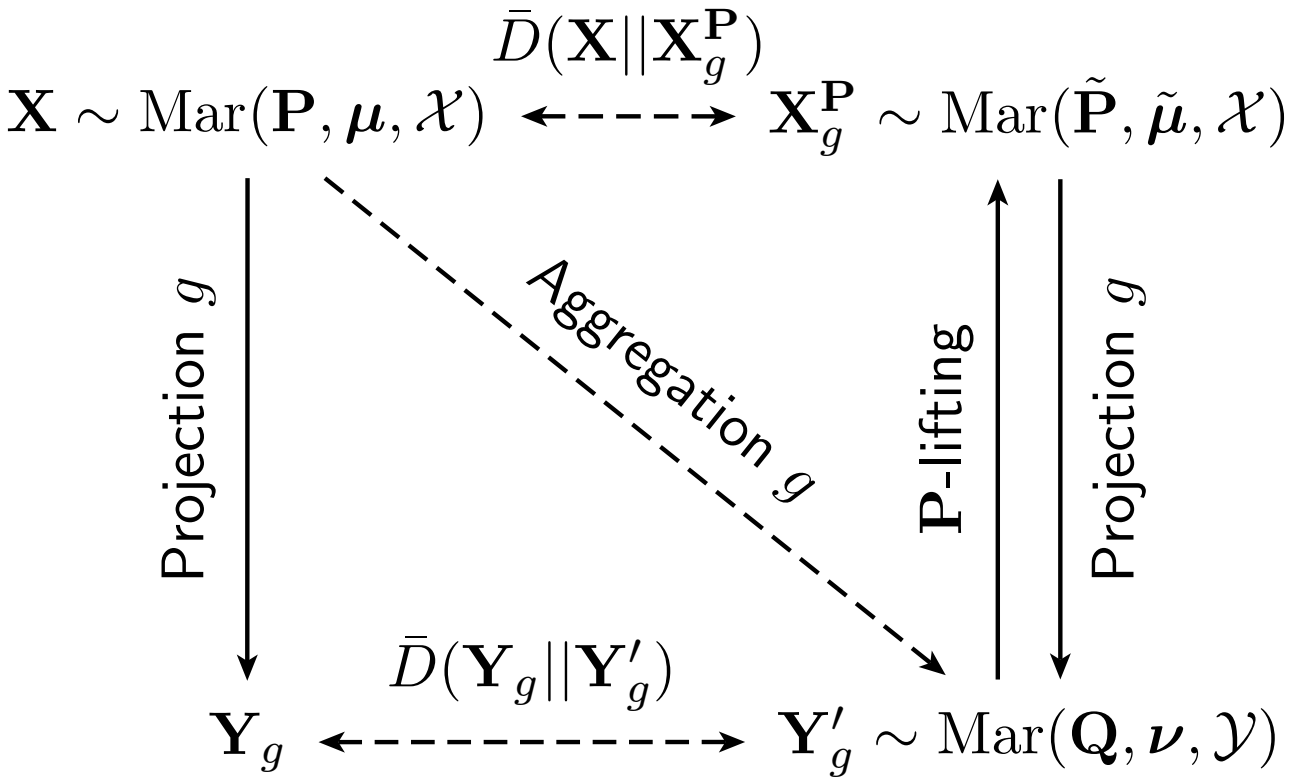
- $\mathbf{Y}'_g = g(\mathbf{X}^{\mathbf{P}}_g)$ (i.e., $\mathbf{X}^{\mathbf{P}}_g$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}^{\boldsymbol{\mu}}_g)$
- (in some sense) best upper bound on $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$
- $\bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g) = H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1)$
  Choose g to make $\mathbf{Y}_g$ "as Markov as possible"

$\mathbf{P}$-Lifting (cont'd)

Properties

- $\mathbf{Y}'_g = g(\mathbf{X}^{\mathbf{P}}_g)$ (i.e., $\mathbf{X}^{\mathbf{P}}_g$ is strongly lumpable w.r.t. g)
- $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g) \leq \bar{D}(\mathbf{X}||\mathbf{X}^{\boldsymbol{\mu}}_g)$
- (in some sense) best upper bound on $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g)$
- $\bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g) = H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1)$
  Choose g to make $\mathbf{Y}_g$ "as Markov as possible"
- tightness, i.e., $\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) = 0 \Leftrightarrow \bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g) = 0$

$\mathbf{P}$-Lifting (cont'd)



The Information Bottleneck

Introduced in [Tishby et al., 1999]; rate-distortion theory with Kullback-Leibler divergence as distortion function as an approach to clustering:

