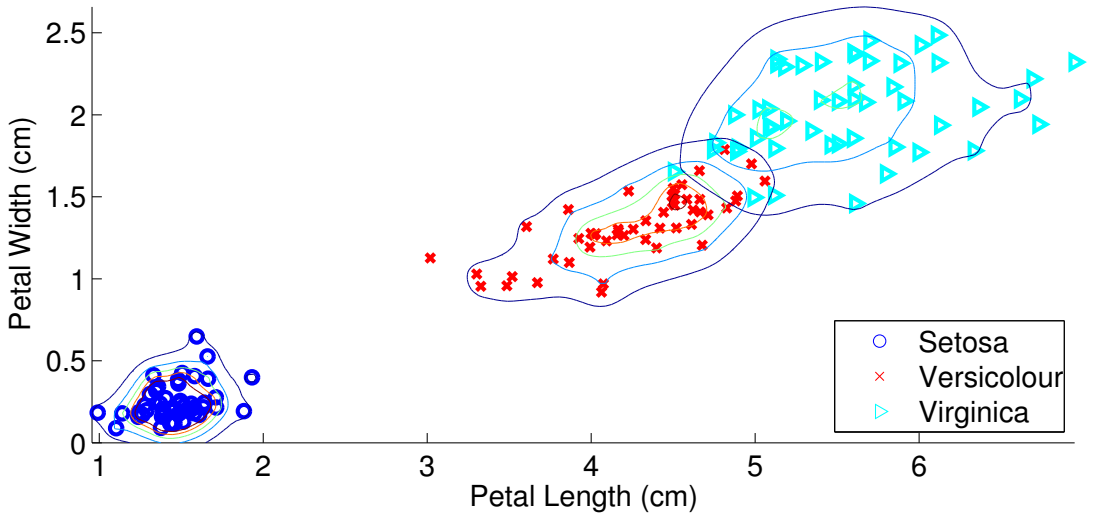
Information Theory in Machine Learning

Kevin Moon(kevin.moon@usu.edu)
STAT 6655

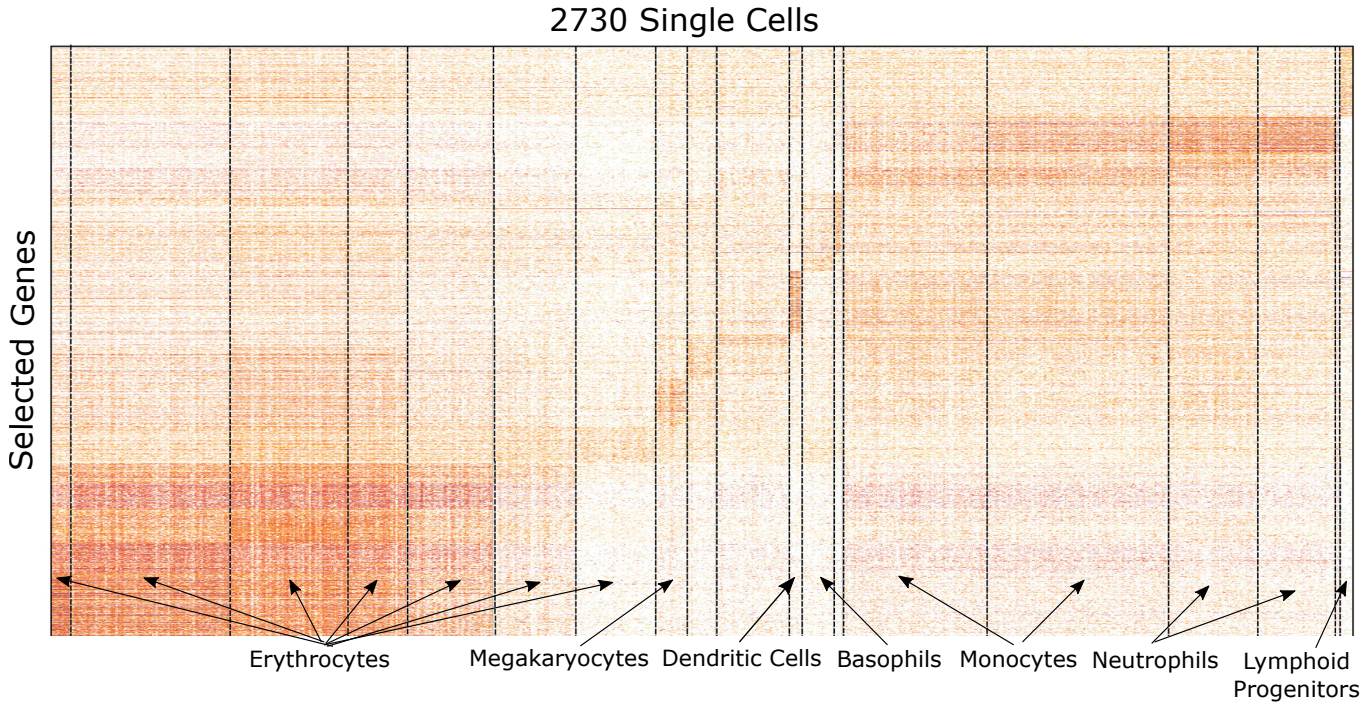
Motivation



Iris data set with added noise and KDE level sets (Fisher, 1936)

1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the best possible error rate any classifier can asymptotically achieve (i.e. the Bayes error)?
4. Are any of the points anomalies?
5. How would you go about building a classifier?

Motivation



- Gene expression matrix of mouse bone marrow single-cell RNA sequencing data (from Paul et al., 2015), divided by cell type
- Total # of genes $>$ 10,000
- Which genes are most relevant for classifying by cell type?
- What dependencies exist between genes?

Solution

1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the Bayes error rate?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?
7. How would you go about building a classifier?

Solution

1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the Bayes error rate?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?
7. How would you go about building a classifier?

Solution: Information Theory!

