Local differences between distributions and distance measures

Alex Cloninger

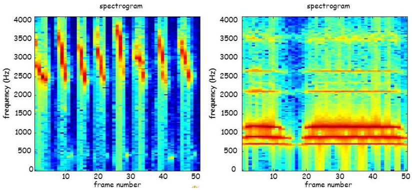
Department of Mathematics
University of California, San Diego

Collaborators

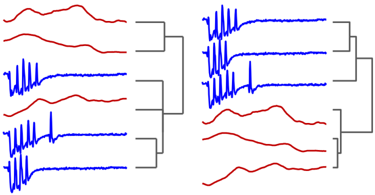
- Srinjoy Das (UCSD)
- Hrushikesh Mhaskar (Claremont Graduate University)
- Xiuyuan Cheng (Duke University)

Motivating Example

- **Goal:** Learn unsupervised clustering of data sets $\{X_i\}_{i=1}^K$
  - Point Cloud
  - Embedded time series windows
  - Embedded image patches
- **Largest issues:**
  - Non-i.i.d. sampling
  - Shift-invariance
  - Saliency of foreground
  - Repeating motifs
  - Subsequence similarity



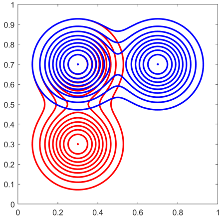
Bird Chirp (Kaggle)

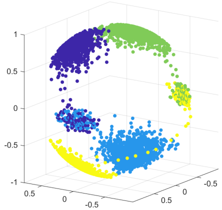


MPDIST (Keogg, et al. 2018)

Partial Overlap of Distributions

- In many situations, distributions don't perfectly match
  - Foreground / background patches
  - Background noise between "chirps"
  - Motivates questions
    Question: How to define a statistic on shared foreground that's independent of background
  - **Sub-question:** How do we define robust statistic for overlap of distributions from finite samples





Outline

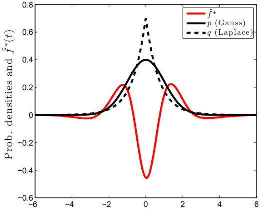
Outline

Importance of Where Distributions Deviate

Problem 1: Detect where two distributions deviate given only finite samples

- Motivation:
  - Want to highlight region(s) that deviate between two samples
  - Determine region of uncertainty where points may be from either distribution
- Goals:
  - Examine stability of deviation detection as $n \to \infty$
  - Build deviation detection that is robust and cautious
- **Initial Solution 1:** Maximum Mean Discrepancy witness function

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



Kernel Differences in Distributions

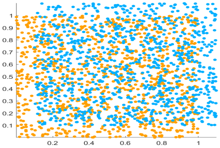
- Take $\mathcal{F}$ as unit ball in Reproducing Kernel Hilbert Space $\mathcal{H}(k)$

$$MMD(p,q;k) := \langle \mathbb{E}_{x\in p}k(\cdot,x) - \mathbb{E}_{y\in q}k(\cdot,y), \mathbb{E}_{x\in p}k(\cdot,x) - \mathbb{E}_{y\in q}k(\cdot,y)\rangle$$

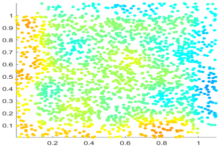
- Witness function maximizes difference

$$\begin{aligned} f^* &= \arg\max_{f\in\mathcal{F}} \left(\int f(x)dp(x) - \int f(x)dq(x) \right) \\ &:= \mathbb{E}_{x\in p}k(\cdot,x) - \mathbb{E}_{y\in q}k(\cdot,y) \end{aligned}$$

- Empirical witness can be very noisy and variable



Two Classes



Witness Function

Empirical Witness Function

Empirical Witness

$$\widehat{f^*} = \frac{1}{N}\sum_{x\in X} k(\cdot, x) - \frac{1}{M}\sum_{y\in Y} k(\cdot, y)$$

- **Question 1:** How does empirical witness converge to f^*?
- **Question 2:** Can we determine a test of whether $f^*(z) \neq 0$?

