Fast Point Cloud Distances and Multi-Sample Testing

Alex Cloninger

Center for Computational Math
University of California, San Diego

Collaborators

- Xiuyuan Cheng (Duke)
- Raphy Coifman (Yale)

Outline

Two Sample Tests

- **Question:** Suppose $X \sim p$ and $Y \sim q$ on $\mathbb{R}^d$

$$H_0 : \quad p = q$$
$$H_1 : \quad p \neq q$$

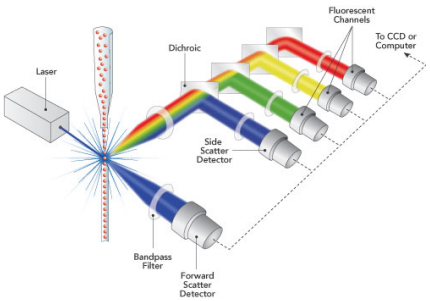
- Statistical in nature
  - Goal is to convince you this touches on
    - Spectral theory
    - Optimization
    - Data science in medicine

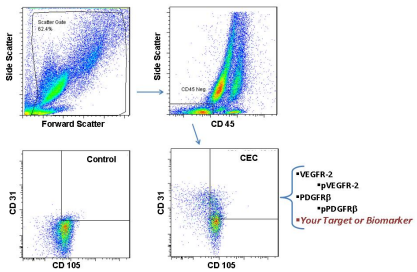
Two Sample Test Applications

- **Question:** Suppose $X \sim p$ and $Y \sim q$ on $\mathbb{R}^d$

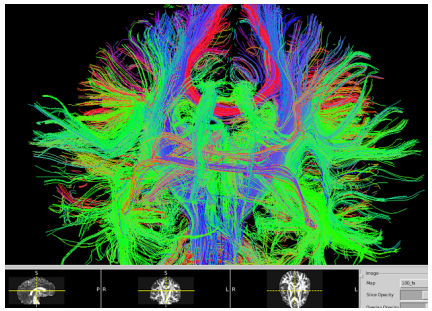
$$H_0 : \quad p = q$$
$$H_1 : \quad p \neq q$$

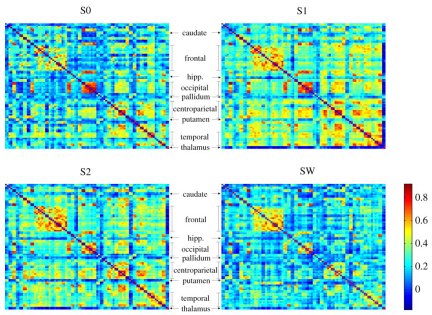
Flow Cytometry



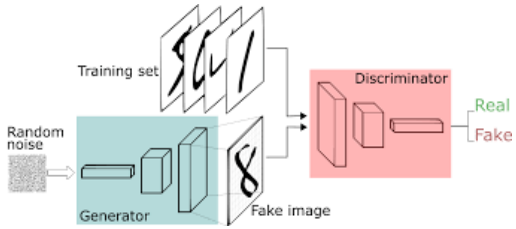


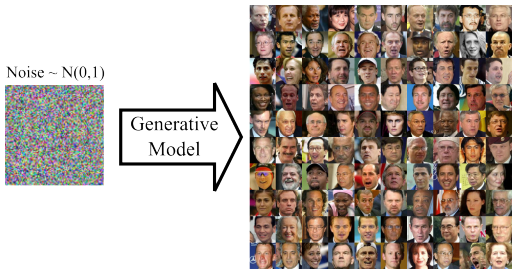
Diffusion MRI





GANs



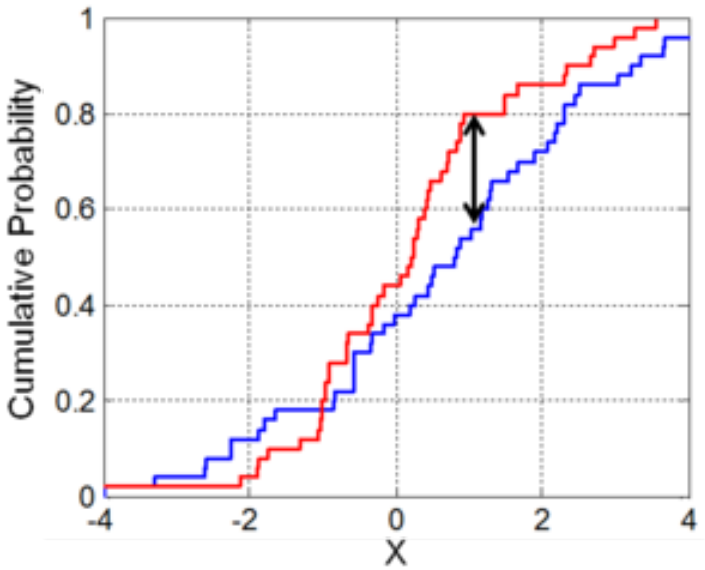


Two Sample Tests in 1D

- **Question:** Suppose $X \sim p$ and $Y \sim q$ on $\mathbb{R}^d$

$$H_0 : \quad p = q$$
$$H_1 : \quad p \neq q$$

- **Answer:**
  - *Easy in 1D:* Kolmogorov-Smirnov



Valores críticos de Kolmogorov-Smirnov

Grados de libertad (N)	Prueba de una muestra* $D_{0,10}$	$D_{0,05}$	$D_{0,01}$	Prueba de dos muestras † $D_{0,05}$	$D_{0,01}$
1	0.950	0.975	0.995		
2	0.776	0.842	0.929		
3	0.642	0.708	0.828		
4	0.564	0.624	0.733	1.000	1.000
5	0.510	0.565	0.669	1.000	1.000
6	0.470	0.521	0.618	0.833	1.000
7	0.438	0.486	0.577	0.857	0.857
8	0.411	0.457	0.543	0.750	0.875
9	0.388	0.432	0.514	0.667	0.778
10	0.368	0.410	0.490	0.700	0.800
11	0.352	0.391	0.468	0.636	0.727
12	0.338	0.375	0.450	0.583	0.667
13	0.325	0.361	0.433	0.538	0.692
14	0.314	0.349	0.418	0.571	0.643
15	0.304	0.338	0.404	0.533	0.600
16	0.295	0.328	0.392	0.500	0.625
17	0.286	0.318	0.381	0.471	0.588
18	0.278	0.309	0.371	0.500	0.556
19	0.272	0.301	0.363	0.474	0.526
20	0.264	0.294	0.356	0.450	0.550
25	0.24	0.27	0.32	0.40	0.48
30	0.22	0.24	0.29	0.37	0.43
35	0.21	0.23	0.27	0.34	0.39
Más de 35	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$	$1.36\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1.63\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

