Nonparametric Distribution Testing

Alexander J. Smola
Alex.Smola@nicta.com.au

National ICT Australia
Statistical Machine Learning Program
CSL RSISE, The Australian National University

Thanks to Arthur Gretton, Bernhard Schölkopf, Olivier Bousquet, Vishy Vishwanathan

Outline

Setting

- Property testing for distributions
- Nonparametric strengthening

Disjoint Support

- Classification problem

Independence

- Necessary and sufficient conditions
- RKHS and Covariance Operators
- Empirical Estimation and Uniform Convergence
- Efficient Computation and Experiments

Identity

- Necessary and sufficient conditions
- Banach spaces, norms, and large deviations
- Empirical Estimation and Uniform Convergence

NATIONAL ICT AUSTRALIA

Tests for distributions

Q1: Support

Given $\{x_1, \ldots, x_m\} \sim \Pr(x)$ and $\{y_1, \ldots, y_n\} \sim \Pr(y)$ test whether $\operatorname{supp} \Pr(x) \cap \operatorname{supp} \Pr(y) = \emptyset$.

Q2: Independence

Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ test whether $\Pr(x, y) = \Pr(x) \Pr(y)$.

Q3: Identity

Given $\{x_1, \ldots, x_m\} \sim \Pr(x)$ and $\{y_1, \ldots, y_n\} \sim \Pr(y)$ test whether $\Pr(x) = \Pr(y)$.

Q4: . . .

Difference to CS-style property testing

- Arbitrary domains $\mathcal{X}, \mathcal{Y}$ (in particular, not finite).
- Given sample size.
- Computational time is $\mathbf{not}$ key issue.

NATIONAL ICT AUSTRALIA

Why?

Independent component analysis

- Given a set of n audio signals, find the independent sources of the recording.
- Denoising of EEG data
- Neuroscience, cosmology, ...

Statistical modeling

- Graphical models
- Variable selection

Database merging

- Merge only if distributions in parts are identical
- Combine experiments from different sources

Fraud detection

- Image manipulations
- Fake art

NATIONAL ICT AUSTRALIA

Key Strategy

Simple linear criterion

- "Linear" witness of dependence, disagreement, etc.
- Easy to compute.
- But no sufficient and necessary implications.

Nonlinearization

- Extend to linear function spaces (e.g. RKHS).
- Efficient parameterization for nonparametric test.
- Efficient empirical approximation.

Statistical analysis

- Compute statistical test based on nonlinear criterion (for finite sample size).

NATIONAL ICT AUSTRALIA

Q1: Disjoint Support

Problem

Given $\{x_1, \ldots, x_m\} \sim \Pr(x)$ and $\{y_1, \ldots, y_n\} \sim \Pr(y)$ test whether $\operatorname{supp} \Pr(x) \cap \operatorname{supp} \Pr(y) = \emptyset$.

Linear witness (sufficient)

If for some linear function $f(x) = w^\top x$

$$f(x) < f(y) \text{ for all } x, y$$

the support of $\Pr(x)$ and $\Pr(y)$ is disjoint.

Nonlinear witness (necessary and sufficient)

If there exists some $[0, 1]$-bounded function f such that

$$f(x) < f(y) \text{ for all } x, y$$

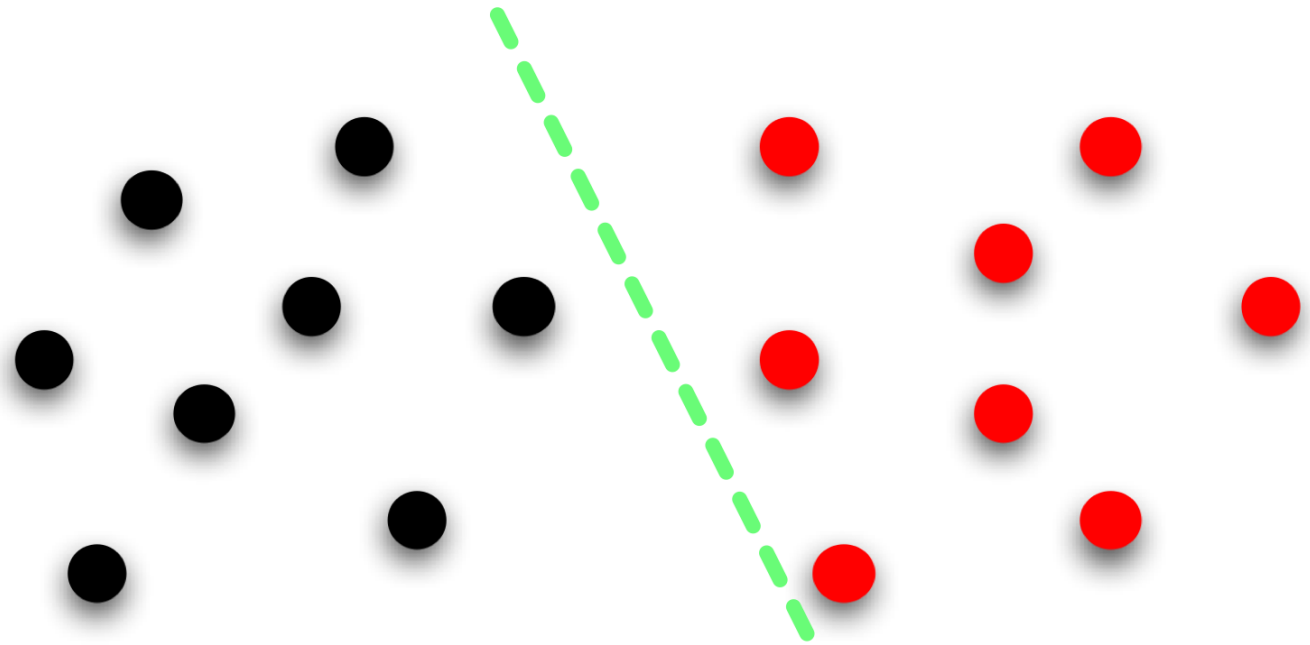
the support of $\Pr(x)$ and $\Pr(y)$ is disjoint.

Finite sample size

We can always find a nonlinear witness!