S	relevant information (class labels)
X	observation (features)
Y	compressed observation (clusters of features)

$$\arg\min_{g} I(X;Y) - \beta I(S;Y)$$

β trades compression vs. preservation of relevant information

Relaxing the Problem (again)

$$\begin{aligned}\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) &\leq \bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g)\\ &= H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1)\\ &= I(Y_{g,2};X_1) - I(Y_{g,2};Y_{g,1})\\ &\leq I(X_2;X_1) - I(X_2;Y_{g,1})\end{aligned}$$

Relaxing the Problem (again)

$$\begin{aligned}\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) &\leq \bar{D}(\mathbf{X}||\mathbf{X}^{\mathbf{P}}_g)\\ &= H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1)\\ &= I(Y_{g,2};X_1) - I(Y_{g,2};Y_{g,1})\\ &\leq I(X_2;X_1) - I(X_2;Y_{g,1})\end{aligned}$$

For $\beta \to \infty$:

$$\begin{aligned}&\arg\min_g I(X_2;X_1) - I(X_2;Y_{g,1})\\ &\qquad = \arg\min_g I(X_1;Y_{g,1}) - \beta I(X_2;Y_{g,1})\end{aligned}$$

Relaxing the Problem (again)

$$\begin{aligned}\bar{D}(\mathbf{Y}_g||\mathbf{Y}'_g) &\leq \bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}}) \\ &= H(Y_{g,2}|Y_{g,1}) - H(Y_{g,2}|X_1) \\ &= I(Y_{g,2};X_1) - I(Y_{g,2};Y_{g,1}) \\ &\leq I(X_2;X_1) - I(X_2;Y_{g,1})\end{aligned}$$

For $\beta \to \infty$:

$$\begin{aligned}&\arg\min_g I(X_2;X_1) - I(X_2;Y_{g,1}) \\ &\qquad = \arg\min_g I(X_1;Y_{g,1}) - \beta I(X_2;Y_{g,1})\end{aligned}$$

Markov aggregation via information bottleneck!

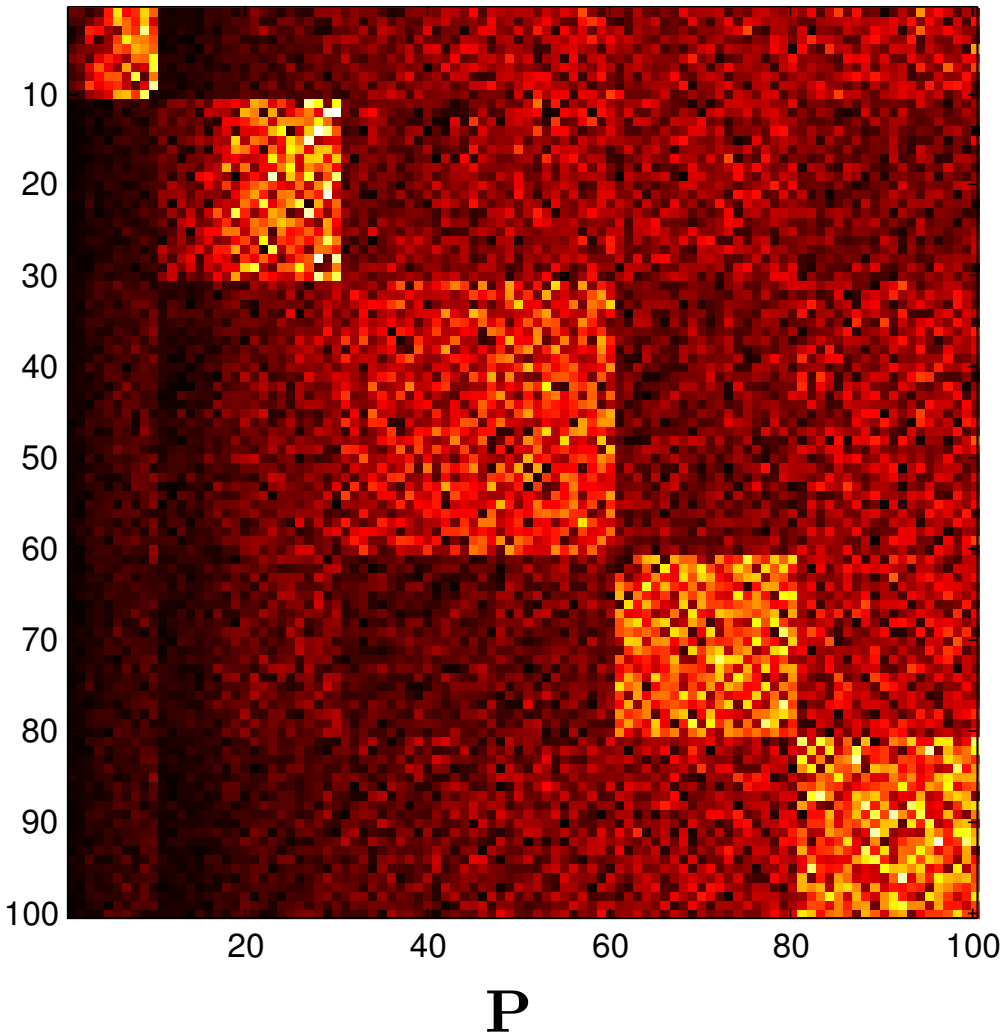
Nearly Completely Decomposable Markov Chains (NCDMCs)