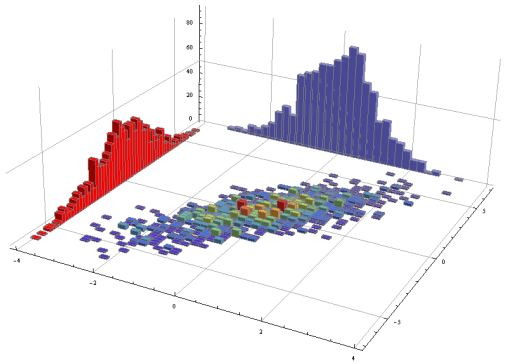
* Usada con objeto de probar la bondad de ajuste de una muestra a una distribución donde N = tamaño de la muestra.

† Usada para determinar si dos muestras provienen de la misma distribución. En el caso de las muestras pequeñas (que llegan hasta 35), $N = n_1 = n_2$.

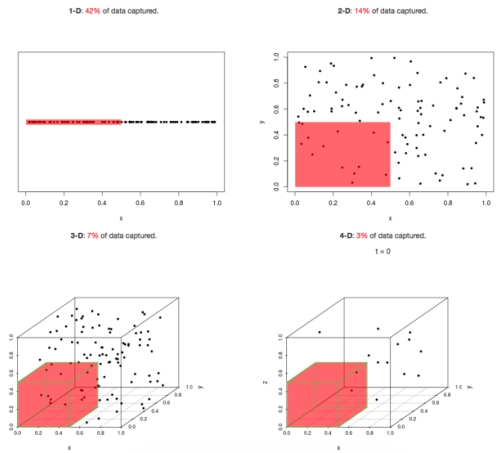
  - *Hard in nD:* Topic of this work

Two Sample Tests in nD

- Why is it hard in higher dimensions:
  - Marginals of distribution are insufficient



  - Difficult to define relevant bins
  - Curse of dimensionality: most bins will have very few points
    - Minimax rate exhibits curse (Arias-Castro et al 2017)



Two Sample Tests in nD

- **More general question:** How do we define a distance between X and Y?

$$d(X, Y) < \epsilon \implies p = q$$
$$d(X, Y) > \epsilon \implies p \neq q$$

  - Exist ways to choose ϵ: permutation test
- **Additional Questions**
  - Not just interested in whether they deviate, but how and where?
  - Can we extend to $k-$samples and use distance matrix between pairwise samples?
  - How without assumption of underlying manifold structure?

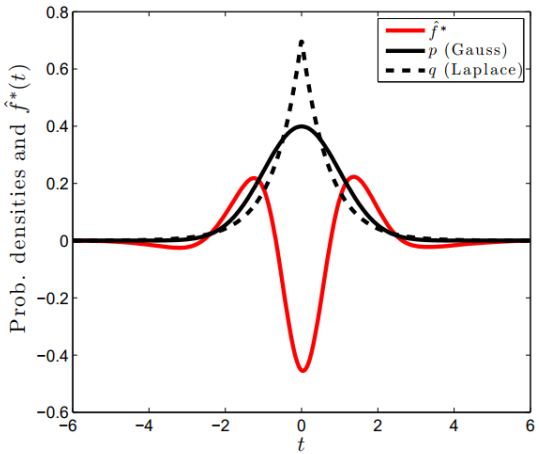
Outline

Bins and Locality

- No point-to-point correspondence, so use bins/histograms
  - Bins too small, lead to high variance
  - Bins too large, lead to poor precision
- More generally by *maximum mean discrepancy* with some set of functions

$$MMD(p, q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \left(\int f(x) dp(x) - \int f(x) dq(x) \right)$$

$$\widehat{MMD}(X, Y; \mathcal{F}) = \sup_{f \in \mathcal{F}} \left(\frac{1}{|X|} \sum_{x \in X} f(x) - \frac{1}{|Y|} \sum_{y \in Y} f(y) \right)$$



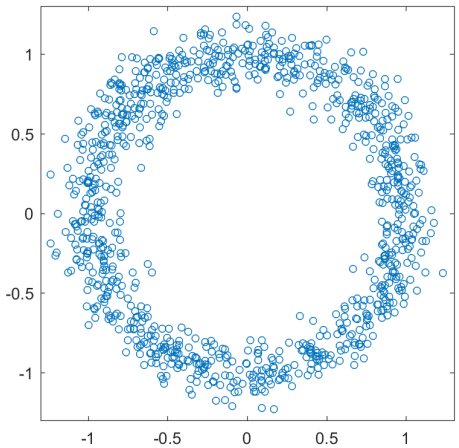
Avoid Optimization with Kernels

- **Problem:** Which function classes are tractable?
- **Possible Solution:** Bins defined by a *kernel* $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$

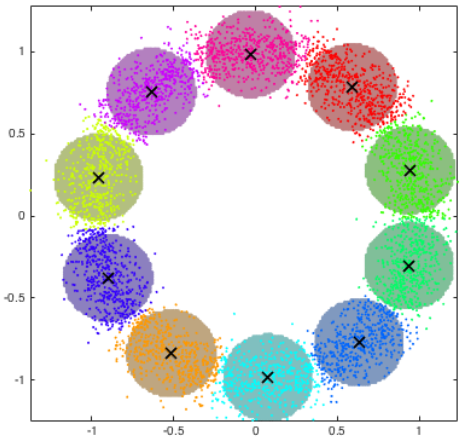
$$k(x, y) = e^{-\|x-y\|^2/\sigma^2} \quad \text{(Example)}$$

