A Hilbert-Schmidt Dependence Maximization Approach to Unsupervised Structure Discovery

Matthew B. Blaschko and **Arthur Gretton**


Max Planck Institute for Biological Cybernetics
Tübingen, Germany

July 5, 2008, MLG Helsinki

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Introduction

- Task: find taxonomies in data
- Simultaneous clustering and taxonomy fitting
  $\longrightarrow$ Numerical Taxonomy Clustering
  - Maximise dependence (HSIC) between data and clusters
- Benefits:
  - Visualization
  - Improved clustering results

Introduction

- Task: find taxonomies in data
- Simultaneous clustering and taxonomy fitting
  $\longrightarrow$ Numerical Taxonomy Clustering
  - Maximise dependence (HSIC) between data and clusters
- Benefits:
  - Visualization
  - Improved clustering results

Introduction

- Task: find taxonomies in data
- Simultaneous clustering and taxonomy fitting
  $\longrightarrow$ Numerical Taxonomy Clustering
  - Maximise dependence (HSIC) between data and clusters
- Benefits:
  - Visualization
  - Improved clustering results

Overview...

- Hilbert-Schmidt Independence Criterion
- Dependence Maximization
- Numerical Taxonomy
- Results

Overview...

- **Hilbert-Schmidt Independence Criterion**
- Dependence Maximization
- Numerical Taxonomy
- Results

Hilbert-Schmidt Independence Criterion (1)

- $\mathcal{F}$ RKHS on $\mathcal{X}$ with kernel k(x,x'), $\mathcal{G}$ RKHS on $\mathcal{Y}$ with kernel l(y,y')
- Covariance operator: $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$$

- HSIC is the Hilbert-Schmidt norm of C_{xy}:

$$\mathrm{HSIC} := \|C_{xy}\|^2_{\mathrm{HS}}$$

- (Biased) empirical HSIC:

$$\widehat{\mathrm{HSIC}} := \frac{1}{n^2}\mathrm{tr}(KHLH)$$

  - K Gram matrix for sample $(x_1, \ldots, x_n)$
  - Centering $H = I - \frac{1}{n}1_n 1_n^\top$

Hilbert-Schmidt Independence Criterion (1)

- $\mathcal{F}$ RKHS on $\mathcal{X}$ with kernel k(x,x'), $\mathcal{G}$ RKHS on $\mathcal{Y}$ with kernel l(y,y')
- Covariance operator: $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$$

- HSIC is the Hilbert-Schmidt norm of C_{xy}:

$$\mathrm{HSIC} := \|C_{xy}\|^2_{\mathrm{HS}}$$

- (Biased) empirical HSIC:

$$\widehat{\mathrm{HSIC}} := \frac{1}{n^2}\mathrm{tr}(KHLH)$$

  - K Gram matrix for sample $(x_1, \ldots, x_n)$
  - Centering $H = I - \frac{1}{n}1_n 1_n^\top$

Hilbert-Schmidt Independence Criterion (1)

- $\mathcal{F}$ RKHS on $\mathcal{X}$ with kernel k(x,x'), $\mathcal{G}$ RKHS on $\mathcal{Y}$ with kernel l(y,y')
- Covariance operator: $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$$

- HSIC is the Hilbert-Schmidt norm of C_{xy}:

$$\mathrm{HSIC} := \|C_{xy}\|^2_{\mathrm{HS}}$$

- (Biased) empirical HSIC:

$$\widehat{\mathrm{HSIC}} := \frac{1}{n^2}\mathrm{tr}(KHLH)$$

  - K Gram matrix for sample $(x_1, \ldots, x_n)$
  - Centering $H = I - \frac{1}{n}1_n 1_n^\top$

Hilbert-Schmidt Independence Criterion (1)

- $\mathcal{F}$ RKHS on $\mathcal{X}$ with kernel k(x,x'), $\mathcal{G}$ RKHS on $\mathcal{Y}$ with kernel l(y,y')
- Covariance operator: $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$$

