Kernel Methods

Lecture 5: Hilbert Schmidt Independence Criterion
Thanks to Arthur Gretton, Le Song, Bernhard Schölkopf, Olivier Bousquet

Alexander J. Smola

Statistical Machine Learning Program
Canberra, ACT 0200 Australia
Alex.Smola@nicta.com.au

Machine Learning Summer School, Taiwan 2006

Course Overview

1. Estimation in exponential families
   - Maximum Likelihood and Priors
   - Clifford Hammersley decomposition
2. Applications
   - Conditional distributions and kernels
   - Classification, Regression, Conditional random fields
3. Inference and convex duality
   - Maximum entropy inference
   - Approximate moment matching
4. Maximum mean discrepancy
   - Means in feature space, Covariate shift correction
5. Hilbert-Schmidt independence criterion
   - Covariance in feature space
   - ICA, Feature selection

Course Overview

1. Estimation in exponential families
   - Maximum Likelihood and Priors
   - Clifford Hammersley decomposition
2. Applications
   - Conditional distributions and kernels
   - Classification, Regression, Conditional random fields
3. Inference and convex duality
   - Maximum entropy inference
   - Approximate moment matching
4. Maximum mean discrepancy
   - Means in feature space, Covariate shift correction
5. Hilbert-Schmidt independence criterion
   - Covariance in feature space
   - ICA, Feature selection

Outline

1. **Measuring Independence**
   - Covariance Operator
   - Hilbert Space Methods
   - A Test Statistic and its Analysis
2. Independent Component Analysis
   - ICA Primer
   - Examples
3. Feature Selection
   - Problem Setting
   - Algorithm
   - Results

Measuring Independence

Problem

- Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ determine whether $\Pr(x, y) = \Pr(x)\Pr(y)$.
- Measure degree of dependence.

Applications

- Independent component analysis
- Dimensionality reduction and feature extraction
- Statistical modeling

Indirect Approach

- Perform density estimate of $\Pr(x, y)$
- Check whether the estimate approximately factorizes

Direct Approach

- Check properties of factorizing distributions
- E.g. kurtosis, covariance operators, etc.

Measuring Independence

Problem

- Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ determine whether $\Pr(x, y) = \Pr(x) \Pr(y)$.
- Measure degree of dependence.

Applications

- Independent component analysis
- Dimensionality reduction and feature extraction
- Statistical modeling

Indirect Approach

- Perform density estimate of $\Pr(x, y)$
- Check whether the estimate approximately factorizes

Direct Approach

- Check properties of factorizing distributions
- E.g. kurtosis, covariance operators, etc.

Measuring Independence

Problem

- Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ determine whether $\Pr(x, y) = \Pr(x)\Pr(y)$.
- Measure degree of dependence.

Applications

- Independent component analysis
- Dimensionality reduction and feature extraction
- Statistical modeling

Indirect Approach

- Perform density estimate of $\Pr(x, y)$
- Check whether the estimate approximately factorizes

Direct Approach

- Check properties of factorizing distributions
- E.g. kurtosis, covariance operators, etc.

Measuring Independence

Problem

- Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ determine whether $\Pr(x, y) = \Pr(x)\Pr(y)$.
- Measure degree of dependence.

Applications

- Independent component analysis
- Dimensionality reduction and feature extraction
- Statistical modeling

Indirect Approach

- Perform density estimate of $\Pr(x, y)$
- Check whether the estimate approximately factorizes

Direct Approach

- Check properties of factorizing distributions
- E.g. kurtosis, covariance operators, etc.

Covariance Operator

Linear Case

For linear functions $f(x) = w^\top x$ and $g(y) = v^\top y$ the covariance is given by

$$\mathrm{Cov}\{f(x), g(y)\} = w^\top C v$$

This is a bilinear operator on the space of linear functions.

Nonlinear Case

Define C to be the operator with $(f, g) \to \mathrm{Cov}\{f, g\}$.

Theorem

C is a bilinear operator in f and g.

Proof.

We only show linearity in f: $\mathrm{Cov}\{\alpha f, g\} = \alpha \mathrm{Cov}\{f, g\}$. Moreover, for $f + f'$ the covariance is additive. □

Covariance Operator

Linear Case

For linear functions $f(x) = w^\top x$ and $g(y) = v^\top y$ the covariance is given by

$$\mathrm{Cov}\{f(x), g(y)\} = w^\top C v$$

This is a bilinear operator on the space of linear functions.

Nonlinear Case

Define C to be the operator with $(f, g) \to \mathrm{Cov}\{f, g\}$.