- Take $\mathcal{F}$ as unit ball in Reproducing Kernel Hilbert Space $\mathcal{H}(k)$

$$f \in \mathcal{H} \text{ if } \mathcal{F}_t[f] = \langle f, k(t, \cdot) \rangle = f(t)$$



Data with local geometry



Example Balls of Affinity > 0.1

Maximum Mean Discrepency

- Want to define density at any $z \in \mathbb{R}^d$

$$\mu_p(z) = \mathbb{E}_{x \sim p}[k(z, x)]$$

- Then avoid optimization

$$\begin{aligned} MMD^2(p, q; \mathcal{F}) &= \left[\sup_{f \in \mathcal{F}} \left(\mathbb{E}_x f(x) - \mathbb{E}_y f(y) \right) \right]^2 \\ &= \left[\sup_{f \in \mathcal{F}} \langle \mu_p - \mu_q, f \rangle \right]^2 \\ &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \end{aligned}$$

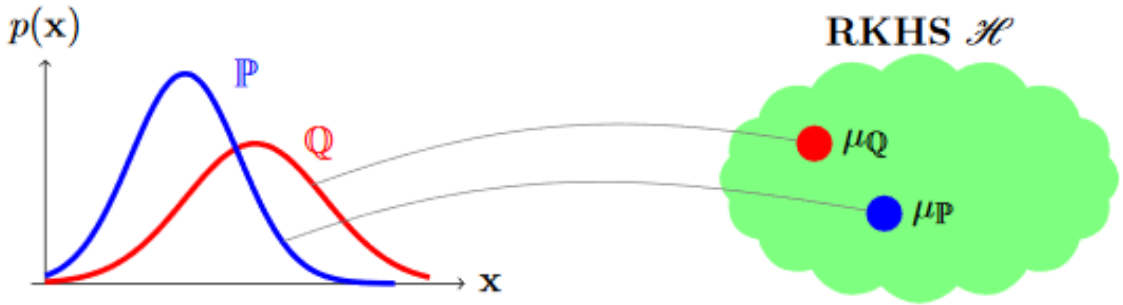


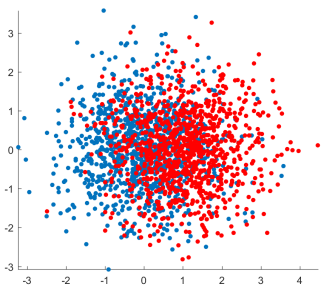
Figure 1.1: Embedding of marginal distributions: Each distribution is mapped into an RKHS via an expectation operation.

Discrete MMD

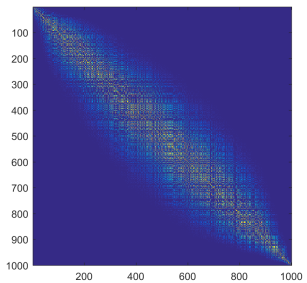
- Still hard to compute, but exists equivalent discrete version

$$MMD(X,Y) = \frac{1}{n(n-1)} \sum_{x,x' \in X} k(x,x') + \frac{1}{m(m-1)} \sum_{y,y' \in Y} k(y,y') - \frac{2}{mn} \sum_{x \in X, y \in Y} k(x,y)$$

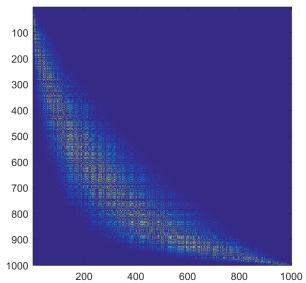
- Avg. Affinity within X, Avg. Affinity within Y
- Avg. Affinity between X and Y



Data X and Y



$K(X,X)$



$K(X,Y)$

Using MMD

Guarantees (Gretton, et al. 2011)

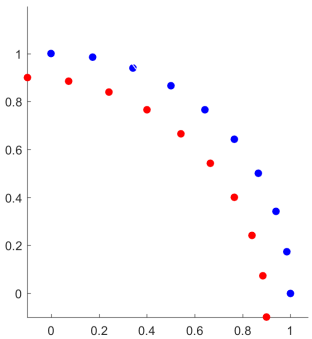
- $MMD(p, q)$ is one-to-one mapping for $\|p - q\|_\infty$
- If $p = q$, $MMD(X, Y) \to 0$ like $\frac{C}{\sqrt{n+m}}$
  - C depends on bandwidth of kernel
- If $p \neq q$, $MMD(X, Y)$ minimum distance detectable is $\|\mu_q - \mu_p\| = \frac{c}{\sqrt{n+m}}$

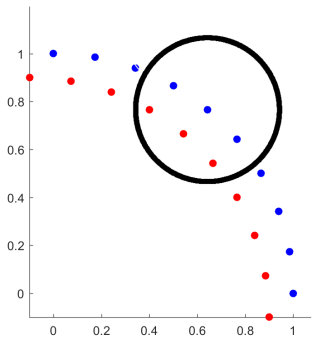
Problems

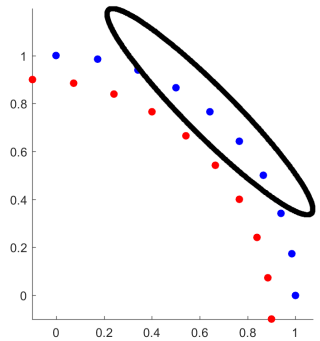
- Kernel is isotropic
  - Treats data on single scale
  - Convergence depends on dimension of ambient space (Wasserman et al 2014)
- $p \neq q$ results don't speak to power of test
- O $\left(n^2\right)$ storage of K
  - Completely intractable for $k-$sample problem and network geometry

Problems

- Kernel is isotropic
  - Treats data on single scale
  - Convergence depends on dimension of ambient space (Wasserman et al 2014)
- $p \neq q$ results don't speak to power of test
- $O\left(n^2\right)$ storage of K
  - Completely intractable for $k-$sample problem and network geometry
- Introduction of local geometry creates data-adaptive test
  - Penalizes moving in normal direction
  - Discounts regions of high-volatility
  - Allows assumptions of local low dimensionality and off manifold deviation