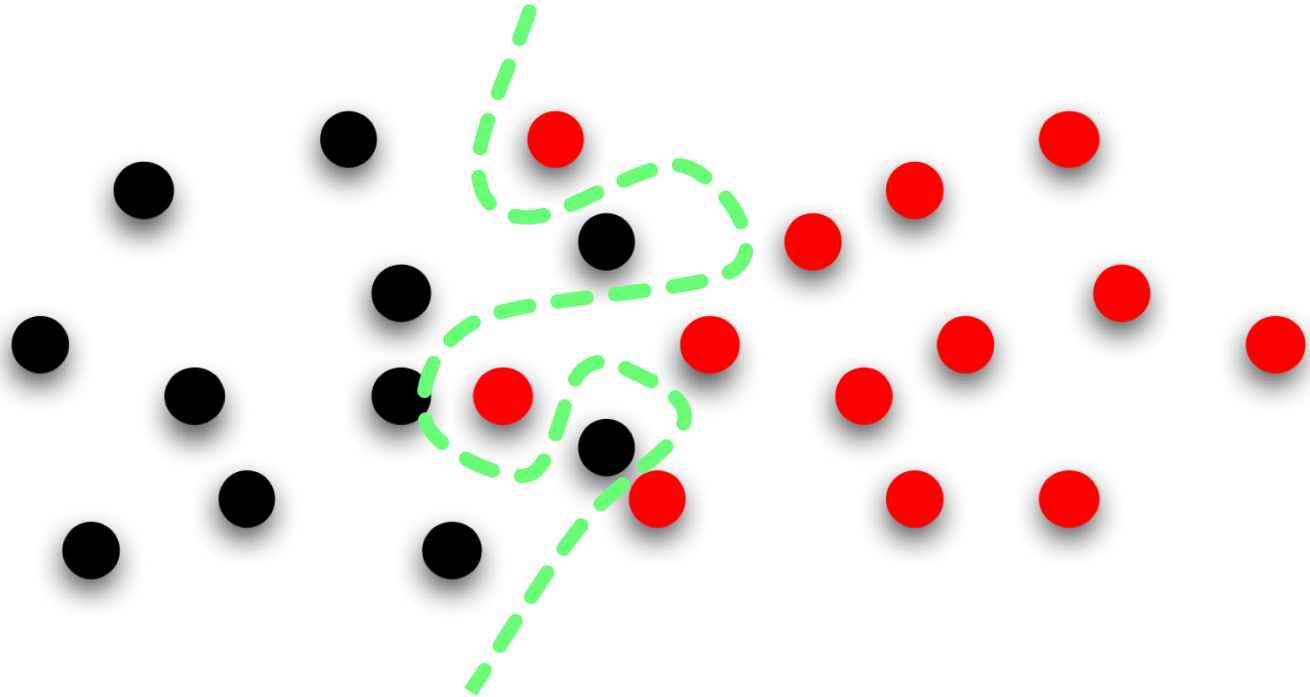
NATIONAL ICT AUSTRALIA

Linear separability



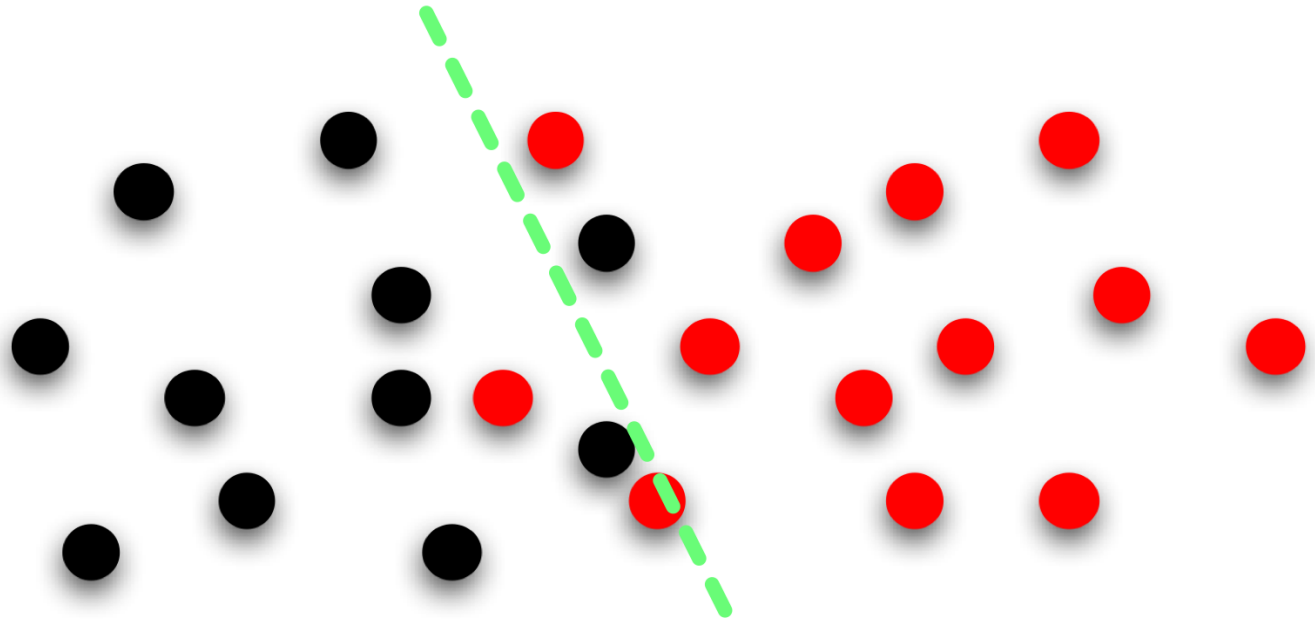
NATIONAL ICT AUSTRALIA

Nonlinear separability?



NATIONAL ICT AUSTRALIA

Nonlinear separability?



NATIONAL ICT AUSTRALIA

Q1: Disjoint Support

Statistical problem

How reliable is our witness?

Binary classification problem

Solve noise-free binary classification problem.

Uniform convergence bounds

Recycle them to test hypothesis of separability.

Efficient computational solution

Use SVM hard-margin optimizer.

NATIONAL ICT AUSTRALIA

Q2: Independence

Problem

Given $\{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \Pr(x, y)$ test whether $\Pr(x, y) = \Pr(x) \Pr(y)$.