Theorem

C is a bilinear operator in f and g.

Proof.

We only show linearity in f: $\mathrm{Cov}\{\alpha f, g\} = \alpha \mathrm{Cov}\{f, g\}$.
Moreover, for $f + f'$ the covariance is additive. □

Covariance Operator

Linear Case

For linear functions $f(x) = w^\top x$ and $g(y) = v^\top y$ the covariance is given by

$$\mathrm{Cov}\{f(x), g(y)\} = w^\top C v$$

This is a bilinear operator on the space of linear functions.

Nonlinear Case

Define C to be the operator with $(f, g) \to \mathrm{Cov}\{f, g\}$.

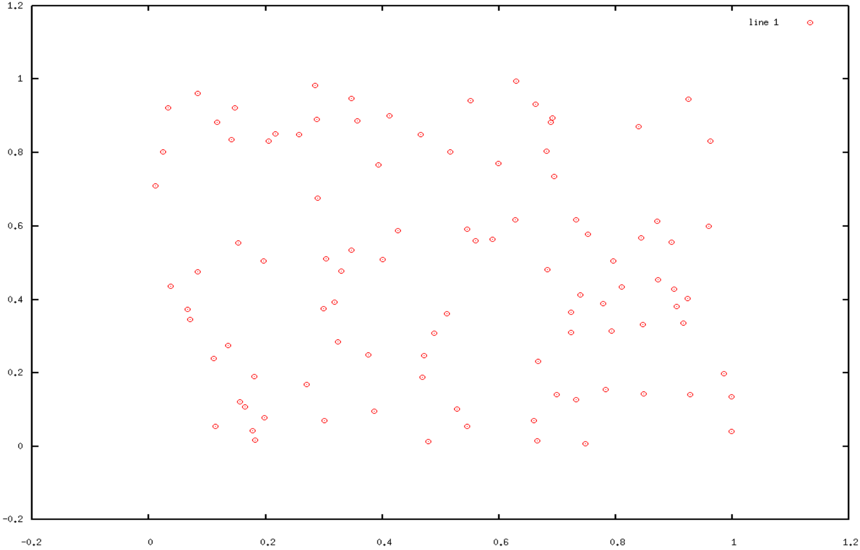
Theorem

C is a bilinear operator in f and g.

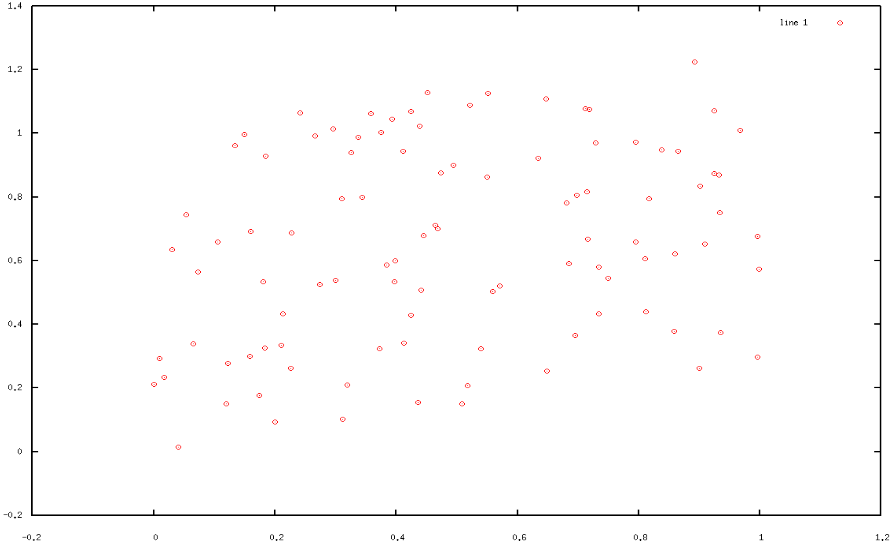
Proof.