Kernel Choice

- For strong convergence guarantees, best to choose kernel as Mehler kernel

$$\Phi_n(x, y) = \sum_{k\in\mathbb{Z}_+^d} H\left(\frac{\sqrt{|k|_1}}{n}\right)\psi_k(x)\psi_k(y)$$

  for $\psi_k(x)$ multi-dimensional Hermite polynomial
- Exists Mahler identity to re-write as weighted exponential kernel
- Has fast decay properties but isn't non-negative $\forall(x, y)$

Stability of Witness Function

Local Concentration Bound (Mhaskar, Cheng, C. 2019)

Difference between empirical witness function $\widehat{f}^*(z) = \frac{1}{n}\sum_{x\in X}\Phi_n(z,x) - \frac{1}{m}\sum_{y\in Y}\Phi_n(z,y)$ with Mehler kernel and true witness f^* satisfies Hoeffding-type concentration for error measured in L^∞_{loc}. In particular, for

$$n \sim \left(\frac{N}{\log N}\right)^{1/(2d+2\gamma)},$$

and $f^* \in W_{\infty,\gamma}(x_0)$ we obtain for $r \geq c_1/n^2$ that

$$\mathrm{Prob}_\tau\left(\left\|\widehat{f^*} - f^*\right\|_{\infty,\mathbb{B}(x_0,r)} \geq c\frac{1+\|f^*\|_{\infty,\gamma,x_0,r}}{n^\gamma}\right) \leq \delta(r/n)^d.$$

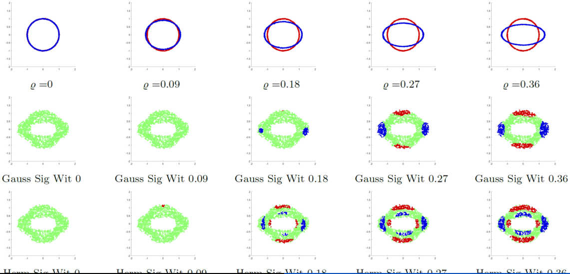
Permutation Test for Stability

- Assess hypothesis $f^*(z) \neq 0$ for $z \in B(x_0, r)$
- Measure through permutation $\pi : \mathbb{Z}_N \to \mathbb{Z}_N$

$$Sig(z) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1} \left[\left| \widehat{f}^*(z) \right| < \left| \widehat{f}_{\pi_i}(x_0) \right| \right], \text{ for}$$

$$\widehat{f}_\pi(x_0) = \frac{1}{N} \sum_{i=1}^{N} \Phi_n(x_0, x_{\pi(i)}) - \frac{1}{M} \sum_{i=M+1}^{N+M} \Phi_n(x_0, y_{\pi(i)})$$

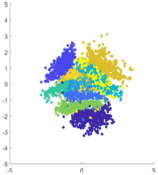
- For multi-class, use gap between largest class and second largest class as statistic



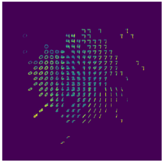
Outline

Variational AutoEncoder Significant Areas

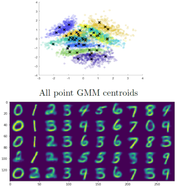
- Variational Autoencoder on MNIST creates 2D latent space
- Suggested model is to sample from $\mathcal{N}(0, I)$ but:
  - exist gaps between classes
  - exist regions where classes blur



Training Data

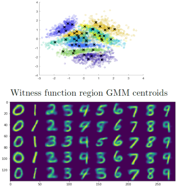


Significant Regions



All point GMM centroids

All point GMM centroids reconstructions

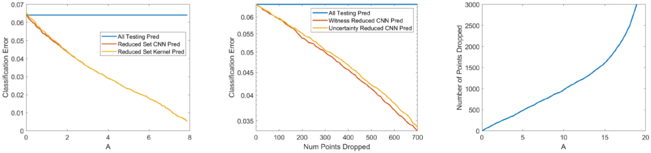


Witness function region GMM centroids

Witness function region GMM centroids reconstructions

CIFAR Uncertainty

- VGG-16 is state-of-the-art net that attains 6% classification on CIFAR10 test set
- Examine last layer for test points that are *significantly within a class*
- Choose not to classify others
    - Remove 7% of points and reduce testing error in half



Outline

Outline

Partial Overlap of Distributions

- Don't always have perfect distribution match
- Don't always have i.i.d. sampling of points
- **Goal:** Create statistic to measure whether distributions match *enough of the time*
- Example: images made into non-i.i.d. point clouds through patches



Kernel Quantile Algorithm

- Related to MP-DIST for time series (Keogg, 2018)
- Let $\mu = (p+q)/2$ and witness

$$f(z) = \left(\mathbb{E}_{x\sim p} k(z,x) - \mathbb{E}_{y\sim q} k(z,y)\right)^2$$