Linear witness (sufficient)

If for some linear functions $f(x) = w^\top x$ and $g(y) = v^\top y$ the covariance deviates from 0:

$$\mathrm{Cov}\{f(x), g(y)\} > 0$$

Nonlinear witness (necessary and sufficient)

If for some $[0, 1]$-bounded functions f, g

$$\mathrm{Cov}\{f(x), g(y)\} > 0$$

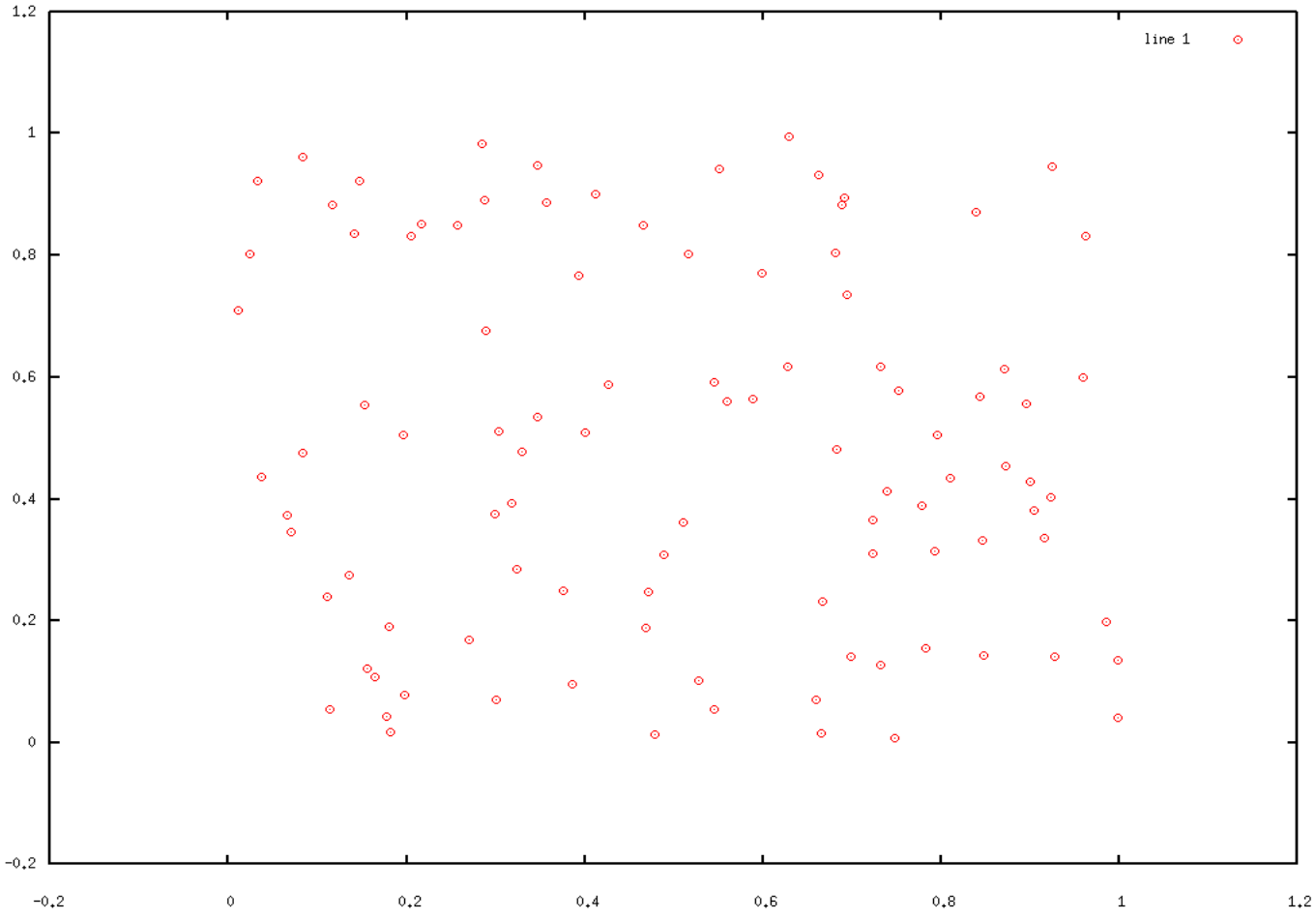
Renyi (1957) proposes this as test for independence.

Finite sample size

We can (almost) always find a nonlinear witness!