We only show linearity in f: $\mathrm{Cov}\{\alpha f, g\} = \alpha \mathrm{Cov}\{f, g\}$.
Moreover, for $f + f'$ the covariance is additive. $\square$

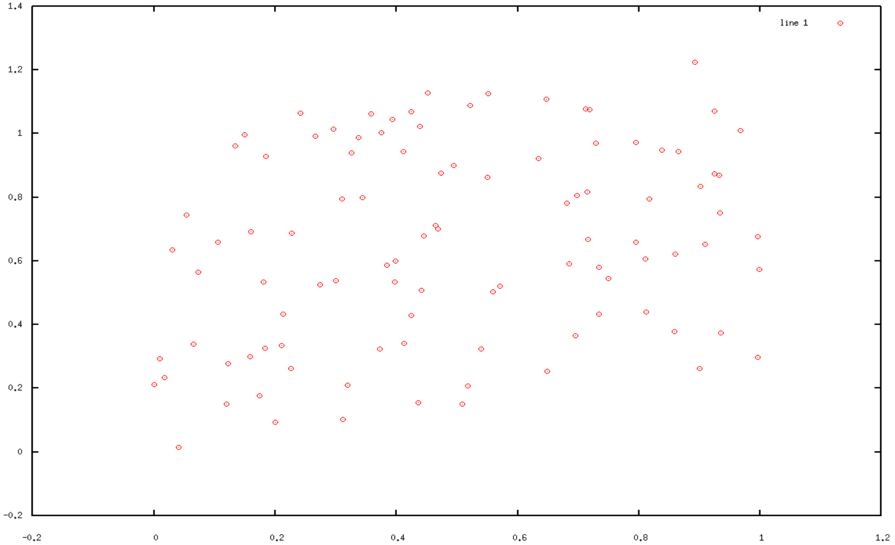
Independent random variables



Dependent random variables



Or are we just unlucky?



Covariance operators

Criterion (Renyi, 1957)

Test for independence by checking whether $C = 0$.

Reproducing Kernel Hilbert Space

- Kernels k, l on $\mathcal{X}, \mathcal{Y}$ with associated RKHSs $\mathcal{F}, \mathcal{G}$.
- Assume bounded k, l on domain.

Mean operator

$$\langle \mu_x, f \rangle = \mathbf{E}_x[f(x)] \text{ and } \langle \mu_y, g \rangle = \mathbf{E}_y[g(y)]$$

Covariance operator

Define covariance operator C via bilinear form

$$f^\top C_{xy} g = \mathrm{Cov}\{f, g\} = \mathbf{E}_{x,y}\left[f(x)g(y)\right] - \mathbf{E}_x\left[f(x)\right]\mathbf{E}_y\left[g(y)\right]$$

Covariance operators

Criterion (Renyi, 1957)

Test for independence by checking whether $C = 0$.

Reproducing Kernel Hilbert Space

- Kernels k, l on $\mathcal{X}, \mathcal{Y}$ with associated RKHSs $\mathcal{F}, \mathcal{G}$.
- Assume bounded k, l on domain.

Mean operator

$$\langle \mu_x, f \rangle = \mathbf{E}_x[f(x)] \text{ and } \langle \mu_y, g \rangle = \mathbf{E}_y[g(y)]$$

Covariance operator

Define covariance operator C via bilinear form

$$f^\top C_{xy} g = \mathrm{Cov}\{f, g\} = \mathbf{E}_{x,y}\left[f(x)g(y)\right] - \mathbf{E}_x\left[f(x)\right]\mathbf{E}_y\left[g(y)\right]$$

Covariance operators

Criterion (Renyi, 1957)

Test for independence by checking whether $C = 0$.

Reproducing Kernel Hilbert Space

- Kernels k, l on $\mathcal{X}, \mathcal{Y}$ with associated RKHSs $\mathcal{F}, \mathcal{G}$.
- Assume bounded k, l on domain.

Mean operator

$$\langle \mu_x, f \rangle = \mathbf{E}_x[f(x)] \text{ and } \langle \mu_y, g \rangle = \mathbf{E}_y[g(y)]$$

Covariance operator

Define covariance operator C via bilinear form

$$f^\top C_{xy} g = \mathrm{Cov}\{f, g\} = \mathbf{E}_{x,y}\left[f(x)g(y)\right] - \mathbf{E}_x\left[f(x)\right]\mathbf{E}_y\left[g(y)\right]$$

Covariance operators

Criterion (Renyi, 1957)

Test for independence by checking whether $C = 0$.

Reproducing Kernel Hilbert Space

- Kernels k, l on $\mathcal{X}, \mathcal{Y}$ with associated RKHSs $\mathcal{F}, \mathcal{G}$.
- Assume bounded k, l on domain.

Mean operator

$$\langle \mu_x, f \rangle = \mathbf{E}_x[f(x)] \text{ and } \langle \mu_y, g \rangle = \mathbf{E}_y[g(y)]$$

Covariance operator

Define covariance operator C via bilinear form

$$f^\top C_{xy} g = \mathrm{Cov}\{f, g\} = \mathbf{E}_{x,y}\left[f(x)g(y)\right] - \mathbf{E}_x\left[f(x)\right]\mathbf{E}_y\left[g(y)\right]$$

Hilbert Space Representation

Theorem

Provided that k, l are universal kernels $\|C_{xy}\| = 0$ if and only if x, y are independent.