- HSIC is the Hilbert-Schmidt norm of C_{xy}:

$$\mathrm{HSIC} := \|C_{xy}\|^2_{\mathrm{HS}}$$

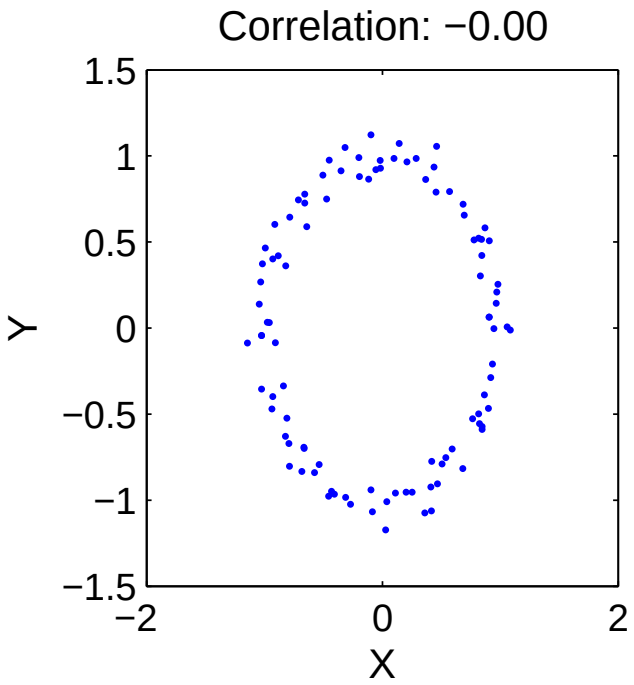
- (Biased) empirical HSIC:

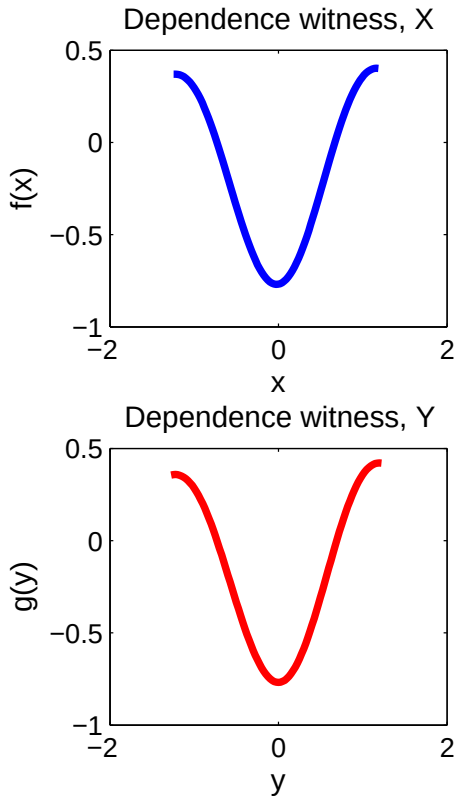
$$\widehat{\mathrm{HSIC}} := \frac{1}{n^2}\mathrm{tr}(KHLH)$$

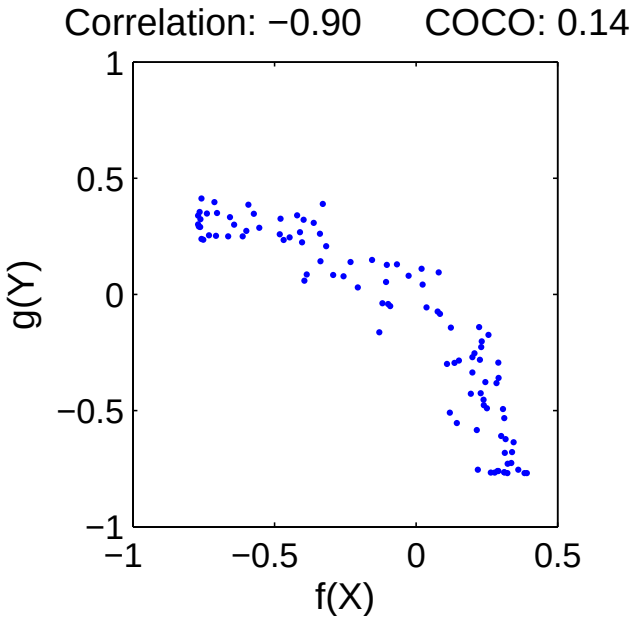
  - K Gram matrix for sample $(x_1, \ldots, x_n)$
  - Centering $H = I - \frac{1}{n}1_n 1_n^\top$

Hilbert-Schmidt Independence Criterion (2)

- Ring-shaped density, correlation approx. zero
- Maximum singular vectors (functions) of C_{xy}







Overview...