Transition matrix of a NCDMC:

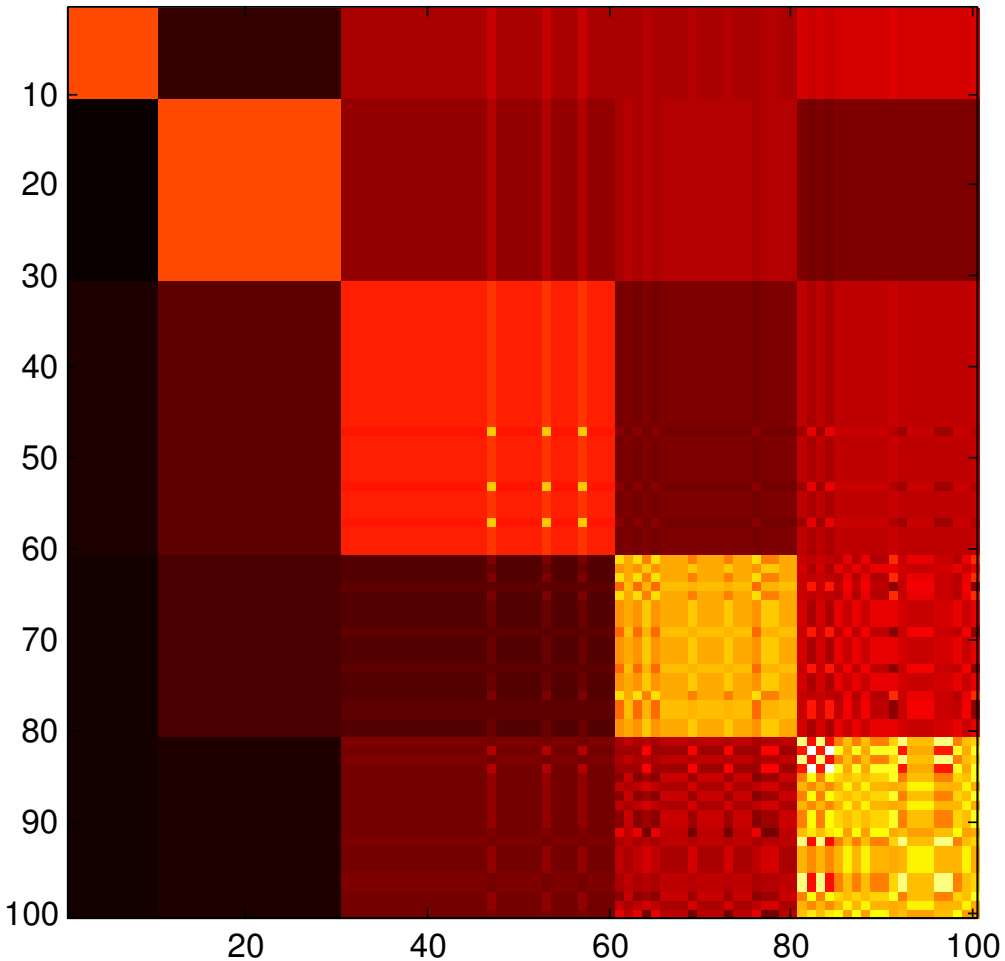
$$\mathbf{P} = (1-\varepsilon)\begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_L \end{bmatrix} + \varepsilon\mathbf{E}$$

- For $M = L$, the aggregation should reveal structure of block-diagonal matrix
- Aggregation via agglomerative information bottleneck (AIB) [Slonim and Tishby, 1999]: iteratively merging states until cardinality is M