Proof.

Step 1: If x, y are dependent then there exist some $[0, 1]$-bounded range f^*, g^* with $\mathrm{Cov}\{f^*, g^*\} = \epsilon > 0$.

Step 2: Since k, l are universal there exist ϵ' approximation of f^*, g^* in $\mathcal{F}, \mathcal{G}$ such that covariance of approximation does not vanish.

Step 3: Hence the covariance operator C_{xy} is nonzero.

$\square$

A Test Statistic

Covariance operator

$$C_{xy} = \mathbf{E}_{x,y}\left[k(x,\cdot)l(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\right]\mathbf{E}_y\left[l(y,\cdot)\right]$$

Operator Norm

Use the norm of C_{xy} to test whether x and y are independent. It also gives us a measure of dependence.

$$\mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|^2$$

where $\|\cdot\|$ denotes the Hilbert-Schmidt norm.

Frobenius Norm

For matrices we can define

$$\|M\|^2 = \sum_{ij} M_{ij}^2 = \operatorname{tr} M^\top M.$$

Hilbert-Schmidt norm is generalization of Frobenius norm.

A Test Statistic

Covariance operator

$$C_{xy} = \mathbf{E}_{x,y}\left[k(x,\cdot)l(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\right]\mathbf{E}_y\left[l(y,\cdot)\right]$$

Operator Norm

Use the norm of C_{xy} to test whether x and y are independent. It also gives us a measure of dependence.

$$\mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|^2$$

where $\|\cdot\|$ denotes the Hilbert-Schmidt norm.

Frobenius Norm

For matrices we can define

$$\|M\|^2 = \sum_{ij} M_{ij}^2 = \operatorname{tr} M^\top M.$$

Hilbert-Schmidt norm is generalization of Frobenius norm.

A Test Statistic

Covariance operator

$$C_{xy} = \mathbf{E}_{x,y}\left[k(x,\cdot)l(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\right]\mathbf{E}_y\left[l(y,\cdot)\right]$$

Operator Norm

Use the norm of C_{xy} to test whether x and y are independent. It also gives us a measure of dependence.

$$\mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|^2$$

where $\|\cdot\|$ denotes the Hilbert-Schmidt norm.

Frobenius Norm

For matrices we can define

$$\|M\|^2 = \sum_{ij} M_{ij}^2 = \operatorname{tr} M^\top M.$$

Hilbert-Schmidt norm is generalization of Frobenius norm.

Computing $\|C_{xy}\|^2$

Rank-one operators

For rank-one terms we have
$\|f \otimes g\|^2 = \langle f \otimes g, f \otimes g\rangle_{HS} = \|f\|^2 \|g\|^2$.

Joint expectation

By construction of C_{xy} we exploit linearity and obtain

$$\begin{aligned}
\|C_{xy}\|^2 =& \langle C_{xy}, C_{xy}\rangle_{HS} \\
=& \left\{\mathbf{E}_{x,y}\mathbf{E}_{x',y'} - 2\mathbf{E}_{x,y}\mathbf{E}_{x'}\mathbf{E}_{y'} + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}\right\} \\
& \left[\langle k(x,\cdot)l(y,\cdot), k(x',\cdot)l(y',\cdot)\rangle_{HS}\right] \\
=& \left\{\mathbf{E}_{x,y}\mathbf{E}_{x',y'} - 2\mathbf{E}_{x,y}\mathbf{E}_{x'}\mathbf{E}_{y'} + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}\right\} \\
& \left[k(x,x')l(y,y')\right]
\end{aligned}$$

This is well-defined if k, l are bounded.

Estimating $\|C_{xy}^2\|$

Empirical criterion

$$\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G}) := \frac{1}{(m-1)^2} \mathrm{tr} KHLH$$

where $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ and $H_{ij} = \delta_{ij} - m^{-2}$.

Theorem

$$\mathbf{E}_Z[\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})] = \mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) + O(1/m)$$

Proof: Sketch only.

Expand tr $KHLH$ into terms of pairs, triples and quadruples of indices of non-repeated terms, which lead to the proper expectations and bound the rest by $O(m^{-1})$. □

Estimating $\|C_{xy}^2\|$

Empirical criterion

$$\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G}) := \frac{1}{(m-1)^2} \mathrm{tr} KHLH$$

where $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ and $H_{ij} = \delta_{ij} - m^{-2}$.

