Adaptive Large-Scale Kernel Two-Sample Testing

Heiko Strathmann
joint work with Arthur Gretton, Bharath Sriperumbudur, Dino Sejdinovic, Massimiliano Pontil and others

Gatsby Computational Neuroscience Unit, UCL London

April 8, 2013

Contents

Adaptive Large-Scale Kernel Two-Sample Testing

Given two probability distributions p, q on a domain $\mathcal{X}$, and two finite sets of iid samples drawn from them

$$X = \{x_i\}_{i=1}^n \sim p \qquad Y = \{y_j\}_{j=1}^m \sim q,$$

can we decide whether $p \neq q$ with high confidence?

Example: t-test – do the confidence intervals of univariate Gaussians overlap?

Adaptive Large-Scale Kernel Two-Sample Testing

Two-sample testing in high dimensional distributions with complex structure:

- Classical methods are often not feasible. [Gretton et al., 2012a]
- Strings/text-data (Websites, primary structure), Graphs (Protein networks), etc
- Mapping to vector space needed. Kernels do this implicitly. E.g. SVM

We also do not want to make assumptions to p and q (like t-test).
$\Rightarrow$ Kernels allow this kind of non-parametric two-sample testing.

Adaptive Large-Scale Kernel Two-Sample Testing

Kernel selection:

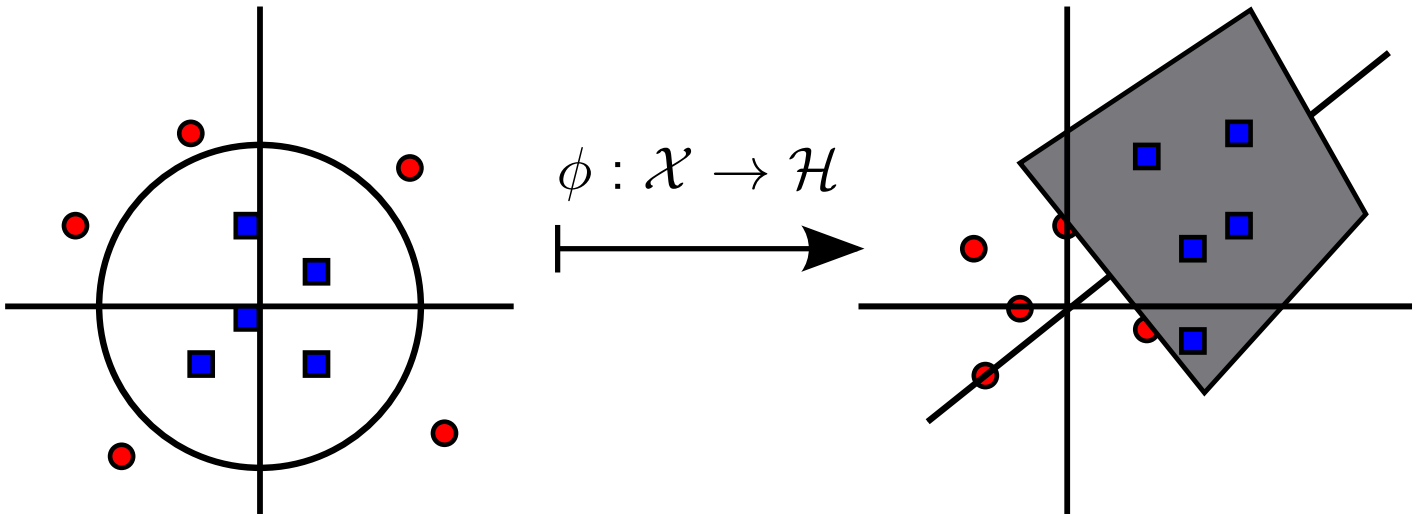
- How to choose the kernel?
- Obviously depends on distributions p and q.
- Usually, performance critically depends on this choice.
- Try to learn from data.
- Difficult problem, no general solution. Naive methods are expensive (cross-validation)

Adaptive Large-Scale Kernel Two-Sample Testing

Large-Scale methods:

- We consider the big-data setting where "infinite" amounts of data are available. Is it possible to exploit this?
- Cannot store data – need a method that touches every datum only once (streaming).

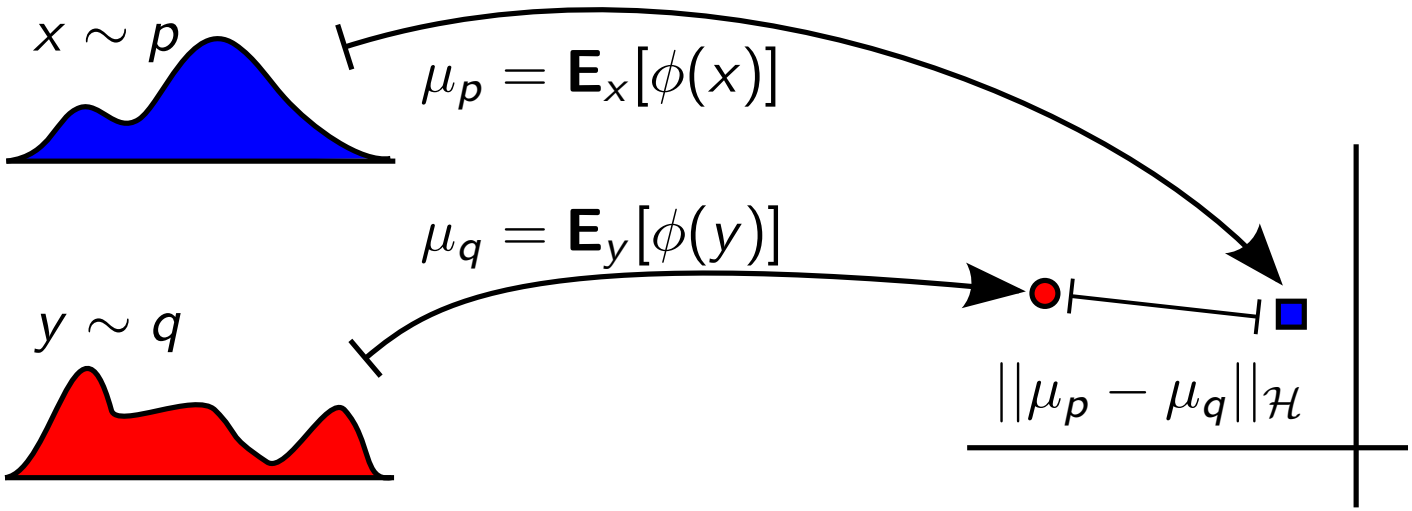
Reproducing Kernel Hilbert Spaces – The Classic Slide