Aggregation of NCDMCs

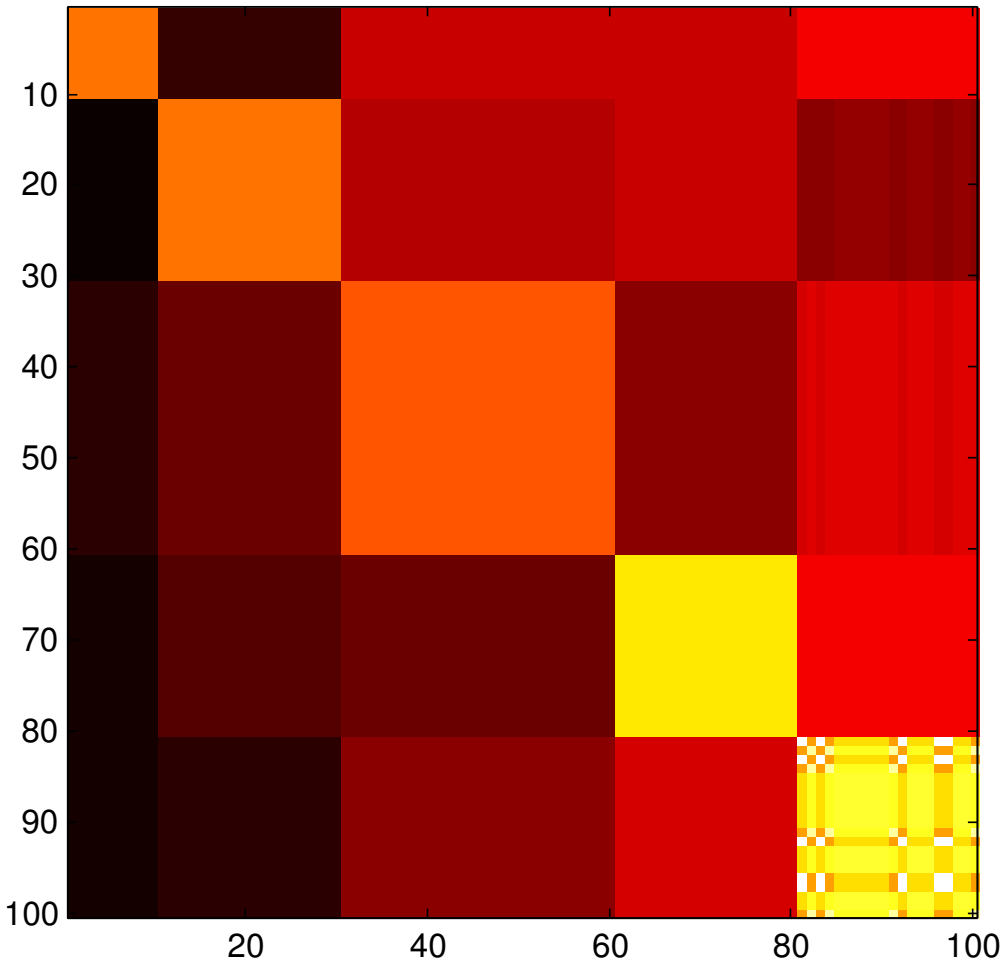


Aggregation of NCDMCs



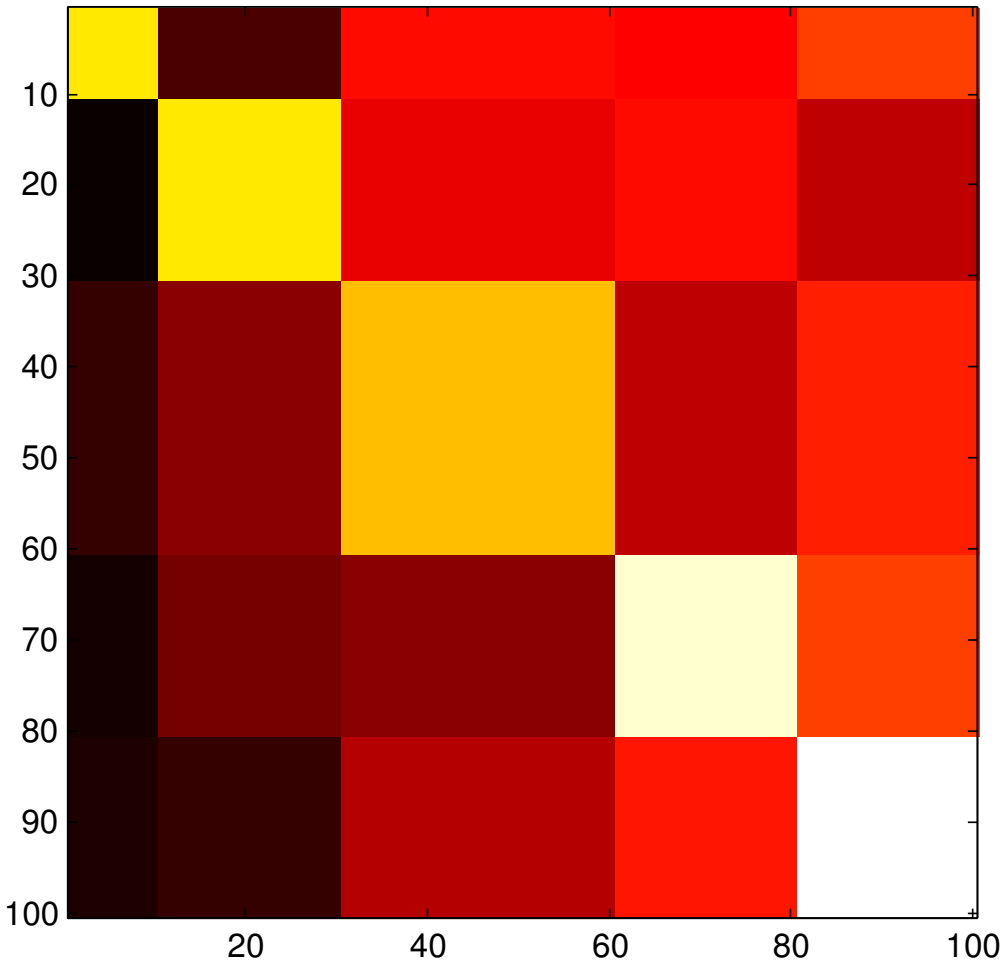
Partition for $M = 12$

Aggregation of NCDMCs



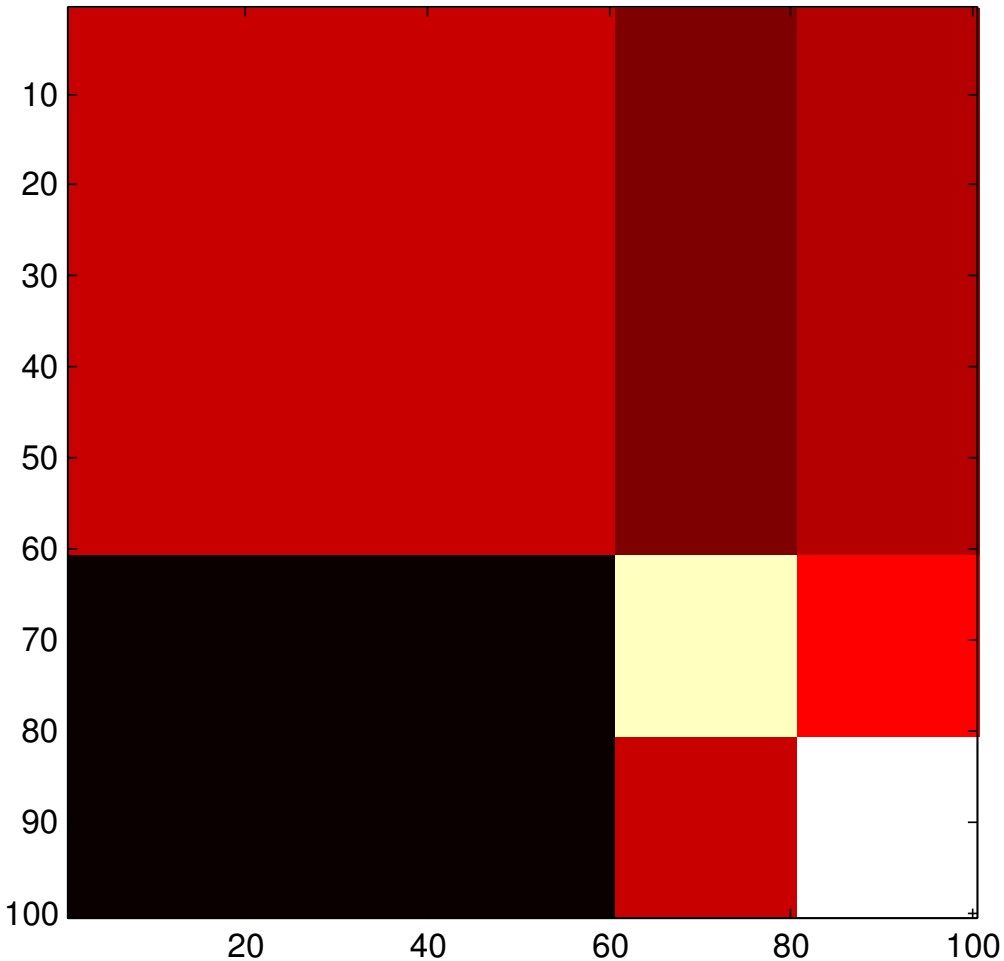
Partition for $M = 7$

Aggregation of NCDMCs



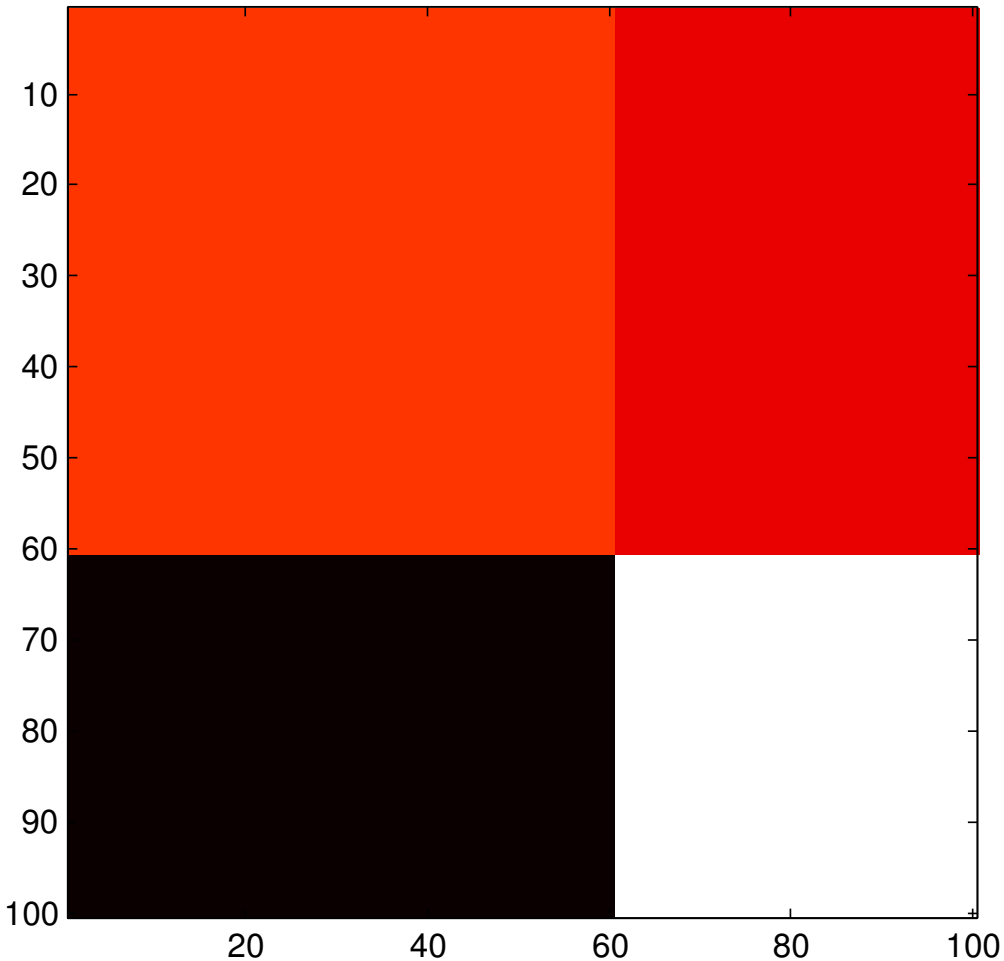
Partition for $M = 5$

Aggregation of NCDMCs



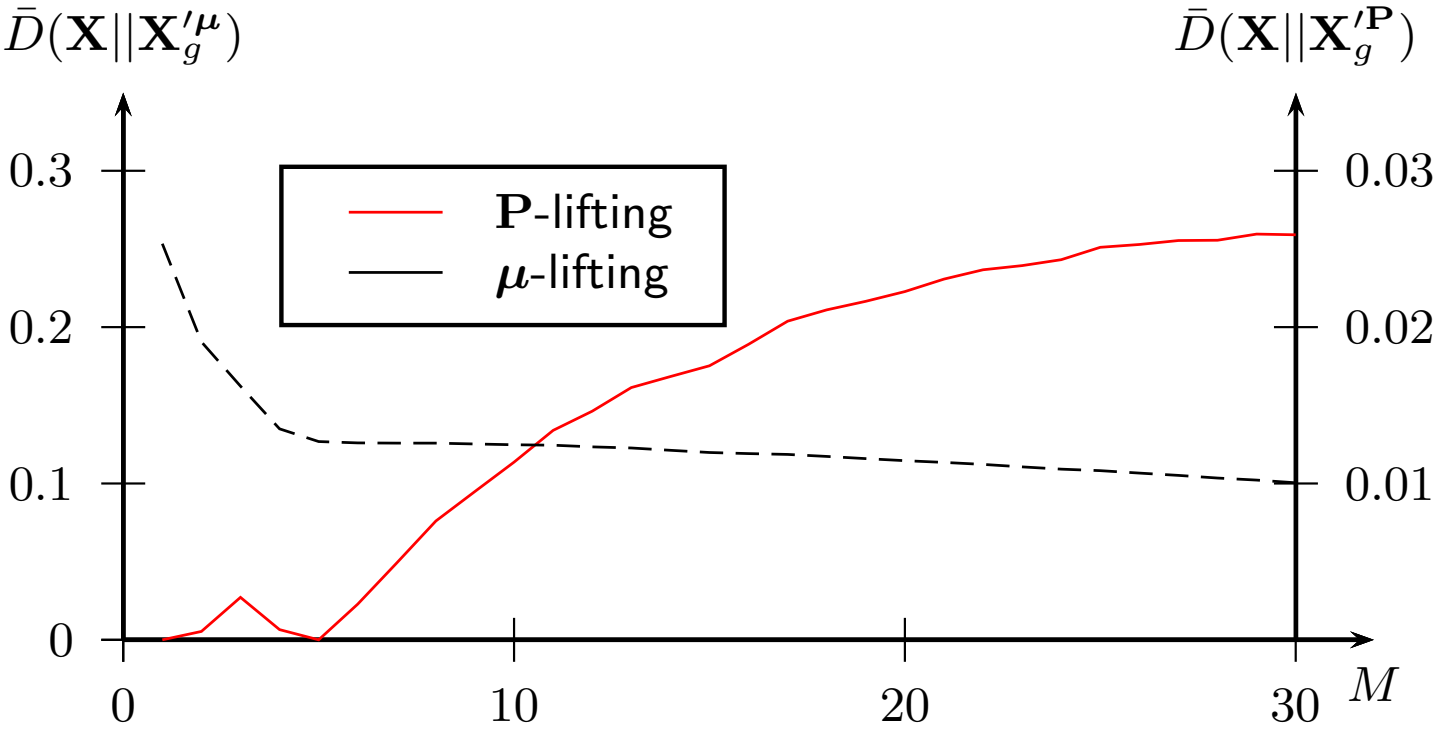