Information Theoretic Measures

- Entropy

$$H(\mathbf{X}) = -\int f_X(x) \log\left(f_X(X)\right) dx,$$
$$H(\mathbf{X}) = -\sum_x p_x \log\left(p_X\right).$$

- Mutual Information

$$I\left(\mathbf{X};\mathbf{Y}\right) = \int f_{XY}(x,y) \log\left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)}\right)$$

- Divergence (most general)
- These can be referred to as distributional functionals

Information Theoretic Measures

1. How similar are each of the classes to each other?
   - Divergence
2. What is the intrinsic dimension of the data?
   - Entropy
3. What is the Bayes error rate?
   - Divergence, Mutual Information
4. Are the variables dependent?
   - Mutual Information
5. Are any of the points anomalies?
   - Entropy
6. Which features are most relevant for classification?
   - Mutual Information
7. How would you go about building a classifier?
   - Divergence, entropy

Entropy

- Shannon Entropy: $H(X) = -\sum_{x \in X} p_x \log(p_X)$
- Differential Entropy: $H(X) = -\int f_X(x) \log(f_X(X))\, dx$
- Can generalize it by replacing $-\log(\cdot)$ with some other functional $g(\cdot)$
  - Renyi entropy has $g(t) = t^\alpha$ for some $\alpha > 0$.
- For both cases, entropy is a measure of uncertainty (although less so for differential entropy)

Properties of Entropy

1. Maximized when X is uniform (discrete, $H(X) = \log N$) or Gaussian (continuous, $\log\left(\sigma\sqrt{2\pi e}\right)$)
   - Uncertainty is highest when all events are equiprobable
2. Continuous
3. Independent of the location of the data
   - Could be good or bad depending on the application
4. $H(X) \geq 0$ (discrete only)
5. $H(X, Y) \leq H(X) + H(Y)$, equality iff X and Y are independent

Group Exercise

1. Compare the differential entropy of a random variable with its variance. What similarities and differences do they have?
2. Give an example of two random variables where they have the same entropy but their variances are different.

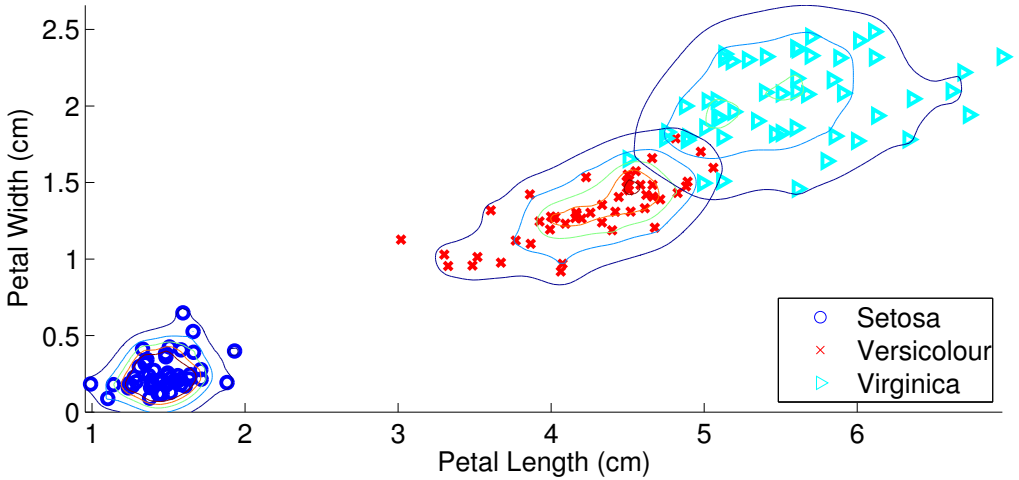
Data Compression

Theorem (Shannon's Source Coding Theorem)

Given N i.i.d random variables with entropy $H(X)$, then the minimum number of bits required for lossless data compression is $NH(X)$.

- Example: Dice rolls
- Example: English alphabet

Anomaly Detection



- Anomaly detection typically reduces to finding a minimum volume (MV) set of the distribution
- The MV set is related to the minimum entropy set of level α, i.e.

$$\min \left\{ H_\nu(A) : \int_A f_o(x)dx \geq 1 - \alpha \right\}$$

where $H_\nu(A)$ is the Renyi-ν entropy of f_0 over the set A

Other Applications of Entropy

- Intrinsic dimension estimation
  - Useful for dimensionality reduction
  - More on this later
- The dual of an unconstrained geometric program is an entropy maximization problem (Boyd, 2012)
- Texture classification and image registration
- Many others

Conditional Entropy

- Conditional Shannon Entropy:

$$\begin{aligned} H(X|Y) &= -\sum_{x\in X} p_x \sum_{y\in Y} p_{x|y} \log(p_{x|y}) \\ &= -\sum_{x\in X}\sum_{y\in Y} p_{xy} \log(p_{x|y}) \end{aligned}$$

- Conditional Differential Entropy:
  $H(X|Y) = \int f_{XY}(x,y) \log(f_{X|Y}(x|y))dxdy$
- The amount of uncertainty left in X if Y is known
- Properties
  1. $H(X|Y) = 0$ iff Y is a deterministic function of X
  2. $H(X|Y) = H(X)$ iff X and Y are independent
  3. $H(X|Y) \leq H(X)$ (knowing Y can only reduce the uncertainty about X)
  4. $H(X|Y) = H(X,Y) - H(X)$

Mutual Information

- Shannon Mutual Information
  - Discrete: $I(X;Y) = -\sum_{y\in Y}\sum_{x\in X} p_{xy} \log\left(\frac{p_x p_y}{p_{xy}}\right)$
  - Continuous: $I(X;Y) = -\int f_{XY}(x,y) \log\left(\frac{f_X(x) f_Y(y)}{f_{XY}(x,y)}\right) dxdy$
- Can generalize it by replacing $-\log(\cdot)$ with some other functional $g(\cdot)$
  - Renyi MI has $g(t) = t^\alpha$ for some $\alpha > 0$.
- Describes the amount of information that X gives about Y and vice versa

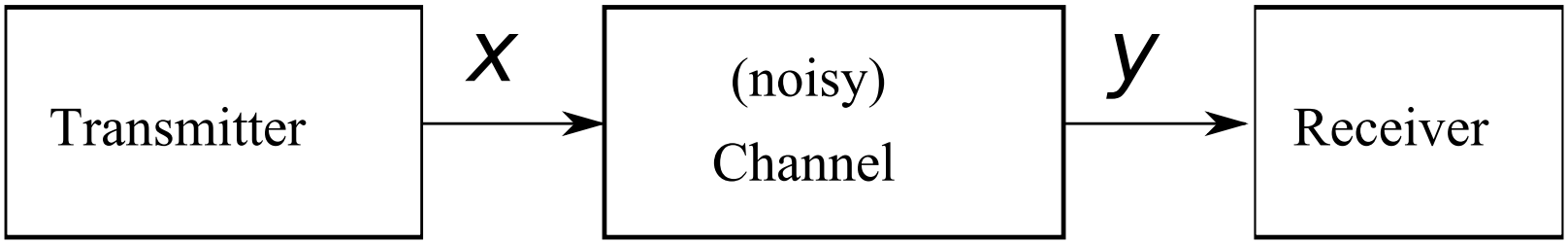
Properties of Mutual Information

$$I(X;Y) = -\int f_{XY}(x,y) \log\left(\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)}\right) dxdy,$$

$$I(X;Y) = -\sum_{y\in Y}\sum_{x\in X} p_{xy} \log\left(\frac{p_x p_y}{p_{xy}}\right)$$

