Anisotropic Diffusion Kernels to Compare Distributions

Xiuyuan Cheng

Duke University

ICIAM 2019

Valencia, Spain

Joint work with *Alex Cloninger* @UCSD

Outline

- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - Testing power analysis
  - Application: Flow Cytometry data
  - Application: Diffusion MRI imaging
- Discussion: by neural network?

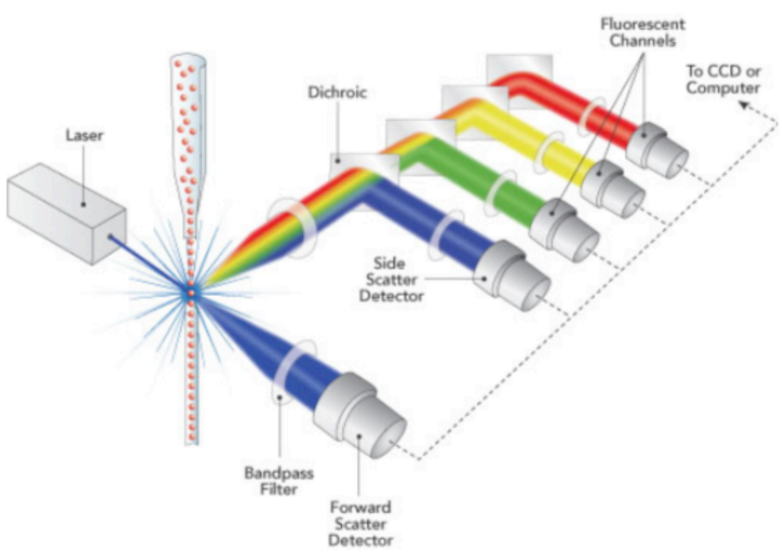
Two-sample Problem

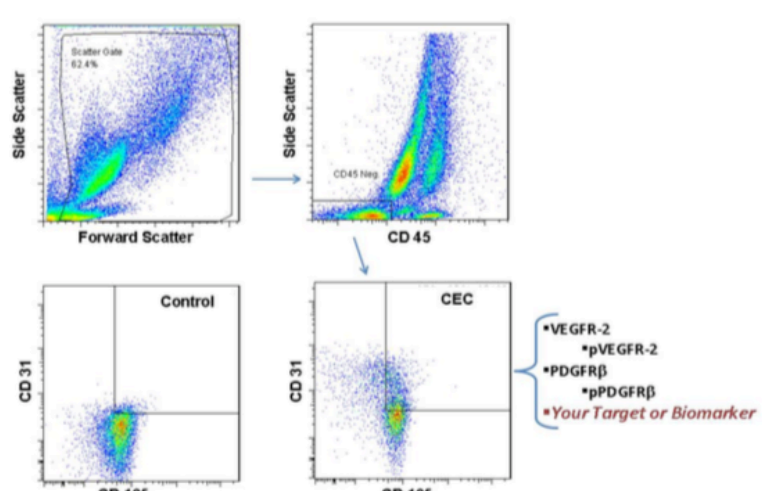
- **Question**: $X_i \sim p$, $Y_j \sim q$, iid., in $\mathbb{R}^D$

$$X = \{X_i\}_{i=1}^{n_X},\ Y = \{X_j\}_{j=1}^{n_Y},\ X \text{ independent from } Y$$

$$\text{Test hypothesis } \mathcal{H}_0 : p = q \text{ against } \mathcal{H}_1 : p \neq q$$

Flow Cytometry



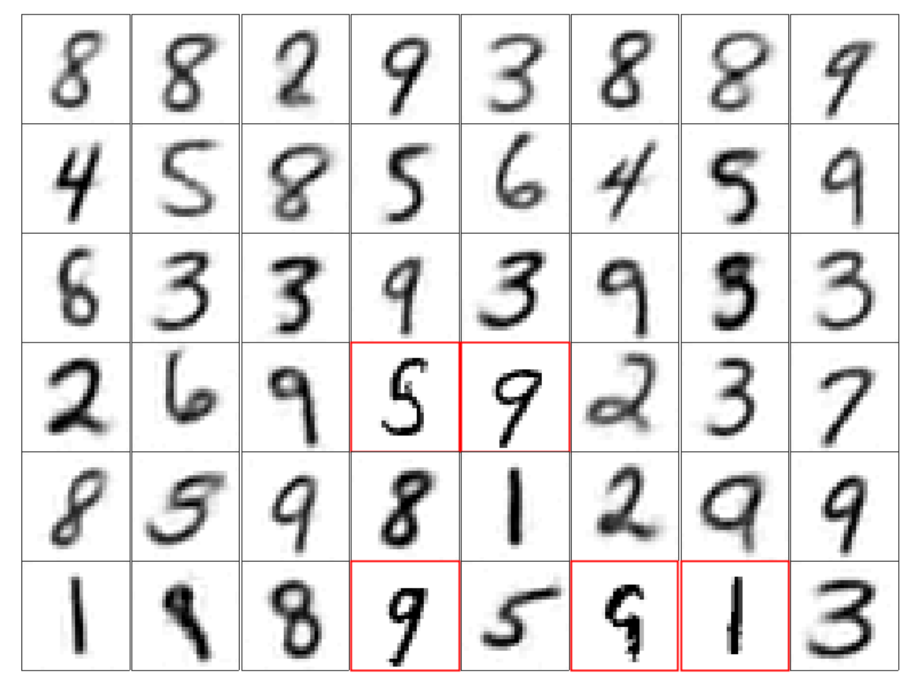


Comparing Groups of Population

Treatment group	NT$_R$	NT$_M$	HT	P value
n	12	14	10	—
Female (%)	58	53	40	0.7
Age, h, mean ± SD	16 ± 5	18 ± 4	15 ± 6	0.7
Weight at start of experiment, kg, mean ± SD	1.557 ± 0.25	1.654 ± 0.27	1.655 ± 0.24	0.6
pH at baseline, mean ± SD	7.47 ± 0.07	7.48 ± 0.10	7.52 ± 0.05	0.3
Lactate at baseline, mmol/l, mean ± SD	2.6 ± 1.2	2.7 ± 1.3	3.2 ± 1.2	0.5
Blood glucose at baseline, mmol/l, mean ± SD	6.3 ± 1.6	7.6 ± 1.4	7.1 ± 1.4	0.2
Duration of LAEEG during insult, min, mean ± SD	36.7 ± 8.2	23.8 ± 2.9	21.8 ± 3.5	<0.0001*
pH at end of insult, mean ± SD	7.04 ± 0.16	7.10 ± 0.14	7.08 ± 0.12	0.5
Lactate at end of insult, mmol/l, mean ± SD	15.0 ± 1.9	19.1 ± 3.5	17.2 ± 2.1	0.006[a]
MABP during insult, mmHg, mean ± SD	37 ± 7	45 ± 7	40 ± 4	0.007[b]

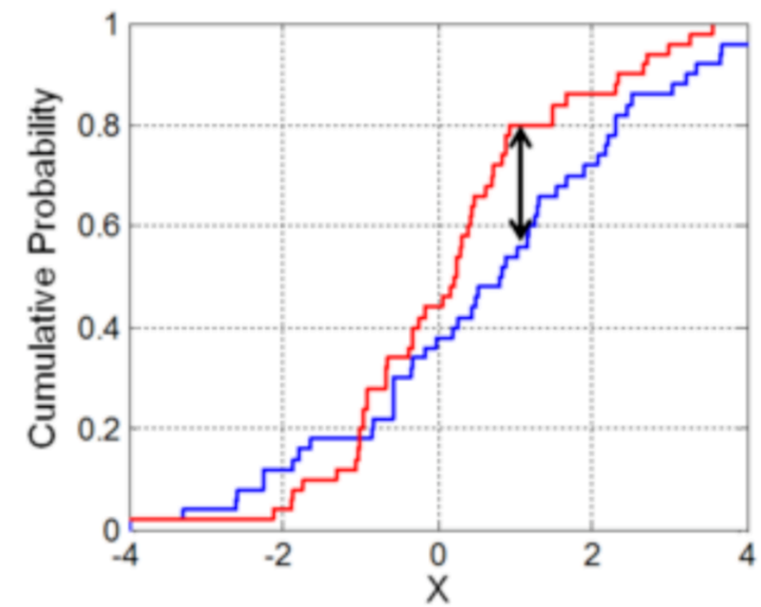
n, number of animals with neuropathology available; HT, hypothermia; LAEEG, low-amplitude electroencephalogram; MABP, mean arterial blood pressure; NT$_M$, matched normothermia; NT$_R$, randomized normothermia; SD, standard deviation.
[a]Lactate at end of insult, NT$_R$ vs. NT$_M$, $P = 0.002$; [b]MABP during insult, NT$_R$ vs. NT$_M$, $P = 0.002$. *Duration of LAEEG during insult: NT$_R$ vs. NT$_M$, $P \leq 0.0001$; NT$_R$ vs. HT, $P \leq 0.0001$.

Authentic and Synthetic images



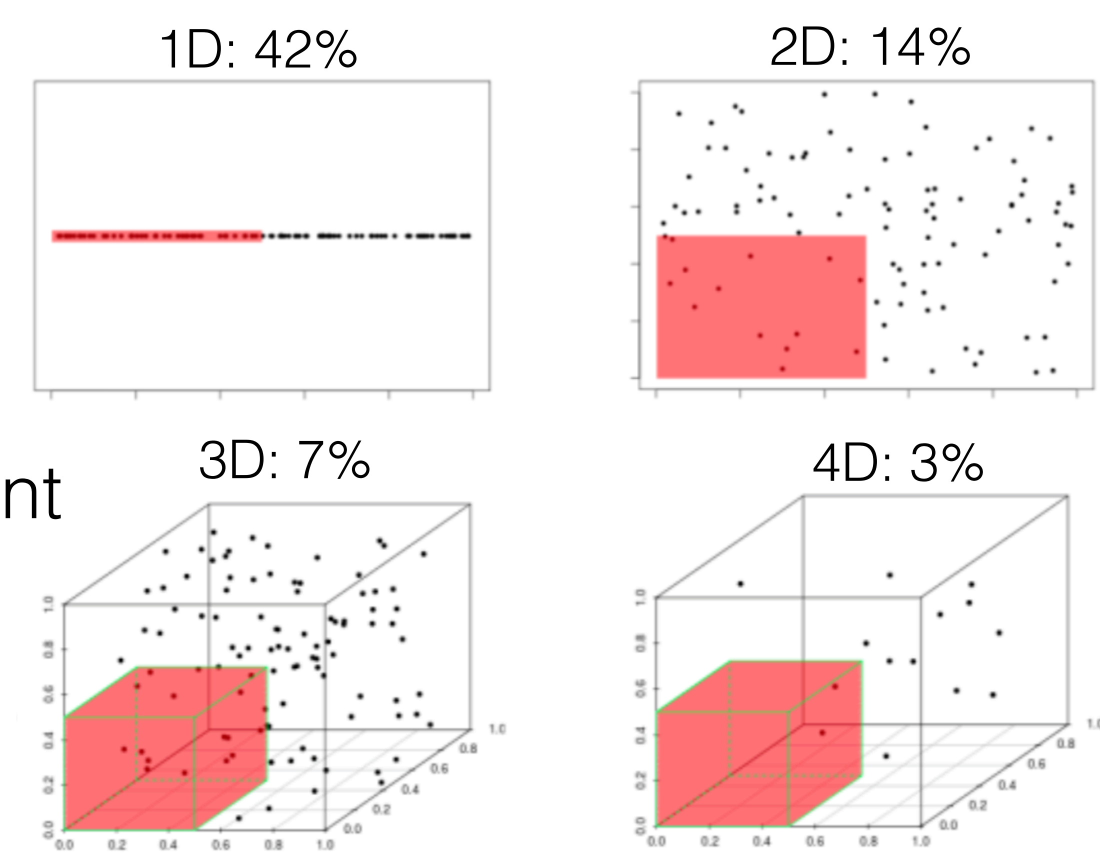
Two-sample Problem

- Standard procedure:
  - Compute test statistic $T(X, Y)$
  - Specify a threshold value τ
  - Accept $\mathcal{H}_0$ if $T(X, Y) < \tau$, reject otherwise



- Traditional solutions in 1D:
  - Kolmogotov-Smirnov Test
  - …

Proportion of samples captured by one box



- Difficulty in higher dimensions:
  - Marginals of distributions are insufficient
  - Most “bins” will have very few points
- Additional question: where $p \neq q$?