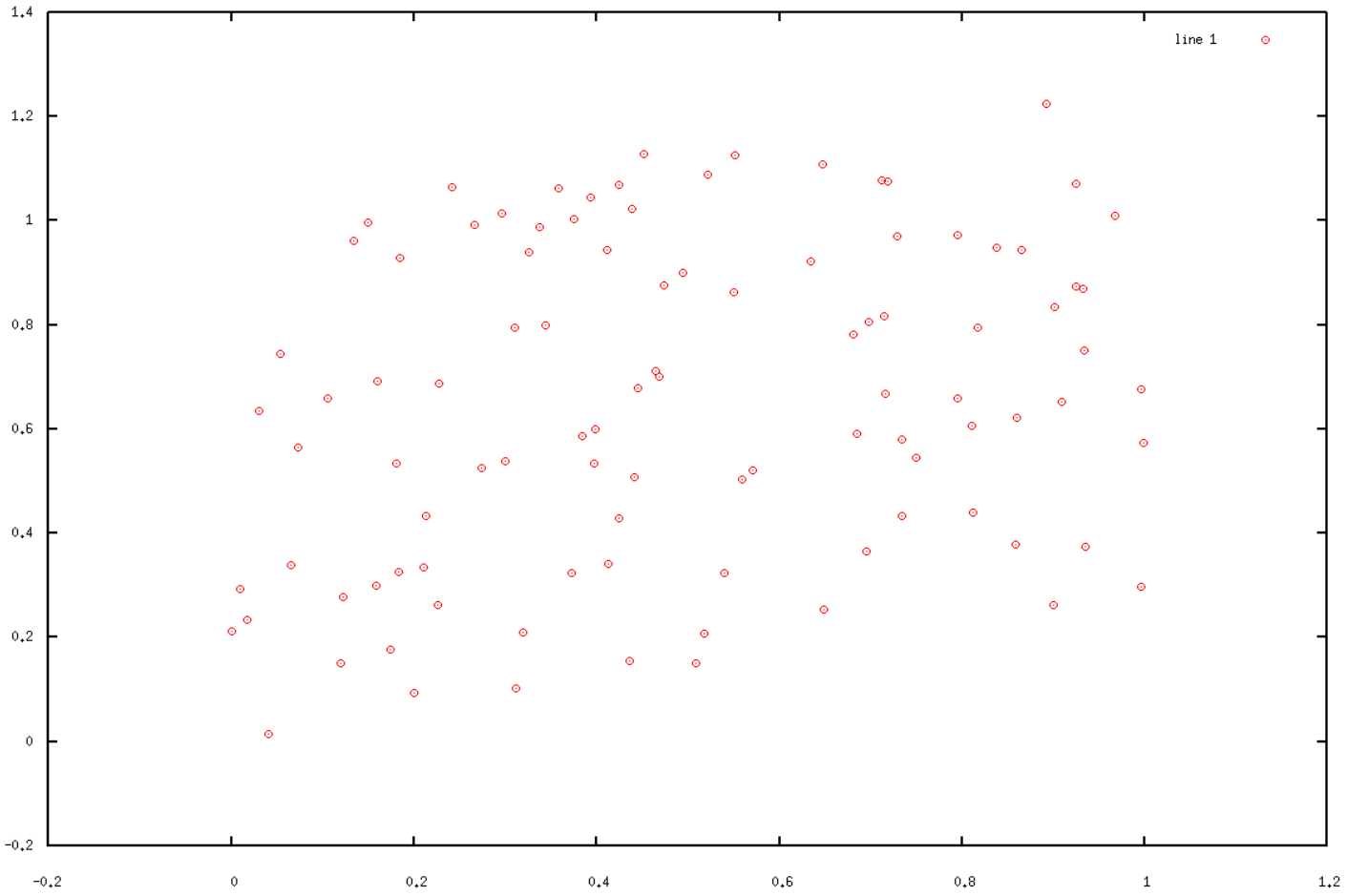
NATIONAL ICT AUSTRALIA

Independent random variables



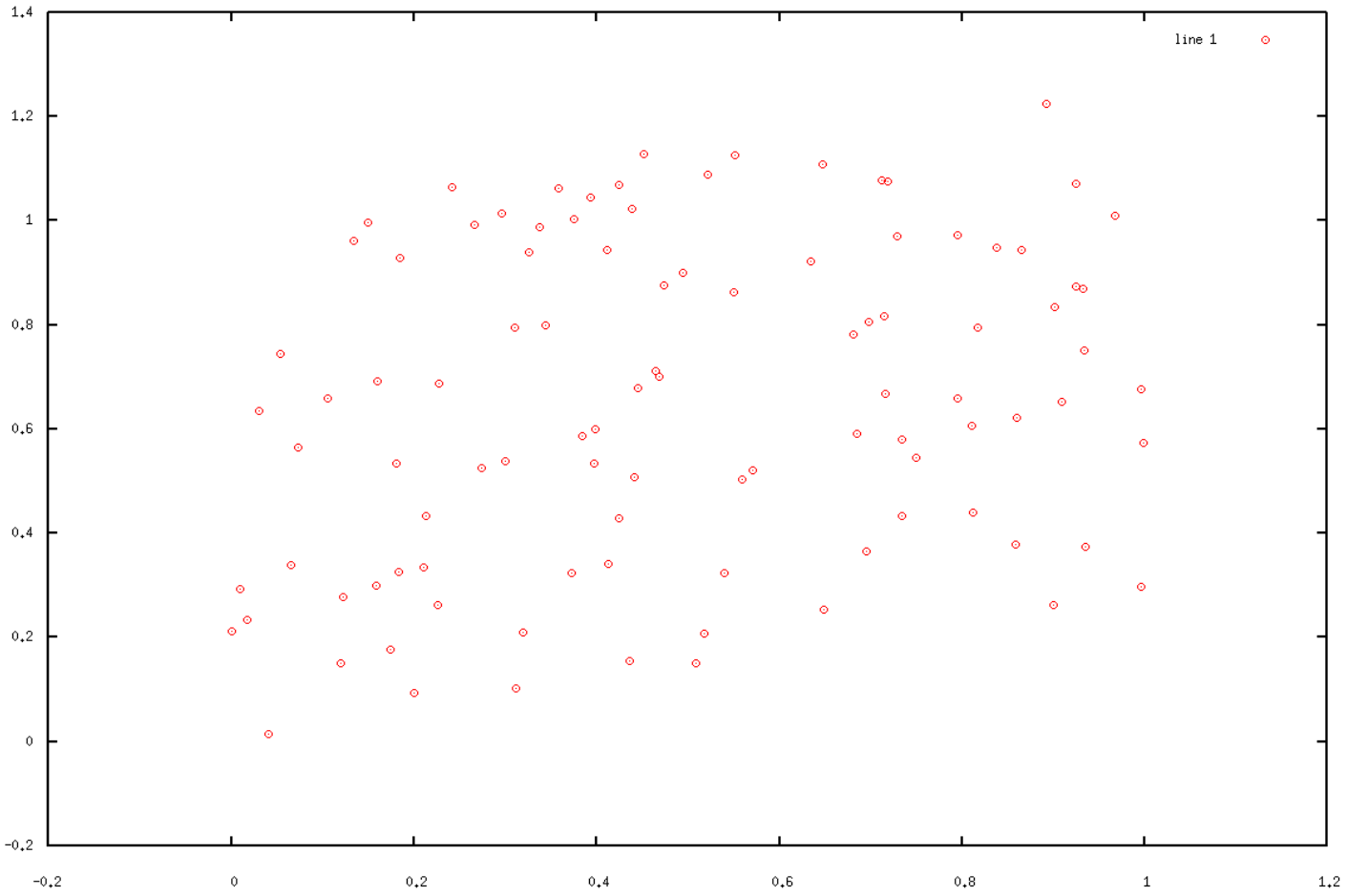
NATIONAL ICT AUSTRALIA

Dependent random variables



NATIONAL ICT AUSTRALIA

Or are we just unlucky?



NATIONAL ICT AUSTRALIA

Covariance operators

Reproducing Kernel Hilbert Space

- Define kernels $k(x, x'), l(y, y')$ on $\mathcal{X}, \mathcal{Y}$ respectively (with associated RKHSs $\mathcal{F}, \mathcal{G}$)
- Evaluation functions $f(x) = \langle f, k(x, \cdot) \rangle$.
- Assume bounded k, l on domain.

Mean operator

Define mean μ_x, μ_y via operators

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

Covariance operator

Define covariance operator C via bilinear form

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

NATIONAL ICT AUSTRALIA

Hilbert Space representation

Theorem

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

Proof sketch

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

Covariance operator

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

Test statistic

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

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

NATIONAL ICT AUSTRALIA

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

Rank-one operators

For rank-one terms we have $\|f \otimes g\|^2 = \|f\|^2 \|g\|^2$.