1. Symmetric
2. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ (the amount of reduction in uncertainty in X by knowing Y and vice versa)
3. $I(X;Y) \geq 0$ with equality iff X and Y are independent

Channel Capacity



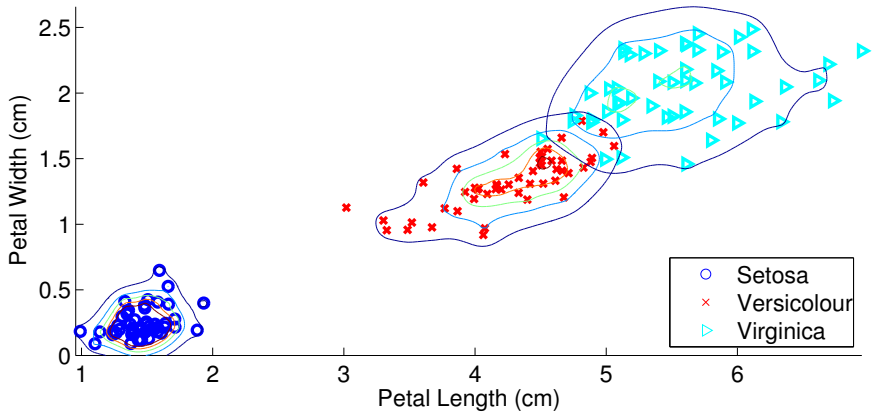
From wikipedia article on channel capacity

- Channel capacity is a "tight upper bound on the rate at which information can be reliably transmitted over a communications channel."
- It is

$$C = \sup_{p_X(x)} I(X;Y).$$

The supremum is taken over all possible choices of $p_X(x)$.

Feature Selection



Iris data set with added noise and KDE level sets

- Y is some outcome variable (e.g. classification labels)
- X is the collection of features
- Choose the subset of X that generally provides the most information about Y
  - Alternatively, throw away the subset of X that provide negligible additional information about Y
- Many, many, many papers have looked at/used this approach

Other Applications of Mutual Information Estimation

- Structure learning (Moon et al, 2017b; Chow, Liu, 1968)
- Intrinsically motivated reinforcement learning (Mohamed and Rezende, 2015; Salge et al., 2014)
- Independent subspace analysis (Pal et al., 2010)
- Forest density estimation (Liu et al., 2012)
- Clustering (Lewi et al., 2006)
- Neuron classification (Schneidman et al., 2003)
- Many more

Divergence

- Kullback-Leibler (KL) Divergence
  - Discrete: $D_{KL}(f_1||f_2) = \sum_{x\in X} f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right)$
  - Continuous: $D_{KL}(f_1||f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx$
- Can obtain f-divergences by replacing $-\log(\cdot)$ with some other functional $g(\cdot)$
  - Renyi divergence has $g(t) = t^\alpha$ for some $\alpha > 0$.
  - Total variation has $g(t) = \frac{1}{2}|t-1|$
- Other generalizations do not require the likelihood ratio
  - E.g. L_2 divergence/distance
- KL divergence also known as relative entropy and other names

Properties of Divergence

1. $D_{KL}(f_1||f_2) \geq 0$, equality iff $f_1 = f_2$ a.e.
2. KL divergence is not symmetric and therefore not a true distance
3. Divergence is the most general information measure
   - $I(X;Y) = D_{KL}\left(f_{XY}(X,Y)||f_X(X)f_Y(Y)\right)$
   - $H(X) = \log N - D_{KL}(f_X(X)||f_U(U))$ where f_U is the pmf of a uniform random variable

Some Applications of Divergence Estimation

- Estimating the decay rates of error probabilities in hypothesis testing (Cover and Thomas, 2006)
- Testing whether two collections of samples come from the same distribution (Moon et al, 2016a)
- Extending machine learning applications (e.g. classification and clustering) to probability distributions as features (Poczos and Schneider, 2011; Oliva et al, 2013; Moon et al, 2016b)
- Text/multimedia clustering (Dhillon et al, 2003)
- Clustering
- Blind source separation
- Steganography
- Estimating the best probability of error for a classification problem
    - More on this later

Principled Classification: An Application of Divergence and Entropy Estimation

Standard Approach

1. Feature selection and/or dimensionality reduction
2. Classification algorithm selection
3. Classifier training
4. Classifier testing

Principled Classification: An Application of Divergence and Entropy Estimation

Standard Approach

1. Feature selection and/or dimensionality reduction
2. Classification algorithm selection
3. Classifier training
4. Classifier testing

Proposed Approach

1. Build a descriptive analysis of the structure of the data
2. Reduce dimensionality according to this analysis
3. Estimate bounds on the Bayes Error
4. Select classifier
5. etc.

Principled Classification: An Application of Divergence Estimation

Standard Approach

1. Feature selection and/or dimensionality reduction
2. Classification algorithm selection
3. Classifier training
4. Classifier testing

Proposed Approach

1. **Build a descriptive analysis of the structure of the data**
2. Reduce dimensionality according to this analysis
3. Estimate bounds on the Bayes Error
4. Select classifier
5. etc.

Relevant questions

1. How many parameters are required to accurately describe the data?
2. Are linear methods sufficient for analysis?
3. Do different sets of features share a similar structure?
4. How complex is the data?

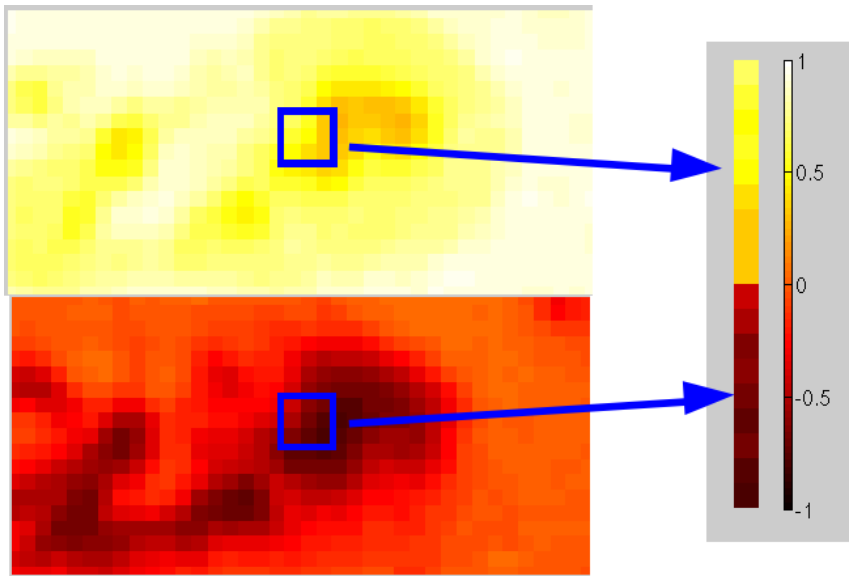
Relevant questions

1. How many parameters are required to accurately describe the data?
2. Are linear methods sufficient for analysis?
3. Do different sets of features share a similar structure?
4. How complex is the data?

