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
Introduction

- **Task:** find **taxonomies** in data

- Simultaneous clustering and taxonomy fitting
  → **Numerical Taxonomy Clustering**
    - Maximise dependence (HSIC) between data and clusters

- **Benefits:**
  - Visualization
  - Improved clustering results
Task: find taxonomies in data

Simultaneous clustering and taxonomy fitting → Numerical Taxonomy Clustering
  ▶ Maximise dependence (HSIC) between data and clusters

Benefits:
  ▶ Visualization
  ▶ Improved clustering results
• Task: find taxonomies in data

• Simultaneous clustering and taxonomy fitting
  → 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: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that
  \[
  \langle f, C_{xy} g \rangle_{\mathcal{F}} = E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]
  \]

- HSIC is the Hilbert-Schmidt norm of $C_{xy}$:
  \[
  \text{HSIC} := \|C_{xy}\|_{\text{HS}}^2
  \]

- (Biased) empirical HSIC:
  \[
  \hat{\text{HSIC}} := \frac{1}{n^2} \text{tr}(KHLH)
  \]

  - $K$ Gram matrix for sample $(x_1, \ldots, x_n)$
  - Centering $H = I - \frac{1}{n} 1_n 1_n^T$
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: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]$$

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

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

- (Biased) empirical HSIC:

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

- $K$ Gram matrix for sample $(x_1, \ldots, x_n)$
- Centering $H = I - \frac{1}{n}1_n1_n^T$
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
  \[
  \langle f, C_{xy}g \rangle_{\mathcal{F}} = E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]
  \]
- HSIC is the Hilbert-Schmidt norm of $C_{xy}$:
  \[
  \text{HSIC} := \|C_{xy}\|_{HS}^2
  \]
- (Biased) empirical HSIC:
  \[
  \hat{\text{HSIC}} := \frac{1}{n^2} \text{tr}(KHLH)
  \]
- $K$ Gram matrix for sample $(x_1, \ldots, x_n)$
- Centering $H = I - \frac{1}{n}1_n1_n^T$
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: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]$$

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

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

- (Biased) empirical HSIC:

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

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

\[ \frac{\text{Tr} \left[ M H \Pi Y \Pi^T H \right]}{\| H \Pi Y \Pi^T H \|_{\text{HS}}} . \]  

1. Centered kernel matrix: \( M = H K H \)
2. \( \Pi \) is \( n \times k \) cluster assignment matrix, \( \Pi 1 = 1, \Pi_{i,j} \in \{0, 1\} \).
3. \( Y \succeq 0 \) Gram matrix between clusters
4. Related cases
   - **CLUHSIC**: fixed \( Y \), optimize \( \text{Tr} \left[ M H \Pi Y \Pi^T H \right] \) (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)
Main objective function:

\[
\frac{\text{Tr} \left[ M H \Pi Y \Pi^T H \right]}{\| H \Pi Y \Pi^T H \|_{\text{HS}}}.
\]  

(1)

- Centered kernel matrix: \( M = HKH \)
- \( \Pi \) is \( n \times k \) cluster assignment matrix, \( \Pi 1 = 1, \Pi_{i,j} \in \{0, 1\} \).
- \( Y \succeq 0 \) Gram matrix between clusters
- Related cases
  - CLUHSIC: fixed \( Y \), optimize \( \text{Tr} \left[ M H \Pi Y \Pi^T H \right] \) (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)
Main objective function:

\[
\frac{\text{Tr} \left[ MH\Pi Y\Pi^T H \right]}{\|H\Pi Y\Pi^T H\|_{\text{HS}}}.
\] (1)

- Centered kernel matrix: \( M = HKH \)
- \( \Pi \) is \( n \times k \) cluster assignment matrix, \( \Pi 1 = 1, \Pi_{i,j} \in \{0, 1\} \).
- \( Y \succeq 0 \) Gram matrix between clusters
- Related cases
  - **CLUHSIC**: fixed \( Y \), optimize \( \text{Tr} \left[ MH\Pi Y\Pi^T H \right] \) (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...