Joint expectation

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

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

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

NATIONAL ICT AUSTRALIA

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

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

NATIONAL ICT AUSTRALIA

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

Hoeffding's Theorem

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

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

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 sketch: Hoeffding and union bound.

NATIONAL ICT AUSTRALIA

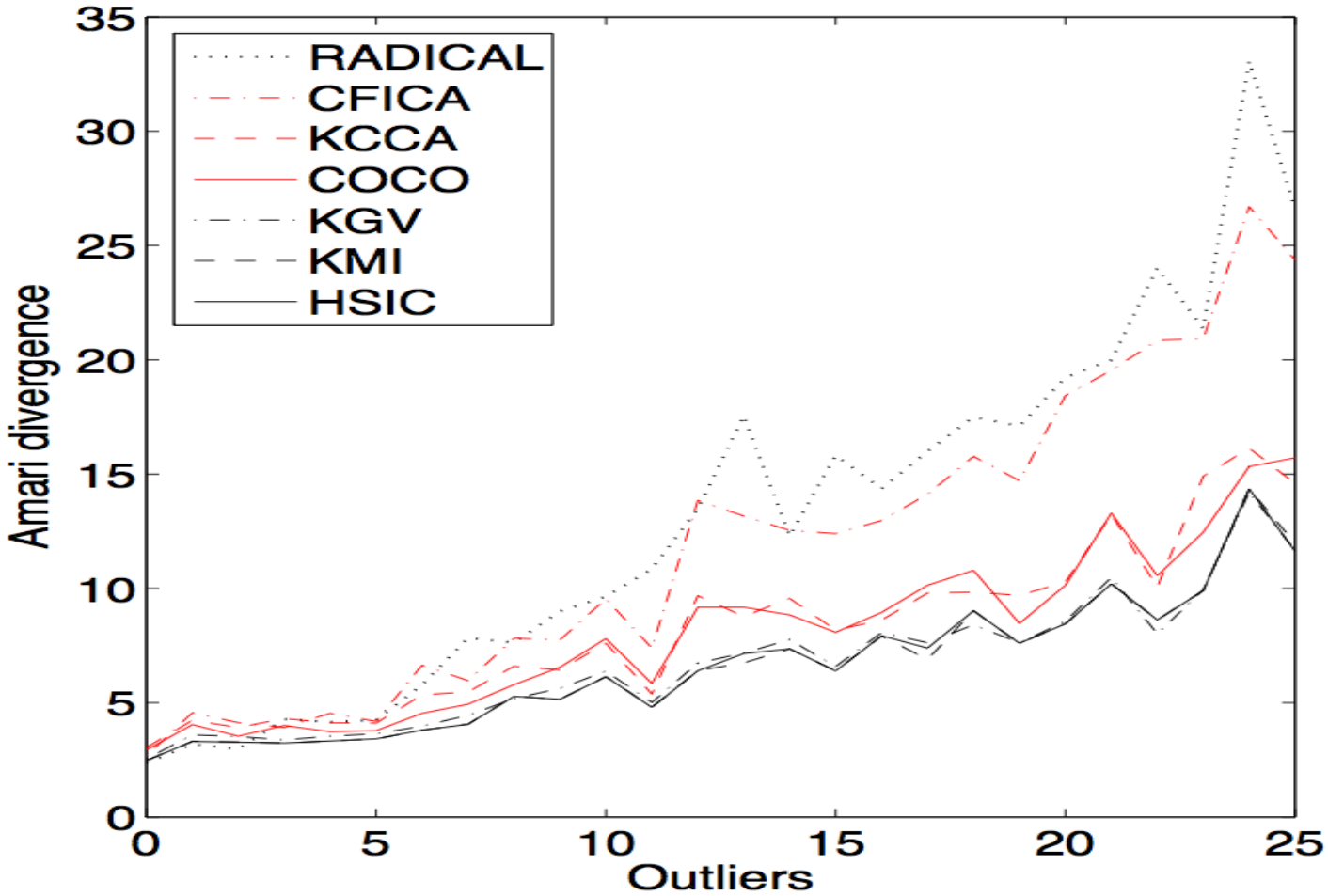
ICA Experiments

n	m	Rep.	FICA	Jade	IMAX	RAD	CFIC	KCC	COg	COI	KGV	KMIg	KMII	HSICg	HSICI
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**

- Linear mixture of independent sources
- Amari divergence of demixing matrix for ICA problem
- Best performing algorithm (Radical) is designed for *linear* ICA, HSIC is a *general purpose criterion*