Intrinsic dimension estimation

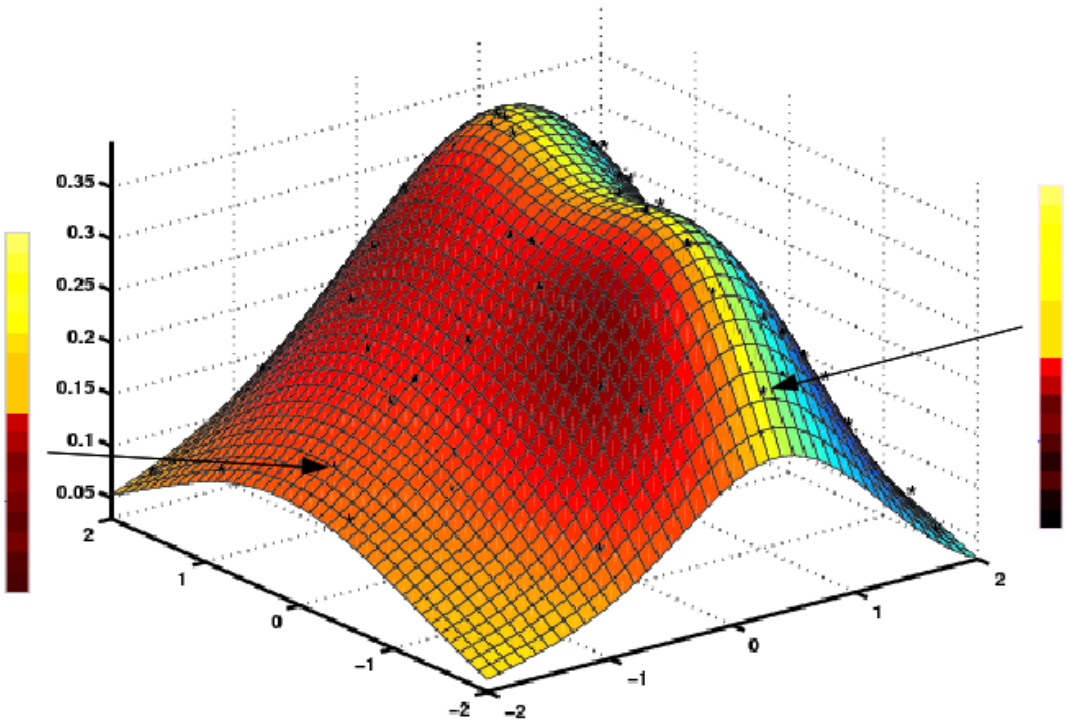
Sunspot & Active Region Image Patch Analysis

- Sunspots and active regions are related to solar flares



- Two image modalities
  - Continuum (white light), top
  - Magnetogram (magnetic field), bottom
- We extract 3×3 patches from each modality

Intrinsic Dimension Estimation



- d-dimensional observations lie on surface of dimension $m < d$.
- For sunspot image patches, $d = 18$ but what is m?

Why Intrinsic Dimension?

- Dimensionality reduction
  - E.g. choosing the number of basis vectors in PCA
  - Reduces computational burden and often improves performance
- Intrinsic dimension = measure of feature dependence
- Data interpretation

Intrinsic Dimension and Entropy

- Intrinsic dimension of the data is related to the Rényi entropy (Costa and Hero, 2006)
  - $H_\alpha(f) = \frac{1}{1-\alpha} \ln \int f^\alpha(x) dx$

Intrinsic Dimension and Entropy

- Intrinsic dimension of the data is related to the Rényi entropy (Costa and Hero, 2006)
  - $H_\alpha(f) = \frac{1}{1-\alpha} \ln \int f^\alpha(x) dx$
- Estimate the intrinsic dimension m using total k-nn graph edge length:

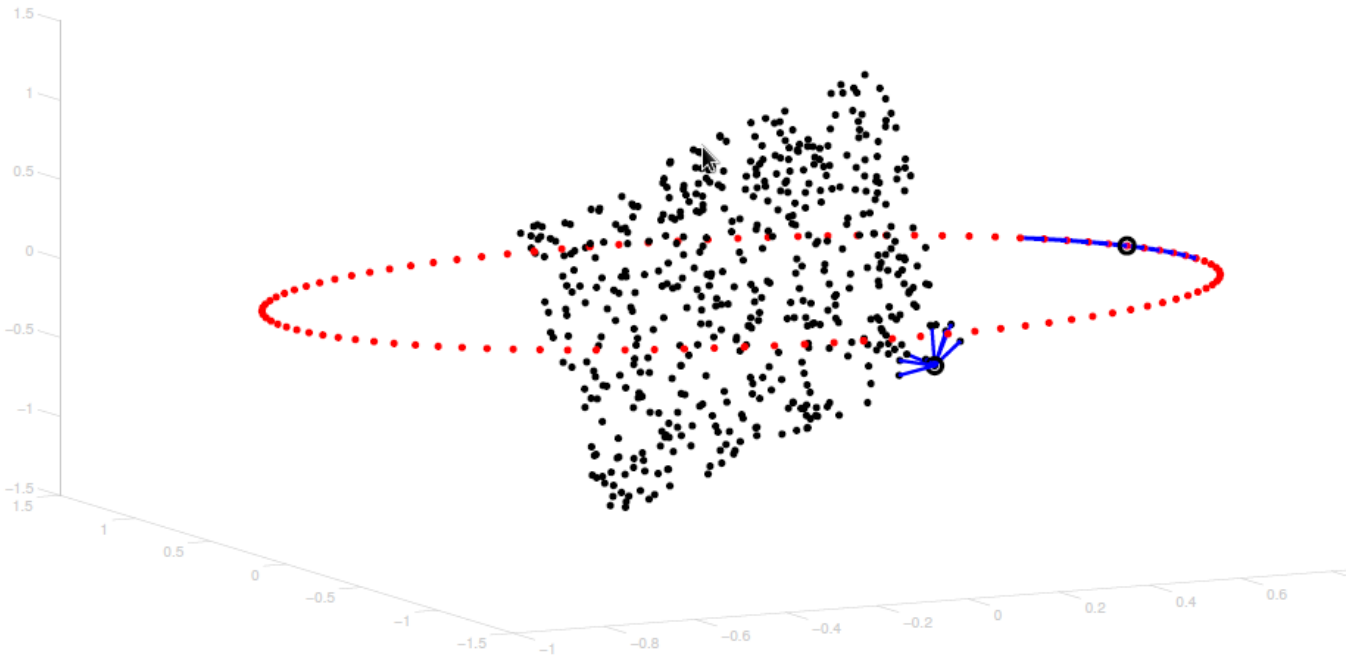
$$L_{\gamma,k}(X_n) = \sum_{i=1}^{n} \sum_{y \in \mathcal{N}_{k,i}} D^\gamma(y, x_i)$$
$$\rightarrow c n^{1-\gamma/m} + \epsilon_n,$$