Outline

- Background
  - Two-sample problem
  - **Kernel MMD and data geometry**
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - Testing power analysis
  - Application: Flow Cytometry data
  - Application: Diffusion MRI imaging
- Discussion: by neural network?

Review: Kernel MMD (Maximum-mean Discrepancy)

- Maximum-mean Discrepancy (MMD)

$$\mathrm{MMD}(p, q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \int f(x)(p(x) - q(x))dx,$$

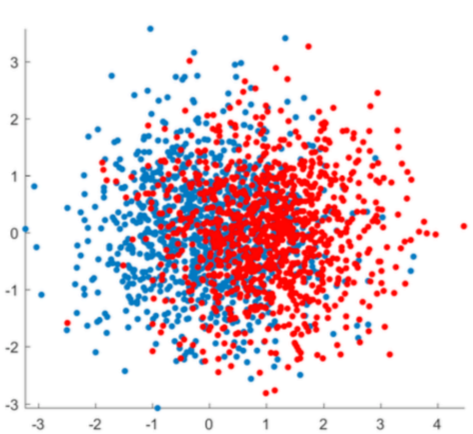
- Reproducing Kernel Hilbert Space (RKHS) MMD: $\mathcal{F} = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1\}$

- Population Kernel MMD

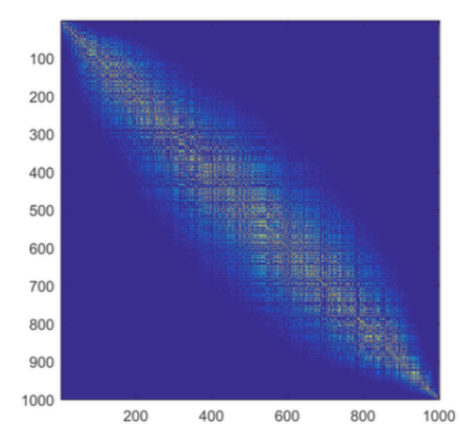
$$\mathrm{MMD}^2(p, q) = \int \int k(x, y)(p(x) - q(x))(p(y) - q(y))dxdy$$

- Discrete Kernel MMD

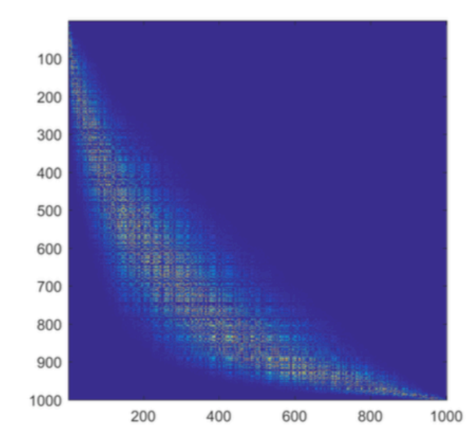
$$\mathrm{MMD}^2(X, Y) = \tfrac{1}{n_X^2} \textstyle\sum_{x,x' \in X} k(x, x') + \tfrac{1}{n_Y^2} \textstyle\sum_{y,y' \in Y} k(y, y') - \tfrac{2}{n_X n_Y} \textstyle\sum_{x \in X, y \in Y} k(x, y).$$



Data X and Y



$K(X, X)$



$K(X, Y)$

[Gretton et al. '12]

Review: Kernel MMD

- Test consistency and test power analysis

Theorem (Gretton '12, Serfling '81). *For fixed p and q, $n := n_X + n_Y$, $n \to \infty$, $\frac{n_X}{n} \to \rho_X \in (0,1)$. Then, under $\mathcal{H}_0$, $MMD^2(X,Y) = O\left(\frac{1}{n}\right)$; Under $\mathcal{H}_1$, $MMD^2(X,Y) = MMD^2(p,q) + O(\frac{1}{\sqrt{n}})$, $MMD^2(p,q) > 0$.*

- Convergence in distribution: Chi-square under $\mathcal{H}_0$, normal under $\mathcal{H}_1$.

- Indicator of density difference

$$\mathrm{MMD}(p,q) = \int f^*(x)(p(x) - q(x))dx$$

$$f^*(x) = \int k(x,y)(p(y) - q(y))dy := w(x)$$ "witness" function

- Empirical witness function

$$\hat{w}(x) = \frac{1}{n_X}\sum_{i=1}^{n_X} k(x, X_i) - \frac{1}{n_Y}\sum_{j=1}^{n_Y} k(x, Y_j)$$

[Gretton et al. '12]

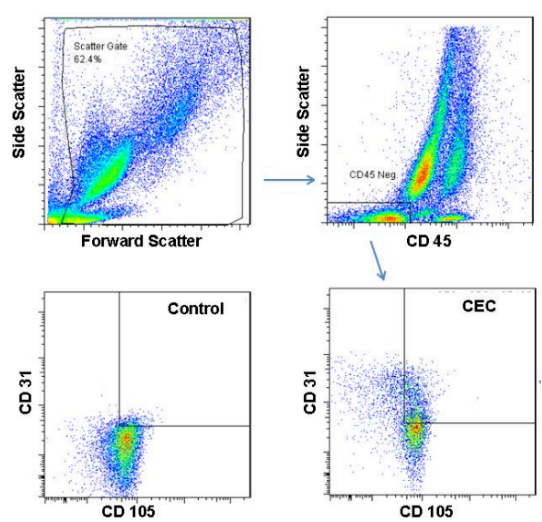
Review: Kernel MMD

Problems with Kernel MMD:

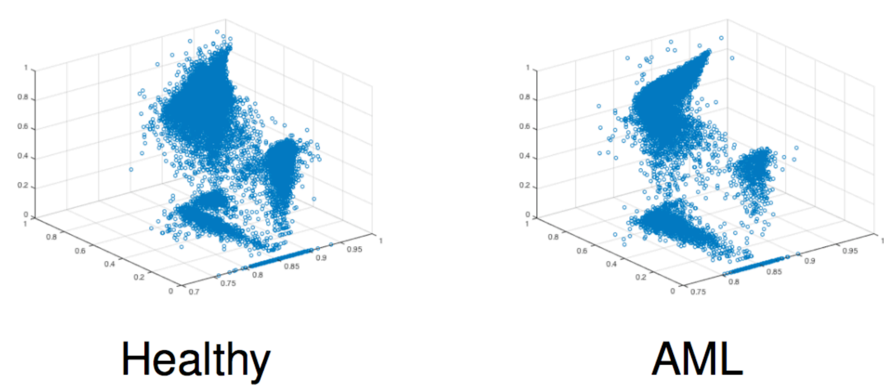
- Isotropic gaussian kernel may not be optimal
  - Potential loss of power in high dimension [Wasserman et al. ’14]
  - Optimization of kernel [Gretton et al. ’12b]
- $O(n^2)$ computation
  - Linear algorithm by decoupling [Gretton et al. ’12]
  - Mean Embedding test [Chwialkowski et al. ’15]

Near-manifold Densities

- The densities lie on or near to low-dimensional manifolds embedded in the ambient space
- Flow cytometry: each patient is represented by a data cloud in 9D



2D slices



First 3 Principal Components

- Authentic and synthetic images: image patch manifold
- **Question**: How manifold geometry helps?

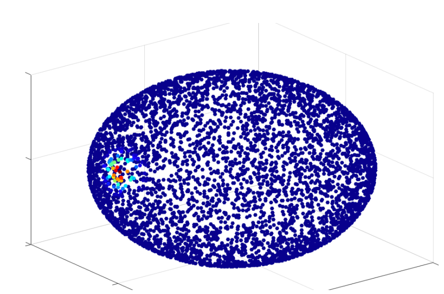
Kernel and Data Geometry

- Observation in view of kernel **spectral decomposition**

$$\mathrm{MMD}^2 = \int\int K(x,y)(p(x)-q(x))(p(y)-q(y))dxdy$$

$$K(x,y) = \sum_k \lambda_k \psi_k(x)\psi_k(y), \quad \mathrm{MMD}^2 = \sum_k \underbrace{\lambda_k}_{\text{weights}} \Big(\underbrace{\overbrace{\int \psi_k(x)(p-q)(x)dx}^{:= c_k}}_{\text{projection to eigenmode}} \Big)^2$$

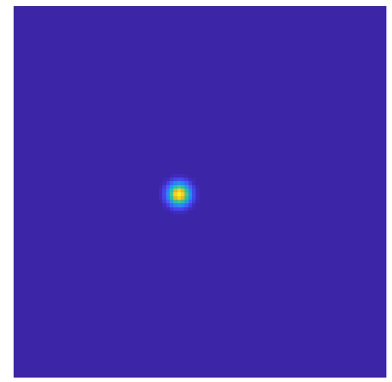
- Idea from traditional manifold learning
  - The eigen-pair $\{\lambda_k, \psi_k\}_k$, when kernel bandwidth $\sigma \to 0$, are determined by intrinsic manifold geometry $\triangle_{\mathcal{M}}$ and data density p.
  - When n large enough, and σ small enough, the kernel matrix spectrally approximate the manifold operator involving $\triangle_{\mathcal{M}}$ and p.
  - The spectrum pattern *persists* with certain near-manifold perturbation of samples in the ambient space.



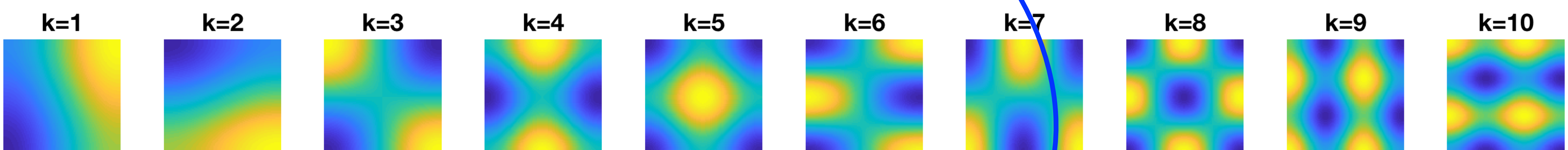
Kernel and Data Geometry

- The effect of **geometry** on $\{\lambda_k, \psi_k\}_k$

$$\mathcal{M} = [0,1]^2$$

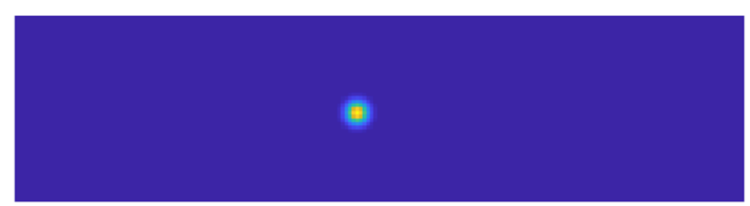


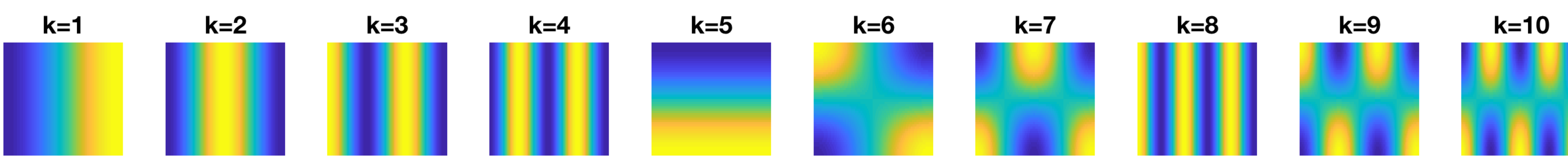
$$k(x,y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$$



Anisotropic kernel

$$\mathcal{M} = [0,2] \times [0,0.5]$$





First 10 non-trivial eigenvectors of the normalized graph laplacian on a uniform grid

Outline

- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - **Test statistic and algorithm**
  - Testing power analysis
  - Application: Flow Cytometry data
  - Application: Diffusion MRI imaging
- Discussion: by neural network?