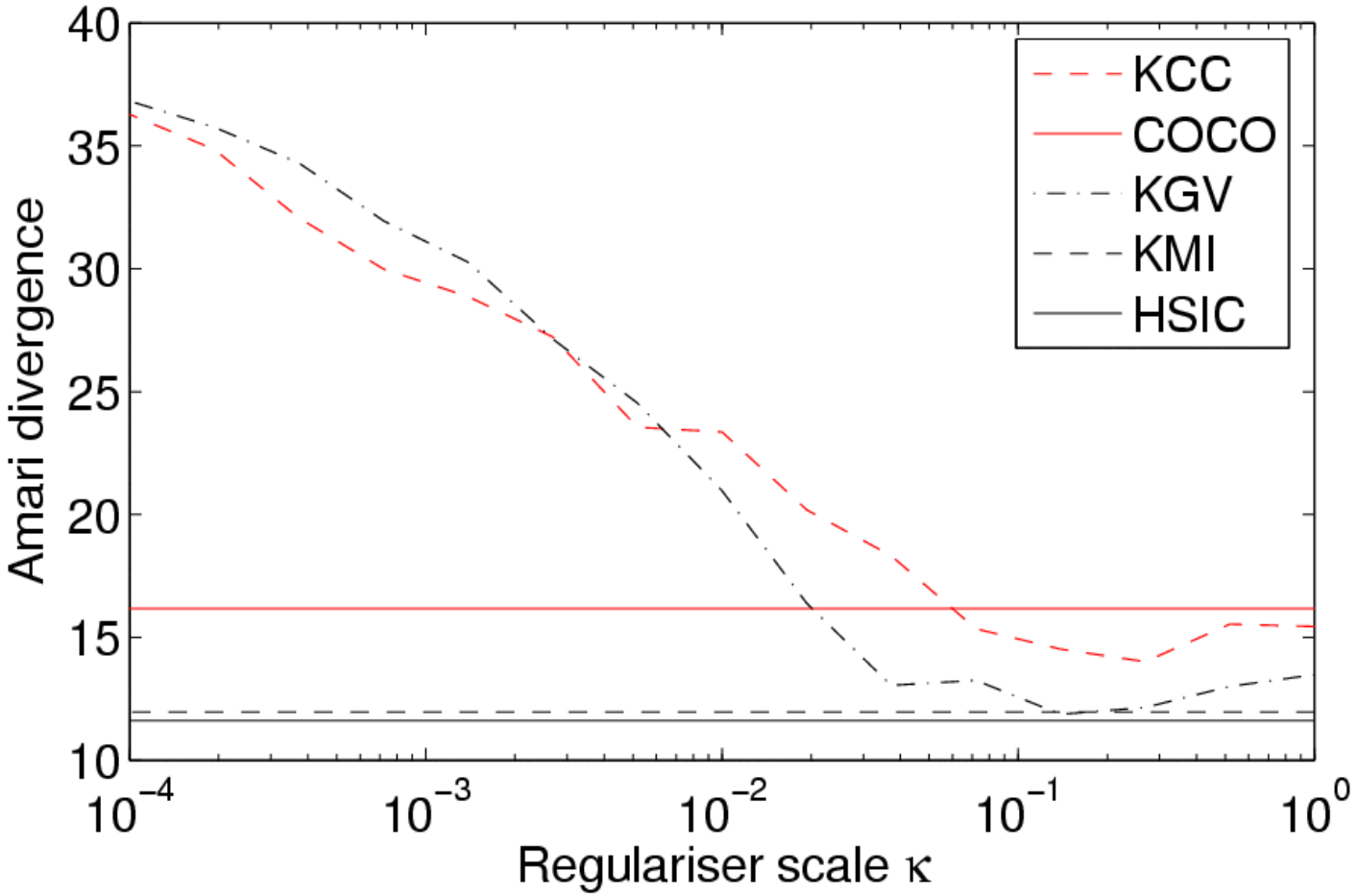
NATIONAL ICT AUSTRALIA

Outlier Robustness



NATIONAL ICT AUSTRALIA

Automatic Regularization



NATIONAL ICT AUSTRALIA

Q3: Identity

Problem

Given $\{x_1, \ldots, x_m\} \sim \Pr(x)$ and $\{y_1, \ldots, y_n\} \sim \Pr(y)$ test whether $\Pr(x) = \Pr(y)$.

Linear witness (sufficient)

If for some linear functions $f(x) = w^\top x$ the means deviate, i.e.

$$\mathbf{E}_x [f(x)] - \mathbf{E}_y [f(y)] > 0$$

Nonlinear witness (necessary and sufficient)

If for some $[0, 1]$-bounded function f

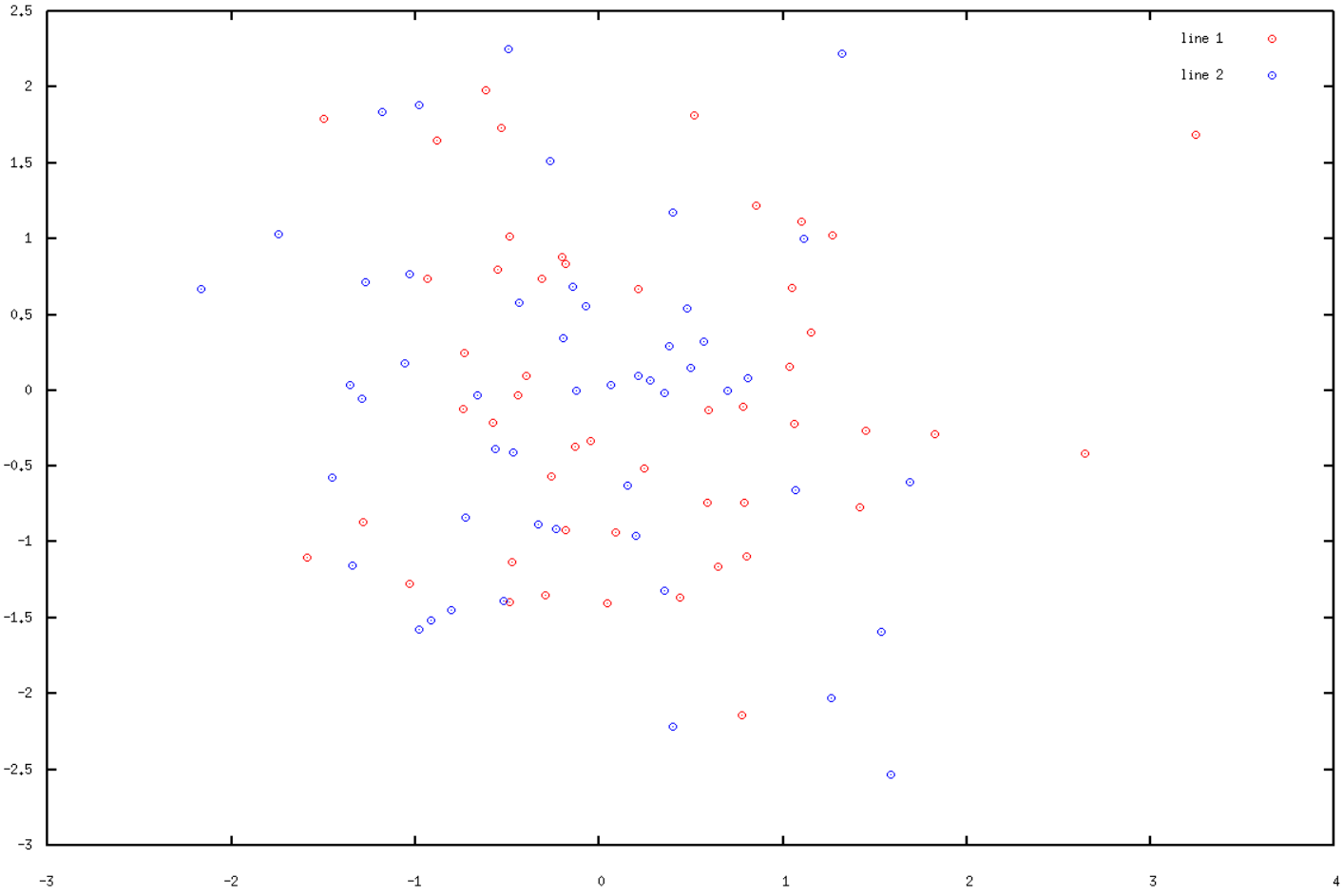
$$\mathbf{E}_x f(x) - \mathbf{E}_y [f(y)] > 0$$

Finite sample size

We can (almost) always find a nonlinear witness!