where X_n is a matrix of samples, $\mathcal{N}_{k,i}$ the k-nn neighborhood of x_i, γ a free parameter, c a constant that depends on the entropy, and ϵ_n an error term.

Local Intrinsic Dimension

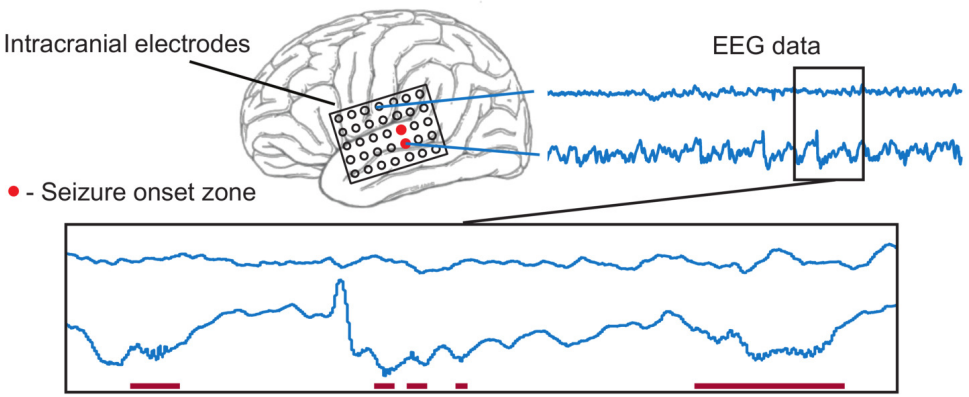
Data points may lie on different manifolds with distinct dimensions

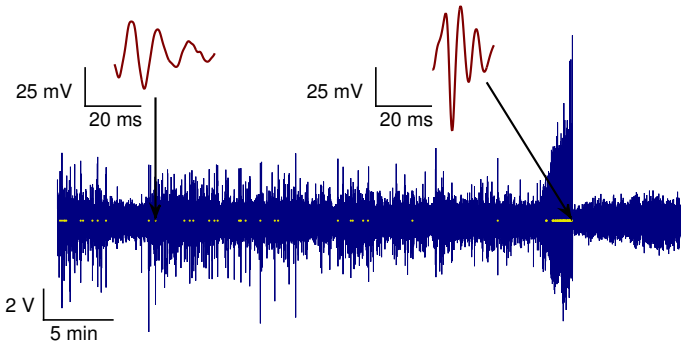
- $d = 3$ but average intrinsic dimension is 1.5
- Local intrinsic dimension is 1 & 2
- Local intrinsic dimension estimated in the neighborhood of a point.



Example: HFO Data in Epileptic Patients

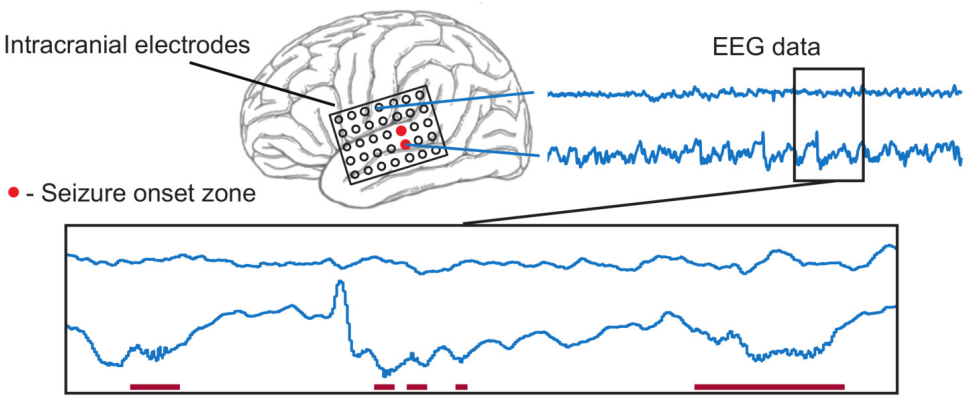
- Goal: improve localization of the seizure onset zone (SOZ)
  - Strong correlation between HFO rate and the SOZ (Gliske et al, 2015)