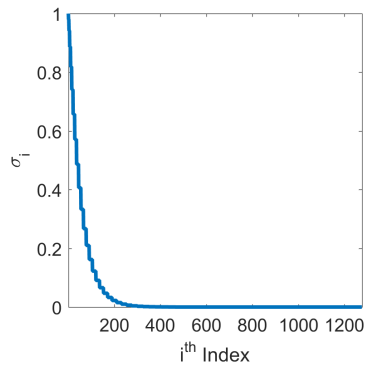
Outline

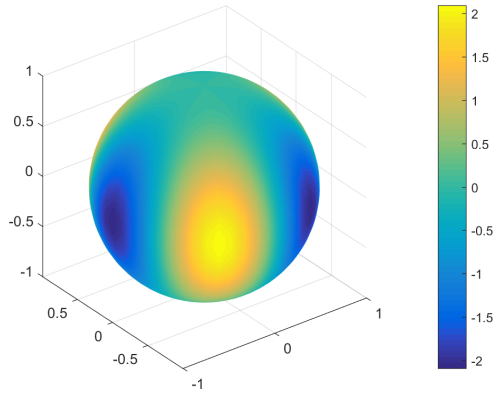
Spectral Convergence

- Easier to work with mean centered kernel $\tilde{k}$ and spectrum

$$\tilde{k}(x,y) = \sum_i \sigma_i \phi_i(x)\phi_i(y), \quad \tilde{k} \in L^2(\mathcal{X} \times \mathcal{X}, p \times p), \quad \sigma_i \to 0$$

- Eigenfunctions have a long history of determining shape
  - Principle Component Analysis
  - Spectral graph theory
- As note, $(D(k) - k)\,\phi = \lambda\phi \to -\Delta f = \lambda f$
  - Implies eigenvalues of $\tilde{k}$ not sufficient to differentiate between datasets



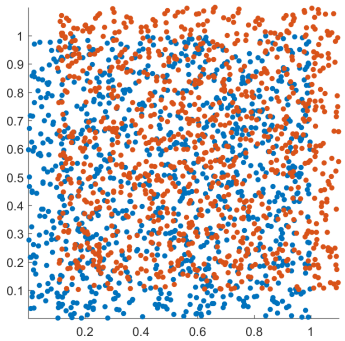


Convergence of Eigenvectors

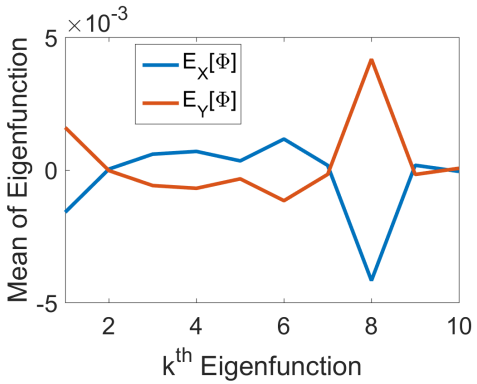
- If $p = q$, goal to show

$$\frac{1}{|X|}\sum_{x\in X}\phi_i(x) \to 0, \qquad \frac{1}{|Y|}\sum_{y\in Y}\phi_i(y) \to 0$$

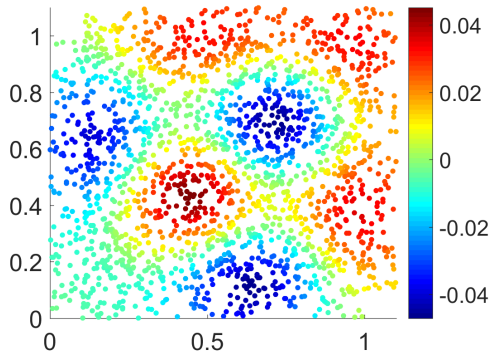
- If $p \neq q$, show convergence to non-zero constants
- Must treat ϕ_i simultaneously b/c samples not independent
  - Multi-dim central limit theorem and spectral decay



X and Y



$\frac{1}{|X|}\sum_{x\in X}\phi_i(x)$ and $\frac{1}{|Y|}\sum_{y\in Y}\phi_i(y)$



$\phi_8(x)$

Eigenvectors are “Redundant”

- Approximate rank of K determined by number of balls needed to cover data (Tygert, Rokhlin 2008; Kühn 2011)
  - For acceptable bandwidth, $|\{i|\sigma_i(K) > \epsilon\}| \ll n$
- Using square matrix is over-redundant
  - Can choose small set of “reference points” with balls that easily cover data

Eigenvectors are “Redundant”

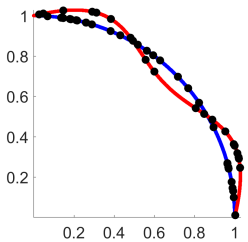
- Approximate rank of K determined by number of balls needed to cover data (Tygert, Rokhlin 2008; Kühn 2011)
  - For acceptable bandwidth, $|\{i|\sigma_i(K) > \epsilon\}| \ll n$
- Using square matrix is over-redundant
  - Can choose small set of “reference points” with balls that easily cover data

Anisotropic Kernels

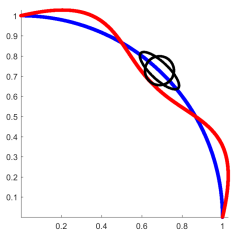
- Assume ambient space has local geometry (r, Σ_r)
  - Sample r from $X \cup Y$
    - Random subsample
    - QR with pivoting
    - **Quadrature methods**
  - Σ_r from covariance of nearest neighbors
- Choose set of "representative points" R to compare distributions
  - Test only as good as reference points

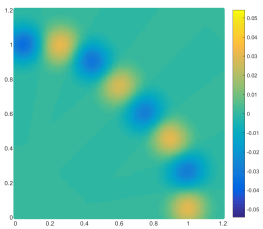
$$A(r, x) = e^{(x-r)^T \Sigma_r^{-1} (x-r)}$$