Positive semi-definite kernel: $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.
Feature itself: $\phi(x) = k(x, \cdot)$ is a real-valued function in $\mathcal{H}$.
Reproducing property: $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \qquad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}$

Reproducing Kernel Hilbert Spaces – Mean Embeddings

Let's do this with probability distributions!



Mean embedding $\mu_p \in \mathcal{H}$ satisfies: $\langle f, \mu_p \rangle_{\mathcal{H}} = \mathbf{E}_x f(x)$

Maximum Mean Discrepancy

Definition (and 2 Lemmas)

Let $\mathcal{F}$ be a unit ball in $\mathcal{H}$, and x, y be random variables with attached probability distributions p, q respectively. Define the kernel Maximum Mean Discrepancy as

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F}, p, q] &:= \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]\right)^2 \\ &= ||\mu_p - \mu_q||^2_{\mathcal{H}} \\ &= \mathbf{E}_{x,x',y,y'}\left[k(x,x') + k(y,y') - k(x,y') - k(x',y)\right]\end{aligned}$$

where x', y' are independent copies of x, y.

MMD is a metric

Lemma

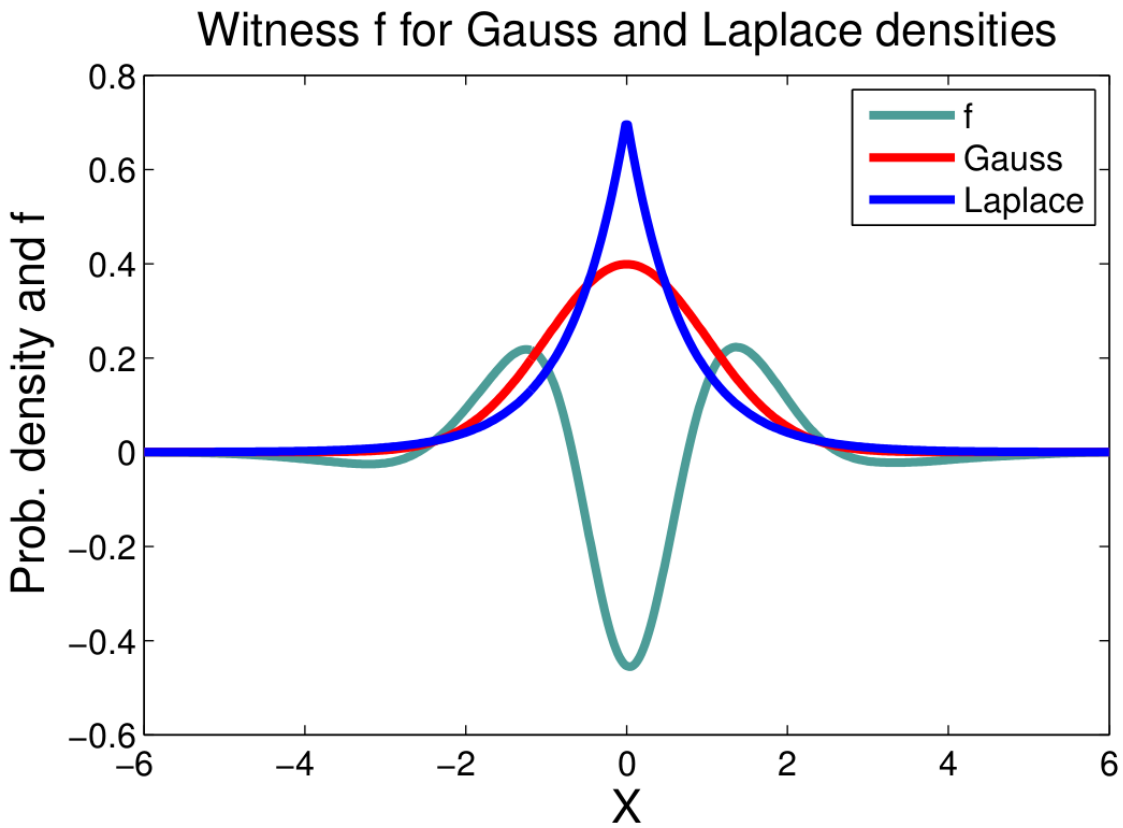
Under certain conditions ($\mathcal{H}$ is universal),

$$\text{MMD}[\mathcal{F}, p, q] = ||\mu_p - \mu_q||_{\mathcal{H}} = 0 \qquad \Leftrightarrow \qquad p = q$$

Consequence: Any pair p, q can be distinguished (!)

Example: RKHS of $k(x, y) = \exp\left(-\frac{||x-y||_2}{2\ell^2}\right)$ is universal.