- $\Pi$ column vector
- $Y$ identity matrix

$$\max_{\Pi} \frac{\text{Tr} \left[ MH\Pi \Pi^T H \right]}{\| H\Pi \Pi^T H \|_{\text{HS}}} = \max_{\Pi} \frac{\Pi^T H MH \Pi}{\Pi^T H \Pi}$$

(2)

Setting the derivative with respect to $\Pi$ to zero we obtain the generalized eigenvalue problem

$$HMH\Pi_i = \rho_i H\Pi_i,$$

or equivalently

$$HMH\Pi_i = \rho_i \Pi_i.$$  

(3)
Relation to Spectral Clustering

Special cases and subproblems...

- \( \Pi \) column vector
- \( Y \) identity matrix

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

Setting the derivative with respect to \( \Pi \) 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 $Y \succeq 0$ Given $\Pi$

Write optimization as constrained problem

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

KKT conditions imply

$$Y^* = \frac{(\Pi^T H\Pi)^\dagger \Pi^T H M H \Pi (\Pi^T H\Pi)^\dagger}{\| \Pi^T H M H \Pi (\Pi^T H\Pi)^\dagger \|_{\text{HS}}}, \quad (5)$$
Plug solution of optimal $Y^*$ back into objective function

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

$Y$ has no prior structure

- Add constraints to $Y$
  - Change $Y^*$ → interpretability
  - Change $\Pi^*$ → improved clustering
Solving for $\Pi$ with the Optimal $Y \succeq 0$

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

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

(6)

$Y$ has no prior structure

- Add constraints to $Y$
  - Change $Y^*$ $\rightarrow$ interpretability
  - Change $\Pi^*$ $\rightarrow$ improved clustering
Solving for $\Pi$ with the Optimal $Y \succeq 0$

- Plug solution of optimal $Y^*$ back into objective function
  \[
  \Pi^* := \max_{\Pi} \left\| \Pi^T H M H \Pi \left( \Pi^T H \Pi \right)^\dagger \right\|_{\text{HS}} .
  \]
  \(6\)

$Y$ has no prior structure

- Add constraints to $Y$
  - Change $Y^* \rightarrow$ interpretability
  - Change $\Pi^* \rightarrow$ improved clustering
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:** \( M \succeq 0 \)

**Ensure:** \((\Pi, Y) \approx (\Pi^*, Y^*)\) that max dependence s.t. 4-point condition

Initialize \( Y = I \)

Initialize \( \Pi \) using the spectral relaxation

**while** Convergence has not been reached **do**

Solve for \( Y \) given \( \Pi \) 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 \( \Pi \) using a normalized version of Song et al. 2007.