Theorem

$$\mathbf{E}_Z \left[\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})\right] = \mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) + O(1/m)$$

Proof: Sketch only.

Expand tr $KHLH$ into terms of pairs, triples and quadruples of indices of non-repeated terms, which lead to the proper expectations and bound the rest by $O(m^{-1})$. □

Estimating $\|C_{xy}^2\|$

Empirical criterion

$$\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G}) := \frac{1}{(m-1)^2} \mathrm{tr} KHLH$$

where $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$ and $H_{ij} = \delta_{ij} - m^{-2}$.

Theorem

$$\mathbf{E}_Z \left[\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})\right] = \mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) + O(1/m)$$

Proof: Sketch only.

Expand tr $KHLH$ into terms of pairs, triples and quadruples of indices of non-repeated terms, which lead to the proper expectations and bound the rest by $O(m^{-1})$. □

Uniform convergence bounds for $\|C_{xy}^2\|$

Theorem (Recall Hoeffding's theorem for U Statistics)

For averages over functions on r variables

$$u := \frac{1}{(m)_r} \sum_{i_r^m} g(x_{i_1}, \ldots, x_{i_r})$$

which are bounded by $a \leq u \leq b$ we have

$$\Pr_u \{u - \mathbf{E}_u[u] \geq t\} \leq \exp\left(-\frac{2t^2 \lceil m/r \rceil}{(b-a)^2}\right)$$

In our statistic we have terms of 2, 3, and 4 random variables.

Uniform convergence bounds for $\left\| C_{xy}^2 \right\|$

Corollary

Assume that $k, l \leq$. *Then at least with probability* $1 - \delta$

$$\left| \mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G}) - \mathrm{HSIC}(\Pr_{xy}, \mathcal{F}, \mathcal{G}) \right| \leq \sqrt{\frac{\log 6/\delta}{0.24m}} + \frac{C}{m}$$

Proof.

Bound each of the three terms separatly via Hoeffding's theorem. $\square$

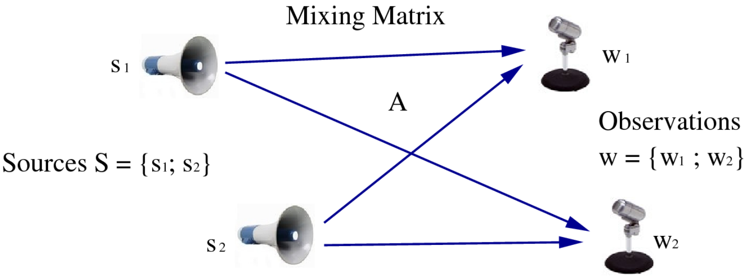
Outline

Blind Source Separation

Data

$w = Ms$, where all s_i are mutually independent.

The Cocktail Party Problem



Task

Recover the sources S and mixing matrix M given W.

Independent Component Analysis

Whitening

Rotate, center, and whiten data before separation. This is always possible.

Optimization

- We cannot recover scale of data anyway.
- Need to find orthogonal matrix U such that $Uw = s$ leads to independent random variables.
- Optimization on the Stiefel manifold.
- Could do this by a Newton method.

Important Trick

- Kernel matrix could be huge.
- Use reduced-rank representation. We get

$$\operatorname{tr} H(AA^\top)H(BB^\top) = \left\|A^\top HB\right\|^2 \text{ instead of } \operatorname{tr} HKHL.$$

Independent Component Analysis

Whitening

Rotate, center, and whiten data before separation. This is always possible.

Optimization

- We cannot recover scale of data anyway.
- Need to find orthogonal matrix U such that $Uw = s$ leads to independent random variables.
- Optimization on the Stiefel manifold.
- Could do this by a Newton method.

Important Trick

- Kernel matrix could be huge.
- Use reduced-rank representation. We get

$$\operatorname{tr} H(AA^\top)H(BB^\top) = \left\|A^\top HB\right\|^2 \text{ instead of } \operatorname{tr} HKHL.$$

Independent Component Analysis

Whitening

Rotate, center, and whiten data before separation. This is always possible.

Optimization

- We cannot recover scale of data anyway.
- Need to find orthogonal matrix U such that $Uw = s$ leads to independent random variables.
- Optimization on the Stiefel manifold.
- Could do this by a Newton method.