A 1D-Example of Subtle Differences



MMD Estimates

Given data X, Y with $|X| = |Y| = m$, a quadratic time estimate is

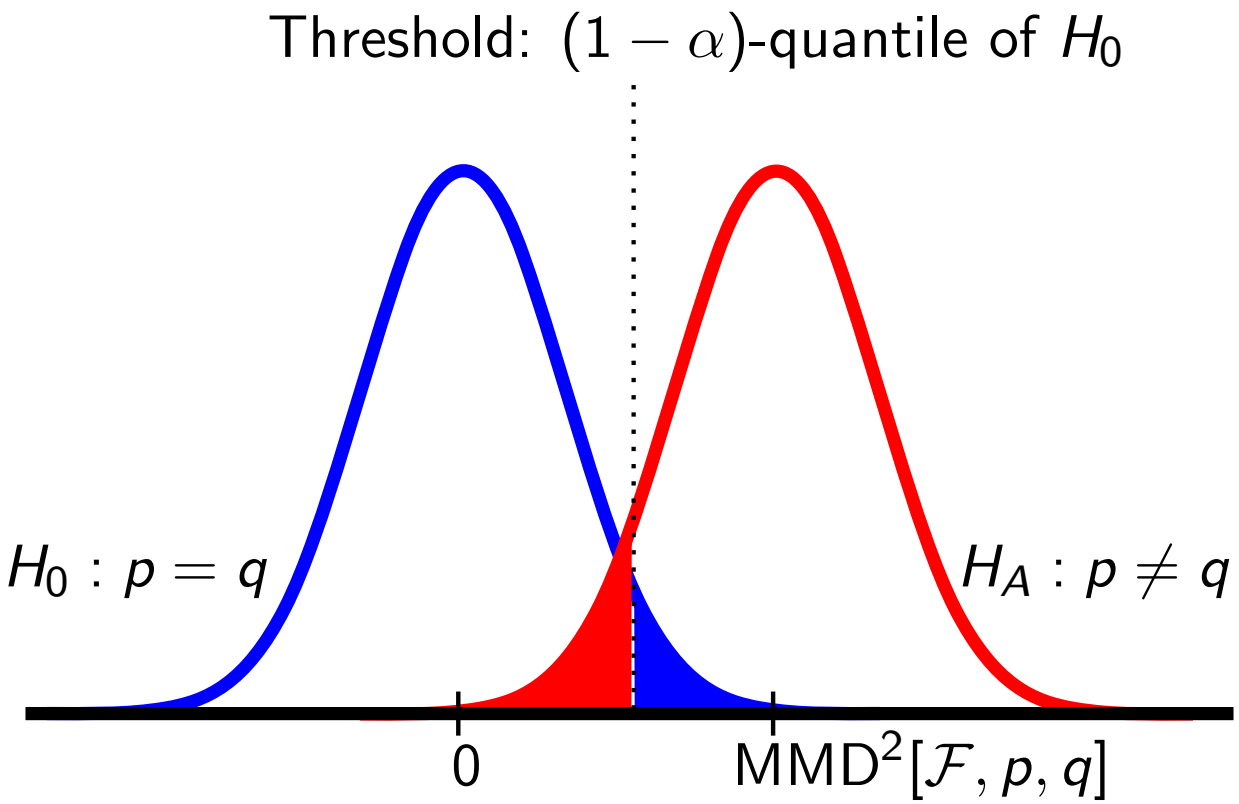
$$\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

Powerful test, null distribution is complicated. An alternative is the linear time estimate: divide data into two halfs and compute

$$\frac{2}{m} \sum_{i=1}^{\frac{m}{2}} k(x_{2i}, x_{2i-1}) + k(y_{2i}, y_{2i-1}) - k(x_{2i}, y_{2i-1}) - k(x_{2i-1}, y_{2i})$$

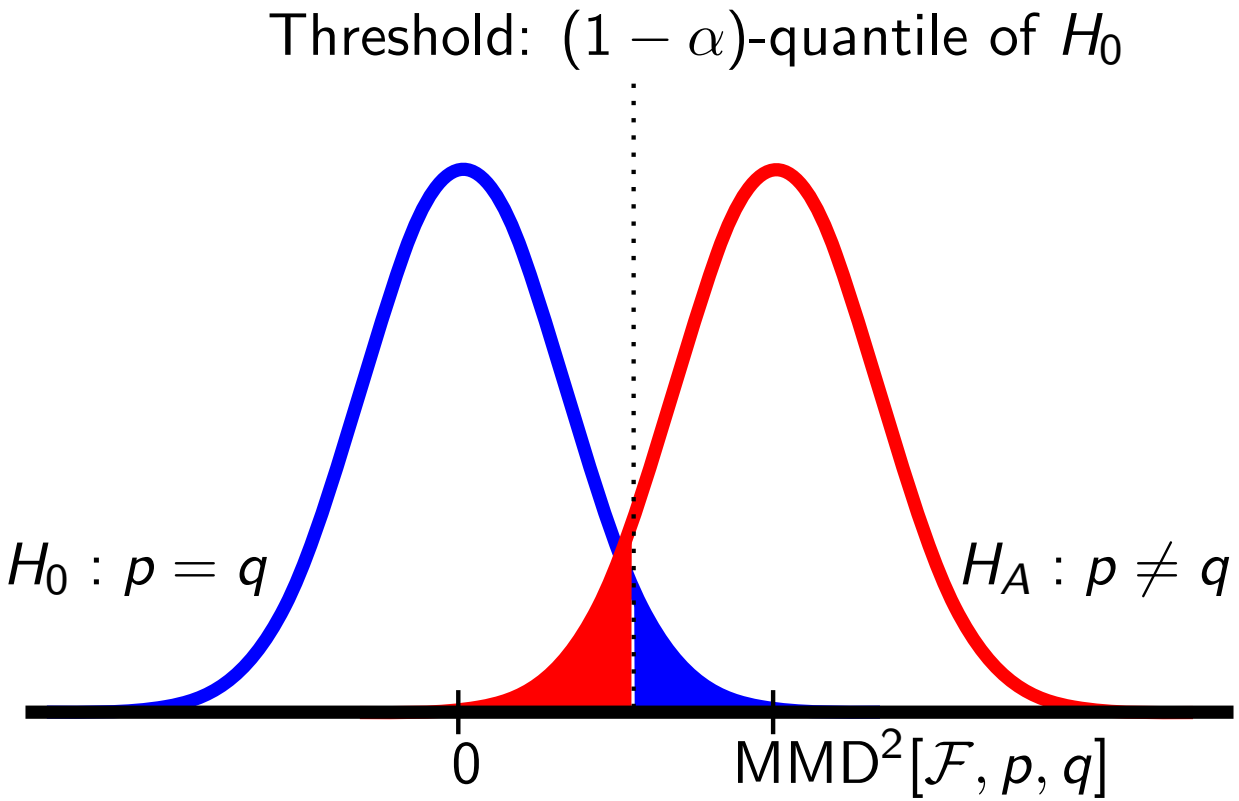
Convenient properties: Possible to stream, null distribution is Gaussian!