Anisotropic Kernel MMD: Formulation

- MMD test statistic
  - Theoretically, assume reference set R and tensor field $\{\Sigma_r\}_{r \in R}$ are given
  - Generally, reference set distribution μ_R
  - Define asymmetric anisotropic kernel

$$a(r,x) = e^{-\|r-x\|^2_{\Sigma_r}} = \exp\left\{-\frac{1}{2}(x-r)^T \Sigma_r^{-1} (x-r)\right\}, \quad \forall r \in R$$

  - Kernel MMD computed with

$$k_{L^2}(x,y) = \int a(r,x) a(r,y) d\mu_R(r)$$

  - Spectral re-weighted kernel

$$k_{\text{spec}}(x,y) = \sum_k f_k \psi_k(x) \psi_k(y)$$

where f_k is sufficiently decaying positive sequence,

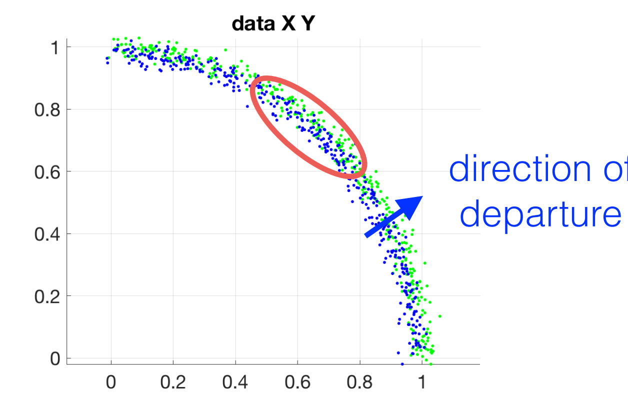
$$a(r,x) = \sum_k \sigma_k \phi_k(r) \psi_k(x), \quad k_{L^2}(x,y) = \sum_k \sigma_k^2 \psi_k(x) \psi_k(y)$$

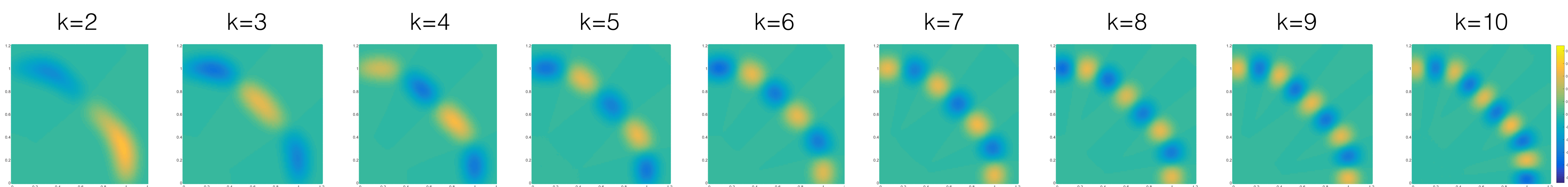
Anisotropic Kernel MMD: Intuition

- Population kernel MMD

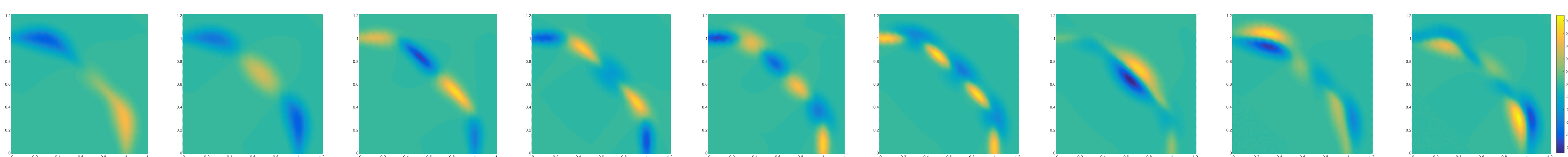
$$\mathrm{MMD}^2 = \sum_k \lambda_k \left(\int \psi_k(x)(p-q)(x)dx \right)^2 = \sum_k \lambda_k {c_k}^2$$

- First 10 eigenfunctions of the isotropic/anisotropic kernel





Isotropic gaussian kernel



Anisotropic gaussian kernel

Anisotropic kernel is more sensitive to the direction of density departure!

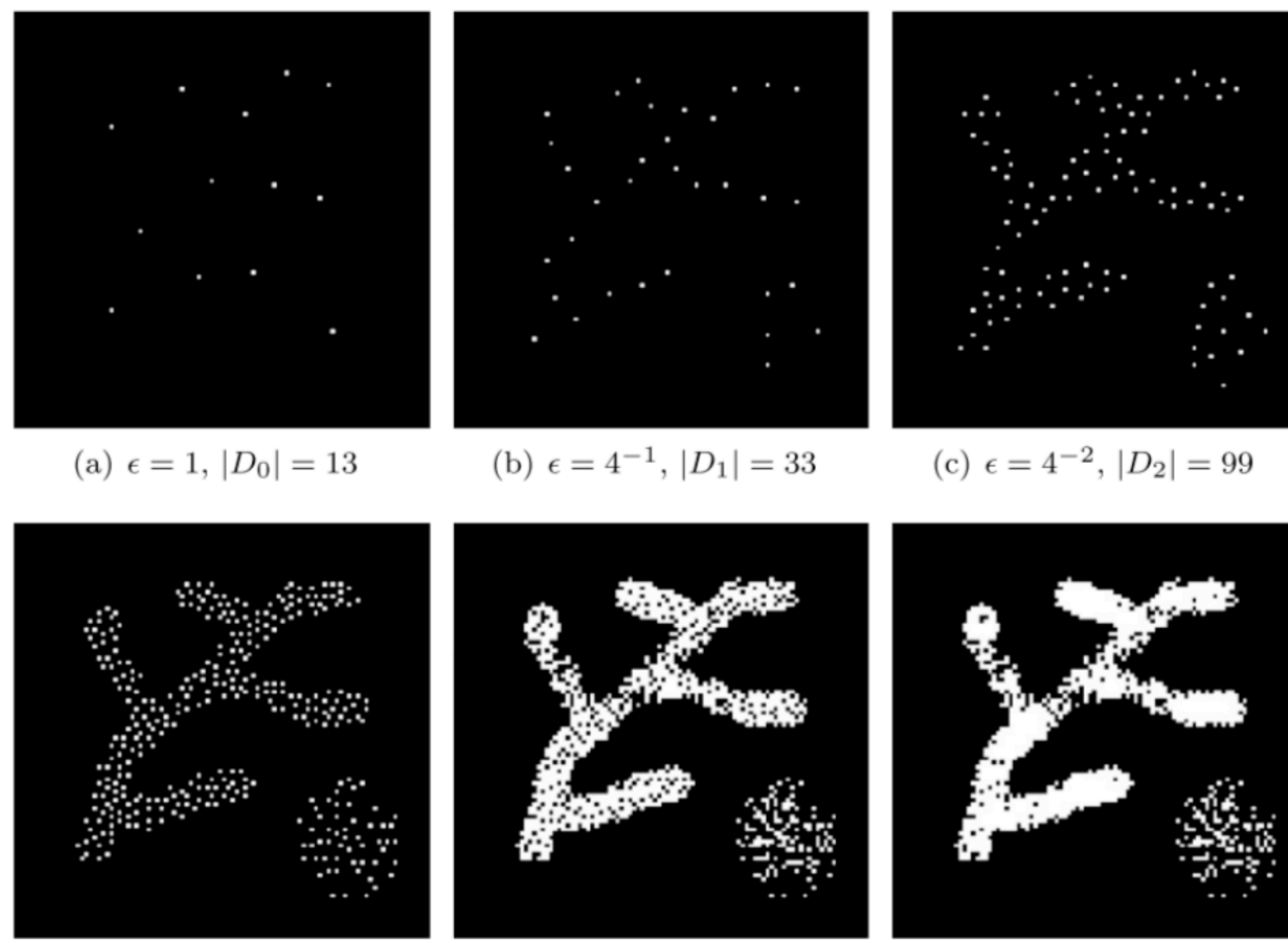
Anisotropic Kernel MMD: Computation

- Algorithm summary
  - Input: two data sets X and Y, reference set R , function handle $a(r,x)$
  - Output: Acceptance/rejection of $\mathcal{H}_0$, the witness function evaluated
  - Choice of threshold τ : by permutation test

- Sampling of R :
  - Random subsample
  - QR with pivoting

- Adaptive construction of $\{\Sigma_r\}_{r \in R}$
  - Local PCA on k nearest neighbors

Square low-rank kernel matrix is over-redundant



(a) $\epsilon = 1$, $|D_0| = 13$ (b) $\epsilon = 4^{-1}$, $|D_1| = 33$ (c) $\epsilon = 4^{-2}$, $|D_2| = 99$

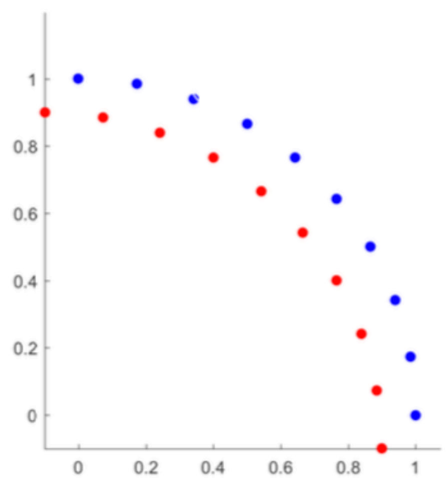
(d) $\epsilon = 4^{-3}$, $|D_3| = 348$ (e) $\epsilon = 4^{-4}$, $|D_4| = 1332$ (f) $\epsilon = 4^{-5}$, $|D_5| = 1469$

Subsample "reference set" without losing accuracy of computing leading eigenvectors

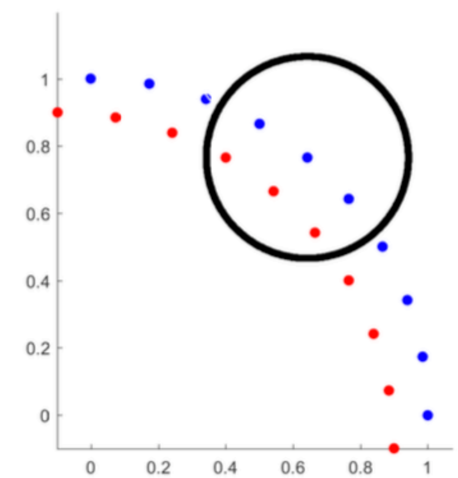
[Tygert et al. '08, Kuhn '11, Bermanis et al. '15]

Anisotropic Kernel MMD: Example

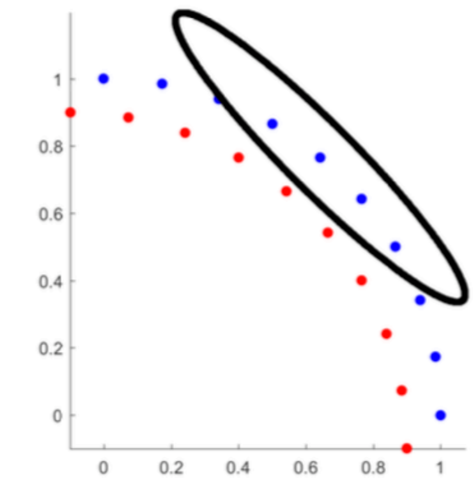
- Near-manifold density setting: Toy example in 2D



Data

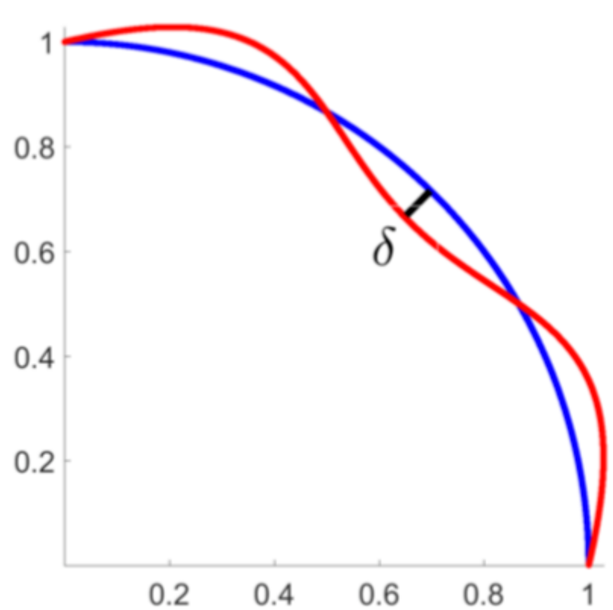


Gretton et al (2011)