- Hilbert-Schmidt Independence Criterion
- **Dependence Maximization**
- Numerical Taxonomy
- Results

Dependence Maximization

Main objective function:

$$\frac{\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]}{\|H\Pi Y\Pi^T H\|_{\mathrm{HS}}}. \tag{1}$$

- Centered kernel matrix: $M = HKH$
- Π is $n \times k$ cluster assignment matrix, $\Pi 1 = 1$, $\Pi_{i,j} \in \{0, 1\}$.
- $Y \succeq \mathbf{0}$ Gram matrix between clusters
- Related cases
  - CLUHSIC: fixed Y, optimize $\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]$ (Song et al., 2007)
  - Normalized cuts: $Y = I$ and $M = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where A is a similarity matrix, $D_{ii} = \sum_j A_{ij}$ (Ng, Weiss, Jordan, 2001)
  - Kernel target alignment (Christianini et al., 2002)

Dependence Maximization

Main objective function:

$$\frac{\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]}{\|H\Pi Y\Pi^T H\|_{\mathrm{HS}}}. \tag{1}$$

- Centered kernel matrix: $M = HKH$
- Π is $n \times k$ cluster assignment matrix, $\Pi 1 = 1$, $\Pi_{i,j} \in \{0, 1\}$.
- $Y \succeq \mathbf{0}$ Gram matrix between clusters
- Related cases
  - CLUHSIC: fixed Y, optimize $\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]$ (Song et al., 2007)
  - Normalized cuts: $Y = I$ and $M = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where A is a similarity matrix, $D_{ii} = \sum_j A_{ij}$ (Ng, Weiss, Jordan, 2001)
  - Kernel target alignment (Christianini et al., 2002)

Dependence Maximization

Main objective function:

$$\frac{\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]}{\|H\Pi Y\Pi^T H\|_{\mathrm{HS}}}. \tag{1}$$

- Centered kernel matrix: $M = HKH$
- Π is $n \times k$ cluster assignment matrix, $\Pi 1 = 1$, $\Pi_{i,j} \in \{0, 1\}$.
- $Y \succeq \mathbf{0}$ Gram matrix between clusters
- Related cases
  - CLUHSIC: fixed Y, optimize $\mathrm{Tr}\left[MH\Pi Y\Pi^T H\right]$ (Song et al., 2007)
  - Normalized cuts: $Y = I$ and $M = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where A is a similarity matrix, $D_{ii} = \sum_j A_{ij}$ (Ng, Weiss, Jordan, 2001)
  - Kernel target alignment (Christianini et al., 2002)

Relation to Spectral Clustering

Special cases and subproblems...

- Π column vector
- Y identity matrix

$$\max_{\Pi} \frac{\mathrm{Tr}\left[MH\Pi\Pi^T H\right]}{\|H\Pi\Pi^\top H\|_{\mathrm{HS}}} = \max_{\Pi} \frac{\Pi^T HMH\Pi}{\Pi^T H\Pi} \tag{2}$$

Setting the derivative with respect to Π to zero we obtain the generalized eigenvalue problem

$$HMH\Pi_i = \rho_i H\Pi_i, \quad \text{or equivalently} \quad HMH\Pi_i = \rho_i \Pi_i. \tag{3}$$

Relation to Spectral Clustering

Special cases and subproblems...