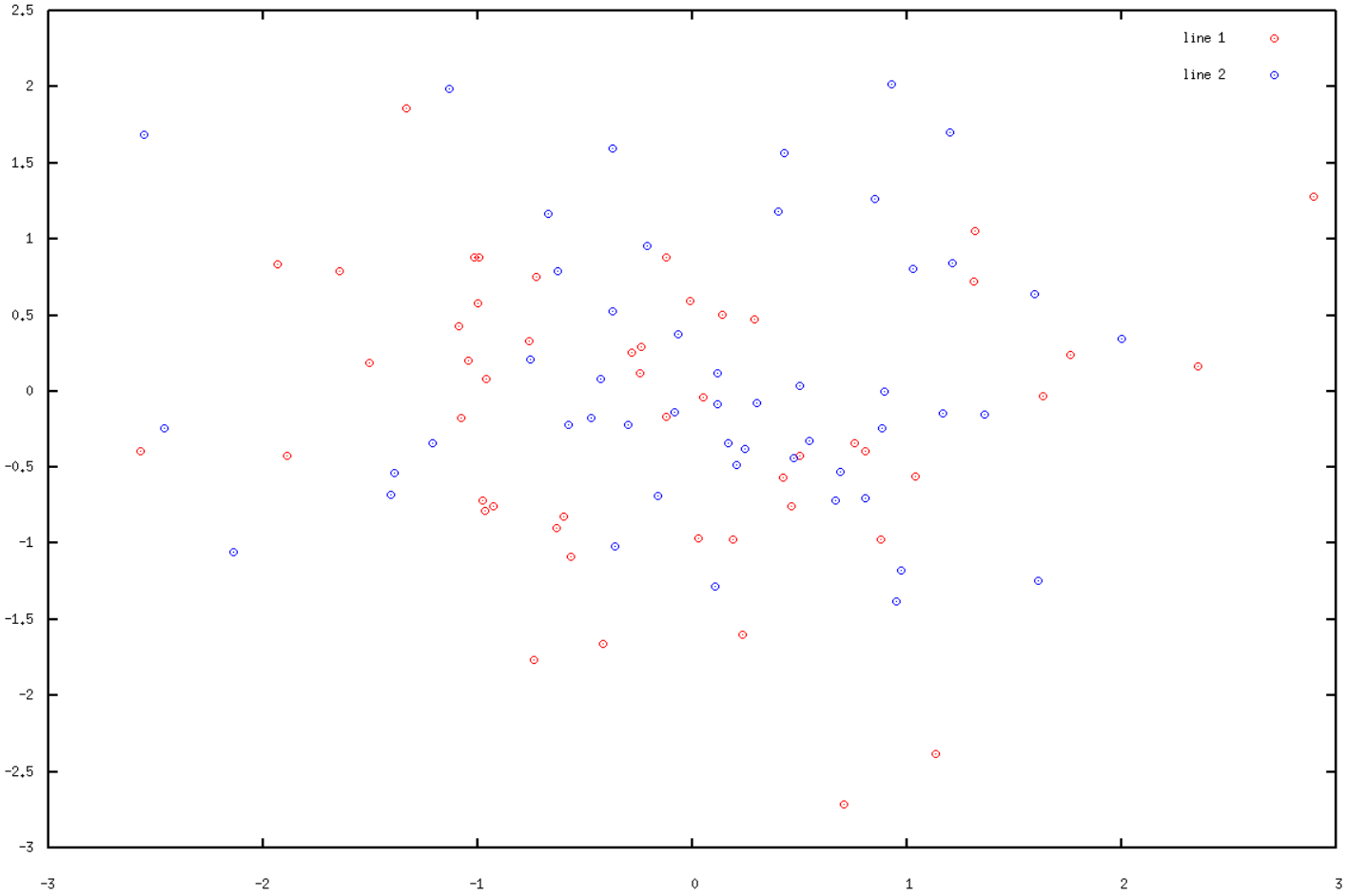
NATIONAL ICT AUSTRALIA

Identical distributions



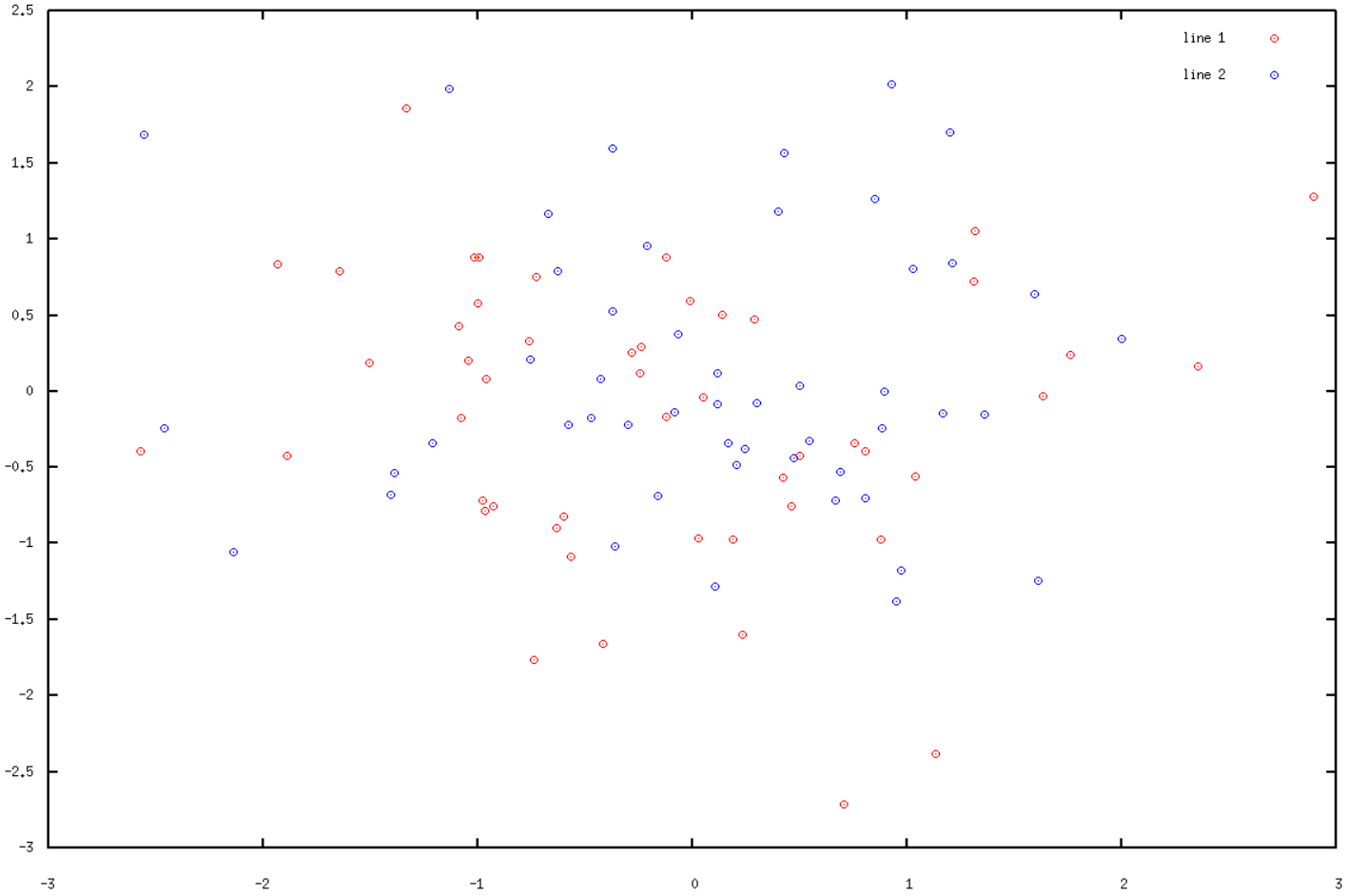
NATIONAL ICT AUSTRALIA

Different distributions



NATIONAL ICT AUSTRALIA

Or are we just unlucky?



NATIONAL ICT AUSTRALIA

Maximum mean discrepancy

Theorem

For $[0,1]$-bounded range functions f the distributions $\Pr_x$ and $\Pr_y$ are identical if $\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] = 0$ for all f.

Proof sketch

Distributions are defined by measure of the sets of the σ-algebra. Hence consider deviations between such sets.

Discrepancy measure

For bounded-range class of functions $\mathcal{F}$ define

$$\mathrm{MMD}(\Pr_x, \Pr_y, \mathcal{F}) := \sup_{f \in \mathcal{F}} [\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]]$$

NATIONAL ICT AUSTRALIA

Empirical estimates and Banach spaces

Empirical criterion

$$\mathrm{MMD}(X, Y, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_i f(x_i) - \frac{1}{n} \sum_i f(y_i) \right]$$

Unit balls in Banach space

Denote by $\mathcal{F}$ the unit ball of functions $f : \mathcal{X} \to \mathbb{R}$ in a Banach space $\mathcal{B}$. Then we have

$$\mathrm{MMD}(X, Y, \mathcal{F}) = \left\| \frac{1}{m} \sum_i \phi(x_i) - \frac{1}{n} \sum_i \phi(y_i) \right\|$$

where $\phi(x)$ are evaluation functionals in the dual Banach space $\mathcal{B}^*$, i.e. $f(x) = \langle f, \phi(x) \rangle$.

Proof sketch

$$\sup_{\|f\| \leq 1} f(x) = \|\phi(x)\|_{\mathcal{B}^*}$$

NATIONAL ICT AUSTRALIA

Computing it

Hilbert Spaces

We exploit that $\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$ is well defined. This yields

$$\left\| \frac{1}{m} \sum_i \phi(x_i) - \frac{1}{n} \sum_i \phi(y_i) \right\|^2$$
$$= m^{-2} \sum_{i,j} k(x_i, x_j) + n^{-2} \sum_{i,j} k(y_i, y_j) - 2m^{-1}n^{-1} \sum_{i,j} k(x_i, y_j)$$

Even cheaper approximate schemes are possible (subsample $O(n)$ terms).

Connection to Parzen windows

For $k(x, x')$ being the inner product between Parzen windows kernel functions this is the L_2 distance between density estimates.