Constructing a Linear Time Two-Sample Test



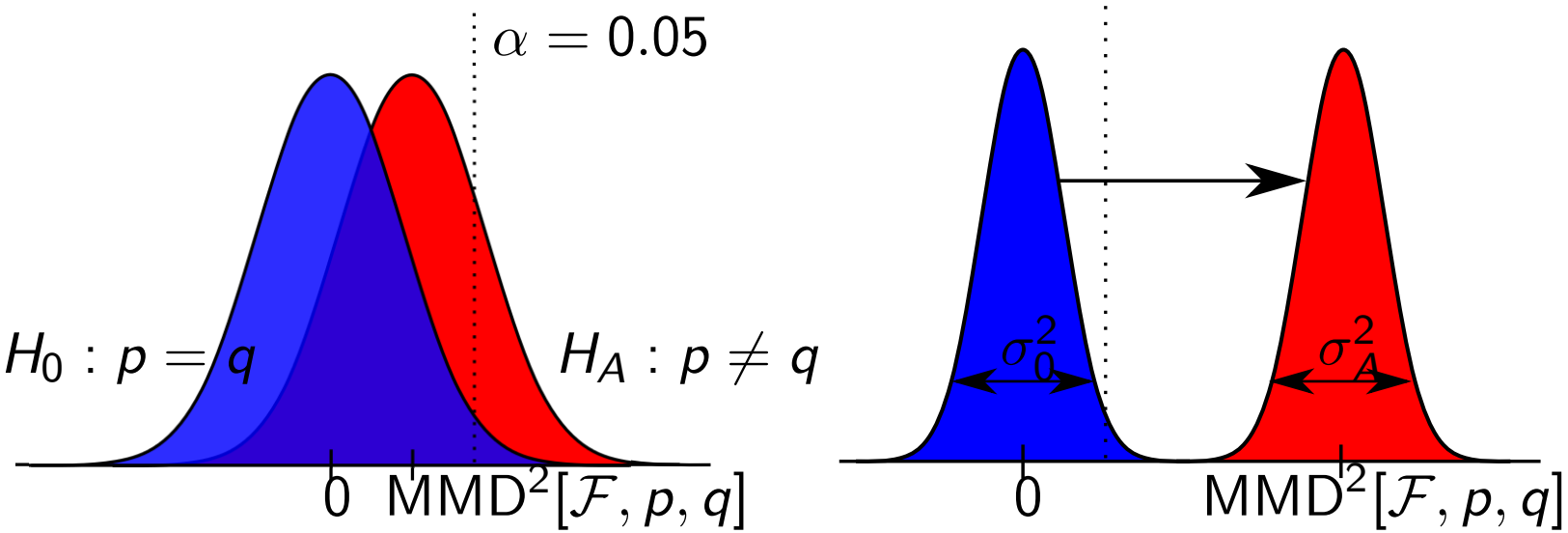
For a fixed type I error (say $\alpha = 0.05$), compute threshold using Gaussian CDF: $t_\alpha = \Phi_0^{-1}(1-\alpha)$.

What to Optimise?



For a fixed type I error (say $\alpha = 0.05$), we want a low type II error.

Optimal Kernel Selection: Idea



Choose kernel to maximise MMD while minimising its variance.
Convenient property of linear time MMD estimate: $\sigma_0^2 = \sigma_A^2$.

Optimal Kernel Selection: Criterion

For a given set of kernels $\mathcal{K}$, compute estimates of MMD $\hat{\eta}$ and its standard deviation $\hat{\sigma}$ and choose kernel such that

$$\hat{k}^* = \arg\sup_{k \in \mathcal{K}} \frac{\hat{\eta}}{\hat{\sigma} + \lambda}$$

where $\lambda > 0$ is a small number. Estimate for σ possible in linear time.

Details and convergence result in [Gretton et al., 2012b].