Partition for $M = 3$

Aggregation of NCDMCs



Partition for $M = 2$

How to choose M?



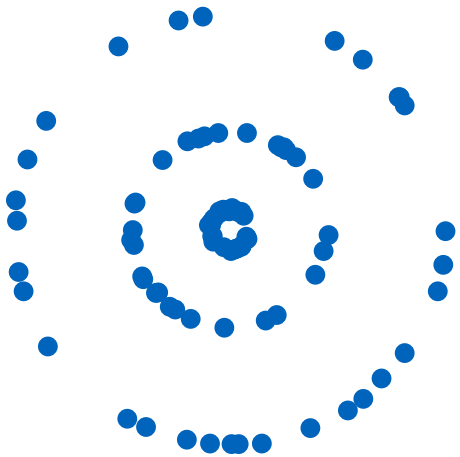
- (Loosely) Related to Markov model order selection
- $\boldsymbol{\mu}$-lifting: Change of slope indicates meaningful partition
- $\mathbf{P}$-lifting: Local minimum indicates meaningful partition

Natural Language Processing

- Letter bi-gram model of "Quo Vadis" by Henryk Sienkiewicz (Markov model of co-occurrence of letters)
- Reduced-size alphabet: space + 26 letters
- Aggregation via AIB

M	Partition
2	[,a,e,i,o,u],[b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z]
3	[ ],[a,e,i,o,u],[b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z]
4	[ ],[a,e,i,o,u],[b,c,h,j,k,l,m,n,p,q,r,v,w,z],[d,f,g,s,t,x,y]
7	[ ],[a,i,o,u],[e],[b,c,h,j,p,q,v,w,z],[d,f,g,s,x,y],[k,l,m,n,r],[t]