- Π column vector
- Y identity matrix

$$\max_{\Pi} \frac{\mathrm{Tr}\left[MH\Pi\Pi^T H\right]}{\|H\Pi\Pi^\top H\|_{\mathrm{HS}}} = \max_{\Pi} \frac{\Pi^T HMH\Pi}{\Pi^T H\Pi} \tag{2}$$

Setting the derivative with respect to Π to zero we obtain the generalized eigenvalue problem

$$HMH\Pi_i = \rho_i H\Pi_i, \quad \text{or equivalently} \quad HMH\Pi_i = \rho_i \Pi_i. \tag{3}$$

Solving for Optimal $Y \succeq \mathbf{0}$ Given Π

Write optimization as constrained problem

$$\max_{Y} \quad \mathrm{Tr}\left[MH\Pi Y\Pi^T H\right], \quad \text{s.t. } \mathrm{Tr}\left[\Pi Y\Pi^T H\Pi Y\Pi^T H\right] = 1 \tag{4}$$

KKT conditions imply

$$Y^* = \frac{\left(\Pi^T H\Pi\right)^\dagger \Pi^T HMH\Pi \left(\Pi^T H\Pi\right)^\dagger}{\|\Pi^T HMH\Pi \left(\Pi^T H\Pi\right)^\dagger\|_{\mathrm{HS}}}, \tag{5}$$

Solving for Π with the Optimal $Y \succeq \mathbf{0}$

- Plug solution of optimal Y^* back into objective function

$$\Pi^* := \max_{\Pi} \left\| \Pi^T H M H \Pi \left(\Pi^T H \Pi \right)^{\dagger} \right\|_{\mathrm{HS}}. \tag{6}$$

Y has no prior structure

- Add constraints to Y
  - Change $Y^* \rightarrow$ interpretability
  - Change $\Pi^* \rightarrow$ improved clustering

Solving for Π with the Optimal $Y \succeq \mathbf{0}$

- Plug solution of optimal Y^* back into objective function

$$\Pi^* := \max_{\Pi} \left\| \Pi^T H M H \Pi \left(\Pi^T H \Pi \right)^{\dagger} \right\|_{\mathrm{HS}}. \tag{6}$$

Y has no prior structure

- Add constraints to Y
  - Change $Y^* \longrightarrow$ interpretability
  - Change $\Pi^* \longrightarrow$ improved clustering

Solving for Π with the Optimal $Y \succeq \mathbf{0}$

- Plug solution of optimal Y^* back into objective function

$$\Pi^* := \max_{\Pi} \left\| \Pi^T H M H \Pi \left(\Pi^T H \Pi \right)^{\dagger} \right\|_{\mathrm{HS}}. \tag{6}$$

Y has no prior structure

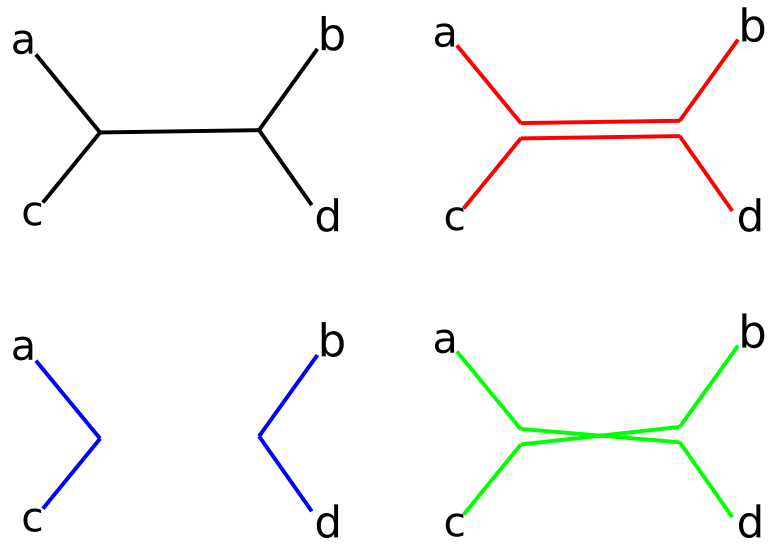
- Add constraints to Y
  - Change $Y^* \longrightarrow$ interpretability
  - Change $\Pi^* \longrightarrow$ improved clustering

Overview...