Important Trick

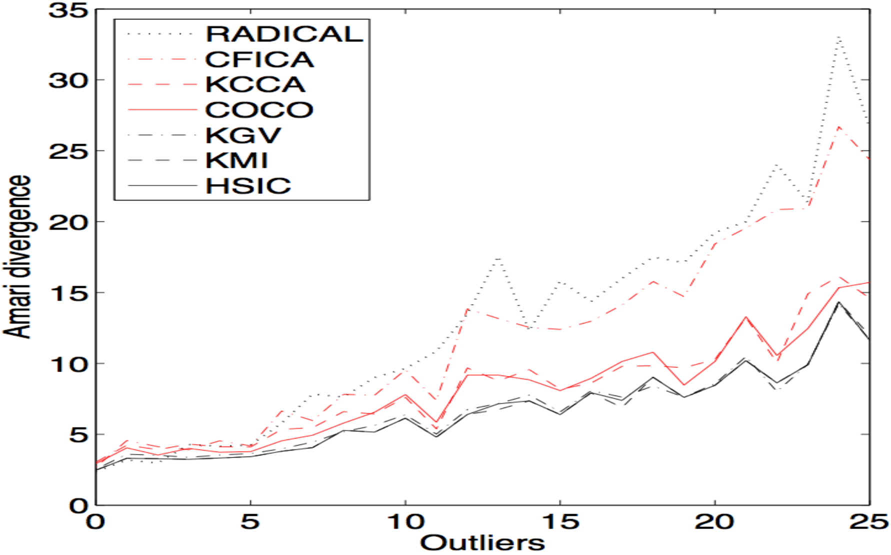
- Kernel matrix could be huge.
- Use reduced-rank representation. We get

$$\operatorname{tr} H(AA^\top)H(BB^\top) = \left\|A^\top HB\right\|^2 \text{ instead of } \operatorname{tr} HKHL.$$