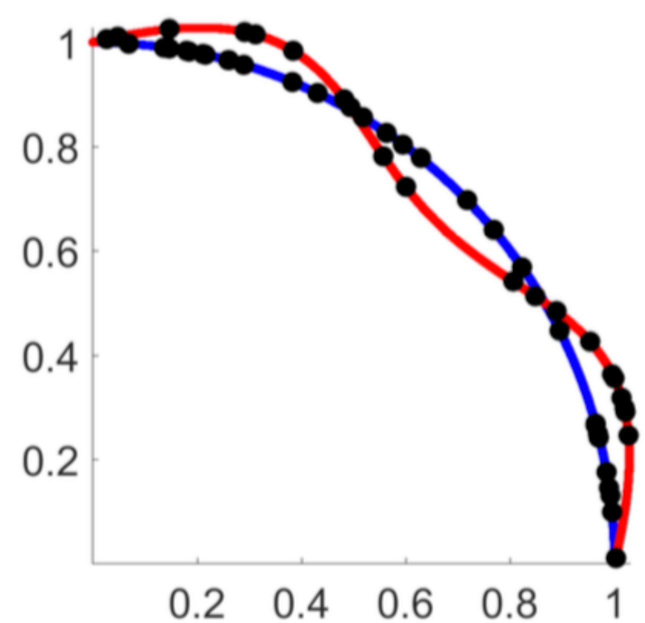


C., Cheng, Coifman (2011)

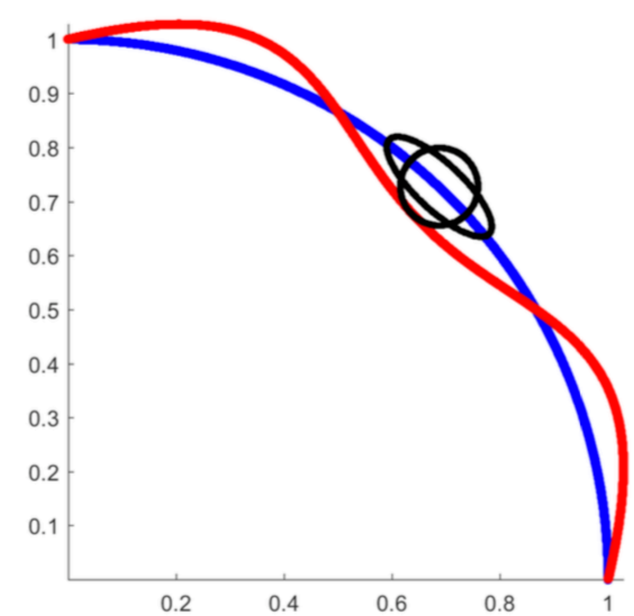
Two densities



Two samples



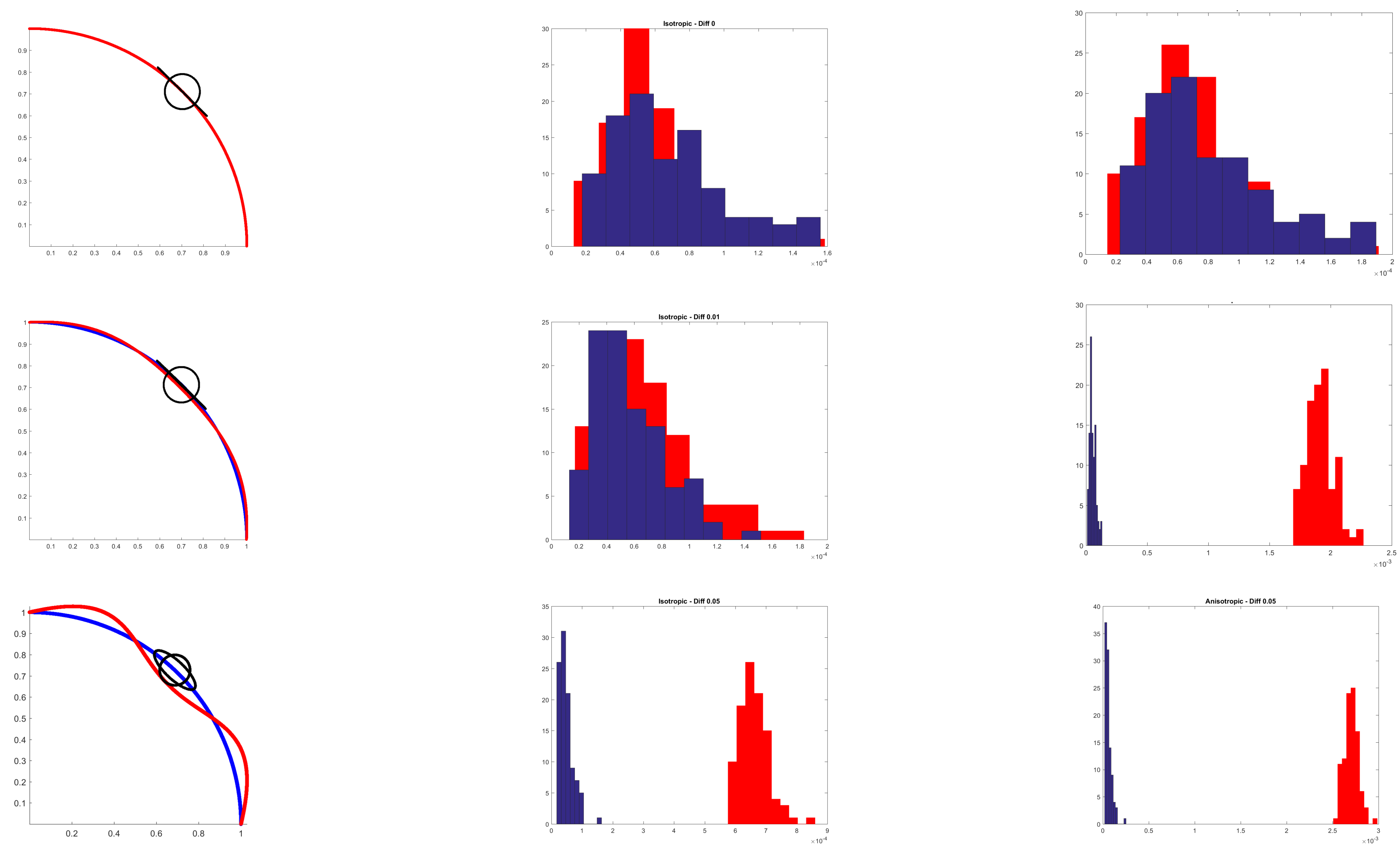
Kernels



Anisotropic Kernel MMD: Example

- Empirical distribution of test statistics

Two densities | Isotropic kernel | Anisotropic kernel



Histograms of test statistics under $\mathcal{H}_0$ (blue) and $\mathcal{H}_1$ (red)

Outline

- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - **Testing power analysis**
  - Application: Flow Cytometry data
  - Application: Diffusion MRI imaging
- Discussion: by neural network?

Limiting Distribution of Test Statistics

$$K(x,y) = \sum_k f_k \psi_k(x) \psi_k(y), \quad f_k \geq 0$$

- Assumptions (informal)

  (A1) The kernel is PSD, continuous, $0 \leq K(x,x) \leq 1$

  (A2) The alternative q belongs to

$$\mathcal{Q} = \left\{ q \middle| \int a(r,x)(p(x) - q(x))dx \neq 0, \text{ a.s. w.r.t } \mu_R \right\}$$

- Define single-parametrized departure $q = p + \tau g$, $c_k := \int \psi_k(y) g(y) dy$
- Limiting distribution of the test statistic: $n = n_X + n_Y \to \infty$, $\frac{n_X}{n} \to \rho_X \in (0,1)$,

Theorem (C, Cloninger, Coifman '17, informal)**.** *All shifts and variance of the test statistic T_n depend on spectral decomposition of the kernel.*

(1) If $\tau = an^{-1/2}$, $0 \leq a < \infty$, then nT_n is asymptotically χ^2.

(2) If $\tau = n^{-1/2+\delta}$, $0 < \delta < \frac{1}{2}$, then T_n is asymptotically normal with $O(n^{-1+2\delta})$ shift and $O(n^{-1+\delta})$ standard deviation.

(3) If $\tau = 1$, then T_n is asymptotically normal with $O(1)$ shift and $O(n^{-1/2})$ standard deviation.

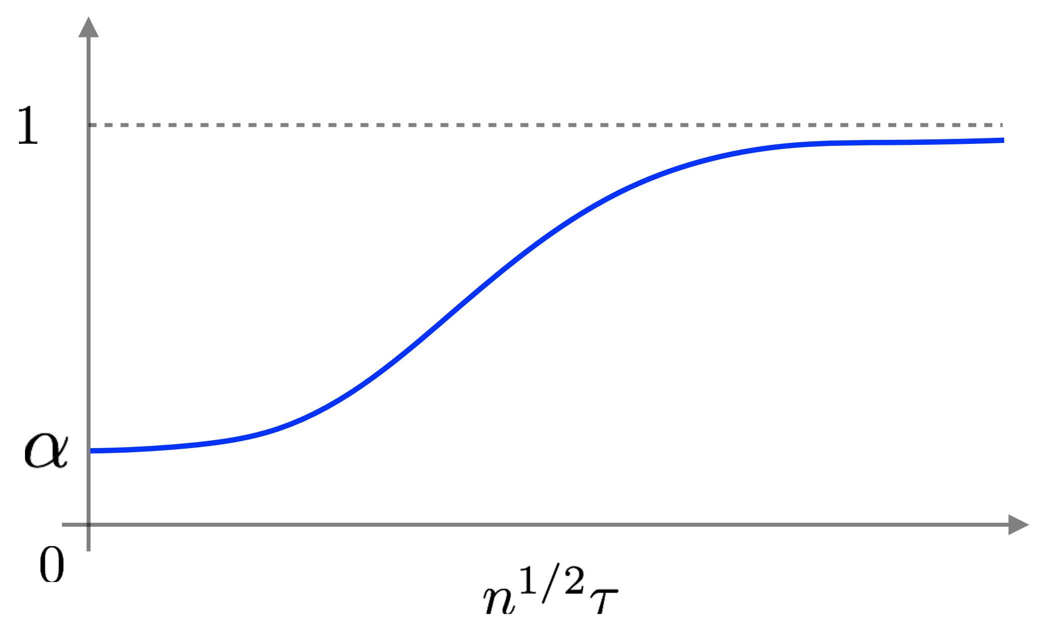
Asymptotic Test Consistency

- Asymptotic test power at/beyond critical regime $\tau \sim n^{-1/2}$

Corollary 1. *Let $\pi_n(q)$ be the test power for controlled type-I error $\leq \alpha$,*

(1) If $\tau = an^{-1/2}$, $0 < a < \infty$, then $\pi_n(q) \to f(a) > \alpha$, where f is a monotonically increasing function.

(2) If $\tau = \Omega(n^{-1/2})$, then $\pi_n(q) \to 1$.