- Hilbert-Schmidt Independence Criterion
- Dependence Maximization
- **Numerical Taxonomy**
- Results

Numerical Taxonomy

- compute distance matrix, D
- $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$

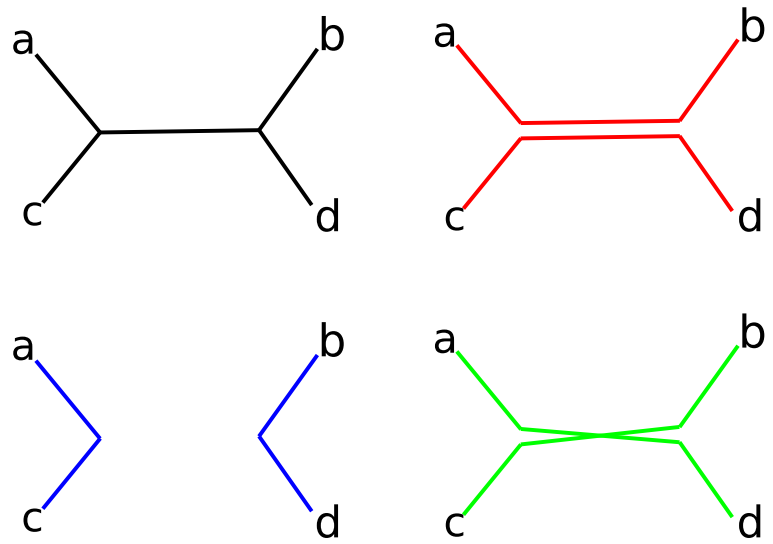


- Four point condition:
- $D_{ab} + D_{cd} \leq \max(D_{ac} + D_{bd}, D_{ad} + D_{bc}) \quad \forall a, b, c, d$

- Numerical taxonomy objective: $\min_{D_T} \|D - D_T\|^2$ where D_T is subject to the four point condition (Harb et al., 2005)
- From D_T to tree (Waterman et al., 1977)

Numerical Taxonomy

- compute distance matrix, D
- $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$



- Four point condition:
- $D_{ab} + D_{cd} \leq \max(D_{ac} + D_{bd}, D_{ad} + D_{bc}) \quad \forall a, b, c, d$

- Numerical taxonomy objective: $\min_{D_T} \|D - D_T\|^2$ where D_T is subject to the four point condition (Harb et al., 2005)
- From D_T to tree (Waterman et al., 1977)