MMD for Combined Kernels

$$\mathcal{K} = \left\{ k = \sum_{i=1}^{d} \beta_i k_i \quad \text{where} \quad \sum_{i=1}^{d} \beta_i \leq D \text{ and } \beta_i \geq 0 \text{ for } 1 \leq i \leq d \right\}$$

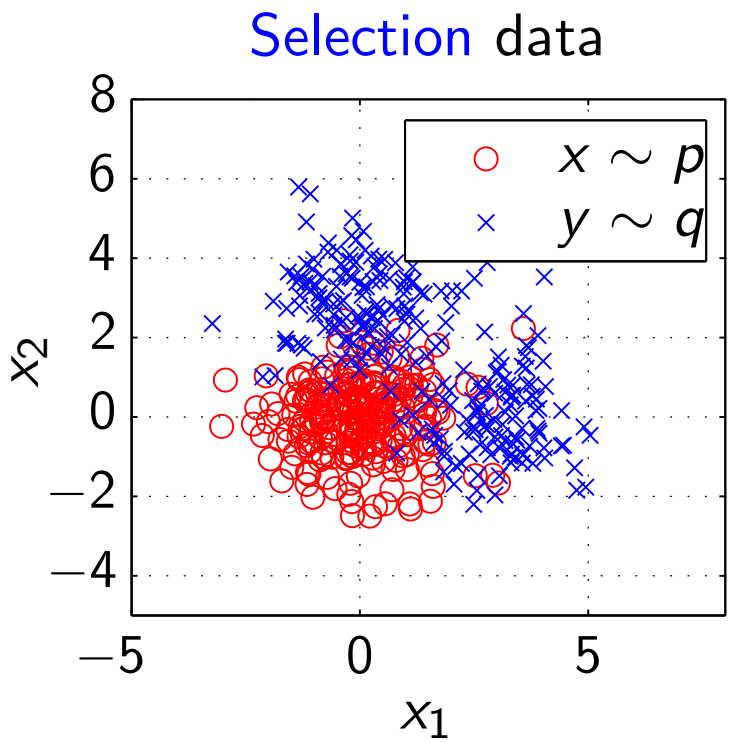
The combined MMD simply is $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^T$. Compute estimates of MMD $\hat{\boldsymbol{\eta}}$ and its covariance $\hat{Q}$. Choose weights:

$$\boldsymbol{\beta}^* = \min\{\boldsymbol{\beta}^T(\hat{Q} + \lambda I)\boldsymbol{\beta} : \boldsymbol{\beta}^T \hat{\boldsymbol{\eta}} = 1, \boldsymbol{\beta} \succeq 0\}$$

Estimate for covariance Q possible in linear time.

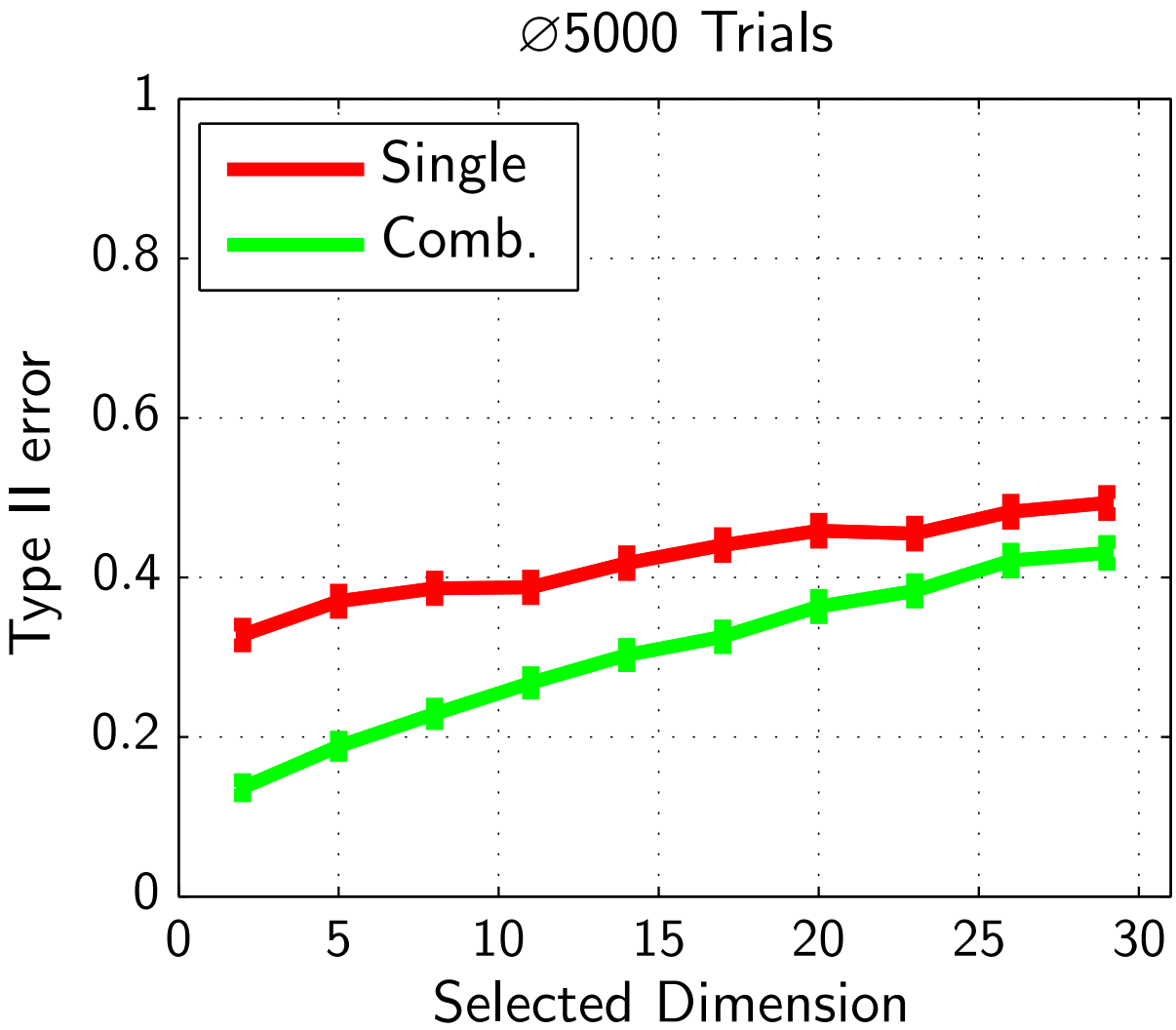
Combined Kernels: Usefulness

Only the first two dimensions are relevant. Rest are equal noise.