NATIONAL ICT AUSTRALIA

Concentration of measure

Large deviation bound

$$\begin{aligned}&\Pr\left[|\mathrm{MMD}(X,Y,\mathcal{F}) - \mathbf{E}_{X,Y}\mathrm{MMD}(X,Y,\mathcal{F})| > \epsilon\right]\\ \leq &2\exp\left(-\frac{mn\epsilon^2}{2(m+n)R^2}\right)\end{aligned}$$

where $\|\phi(x)\| \leq R$.

Proof sketch

Use McDiarmid on centered norm. Swapping one summand changes terms by $2R/m$ or $2R/n$ depending on which term is removed.

NATIONAL ICT AUSTRALIA

Value of norm discrepancy

General case

Norm discrepancy can be bounded by Rademacher average for associated function class.

Hilbert space case

By Jensen's inequality we have

$$\begin{aligned}
&\mathbf{E}\left[\left\|\frac{1}{m}\sum_i \phi(x_i) - \frac{1}{n}\sum \phi(y_i)\right\|\right]^2 \\
\leq &\mathbf{E}\left[\left\|\frac{1}{m}\sum_i \phi(x_i) - \frac{1}{n}\sum \phi(y_i)\right\|^2\right] \\
= &\frac{m+n}{mn}\left(\mathbf{E}_x k(x,x) - \mathbf{E}_{x,x'} k(x,x')\right)
\end{aligned}$$

The norm discrepancy vanishes with $O(\sqrt{(m+n)/mn})$.

NATIONAL ICT AUSTRALIA

Consequences

Sensitivity of test

Can detect $O(1/\sqrt{m})$ deviation from identity

Tightness

Rate is tight, as can be seen when distinguishing two Gaussians with same variance: the empirical means will deviate with variance $(1/n + 1/m) \operatorname{tr} \Sigma$.

KL-divergence

For exponential families this also bounds the KL-divergence.

NATIONAL ICT AUSTRALIA

Summary

- Simple linear test for property of distribution
- Simple algorithm associated with it
- Strengthen test by use of nonlinear functions
- In Hilbert space this is still easy to do
- Examples
  - Support of distributions
  - Independence
  - Identity
- Applications
  - ICA
  - verification and fraud detection
  - Database merging

NATIONAL ICT AUSTRALIA

Shameless Plugs

We are hiring. For details contact

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

Machine Learning Summer School

- Canberra, February 6-17, 2005
- http://canberra06.mlss.cc

NATIONAL ICT AUSTRALIA