Test Power Lower Bound with Finite Samples

- Non-asymptotic **lower bound** of test power

Theorem (C, Cloninger, Coifman '17, informal). *Define* $T_1 := \sum_k \lambda_k c_k^2 > 0$. *If* $n > \frac{16}{0.1}(\frac{1}{\rho_{X,n}^3} + \frac{4}{\rho_{Y,n}^3})$, *and* $(\tau^2 n) T_1 > C_4 + \sqrt{\frac{C_3+0.1}{\alpha}}$, *then*

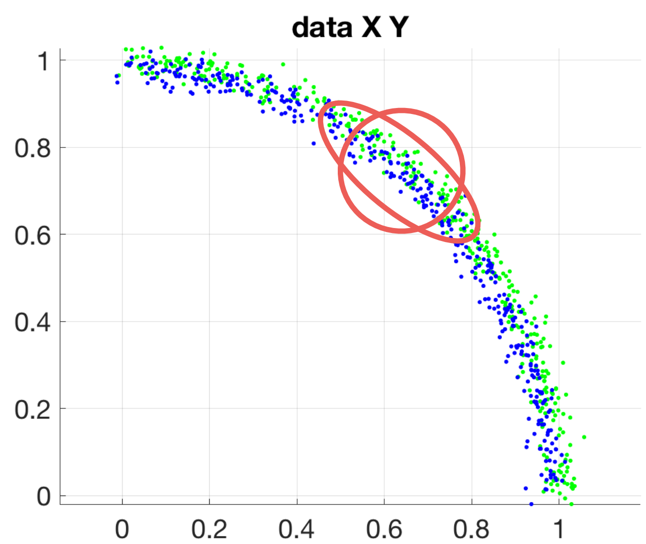
$$1 - \pi_n(q) \leq \frac{(\tau^2 n) C_1 + \tau C_2 + C_3 + 0.1}{\left((\tau^2 n) T_1 - (C_4 + \sqrt{\frac{C_3+0.1}{\alpha}}) \right)^2} \sim \frac{C_1}{T_1^2} \frac{1}{\tau^2 n},$$

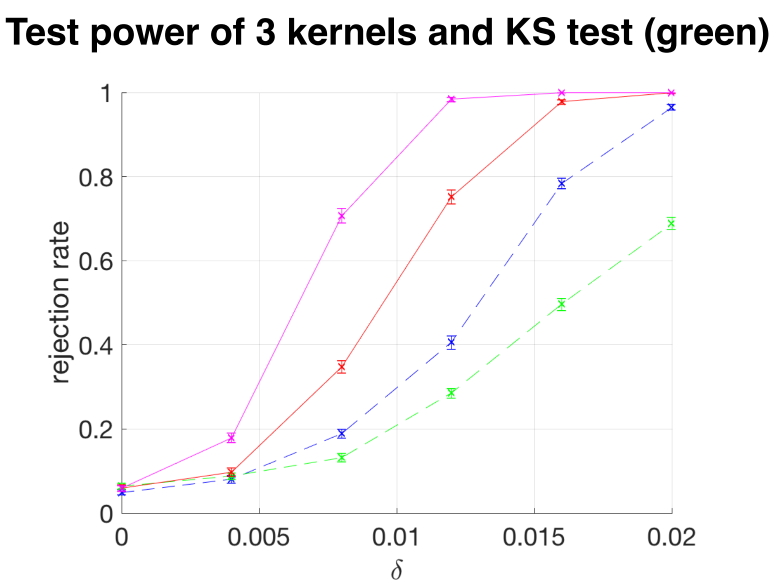
where $C_1 := 4\left(\frac{1}{\rho_{X,n}} \sum_k \lambda_k^2 c_k^2 + \frac{16}{\rho_{Y,n}}\right)$, $C_2 := 128\left(\frac{1}{\rho_{X,n}^2} + \frac{1}{\rho_{Y,n}^2}\right)$, $C_3 := \frac{32}{(\rho_{X,n}\rho_{Y,n})^2}$, $C_4 := \frac{1}{\rho_{X,n}\rho_{Y,n}} \sum_k \lambda_k$.

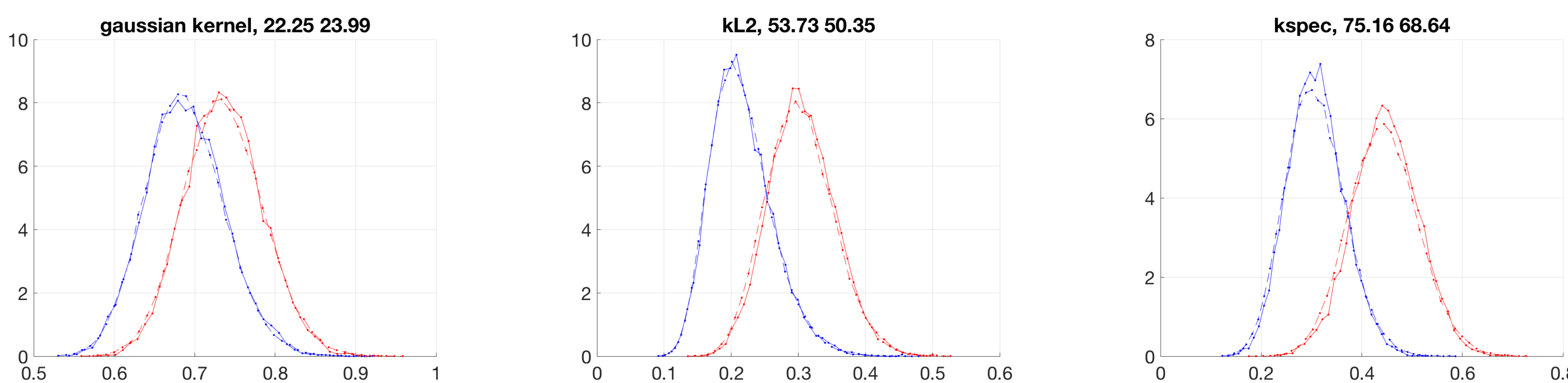
- Proof by Chebyshev.

Comparison of Kernels

- Numerical simulation on 2D synthetic example

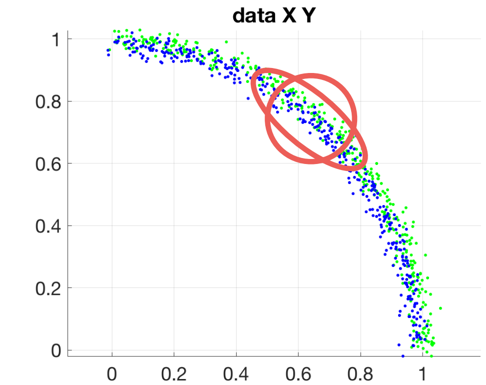






Histograms of test statistics under $\mathcal{H}_0$ (blue) and $\mathcal{H}_1$ (red), $\delta = 0.02$

Comparison of Kernels



- Mean and variance of test statistic
  - 1st row: estimated value by Monte-Carlo
  - 2nd row: theoretical value by limiting distribution
  - The larger ratio, the more discriminative the test

$$\theta_0 = \mathbb{E}[T_n|\mathcal{H}_0],\ \theta_1 = \mathbb{E}[T_n|\mathcal{H}_1],\ \sigma_0^2 = \mathrm{Var}(T_n|\mathcal{H}_0),\ \sigma_1^2 = \mathrm{Var}(T_n|\mathcal{H}_1),\ r = \frac{\theta_1 - \theta_0}{\sigma_1 + \sigma_0}$$

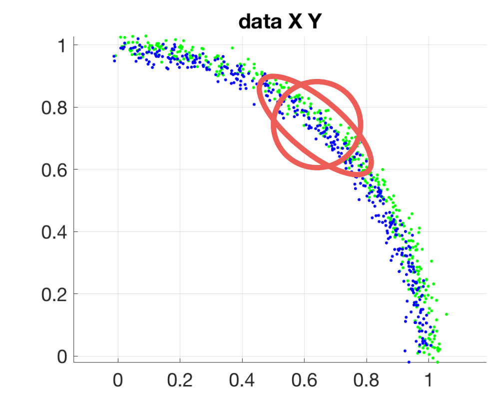
$$\bar{\theta}_0 = \frac{2}{n}\sum_k \lambda_k,\ \bar{\theta}_1 = \sum_k \lambda_k \tau^2 c_k^2 + \frac{2}{n}\sum_k \lambda_k,$$

$\lambda_k,\ c_k,\ \tau$

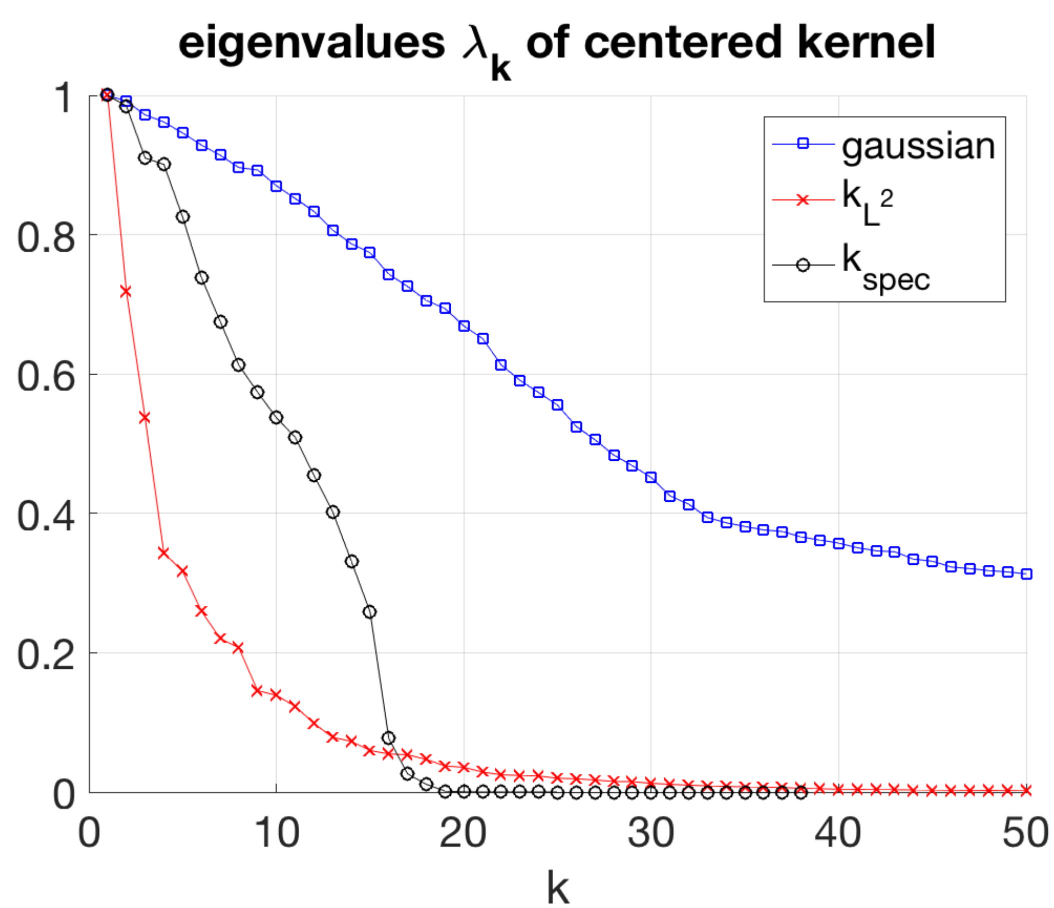
$$\bar{\sigma}_0^2 = \mathrm{Var}(\sum_k \lambda_k \frac{1}{n}(h_k - g_k)^2),\ \bar{\sigma}_1^2 = \mathrm{Var}(\sum_k \lambda_k(\tau^2 c_k + \frac{1}{\sqrt{n}}(h_k - g_k))^2),\ h_k, g_k \sim N(0,1) \text{ i.i.d}$$