12	[ ],[a],[e],[i,u],[o],[b,j,p,q],[c,w],[d,f,g,s,x,y],[h,v,z],[k,l,m,r],[n],[t]

Pairwise Clustering (Three-Circles Data)



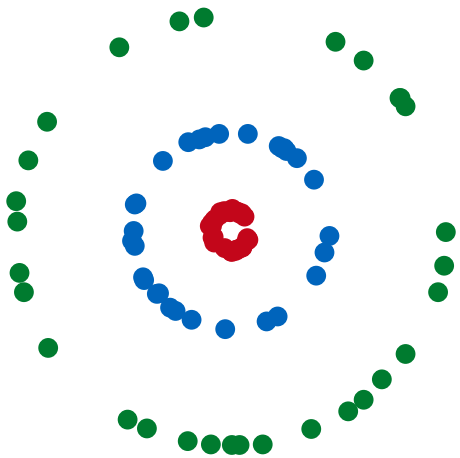
Pairwise distance matrix converted to transition probability matrix

$$P_{i,j} = \frac{e^{-0.05||x_i - x_j||^2}}{\sum_{j \in \mathcal{X}} e^{-0.05||x_i - x_j||^2}}$$

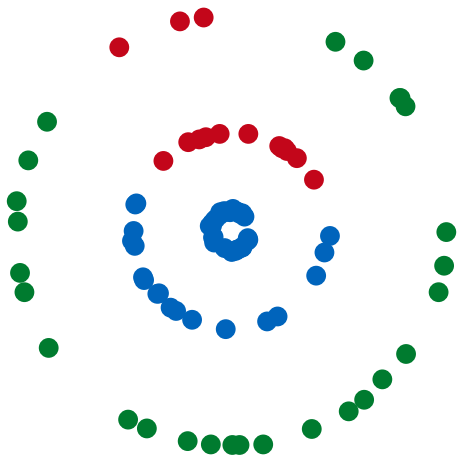
Three agglomerative algorithms:

- minimize $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}})$ (Lump)
- maximize $I(X_2; Y_{g,1})$ (AIB)
- minimize $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\boldsymbol{\mu}})$ (MI)

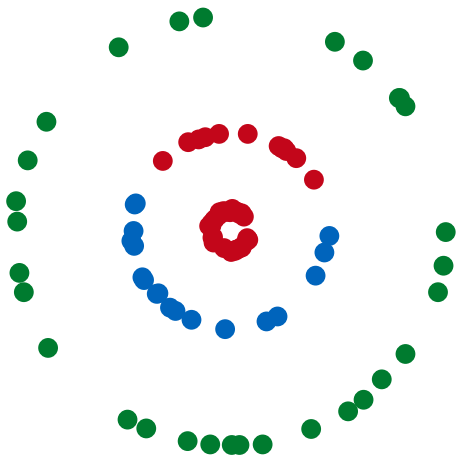
Pairwise Clustering (Three-Circles Data; cont'd)



Lump



AIB



MI

Conclusion

- Markov aggregation using information-theoretic cost functions linked to spectral theory/lumpability
- Markov aggregation using information bottleneck method
- Works for nearly completely decomposable chains, surprisingly problems with quasi-lumpable chains
- Seems to work for random walk-based clustering (three-circles data set)

Outlook?

- Is there a connection between $\mathbf{P}$-lifting and spectral graph theory?
- Generalization to "stochastic" aggregations
- Agglomeratively merging states to directly minimize $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}})$, applying IB iteratively, M-partition by successive bi-partitions, etc.
- Killer Apps?

Outlook?

- Is there a connection between $\mathbf{P}$-lifting and spectral graph theory?
- Generalization to "stochastic" aggregations
- Agglomeratively merging states to directly minimize $\bar{D}(\mathbf{X}||\mathbf{X}_g^{\mathbf{P}})$, applying IB iteratively, M-partition by successive bi-partitions, etc.
- Killer Apps?

Thanks for your attention!

Bibliography I

Birch, J. J. (1962).

Approximation for the entropy for functions of Markov chains.

Ann. Math. Statist., 33:930–938.

Cover, T. M. and Thomas, J. A. (2006).

Elements of Information Theory.

Wiley Interscience, Hoboken, NJ, 2 edition.

Deng, K., Mehta, P. G., and Meyn, S. P. (2011).

Optimal Kullback-Leibler aggregation via spectral theory of Markov chains.

56(12):2793–2808.

Dhillon, I., Mallela, S., and Modha, D. (2003).

Information-theoretic co-clusterings.

In *Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, pages 89–98, Washington, D.C.

Bibliography II

Friedman, A. and Goldberger, J. (2013).

Information theoretic pairwise clustering.

In Hancock, E. and Pelillo, M., editors, *Proc. Similarity-Based Pattern Recognition*, volume 7953 of *LNCS*, pages 106–119. Springer, Berlin.

Geiger, B. C., Petrov, T., Kubin, G., and Koeppl, H. (2013).

Optimal Kullback-Leibler aggregation via information bottleneck.

accepted for publication in *IEEE Trans. Autom. Control*; preprint available: `arXiv:1304.6603 [cs.SY]`.

Geiger, B. C. and Temmel, C. (2012).

Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability.

accepted for publication in *J. Appl. Prob.*; preprint available: `arXiv:1212.4375 [cs.IT]`.

Bibliography III

Goldberger, J., Erez, K., and Abeles, M. (2007).

A Markov clustering method for analyzing movement trajectories.

In *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, pages 211–216, Thessaloniki.

Gray, R. M. (1990).

Entropy and Information Theory.

Springer, New York, NY.

Rached, Z., Alajaji, F., and Campbell, L. L. (2004).

The Kullback-Leibler divergence rate between Markov sources.

50(5):917–921.

Slonim, N. and Tishby, N. (1999).

Agglomerative information bottleneck.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 617–623, Denver, CO.

Bibliography IV


Tishby, N., Pereira, F. C., and Bialek, W. (1999).
The information bottleneck method.
In *Proc. Allerton Conf. on Communication, Control, and Computing*, pages 368–377, Monticello, IL.


Tishby, N. and Slonim, N. (2000).
Data clustering by Markovian relaxation and the information bottleneck method.
In *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO.


Vidyasagar, M. (2010).
Reduced-order modeling of Markov and hidden Markov processes via aggregation.
In *Proc. IEEE Conf. on Decision and Control (CDC)*, pages 1810–1815, Atlanta, GA.