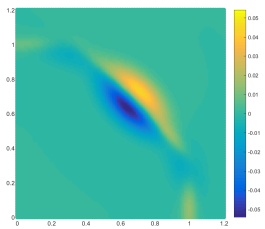
- Still have mean embedding $\mu_X(r) = \mathbb{E}_{x \in X}[A(r, x)]$



Data/Ref

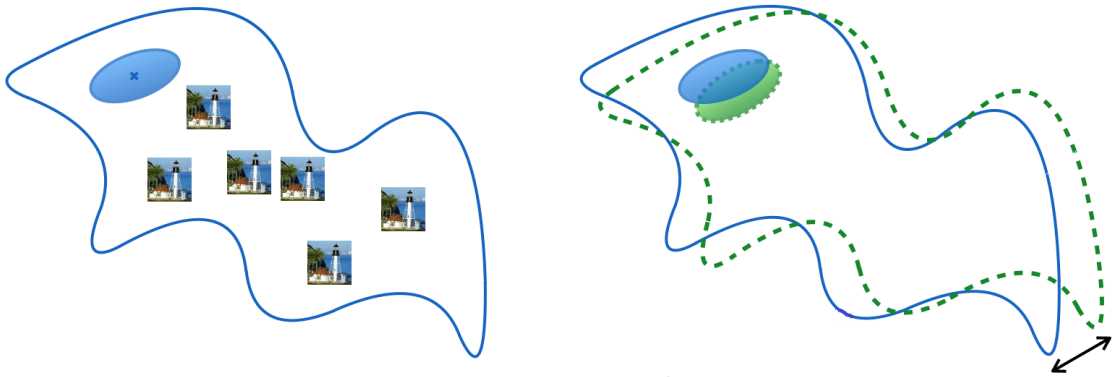

Neighborhood


Gauss Eig


Mahal Eig

Adaptive Kernel MMD

- $AMMD(X, Y; \{r, \Sigma_r\}) = \|\mu_X(r) - \mu_Y(r)\|_2$
  - Distance of distribution proj. to lighthouses
  - Lighthouses contain directional and dimension information
- Calculation is $O(N|R|d)$
  - Never calculating eigendecomposition



Theory for Anisotropic Kernels

Assumption: R sufficiently span column space of K

- Parallel to kernel "having non-vanishing Fourier transforms on any interval" for KDE
- $q = p + \tau g$

Informal Theorems (C., Cheng, Coifman 2017)

All shifts and variance depend on spectrum and how quickly $\int \psi_k(y) g(y) dy$ detect deviation

1. If $\tau = O(n^{-1/2})$, then nT_n is χ^2
2. If $\tau = O(n^{-1/2+\delta})$, then $n^{1-\delta} T_n$ is normal with shift "larger than variance" as $n \to \infty$
3. If $\tau = 1$, $\sqrt{n} T_n$ is normal with shift "larger than variance"

Power of the Test

- Need to know if deviation is enough in comparison to H_0
  - Either true H_0 or permutation null
- Important to know not just threshold but for fixed deviation
  - Power is prob to detect deviation when it exists

Power (C., Cheng, Coifman 2017)

Notations as above, under Assumption, for specific $g = q - p$ fixed, if $\tau_n = O(n^{-1/2+\delta})$ where $0 < \delta \leq \frac{1}{2}$ ($\delta = \frac{1}{2}$ means that $\tau = 1$), the test power $\pi_n(p + \tau_n g) \to 1$ as $n \to \infty$.

Also extends for MMD results of Gretton et al

Outline

Bounded Diffusion Subsampling

- Choosing (r, Σ_r) currently done by random sampling of $X \cup Y$
  - Effectively Monte Carlo integration
- Reframe as problem of efficiently learning eigenfunction means in CLT

Bounded Diffusion Subsampling

- Choosing (r, Σ_r) currently done by random sampling of $X \cup Y$
  - Effectively Monte Carlo integration
- Reframe as problem of efficiently learning eigenfunction means in CLT

Near Perfect Spherical Designs; (Steinerberger, Linderman, 2018)

For $a_r \geq 0$ and $\sum a_r = 1$, then for any f expressible in terms of low-freq eigenfunctions of graph $G = (X, K)$,

$$\left| \frac{1}{|R|} \sum_{r \in R} a_r f(r) - \mathbb{E}_{x \in X}[f(x)] \right| \leq \frac{\|f\|_{X_\lambda}}{\lambda^t} \left(\left\| K^t \sum_{r \in R} a_r \delta_r \right\|_2^2 - \frac{1}{|X|} \right)^{1/2}$$

Geometric interpretation: R distributed to diffuse sufficiently quickly for small t.

Reference Point MMD

Reference points subsample

$$\int_{z \in X \cup Y} f(z)dz = \int_{z \in X \cup Y} |\mu_X(z) - \mu_Y(z)|^2 dz$$

Non-Asymptotic MMD Reference Error; (C. 2018)

Assuming $\mu_X - \mu_Y$ has energy τ and projects mostly projects "mostly" $(1 - \epsilon)$ onto eigenfunctions $\lambda > \nu$, for fixed reference set R,

$$\left|AMMD(X, Y; a_r) - MMD(X, Y; K)\right|^2 < \frac{\tau^2}{\nu^2}\left(\left\|K\sum_{r \in R} a_r \delta_r\right\|_2 - \frac{1}{N}\right) + \epsilon^2 \tau^2 \left(\sum_{\lambda < \nu} \lambda^2\right)$$

Effectively dominated by finite error $\frac{c}{\sqrt{N}}$

Currently working on fast preprocessing optimization scheme to minimize over choice of R and a_r