**end while**
Overview...

- Hilbert-Schmidt Independence Criterion
- Dependence Maximization
- Numerical Taxonomy
- Results
Face dataset and the resulting taxonomy that was discovered by the algorithm
Conditional entropy scores for spectral clustering (NJW 2001), the clustering algorithm of Song et al. 2007, and the method presented here (last column).
The taxonomy discovered for the NIPS dataset.
<table>
<thead>
<tr>
<th>neurosci.</th>
<th>hardware</th>
<th>misc.</th>
<th>train-neural</th>
<th>app.-neural</th>
<th>reinforcement</th>
<th>discriminative</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>error</td>
<td>cells</td>
<td>learning</td>
<td>training</td>
<td>data</td>
<td>chip</td>
<td>neural</td>
<td>network</td>
</tr>
<tr>
<td>training</td>
<td>model</td>
<td>state</td>
<td>recognition</td>
<td>model</td>
<td>circuit</td>
<td>networks</td>
<td>units</td>
</tr>
<tr>
<td>algorithm</td>
<td>visual</td>
<td>policy</td>
<td>network</td>
<td>models</td>
<td>analog</td>
<td>function</td>
<td>learning</td>
</tr>
<tr>
<td>function</td>
<td>neurons</td>
<td>action</td>
<td>set</td>
<td>gaussian</td>
<td>neuron</td>
<td>matrix</td>
<td>input</td>
</tr>
<tr>
<td>learning</td>
<td>cell</td>
<td>reinforce.</td>
<td>speech</td>
<td>distribution</td>
<td>voltage</td>
<td>functions</td>
<td>hidden</td>
</tr>
<tr>
<td>set</td>
<td>activity</td>
<td>control</td>
<td>performance</td>
<td>parameters</td>
<td>current</td>
<td>theorem</td>
<td>unit</td>
</tr>
<tr>
<td>generaliz.</td>
<td>response</td>
<td>optimal</td>
<td>neural</td>
<td>likelihood</td>
<td>figure</td>
<td>dynamics</td>
<td>networks</td>
</tr>
<tr>
<td>examples</td>
<td>synaptic</td>
<td>time</td>
<td>word</td>
<td>mixture</td>
<td>vlsi</td>
<td>threshold</td>
<td>output</td>
</tr>
<tr>
<td>functions</td>
<td>cortex</td>
<td>function</td>
<td>features</td>
<td>em</td>
<td>output</td>
<td>network</td>
<td>weights</td>
</tr>
<tr>
<td>vector</td>
<td>stimulus</td>
<td>states</td>
<td>image</td>
<td>algorithm</td>
<td>circuits</td>
<td>hopfield</td>
<td>training</td>
</tr>
<tr>
<td>class</td>
<td>firing</td>
<td>actions</td>
<td>classification</td>
<td>probability</td>
<td>system</td>
<td>proof</td>
<td>error</td>
</tr>
<tr>
<td>data</td>
<td>spike</td>
<td>algorithm</td>
<td>trained</td>
<td>bayesian</td>
<td>signal</td>
<td>neurons</td>
<td>weight</td>
</tr>
<tr>
<td>case</td>
<td>neuron</td>
<td>reward</td>
<td>system</td>
<td>density</td>
<td>neural</td>
<td>energy</td>
<td>time</td>
</tr>
<tr>
<td>linear</td>
<td>cortical</td>
<td>agent</td>
<td>test</td>
<td>posterior</td>
<td>synapse</td>
<td>equations</td>
<td>layer</td>
</tr>
<tr>
<td>weight</td>
<td>orientation</td>
<td>sutton</td>
<td>networks</td>
<td>log</td>
<td>time</td>
<td>polynomial</td>
<td>recurrent</td>
</tr>
<tr>
<td>optimal</td>
<td>direction</td>
<td>dynamic</td>
<td>feature</td>
<td>prior</td>
<td>pulse</td>
<td>points</td>
<td>neural</td>
</tr>
<tr>
<td>regression</td>
<td>motion</td>
<td>goal</td>
<td>data</td>
<td>variables</td>
<td>neurons</td>
<td>fixed</td>
<td>net</td>
</tr>
<tr>
<td>bound</td>
<td>frequency</td>
<td>robot</td>
<td>images</td>
<td>estimation</td>
<td>silicon</td>
<td>neuron</td>
<td>propagation</td>
</tr>
<tr>
<td>algorithms</td>
<td>spatial</td>
<td>step</td>
<td>layer</td>
<td>matrix</td>
<td>implem.</td>
<td>equation</td>
<td>back</td>
</tr>
<tr>
<td>loss</td>
<td>eye</td>
<td>algorithms</td>
<td>classifier</td>
<td>markov</td>
<td>digital</td>
<td>stable</td>
<td>architecture</td>
</tr>
</tbody>
</table>
Perturbing Spectrum of $M$

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

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

$\kappa = 10^1$

$\kappa = 10^{-1}$

$\kappa = 10^{-3}$

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 $\Pi$ given $Y$