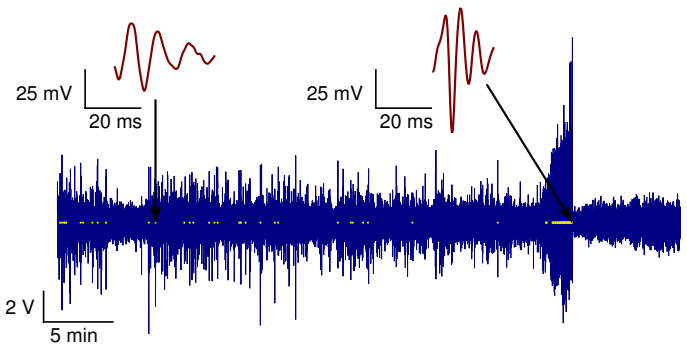




Example: HFO Data in Epileptic Patients

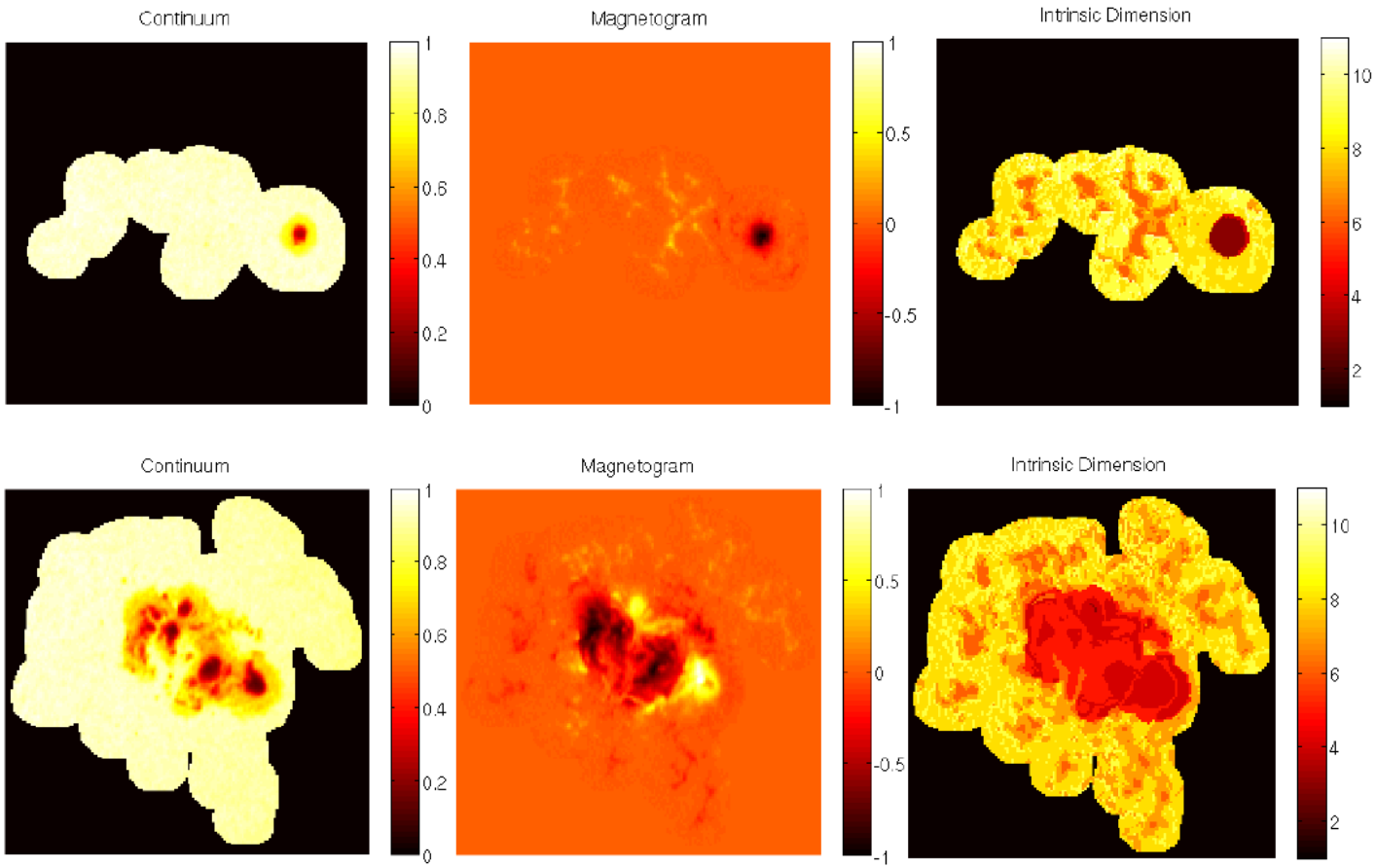
- Goal: improve localization of the seizure onset zone (SOZ)
  - Strong correlation between HFO rate and the SOZ (Gliske et al, 2015)





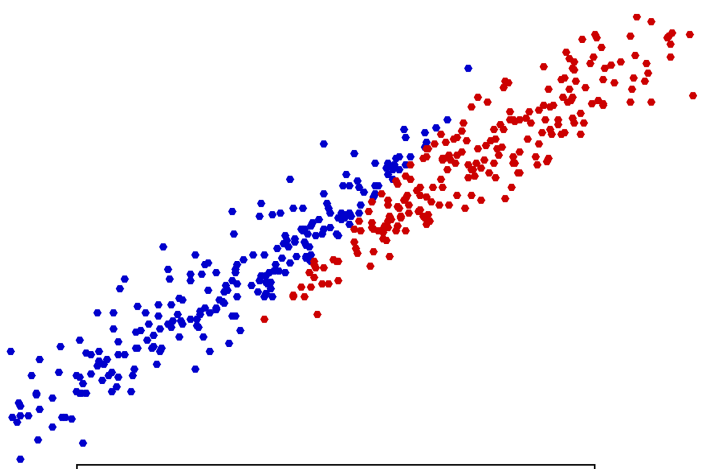
- Intrinsic dimension can vary significantly across different channels (Gliske, Moon et al, 2016)
  - Contradicts common practice

Intrinsic Dimension of Sunspot Image Patches

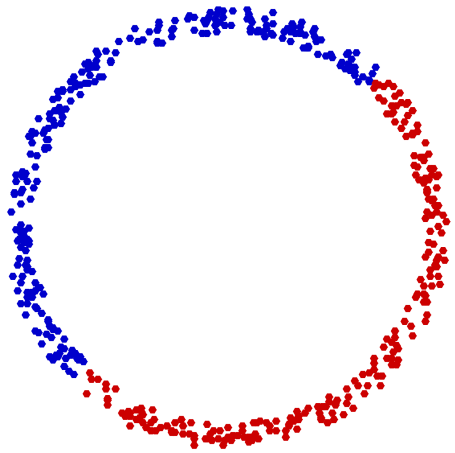


A "simple" sunspot (top) and "complex" sunspot

Linear vs. Nonlinear Topology



1D = 90.3% of variance



1D = 50.7% of variance

- Compare the non-linear estimated intrinsic dimension with a linear estimate (e.g. PCA)
  - Calculate the fraction of variance accounted for using the nonlinear estimate
  - Linear methods are sufficient if the fraction is close to 1 (dependent on the noise)

Linear vs Nonlinear Methods in Sunspots and HFO Data

- Sunspot data
  - Nonlinear estimate accounts for approximately 97% of the variance (Moon et al, 2016a)
- HFO data
  - Nonlinear estimate accounts for approximately 90% of the variance (Gliske et al, 2016)
- Based on noise levels, linear methods are sufficient for both data types

Principled Classification: An Application of Divergence Estimation

Standard Approach

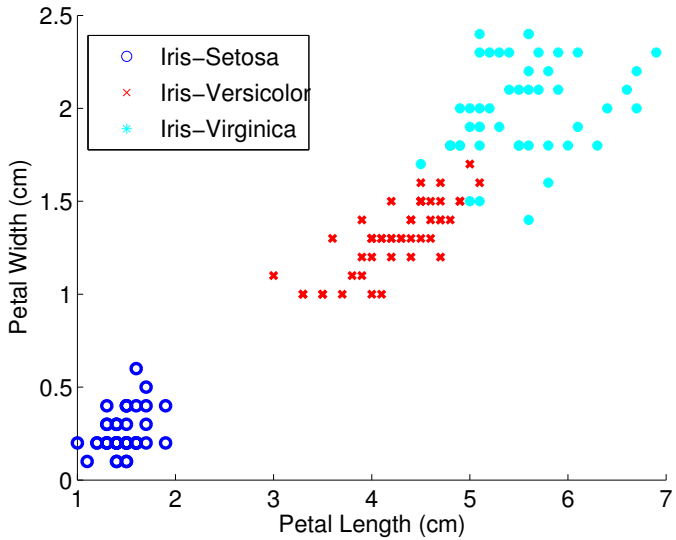
1. Feature selection and/or dimensionality reduction
2. Classification algorithm selection
3. Classifier training
4. Classifier testing

Proposed Approach

1. Build a descriptive analysis of the structure of the data
2. Reduce dimensionality according to this structure
3. **Estimate bounds on the Bayes Error**
4. Select classifier
5. etc.