Complexity

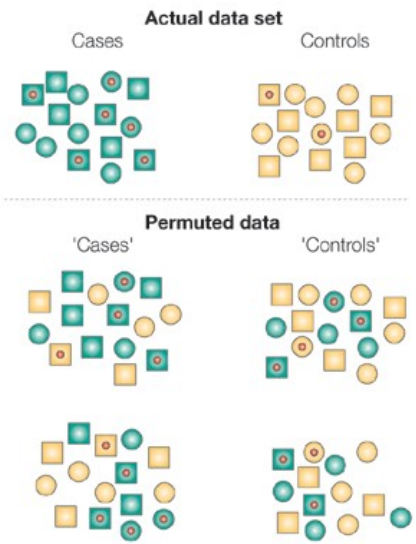
- For $k-$sample problem, only need comparison to fixed reference set
- Requires one loop through data for AMMD, and one loop over $\mu_p(r)$

$k-$sample test from $\mathrm{O}\left(\binom{K}{2} N^2 \cdot D\right)$ to $\mathrm{O}\left(\binom{K}{2} |R| + K \cdot N |R| D\right)$

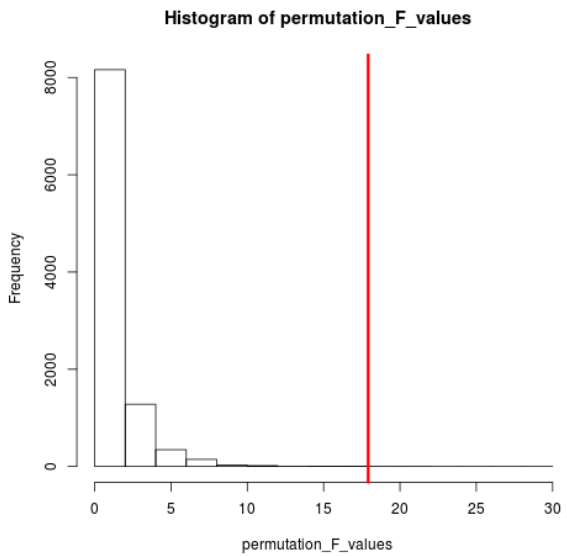
IsotropicMMD *RefSetAMMD*

Using MMD

- In practice, can use permutation test
  1. Set $Z = X \cup Y$ and define permutation $p = \begin{bmatrix} p_1 & p_2 \end{bmatrix}$ of $\{1, ..., n+m\}$
  2. Compute $MMD(Z_{p_1}, Z_{p_2})$
  3. Create histogram over N iterations of p for null hypothesis
  4. If $\#\{p : MMD(X, Y) > MMD(Z_{p_1}, Z_{p_2})\}/N > 0.95$, reject H_0