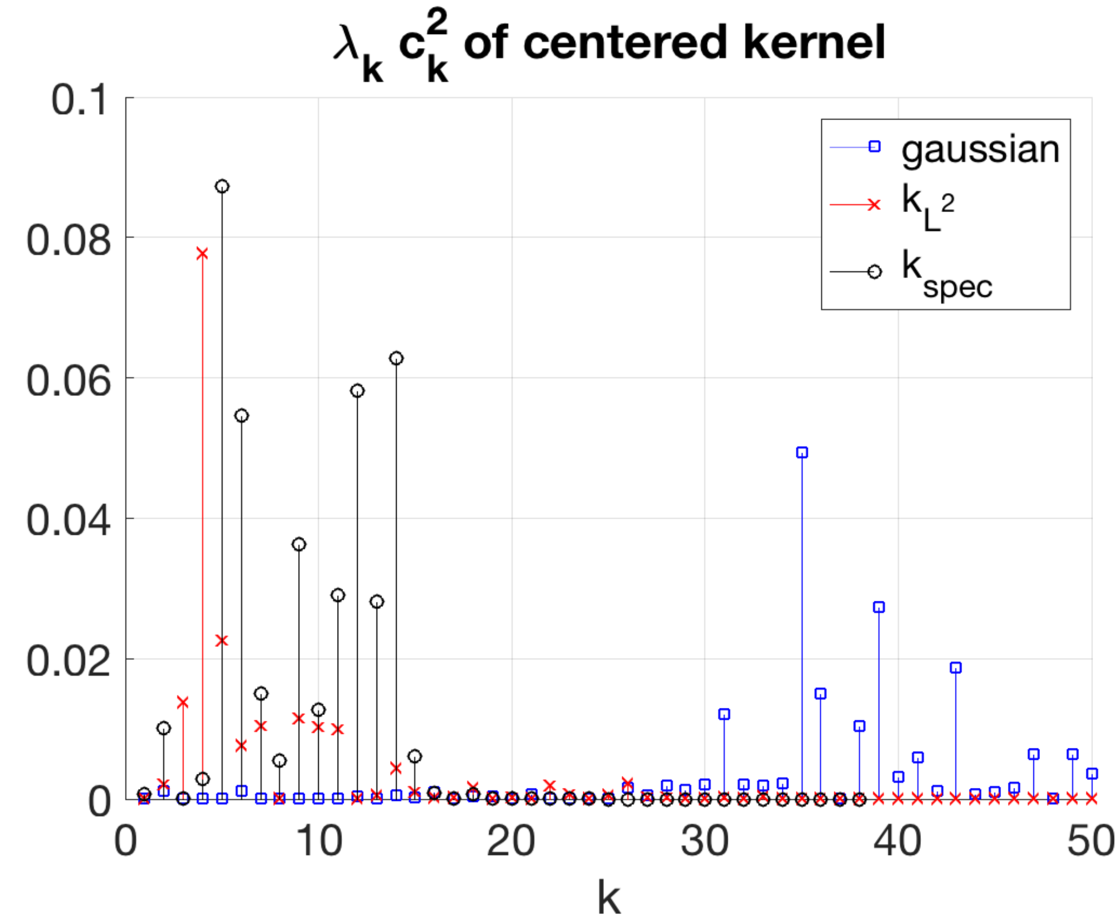
	θ_0	θ_1	σ_0	σ_1	r
$n = 200, \tau = 0.5$					
Gaussian	0.4771	0.5444	0.0677	0.0704	0.4874
	0.4754	0.5439	0.0676	0.0736	0.4848
k_{L^2}	0.0489	0.0958	0.0214	0.0306	0.9000
	0.0488	0.0939	0.0214	0.0312	0.8573
k_{spec}	0.0985	0.2046	0.0348	0.0587	1.1351
	0.0983	0.2013	0.0374	0.0620	1.0354
$n = 400, \tau = 0.5/\sqrt{2}$					
Gaussian	0.2381	0.2720	0.0334	0.0359	0.4885
	0.2379	0.2722	0.0339	0.0368	0.4850
k_{L^2}	0.0243	0.0477	0.0107	0.0153	0.8972
	0.0244	0.0471	0.0106	0.0157	0.8616
k_{spec}	0.0490	0.1036	0.0177	0.0305	1.1343
	0.0490	0.1003	0.0188	0.0310	1.0290

Comparison of Kernels



- Contribution per eigenmode





k_{L^2} anisotropic kernel

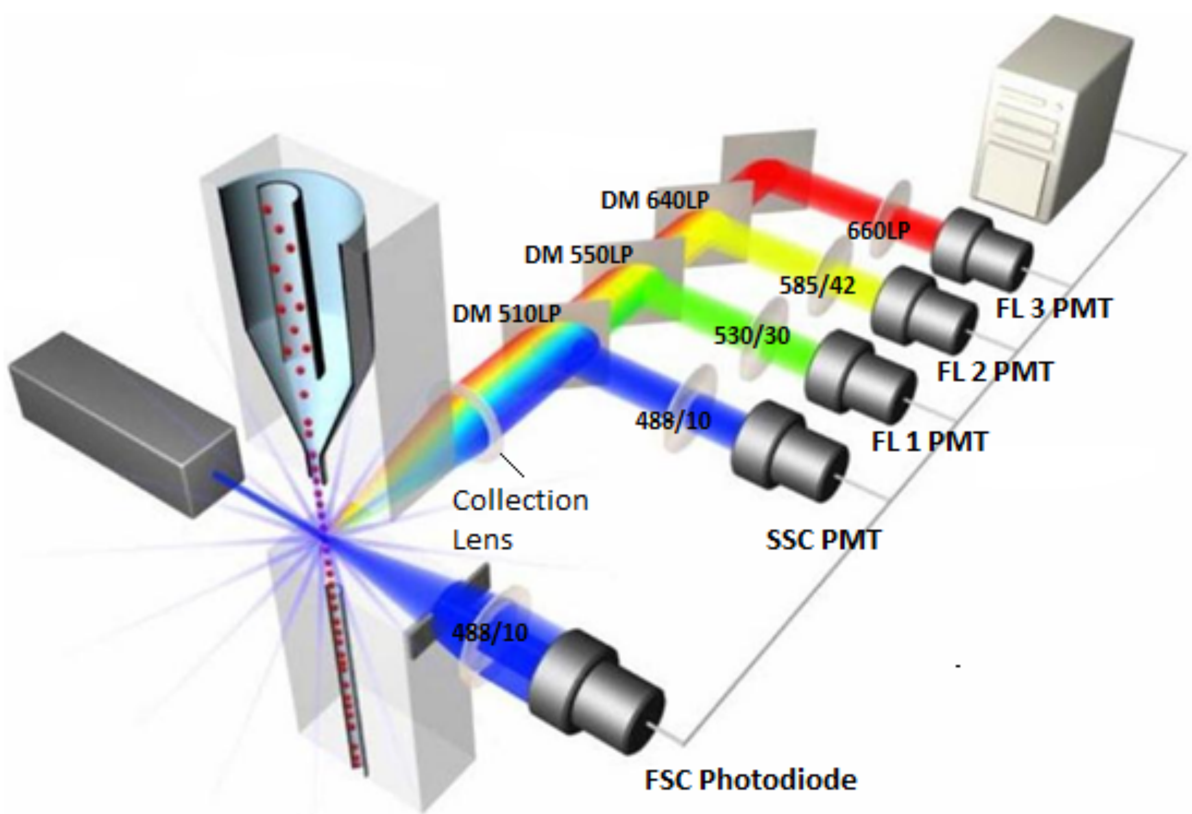
k_{spec} anisotropic kernel with spectral re-weighting

Outline

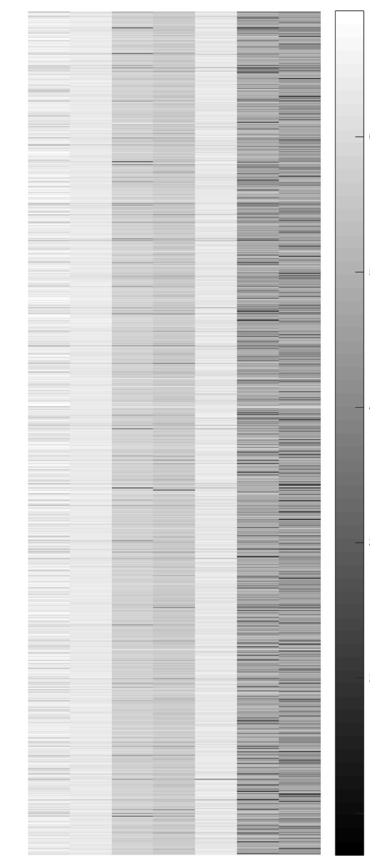
- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - Testing power analysis
  - **Application: Flow Cytometry data**
  - Application: Diffusion MRI imaging
- Discussion: by neural network?

Application: Flow-cytometry Data

- Flow Cytometry technology for single-cell analysis



Number of cells
$\sim 10^5$

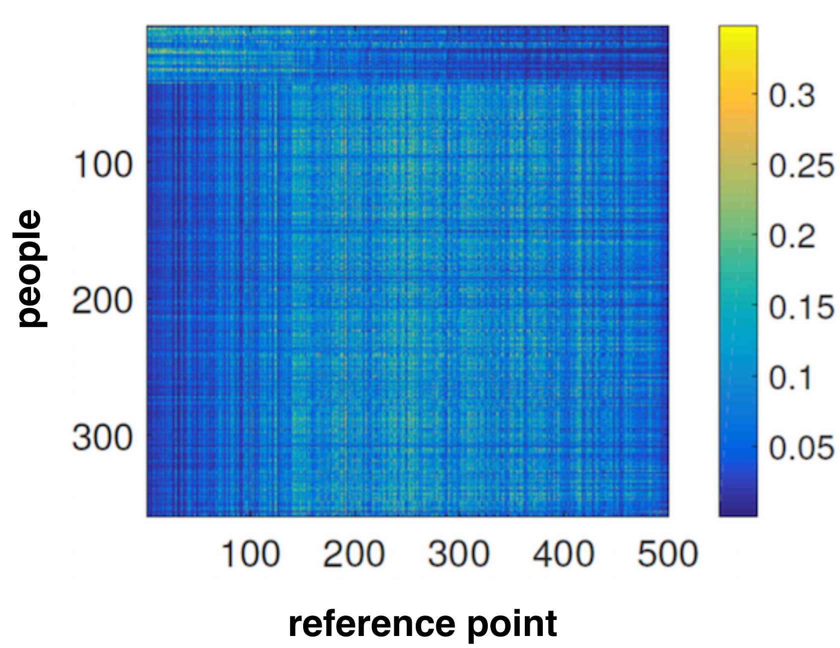


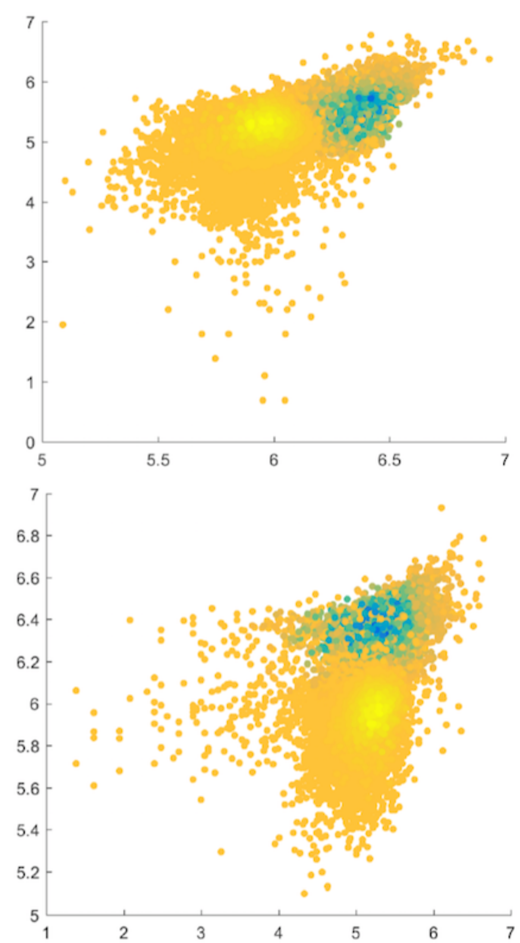
Collaboration with Kluger's Lab @ Yale School of Medicine

Application: Flow-cytometry Data

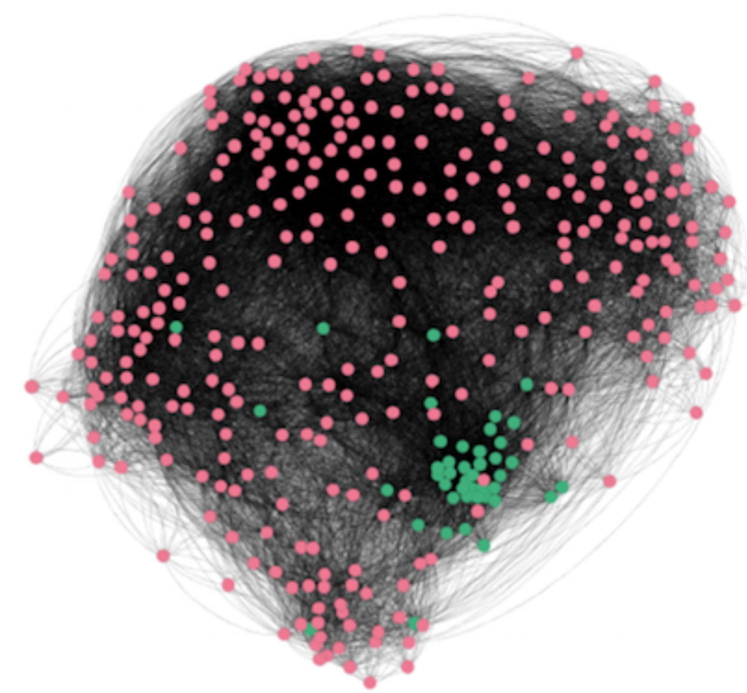
- Flow-Cap(I) Competition AML Datasets
  - 359 people: 43 affected, 316 healthy
  - 7 markers