The Iris Data Set

- Iris data set (Fisher, 1936)
- 3 classes, 4 features (petal length and width, sepal length and width), 50 samples per class



- What is the optimal probability of classification error?

The Bayes Error

- Two classes C_1 and C_2 and observation $x \in \mathbb{R}^d$
  - Prior class probabilities $q_1 = \Pr(C_1)$ and $q_2 = \Pr(C_2) = 1 - q_1$
  - Class probability densities $f_1(x) = p(x|C_1)$ and $f_2(x) = p(x|C_2)$
  - $p(x) = q_1 f_1(x) + q_2 f_2(x)$

The Bayes Error

- Two classes C_1 and C_2 and observation $x \in \mathbb{R}^d$
  - Prior class probabilities $q_1 = \Pr(C_1)$ and $q_2 = \Pr(C_2) = 1 - q_1$
  - Class probability densities $f_1(x) = p(x|C_1)$ and $f_2(x) = p(x|C_2)$
  - $p(x) = q_1 f_1(x) + q_2 f_2(x)$
- We wish to assign x to either C_1 or C_2
  - Divide $\mathbb{R}^d$ into decision regions R_1 and R_2
  - Assign x to C_1 if $x \in R_1$, etc.
- How do we choose R_1 and R_2?

The Bayes Error

- Minimize the average probability of error

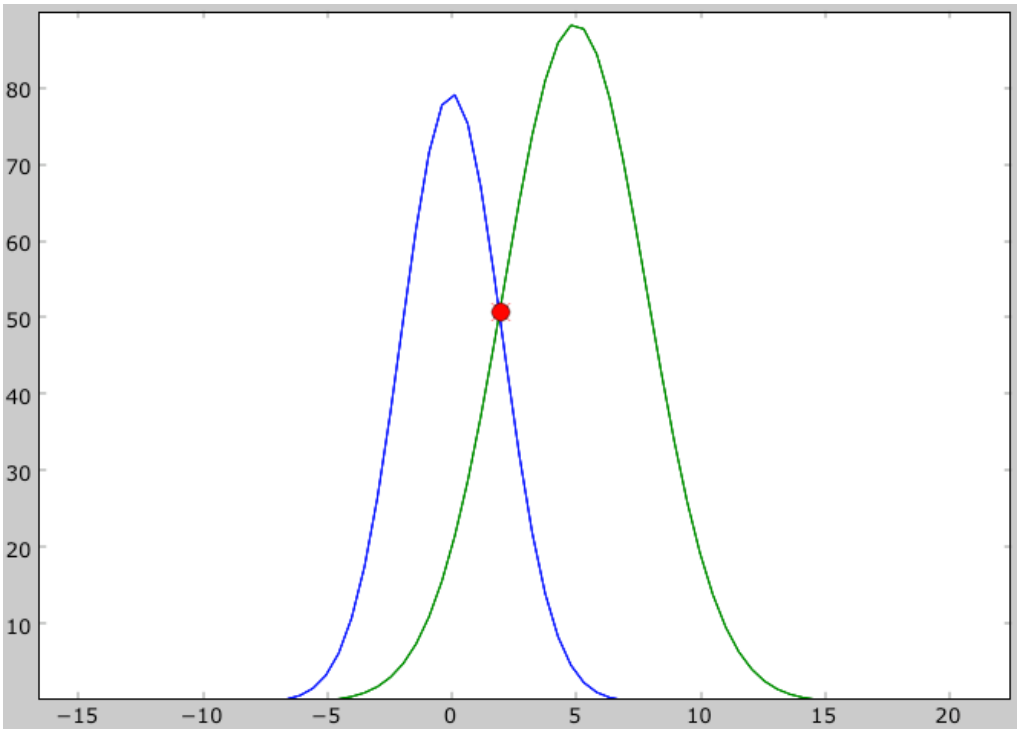
$$P_e = \int_{R_1} \Pr(C_2|x)p(x)dx + \int_{R_2} \Pr(C_1|x)p(x)dx$$

- Corresponds to assigning x to C_1 if and only if $q_1 f_1(x) > q_2 f_2(x)$
  - Bayes decision rule (derived from Bayes Rule)
- Corresponding minimum average probability of error (the Bayes error):

$$P_e^* = \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx$$

The Bayes Error of 2 Gaussians

$$P_e^* = \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx$$

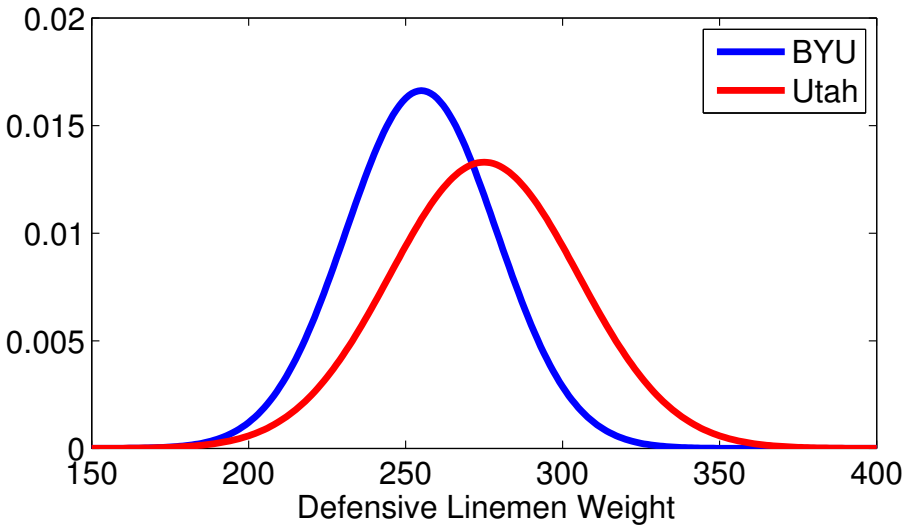


Corresponds to the area under the curves in the center.