Hirshorn, Daly

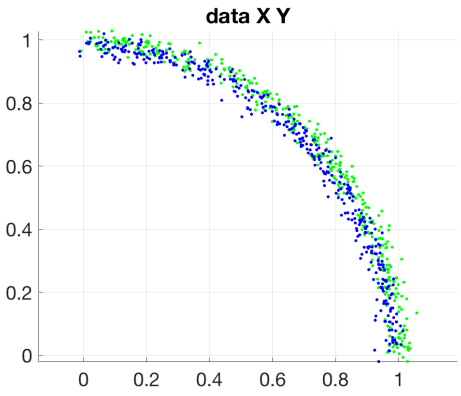


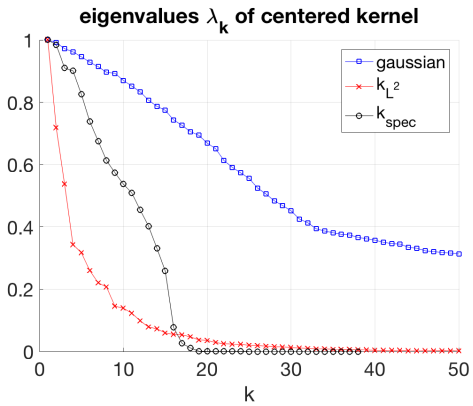
Perm Test Histograms

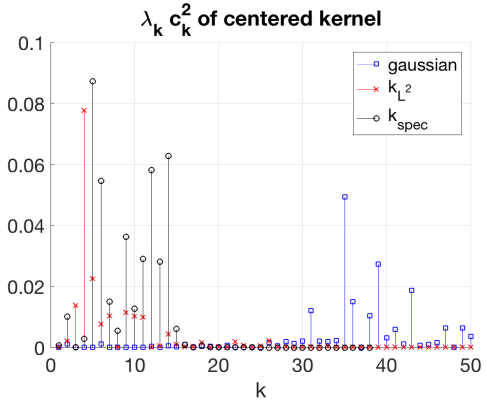
Outline

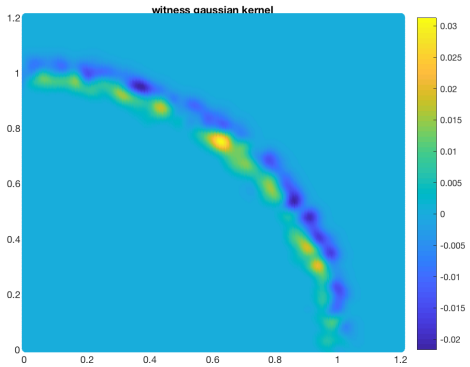
Comparison to Isotropic Gaussian

- At the moment, mostly empirical comparison of spectrum and deviation

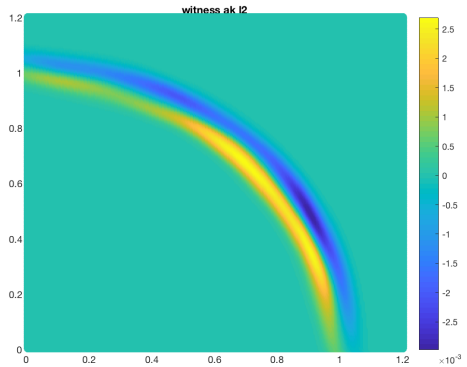




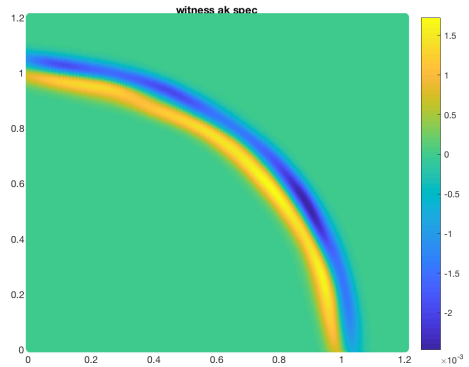




Gauss Witness



w_{L2}



w_{spec}

Adaptive Kernels Example

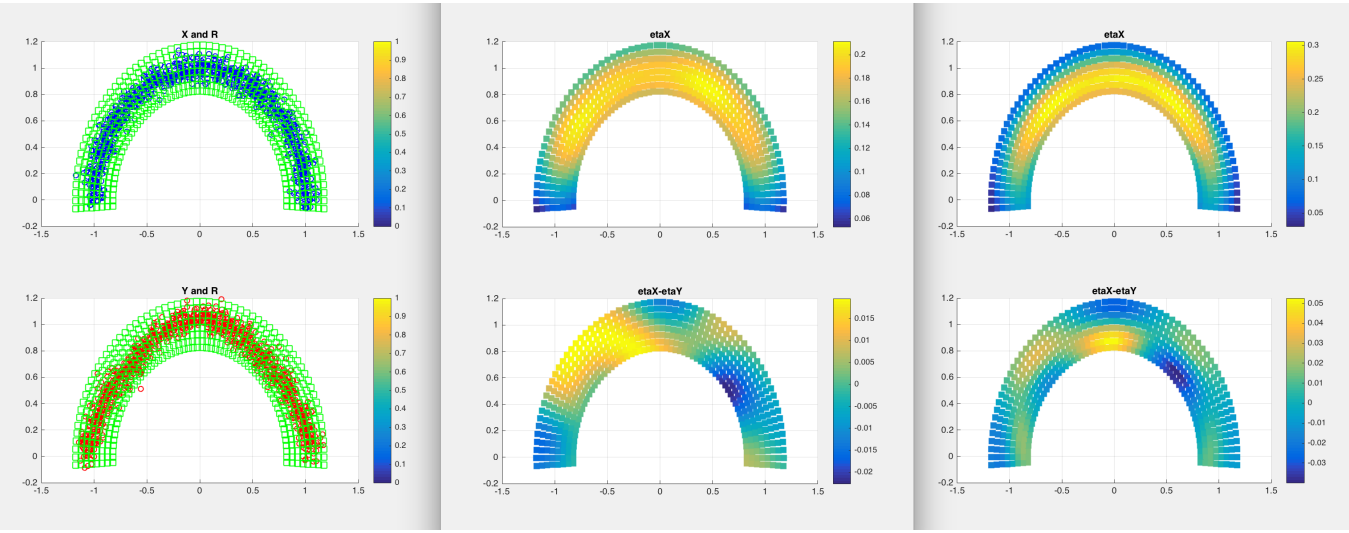
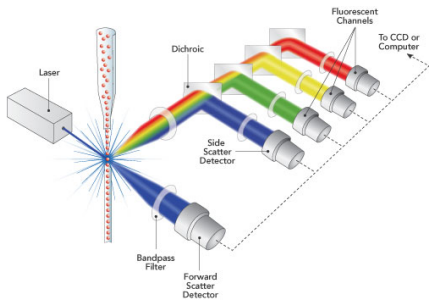
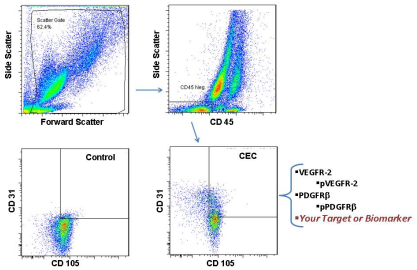


Figure: Estimated $\xi_p(r)$ and $\xi_q(r)$ for two distributions p and q with 1000 samples each (left) with gaussian kernel (middle) and anisotropic kernel (right) respectively.

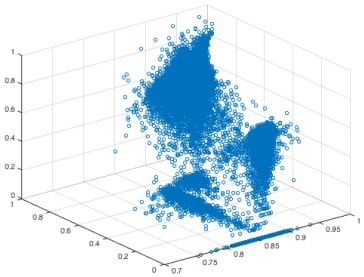
Real Example

- Flow cytometry: each patient is represented by 9D point cloud of cells

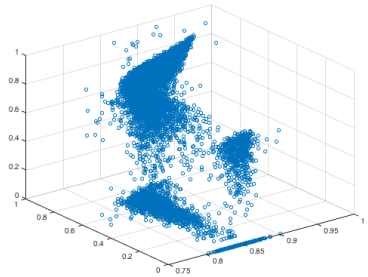




- Used to tell if people have blood disease
  - Medical test is to look at every 2D slice



Healthy

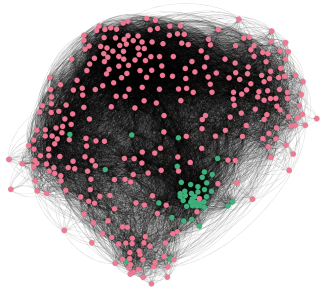


AML

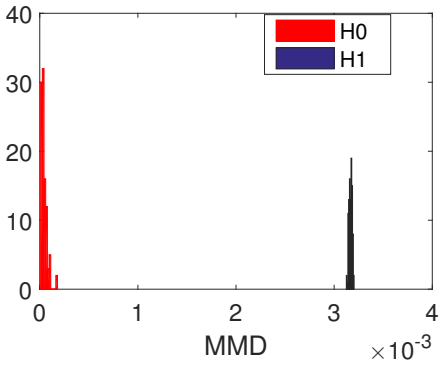
Unsupervised Organization:

- Measure distance between every two people X_i and X_j
  - Form network of people
  - Use eig. fctns. of distance matrix for unsupervised clustering