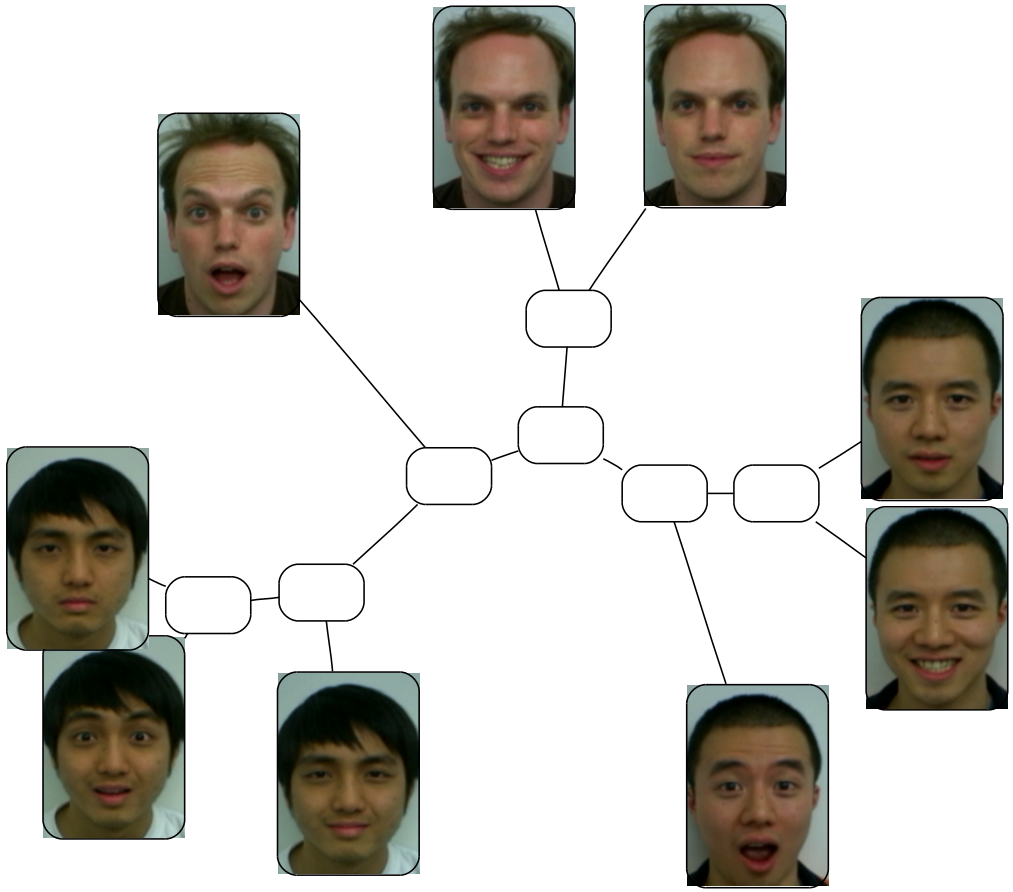
Numerical Taxonomy Clustering

Require: $M \succeq \mathbf{0}$
Ensure: $(\Pi, Y) \approx (\Pi^*, Y^*)$ that max dependence s.t. 4-point condition
  Initialize $Y = I$
  Initialize Π using the spectral relaxation
  while Convergence has not been reached **do**
    Solve for Y given Π using closed form solution
    Construct D such that $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$
    Solve for $\min_{D_T} \|D - D_T\|^2$
    Assign $Y = -\frac{1}{2} H (D_T \odot D_T) H$
    Update Π using a normalized version of Song et al. 2007.
  end while

Overview...

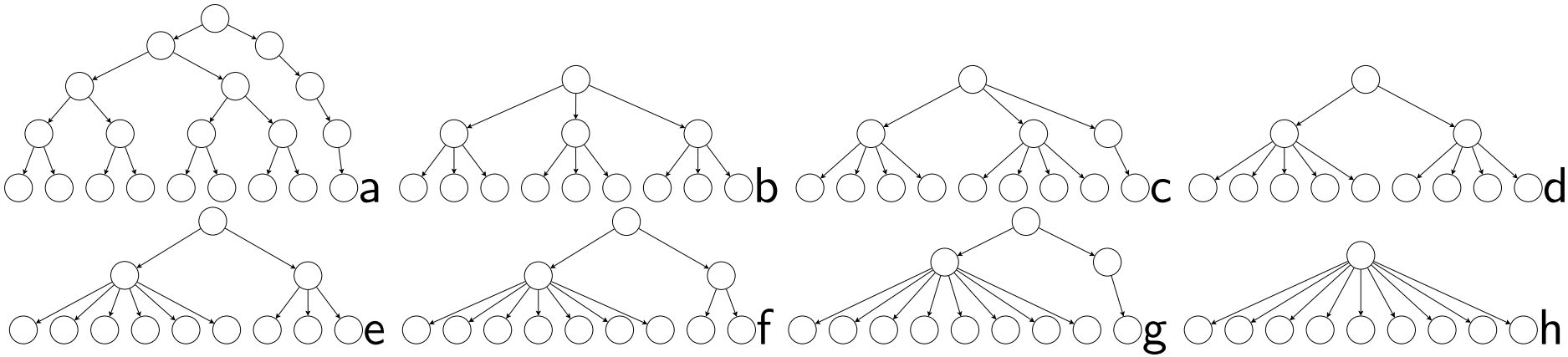
- Hilbert-Schmidt Independence Criterion
- Dependence Maximization
- Numerical Taxonomy
- **Results**