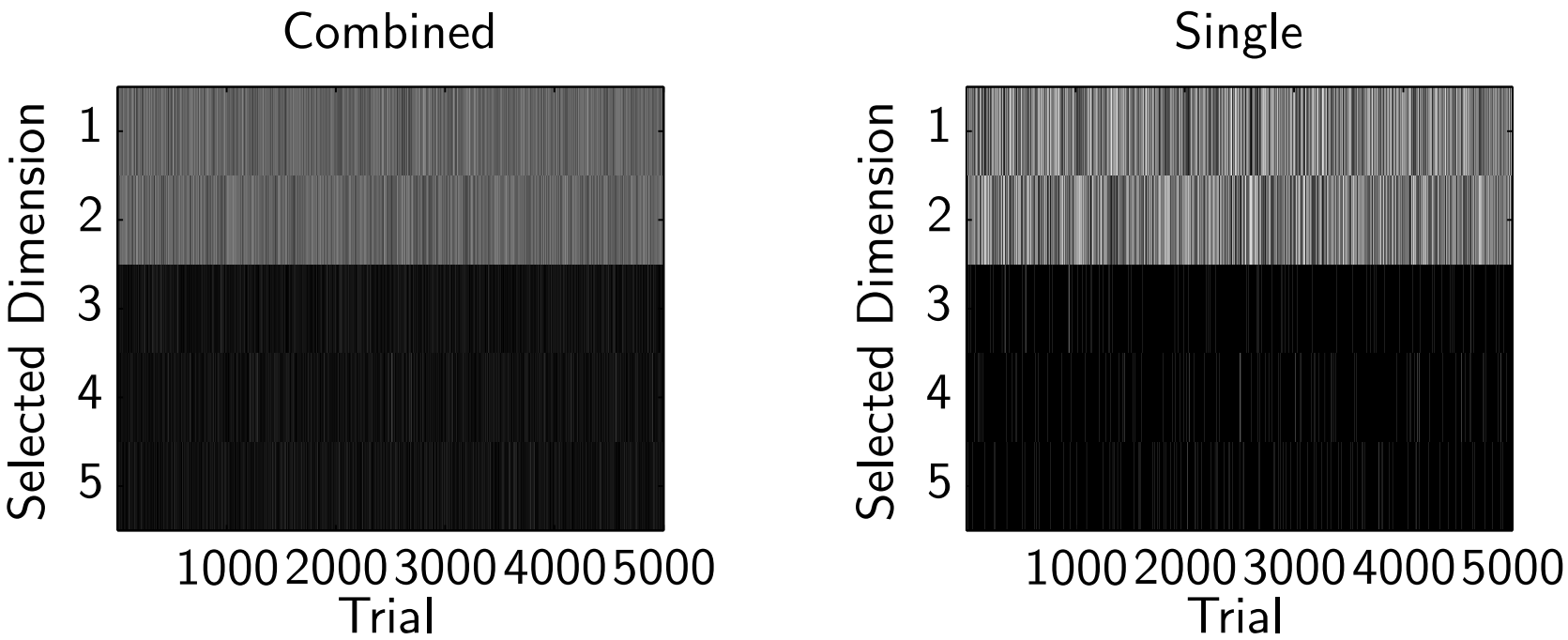


One univariate kernel (fixed) per dimension: feature selection.

Combined Kernels: Results – Type II Errors



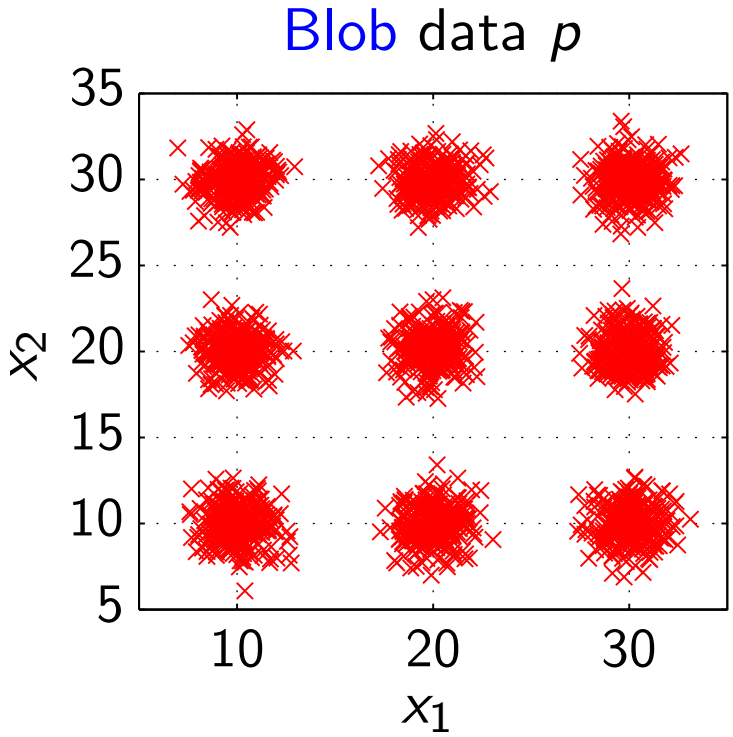
Combined Kernels: Results – Kernel Weights