Distance Between Flow Cytometry Measurements

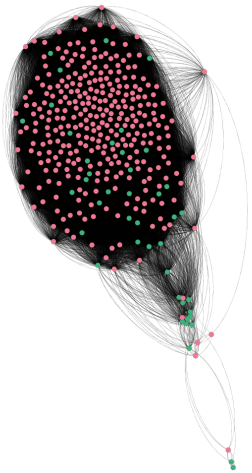


Mahalanobis Network

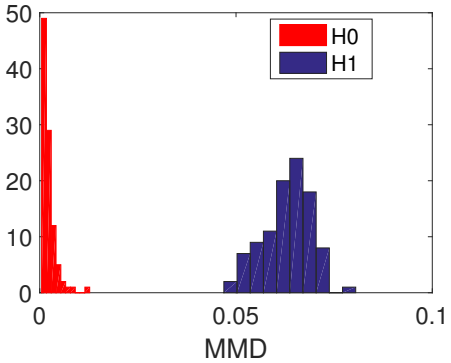


Perm. Test

$\sim$ 10 minutes



Gauss Network



Gauss Perm. Test

$\sim$ 1 day

Interpretability and Classification

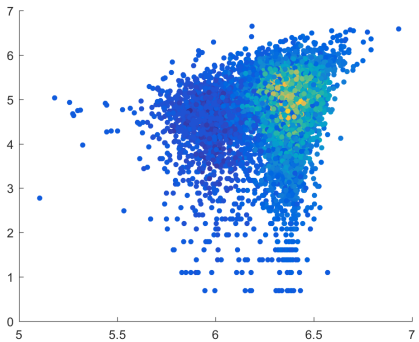
Supervised Information:

- Kernel MMD has dual form

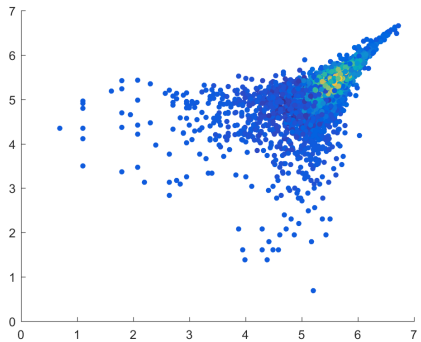
$$MMD(p, q, \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{x \in p} f(x) - \mathbb{E}_{y \in q} f(y)|$$

- Arg max f is called *witness function*
  - Interprets regions of "importance" to differentiate p and q
- Computable as

$$f(x) = \int \left(\int a(r, x) a(r, y) p_R(r) dr \right) (p(y) - q(y)) dy$$



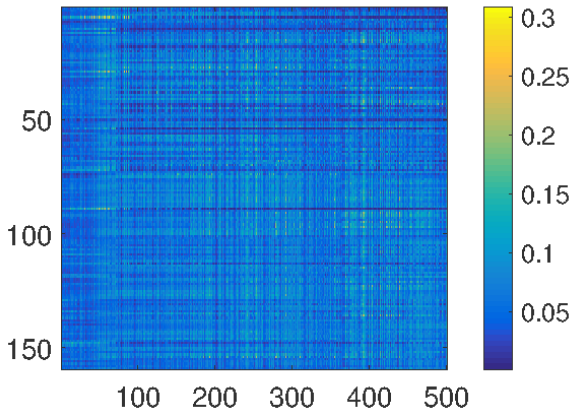
2D Slice Witness



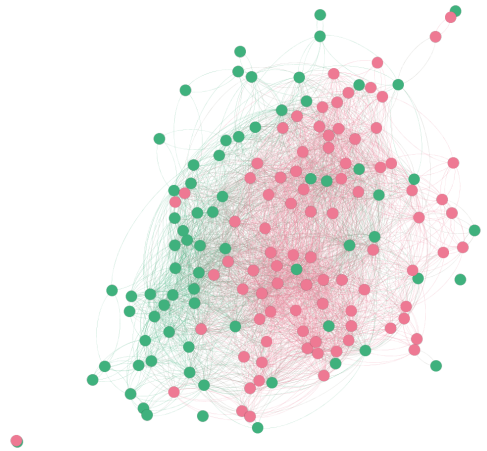
2D Slice Witness

Flow Cytometry from MDS

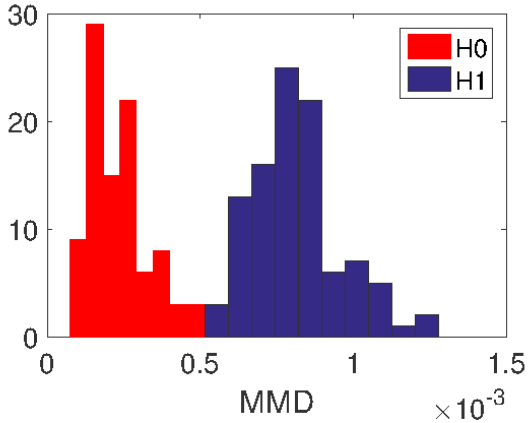
- MDS is a more general class of blood cancers
- Much more difficult to detect than AML



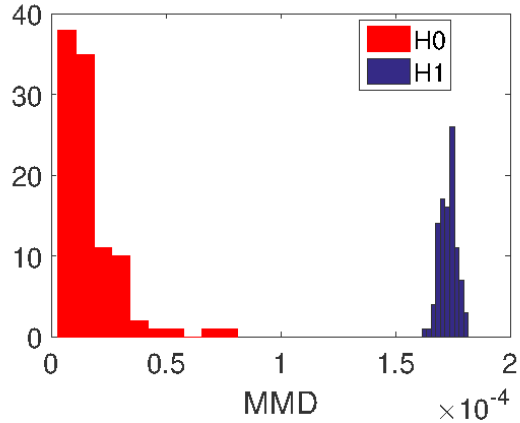
$[h_i(r)]_{i=1}^{159}$



Network pairwise distances $[h_i(r)]_{i=1}^{159}$



Permutation Test Gaussian Kernel



Permutation Test Anisotropic Kernel

General Take-Aways

- Adaptive kernels are necessary tool for data science
- Kernels allow extreme flexibility for defining point-to-point correspondence
  - Multi-scale affinity kernels
  - Function model driven kernels
  - Features generated from any transform

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

- Extensions:
  - Multi-bandwidth generalization
  - Local significance of deviation
  - Generalization to non-iid samples (time series)
  - GANs

Thank you!

Questions?