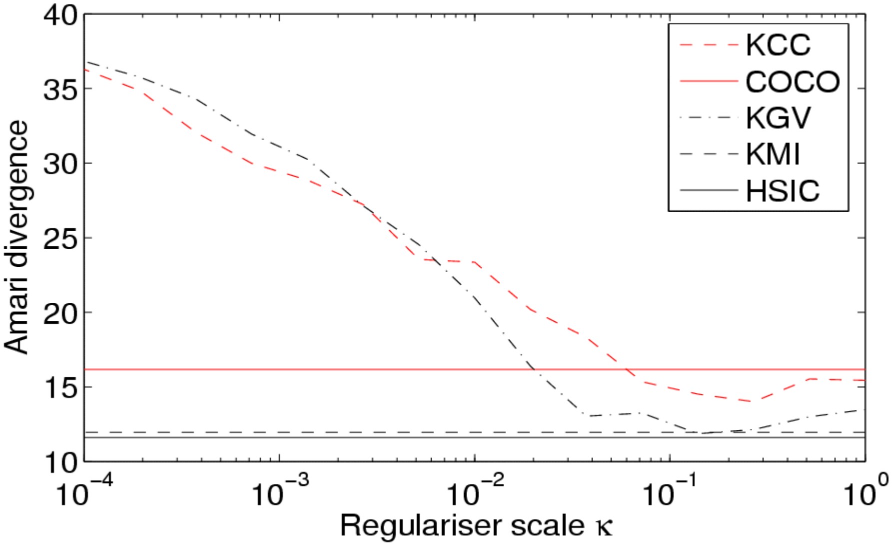
ICA Experiments

n	m	Rep.	FICA	Jade	IMAX	RAD	CFIC	KCC	COg	COl	KGV	KMIg	KMIl	HSICg	HSICl
2	250	1000	10.5 ± 0.4	9.5 ± 0.4	44.4 ± 0.9	**5.4 ± 0.2**	7.2 ± 0.3	7.0 ± 0.3	7.8 ± 0.3	7.0 ± 0.3	**5.3 ± 0.2**	6.0 ± 0.2	5.7 ± 0.2	5.9 ± 0.2	5.8 ± 0.3
2	1000	1000	6.0 ± 0.3	5.1 ± 0.2	11.3 ± 0.6	**2.4 ± 0.1**	3.2 ± 0.1	3.3 ± 0.1	3.5 ± 0.1	2.9 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.4 ± 0.1**
4	1000	100	5.7 ± 0.4	5.6 ± 0.4	13.3 ± 1.1	**2.5 ± 0.1**	3.3 ± 0.2	4.5 ± 0.4	4.2 ± 0.3	4.6 ± 0.6	3.1 ± 0.6	4.0 ± 0.7	3.5 ± 0.7	2.7 ± 0.1	**2.5 ± 0.2**
4	4000	100	3.1 ± 0.2	2.3 ± 0.1	5.9 ± 0.7	**1.3 ± 0.1**	1.5 ± 0.1	2.4 ± 0.5	1.9 ± 0.1	1.6 ± 0.1	1.4 ± 0.1	1.4 ± 0.05	**1.2 ± 0.05**	**1.3 ± 0.05**	**1.2 ± 0.05**
8	2000	50	4.1 ± 0.2	3.6 ± 0.2	9.3 ± 0.9	**1.8 ± 0.1**	2.4 ± 0.1	4.8 ± 0.9	3.7 ± 0.9	5.2 ± 1.3	2.6 ± 0.3	2.1 ± 0.1	**1.9 ± 0.1**	**1.9 ± 0.1**	**1.8 ± 0.1**
8	4000	50	3.2 ± 0.2	2.7 ± 0.1	6.4 ± 0.9	**1.3 ± 0.05**	1.6 ± 0.1	2.1 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	**1.4 ± 0.1**	**1.3 ± 0.05**	**1.4 ± 0.05**	**1.3 ± 0.05**
16	5000	25	2.9 ± 0.1	3.1 ± 0.3	9.4 ± 1.1	**1.2 ± 0.05**	1.7 ± 0.1	3.7 ± 0.6	2.4 ± 0.1	2.6 ± 0.2	1.7 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	**1.3 ± 0.05**	**1.3 ± 0.05**

Outlier Robustness



Automatic Regularization



Mini Summary

Linear mixture of independent sources

- Remove mean and whiten for preprocessing
- Use HSIC as measure of dependence
- Find best rotation to demix the data

Performance

- HSIC is very robust to outliers
- General purpose criterion
- Best performing algorithm (Radical) is designed for *linear* ICA, HSIC is a *general purpose criterion*
- Low rank decomposition makes optimization feasible

Mini Summary

Linear mixture of independent sources

- Remove mean and whiten for preprocessing
- Use HSIC as measure of dependence
- Find best rotation to demix the data

Performance

- HSIC is very robust to outliers
- General purpose criterion
- Best performing algorithm (Radical) is designed for *linear* ICA, HSIC is a *general purpose criterion*
- Low rank decomposition makes optimization feasible

Outline

Feature Selection

The Problem

- Large number of features
- Select a small subset of them

Basic Idea

- Find features such that the distributions $p(x|y = 1)$ and $p(x|y = -1)$ are as different as possible.
- Use a two-sample test for that.

Important Tweak

We can find a similar criterion to measure dependence between data and labels (by computing the Hilbert-Schmidt norm of covariance operator).

Recursive Feature Elimination

Algorithm

- Start with full set of features
- Adjust kernel width to pick up maximum discrepancy
- Find feature which decreases dissimilarity the least
- Remove this feature
- Repeat

Applications

- Binary classification (standard MMD criterion)
- Multiclass
- Regression

Algorithm 1 Feature Selection via Backward Elimination

Input: The full set of features $\mathcal{S}$

Output: An ordered set of features $\mathcal{S}^\dagger$

1: $\mathcal{S}^\dagger \leftarrow \varnothing$

2: **repeat**

3: $\quad \sigma_0 \leftarrow \arg\max_\sigma \ \mathrm{HSIC}(\sigma, \mathcal{S})$

4: $\quad i \leftarrow \arg\max_i \ \mathrm{HISC}(\sigma_0, \mathcal{S} \setminus \{i\}), \ \ i \in \mathcal{S}$

5: $\quad \mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$

6: $\quad \mathcal{S}^\dagger \leftarrow \mathcal{S}^\dagger \cup \{i\}$

7: **until** $\mathcal{S} = \varnothing$

Comparison to other feature selectors

Synthetic Data

Table 1: Classification error (%) after selecting features using BAHSIC and 6 other methods.