Face Dataset



Face dataset and the resulting taxonomy that was discovered by the algorithm

Structure Selection

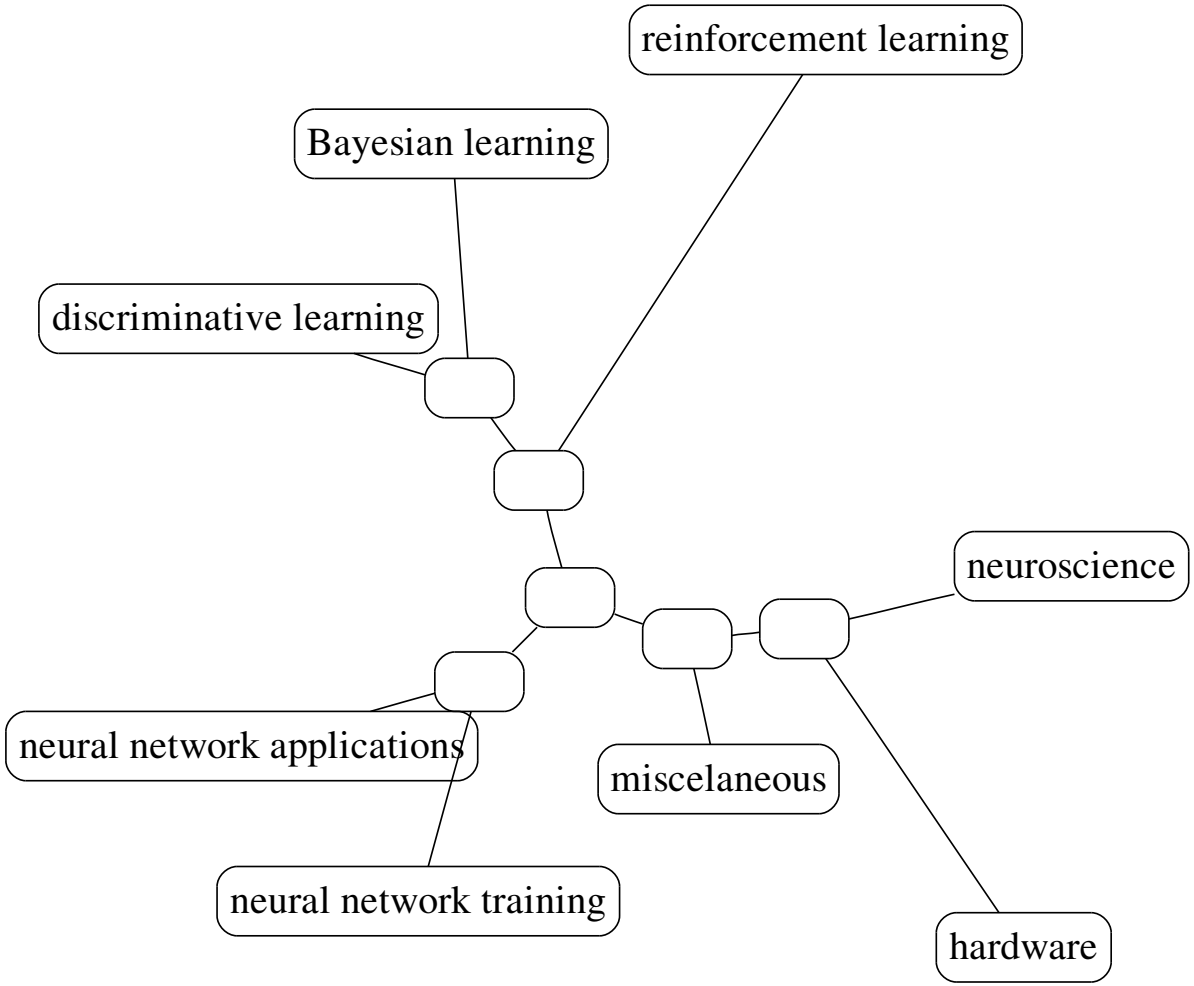


Structures used in the structure selection experiments

spectral	a	b	c	d	e	f	g	h	taxonomy
0.5443	0.7936	0.4970	0.6336	0.8652	1.2246	1.1396	1.1325	0.5180	**0.2807**

Conditional entropy scores for spectral clustering (NJW 2001), the clustering algorithm of Song et al. 2007, and the method presented here (last column).