- Maximum mean discrepancy is average $\mathbb{E}_\mu f(z)$
  - Only unbiased if $p = q$
- Instead consider CDF and quantile measure

$$\begin{aligned} \mathit{CDF}: &\qquad \lambda_f(t) = \mu\left(\{z : f(z) < t\}\right) \\ \mathit{Quantile}: &\qquad Q_{p,q}(\alpha) = \sup_t \{t : \lambda_f(t) < \alpha\} \end{aligned}$$

- Quantile unbiased if p and q agree on α percent of their mass

Theoretical Toy Example

Small Commonalities (Das, Mhaskar, C., 2019)

Consider mixed distributions,

$$
\begin{aligned}
p_1 &= \delta p + (1-\delta) b_1 \\
p_2 &= \delta p + (1-\delta) b_2 \\
X_1 &\sim p_1, X_2 \sim p_2,
\end{aligned}
$$

for p, b_1, b_2 with disjoint support. Then for $x, x' \sim p$ and $y \sim b_1$ and $y' \sim b_2$, if $\|y - y'\|$ stocastically dominates $\|x - x'\|$ then there exits $\alpha > 0$ for Gaussian or Mehler kernel such that

$$
\begin{aligned}
Q_{X_1, X_2}(\delta\alpha) &\to 0 \\
MMD(X_1, X_2) &\to (1-\delta)^2 MMD(b_1, b_2).
\end{aligned}
$$

Similarly for $p_3 = \delta q + (1-\delta) b_3$ and $X_3 \sim p_3$, $Q_{X_1, X_2}(\delta\alpha)$ nonzero (greater than min of four quantiles).

Convergence Under non-i.i.d. Sampling

Barry-Essen Convergence (Das, Mhaskar, C., 2019)

Let $X \sim p$ and $Y \sim q$ for compact support p, q with exponential strong mixing, and X independent of Y. Then for the Mehler kernel,

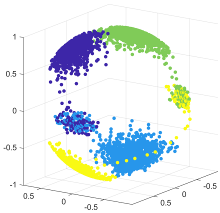
$$\sup_{x \in \mathbb{R}} \left| P(\sqrt{N}(Q_{X,Y}(\alpha) - Q_{p,q}(\alpha)) < x) - \Phi(x) \right| \leq \frac{C}{\sqrt{N}}.$$

- Requires three independent parts:
  1. Need $\widehat{f}^* \to f^*$ uniformly (augment Mhaskar, Cheng, C. 2018 with strong mixing Hoeffding inequality)
  2. Uniform convergence of witness gives convergence of empirical CDFs $\widehat{\lambda}_{\widehat{f}*} \to \widehat{\lambda}_{f*}$
  3. Need convergence of quantile from empirical CDF to true quantile under strong mixing (Lahiri, Sun, 2009)

Outline

Time Series Clustering

- Seek when time series behave similarly for given fraction of time
- AR(5) process that at random time jumps to new state
  - Anomalous states overlap, start state unique
  - Two instantiations of each stochastic process
- Window of length 20 in $3D$, Euclidean norm across all channels
- Quantile of $\alpha = 0.05$



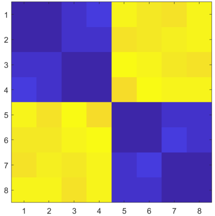
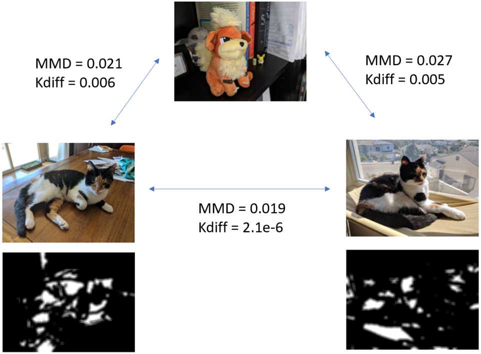


Image Foreground Similarity

- Took 5x5 patches of pixels and of edge extracted image (texture)
- Quantile of $\alpha = 0.1$



Conclusions

- Using witness function allows ability to create local statistics
- Important to answer questions beyond whole distribution matching
- Witness function attains similar statistical guarantees to global statistic
- Ongoing work of time series clustering on:
  - bird chirp clustering
  - weekly HSI series clustering for agriculture

Papers

- Mhaskar, Cloninger, Cheng. “A witness function based construction of discriminative models using Hermite polynomials”. Submitted, 2019.
- Das, Mhaskar, Cloninger. “Kernel distance measure for partial overlap of non-i.i.d. sampled distributions”. Preprint, 2019.

Thank you!

Questions?