- MMD with anisotropic kernel





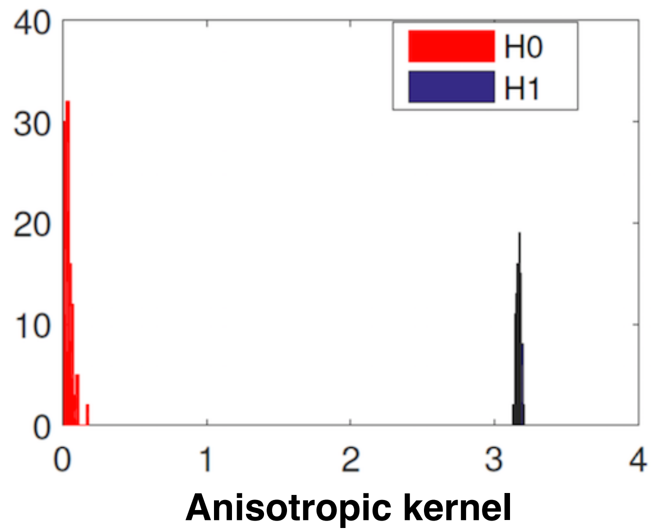
Witness function visualized on point cloud (sliced 2D)

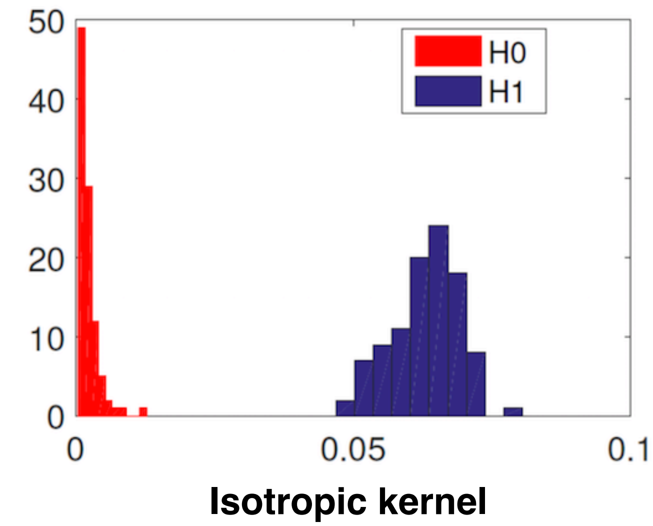


graph embedding based on computed sample distances

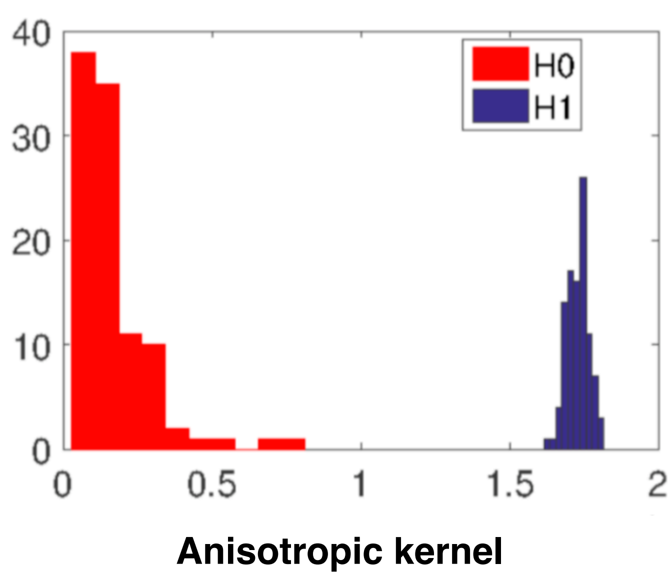
Application: Flow-cytometry Data

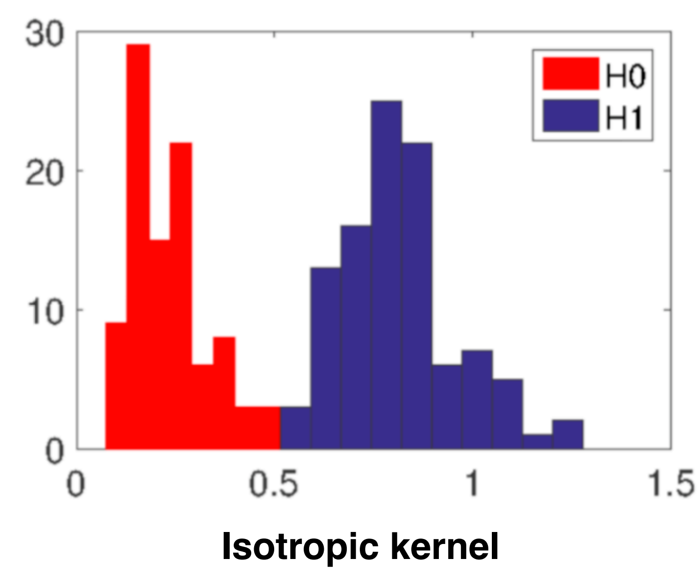
- Comparison to isotropic kernel





- On MDS dataset
  - 72 patients, 87 healthy subjects,
  - Each sample: 25,000 cells in 8 dimensions



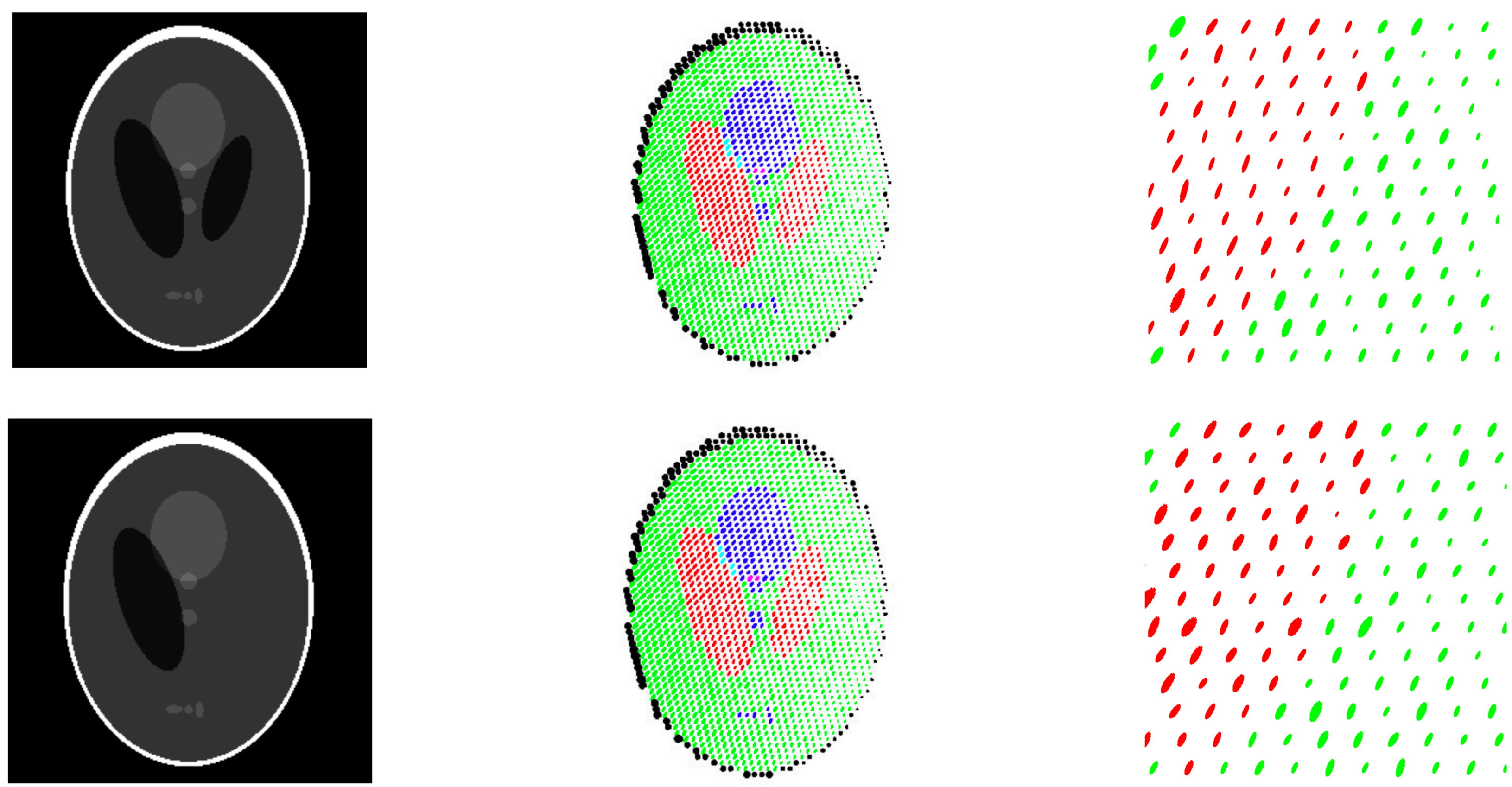


Outline

- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - Testing power analysis
  - Application: Flow Cytometry data
  - **Application: Diffusion MRI imaging**
- Discussion: by neural network?

Application: Diffusion MRI Data

- 3D brain Image of size 200x200x200
- Each pixel ~ a 3x3 tensor (local flow of water molecules)
- Want to identify regions of brain that differ between healthy/sick individuals



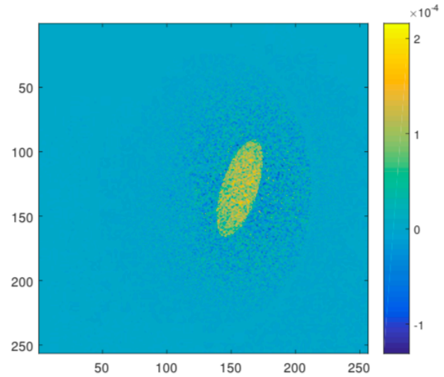
brain template　　brain diffusion tensors　　zoomed-in image

- Formulate as two-sample problem:

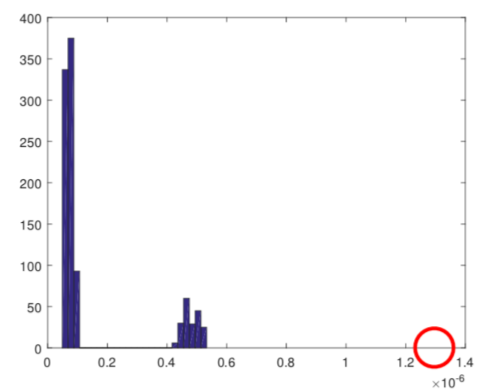
  comparing the distribution of diffusion tensors in various regions

Application: Diffusion MRI Data

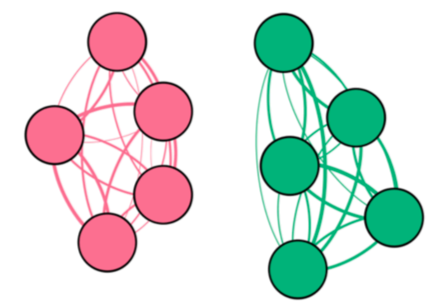
- Downsample pixels by a factor of 5 for reference points
- Synthetic datasets: 5 healthy, 5 sick subjects



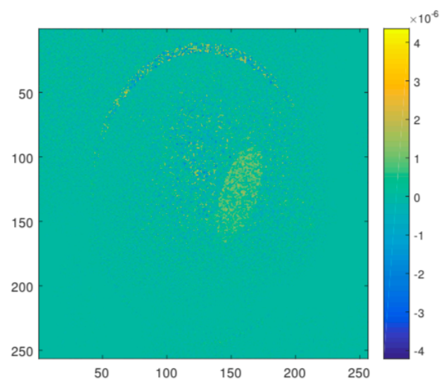
H_1 Witness (Anisotropic)



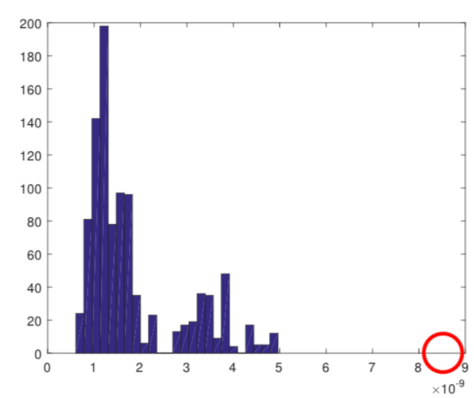
H_1 Perm. Test (Anisotropic)