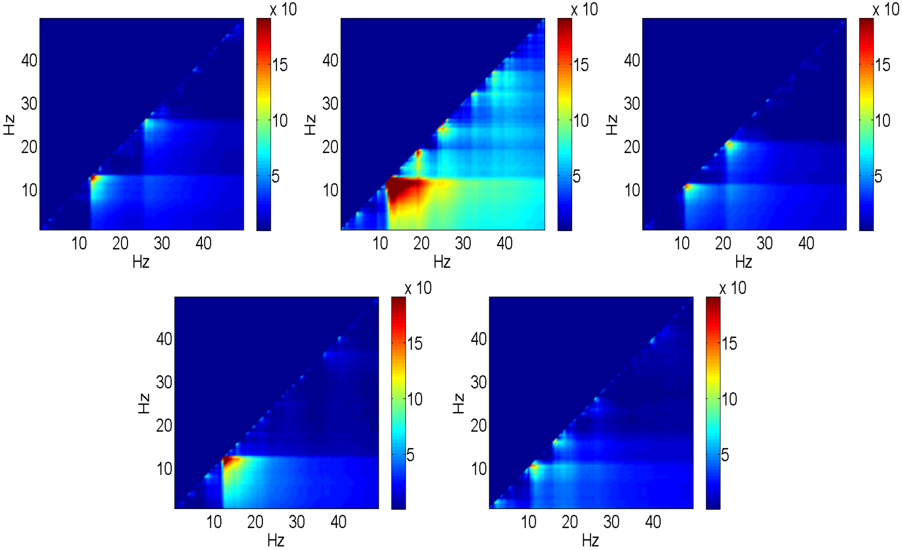
Method	Fisher	FSV	L0	MI	R2W2	RFE	BAHSIC
WL-6	10.0±4.5	*2.0±2.0*	**0.0±0.0**	6.0±3.1	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**
WN-2	57.0±3.7	58.0±5.3	*2.0±1.3*	18.0±2.9	54.0±6.5	*2.0±1.3*	**1.0±1.0**

Brain Computer Interface Data

Table 2: Classification errors (%) on BCI data after selecting a frequency range.

Subject	aa	al	av	aw	ay
CSP(8-40Hz)	17.5±2.5	3.1±1.2	32.1±2.5	7.3±2.7	**6.0±1.6**
CSSP	14.9±2.9	2.4±1.3	33.0±2.7	**5.4±1.9**	*6.2±1.5*
CSSSP	**12.2±2.1**	*2.2±0.9*	*31.8±2.8*	6.3±1.8	12.7±2.0
BAHSIC	*13.7±4.3*	**1.9±1.3**	**30.5±3.3**	*6.1±3.8*	9.0±6.0

Frequency Band Selection



Microarray Feature Selection

Goal

- Obtain small subset of features for estimation
- Reproducible feature selection

Results

Table 3: Comparison between SVM-RFE and BAHSIC for bioinformatics data. From top to bottom: data set description, classification errors in (%), and feature stability.

Dataset	Lymphoma		Yeast		Colon	Berchuck
Dim	4026		79		2000	22283
Sample	42/11/9		121/35/27/14/11		40/22	30/24
	MC	OVR	MC	OVR		
SVM	32.4±6.9	*32.4±6.9*	**5.3±2.1**	*5.8±1.8*	*17.6±5.1*	43.3±6.9
RFE	*12.86±3.30*	**0.00±0.00**	30.36±2.39	6.76±2.10	22.38±6.05	*30.00±7.57*
BAHSIC	**0.00±0.00**	**0.00±0.00**	*5.79 ±1.99*	**4.81±1.59**	**15.71±5.27**	**19.33±6.30**
RFE	0.77±0.09	0.46±0.28	0.41±0.31	0.39±0.32	0.38±0.11	0.57±0.28
BAHSIC	**0.96±0.03**	**0.96±0.03**	**0.82±0.14**	**0.82±0.14**	**0.90±0.06**	**0.73±0.19**

Table 4: Root mean square error (RMSE) of support vector regression with and without HSIC

Method	Sample	Dim	Feature	ϵ-SVR	RAND	BAHSIC
Pyrim	55	27	5	0.112±0.067	0.092±0.073	**0.085±0.066**
Triaz	186	60	2	0.147±0.027	0.157±0.036	**0.144±0.033**
Bodyfat	227	14	7	0.0019±0.0026	0.0019±0.0026	**0.0019±0.0024**

Summary

1. **Measuring Independence**
   - Covariance Operator
   - Hilbert Space Methods
   - A Test Statistic and its Analysis
2. **Independent Component Analysis**
   - ICA Primer
   - Examples
3. **Feature Selection**
   - Problem Setting
   - Algorithm
   - Results

Shameless Plugs

Looking for a job . . . talk to me!

- Alex.Smola@nicta.com.au (http://www.nicta.com.au)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

- **http://sml.nicta.com.au**
- **http://www.kernel-machines.org**
- **http://www.learning-with-kernels.org**
  Schölkopf and Smola: Learning with Kernels