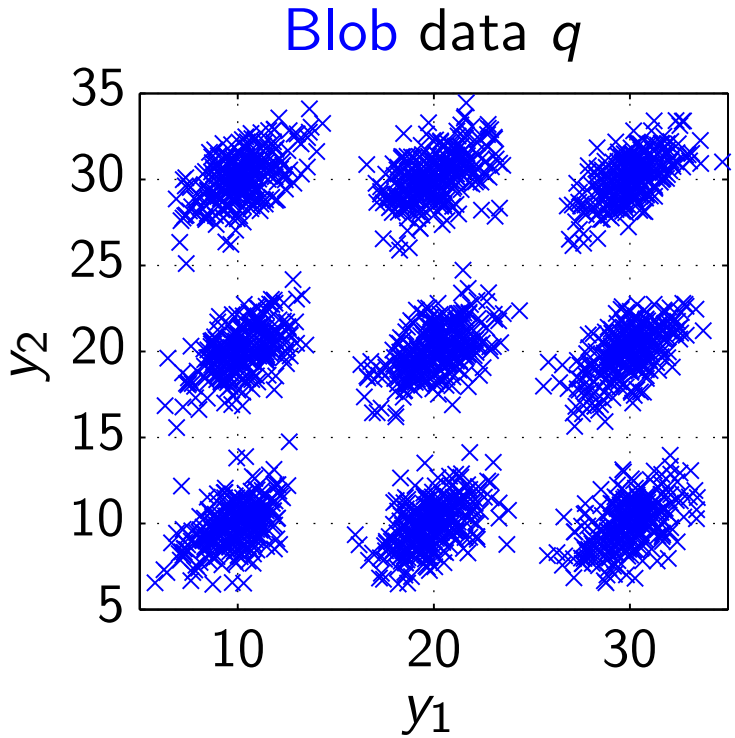


Combines kernels in a meaningful way. Useful for biological data?

A Very Hard Problem: Gaussian Blobs

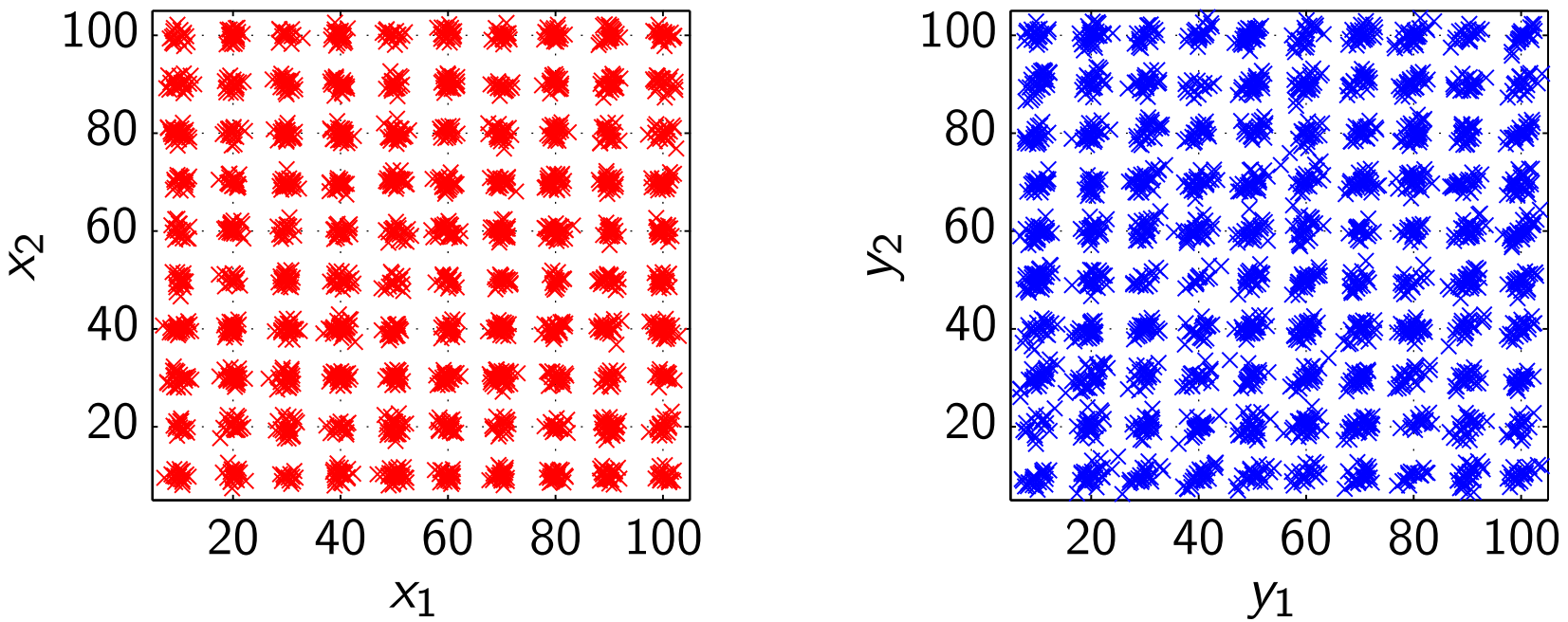
Idea: Stretch first Eigenvalue of rotated Gaussians.