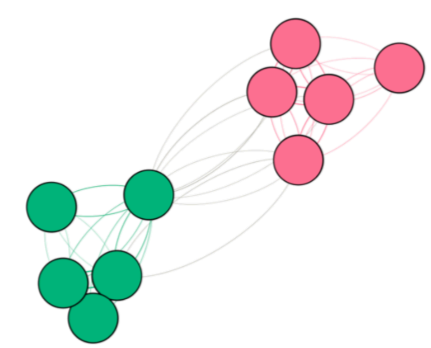
H_1 Graph (Anisotropic)



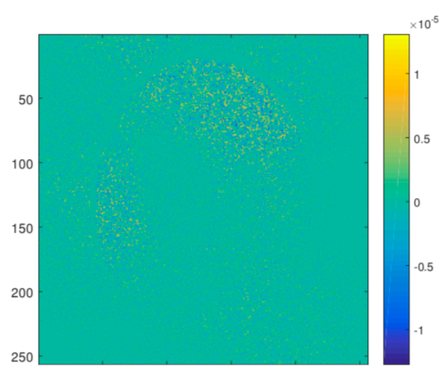
H_1 Witness (Isotropic)



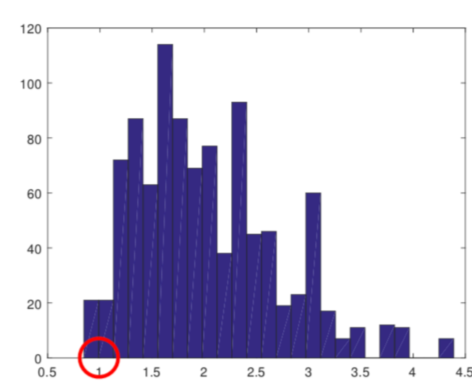
H_1 Perm. Test(Isotropic)



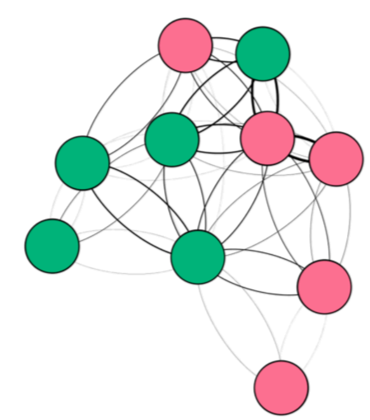
H_1 Graph (Isotropic)



H_0 Witness (Anisotropic)



H_0 Perm. Test (Anisotropic)



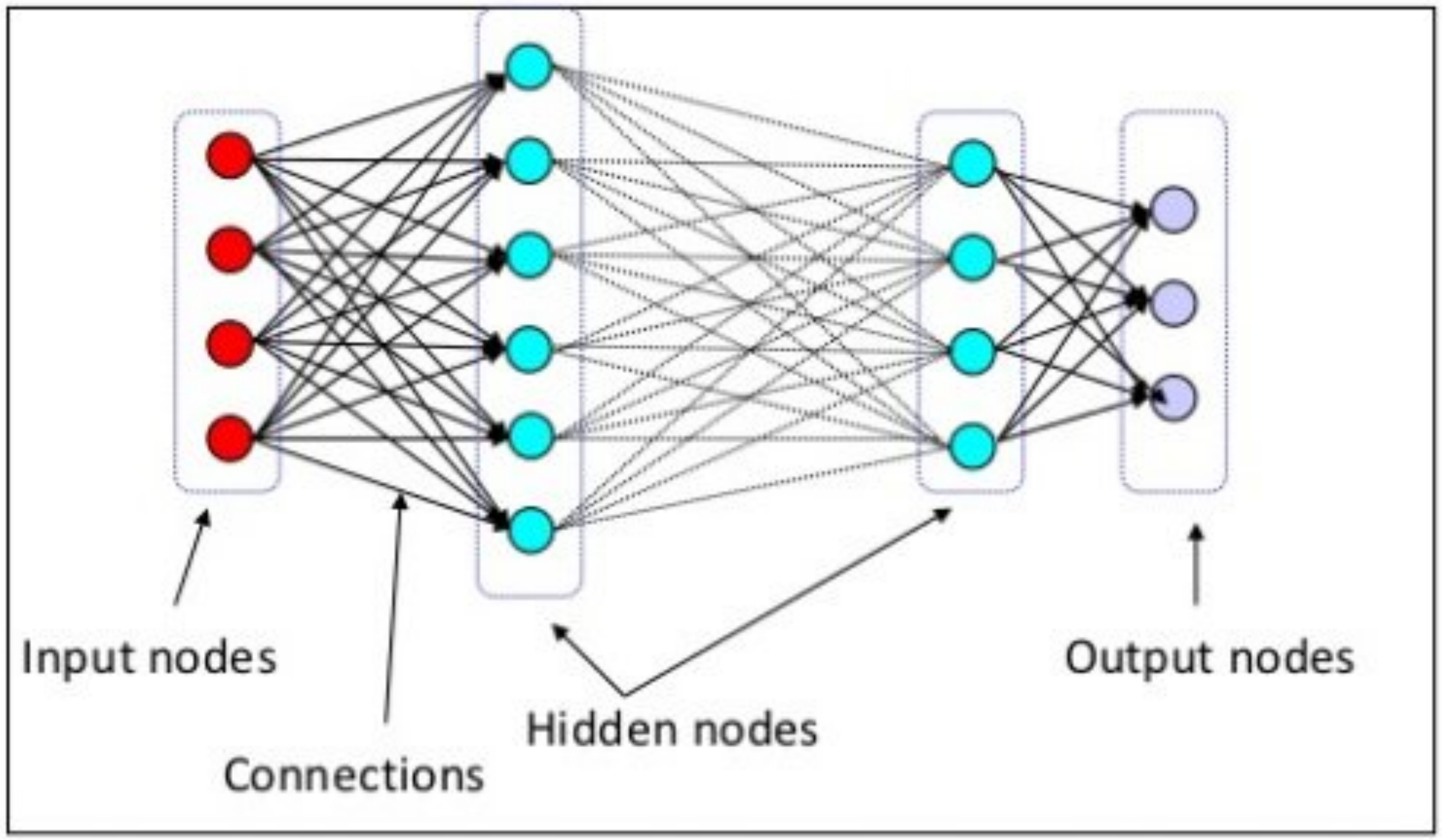
H_0 Graph (Anisotropic)

Outline

- Background
  - Two-sample problem
  - Kernel MMD and data geometry
- Anisotropic kernel MMD test
  - Test statistic and algorithm
  - Testing power analysis
  - Application: Flow Cytometry data
  - Application: Diffusion MRI imaging
- **Discussion: by neural network?**

Neural Network Classifier

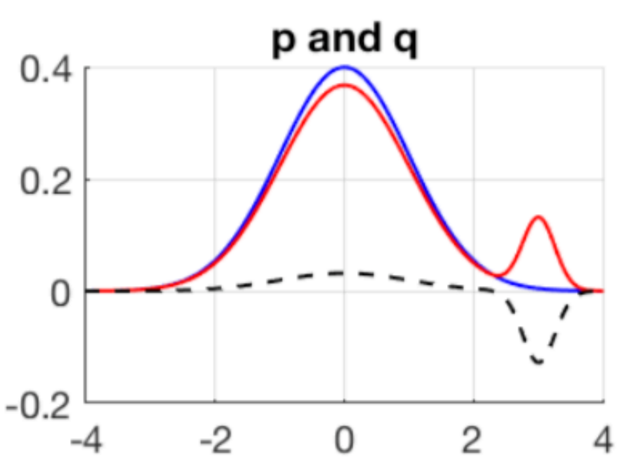
- Network classifier two-sample test

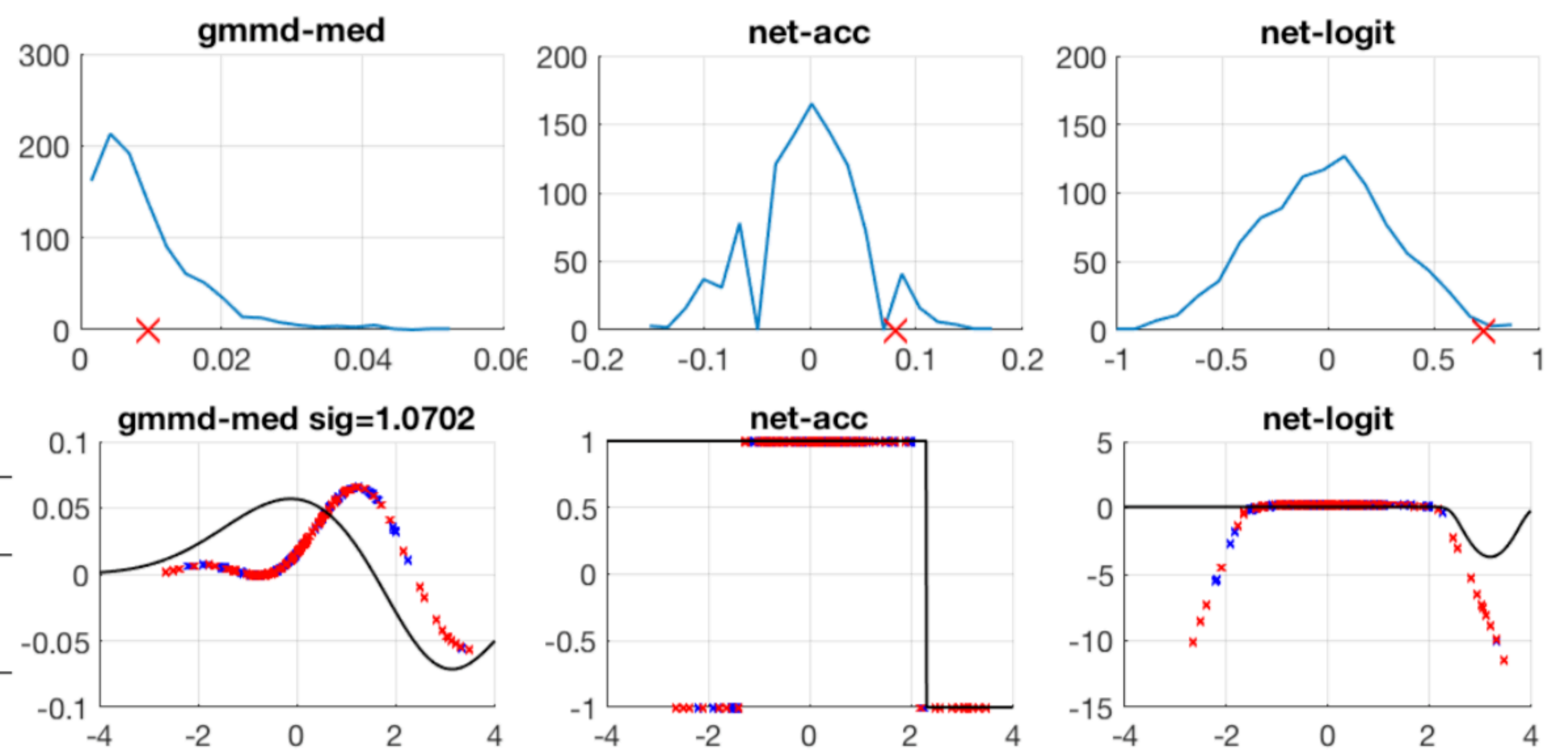


- Test statistic:
  - Classification accuracy [Lopez-Paz et al. '16]
  - Net-logit (ours)

Neural Network Classifier

- 1D toy example
  - Fully-connected network with 32 nodes in 2 hidden layers
  - Gaussian mmd with median distance as kernel bandwidth





	gmmd	*net-acc*	*net-logit*
mean	19.14	19.98	78.09
std	1.95	10.43	20.56
median	19.63	17.63	84.13

Test Power

Papers & Preprints

- X. Cheng, A. Cloninger and R. R. Coifman. “Two-sample statistics based on anisotropic kernels”. To appear at *Information and Inference: A Journal of the IMA* (2019). [arXiv:1709.05006]
- X. Cheng, A. Cloninger, “Neural network classifier log-ratio two-sample tests for densities on manifolds”, in preparation.

Questions?

Thank You!