NIPS Articles



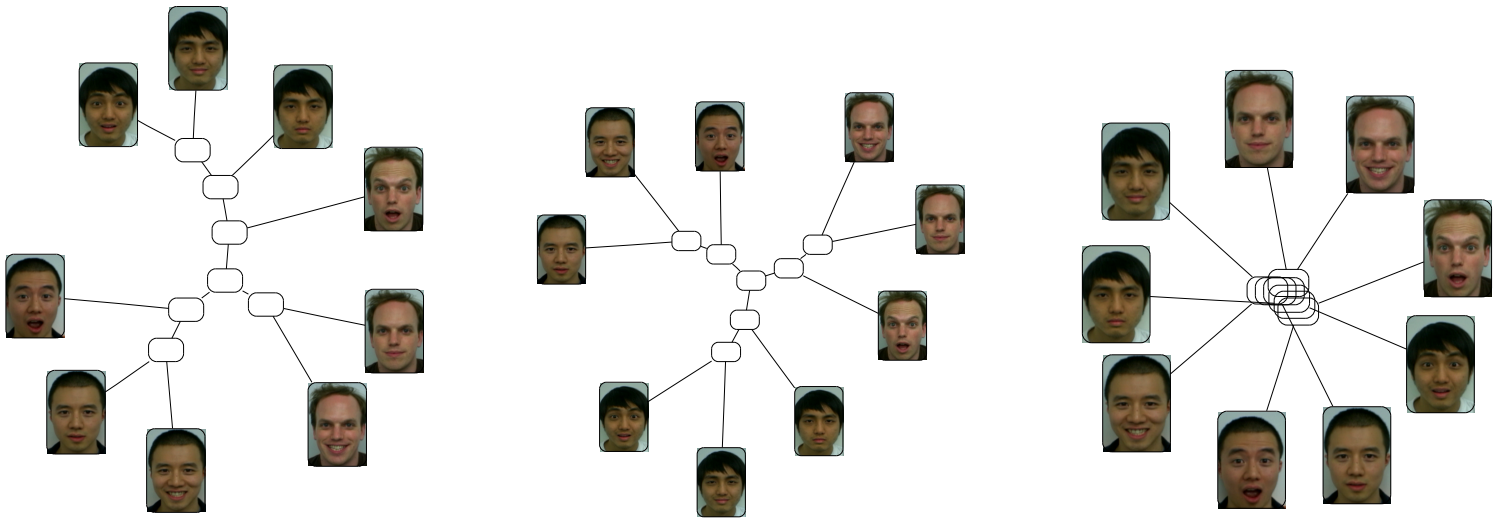
The taxonomy discovered for the NIPS dataset.

NIPS Articles - Categories

neurosci.	hardware	misc.	train-neural	app.-neural	reinforcement	discriminative	Bayesian
error	cells	learning	training	data	chip	neural	network
training	model	state	recognition	model	circuit	networks	units
algorithm	visual	policy	network	models	analog	function	learning
function	neurons	action	set	gaussian	neuron	matrix	input
learning	cell	reinforce.	speech	distribution	voltage	functions	hidden
set	activity	control	performance	parameters	current	theorem	unit
generaliz.	response	optimal	neural	likelihood	figure	dynamics	networks
examples	synaptic	time	word	mixture	vlsi	threshold	output
functions	cortex	function	features	em	output	network	weights
vector	stimulus	states	image	algorithm	circuits	hopfield	training
class	firing	actions	classification	probability	system	proof	error
data	spike	algorithm	trained	bayesian	signal	neurons	weight
case	neuron	reward	system	density	neural	energy	time
linear	cortical	agent	test	posterior	synapse	equations	layer
weight	orientation	sutton	networks	log	time	polynomial	recurrent
optimal	direction	dynamic	feature	prior	pulse	points	neural
regression	motion	goal	data	variables	neurons	fixed	net
bound	frequency	robot	images	estimation	silicon	neuron	propagation
algorithms	spatial	step	layer	matrix	implem.	equation	back
loss	eye	algorithms	classifier	markov	digital	stable	architecture

Perturbing Spectrum of M

- $M = HKH(HKH + \varepsilon_k I)^{-1}$
- $\varepsilon_k = n\kappa$ where n is the number of samples



$\kappa = 10^{1}$ $\kappa = 10^{-1}$ $\kappa = 10^{-3}$

The effect of varying the regularization parameter in HSNIC. Smaller values tend towards a star topology.

Conclusions

- Maximize dependence (normalized HSIC) between data and clustering
- Learn cluster Gram matrix Y corresponding to taxonomy
- Numerical taxonomy clustering useful for
  - visualizations
  - improved clustering
- Further work: better solution method for Π given Y