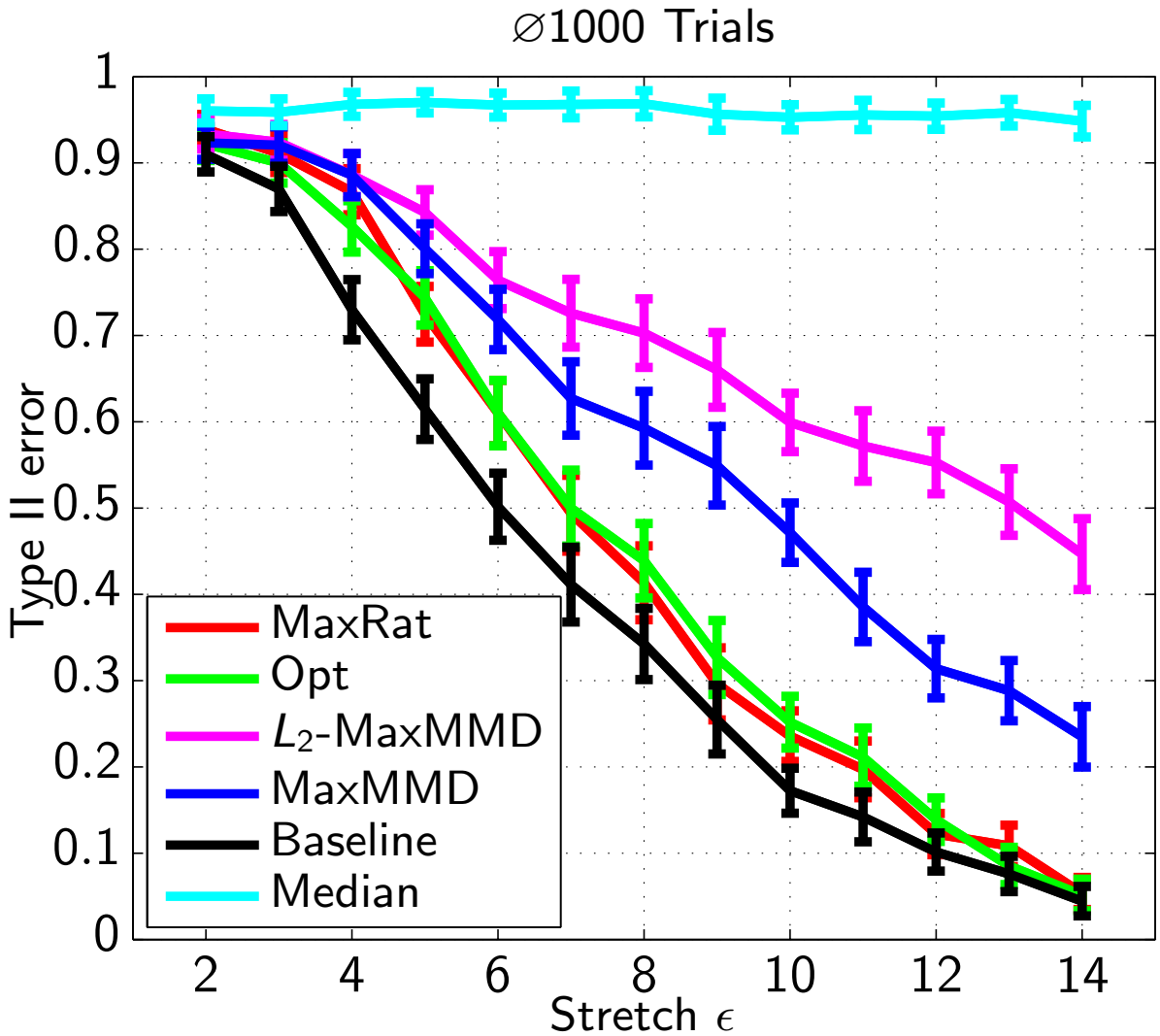


Gaussian Blobs: Why is it hard?

Which (Gaussian) kernel size $\ell \in \{2^{-5}, \ldots, 2^{15}\}$ is best?



Gaussian Blobs: Results



The Power of "Large-Scale" on Gaussian Blobs

12×12 Gaussians with scaling $\epsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel. Infinite data available.

	m **per trial**	**Type II error**	**Trials**
Quadratic	5000	$[0.7996, 0.8516]$	820
	10000	$[0.5161, 0.6175]$	367
	> 10000	Buy more RAM!	
Linear	$\varnothing$119580000	$[0.2250, 0.3049]$	468
	$\varnothing$185130000	$[0.1873, 0.2829]$	302
	$\vdots$	$\vdots$	$\vdots$
	$\varnothing$502430000	$\mathbf{0.0270 \pm 0.0302}$	111

Summary

- Non-parametric two-sample testing on huge amounts of data
- "Black-box test" – kernel is selected automagically
- For details, Maths, and experiments, see NIPS and JLMR in references

In the future:

- Kernel selection for quadratic time MMD
- How much data do we need to tell $p \neq q$?
- Do you have interesting data? (Kernel combinations?)

Efficient open-source implementation available under
`http://www.shogun-toolbox.org`

Thank you for your attention!

Questions?


Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006).
Integrating structured biological data by Kernel Maximum Mean Discrepancy.
Bioinformatics (Oxford, England), 22(14):e49–57.


Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a).
A Kernel Two-Sample Test.
Journal of Machine Learning Research, 13:671–721.


Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., and Fukumizu, K. (2012b).
Optimal kernel choice for large-scale two-sample tests.
In *Advances in Neural Information Processing Systems*.


Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., and Schölkopf, B. (2009).
Kernel choice and classifiability for RKHS embeddings of probability distributions.
In *Advances in Neural Information Processing Systems*.


Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007).
Covariate shift adaptation by importance weighted cross validation.
The Journal of Machine Learning Research, 8:985–1005.


Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011).
Least-squares two-sample test.
Neural networks : the official journal of the International Neural Network Society, 24(7):735–51.