Utility of the Bayes error

- The Bayes error tells us what the best (minimum) average probability of error **any** classifier can achieve on a given feature space
  - Provides a benchmark for classification
  - Informs us of the discriminating/predictive power of the given feature space
  - Therefore, it is useful for feature selection

Toy Example: Utah vs. BYU Football



- Suppose these are the distributions for the weight of each team's defensive linemen (**they're not**)
- Statistically different means (t-test, $\alpha = 0.05$)
- But the Bayes error = 0.35
  - Weight (alone) is not a very good predictor for team affiliation

Problems with calculating the Bayes error

- Even when the distributions are known, the Bayes error often cannot be analytically computed
- Most of the time we do not know the distributions
  - Especially in high dimensions
- What do we do in those cases?

Bounding the Bayes error

- $P_e^* = \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx$
- $P_e^* \leq \min_{\alpha \in (0,1)} q_1^{\alpha} q_2^{1-\alpha} \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx$ (Chernoff bound)

Bounding the Bayes error

- $P_e^* = \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx$
- $P_e^* \leq \min_{\alpha \in (0,1)} q_1^{\alpha} q_2^{1-\alpha} \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx$ (Chernoff bound)

- Same form as the f-divergence functional $G(f_1||f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$!
  - $g(t) = t^{\alpha}$

Bounding the Bayes error

- $P_e^* = \int \min\left(q_1 f_1(x), q_2 f_2(x)\right) dx$
- $P_e^* \leq \min_{\alpha \in (0,1)} q_1^{\alpha} q_2^{1-\alpha} \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx$ (Chernoff bound)
- Same form as the f-divergence functional $G(f_1||f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$!
  - $g(t) = t^{\alpha}$
- If we can estimate $G(f_1||f_2)$ for the appropriate g, then we can estimate this upper bound

Bounds on the Bayes error

- Chernoff bound
  - $P_e^* \leq \min_{\alpha \in (0,1)} q_1^{\alpha} q_2^{1-\alpha} \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx$
  - $g(t) = t^{\alpha}$

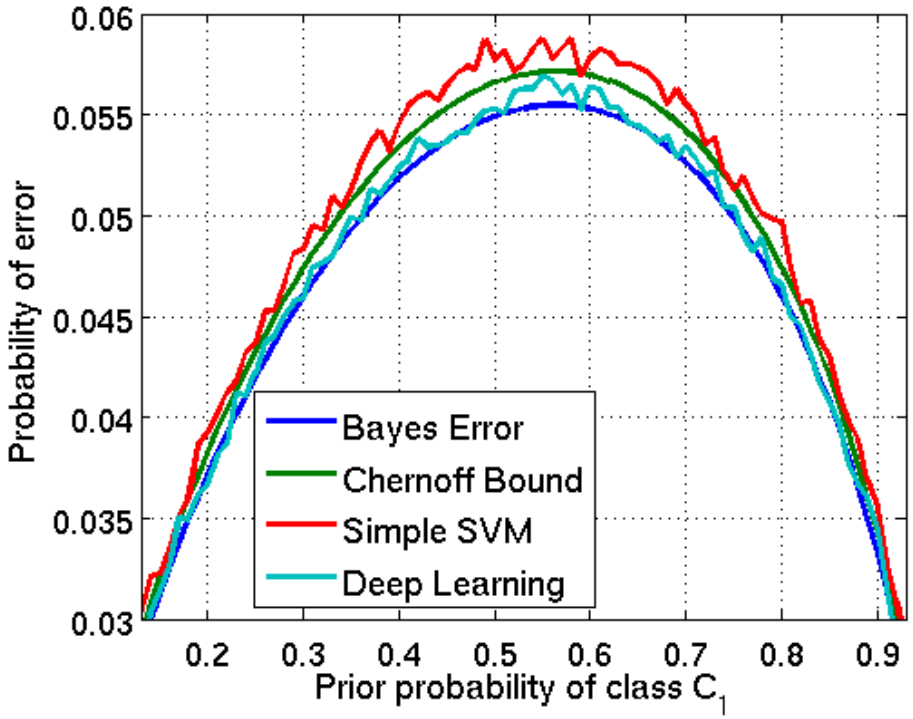
Bounds on the Bayes error

- Chernoff bound
  - $P_e^* \leq \min_{\alpha \in (0,1)} q_1^\alpha q_2^{1-\alpha} \int f_1^\alpha(x) f_2^{1-\alpha}(x) dx$
  - $g(t) = t^\alpha$
- Henze-Penrose divergence bound (Berisha et al, 2014)
  - $.5 - .5\sqrt{\tilde{D}_{q_1}(f_1, f_2)} \leq P_e^* \leq .5 - .5\tilde{D}_{q_1}(f_1, f_2)$
  - $\tilde{D}_{q_1}(f_1, f_2) = \int \frac{(q_1 f_1(x) - q_2 f_2(x))^2}{q_1 f_1(x) + q_2 f_2(x)} dx$
  - $g(t) = \frac{(q_1 t - q_2)^2}{q_1 t + q_2}$

Bounds on the Bayes error

- Chernoff bound
  - $P_e^* \leq \min_{\alpha \in (0,1)} q_1^\alpha q_2^{1-\alpha} \int f_1^\alpha(x) f_2^{1-\alpha}(x) dx$
  - $g(t) = t^\alpha$
- Henze-Penrose divergence bound (Berisha et al, 2014)
  - $.5 - .5\sqrt{\tilde{D}_{q_1}(f_1, f_2)} \leq P_e^* \leq .5 - .5\tilde{D}_{q_1}(f_1, f_2)$
  - $\tilde{D}_{q_1}(f_1, f_2) = \int \frac{(q_1 f_1(x) - q_2 f_2(x))^2}{q_1 f_1(x) + q_2 f_2(x)} dx$
  - $g(t) = \frac{(q_1 t - q_2)^2}{q_1 t + q_2}$
- Other bounds (e.g. Avi-Itzhak and Diep, 1996)

The Bayes error as a classification benchmark



- **Hypothetical** classification problem
- Overfitting occurs when the classifiers perform below the Bayes error

Divergence Estimation

Hölder Class of functions

Mean squared error (MSE) convergence rates are typically derived in terms of a smoothness condition

Definition (Hölder Class)

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. For $r = (r_1, \ldots, r_d)$, $r_i \in \mathbb{N}$, define $|r| = \sum_{i=1}^d r_i$ and $D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \ldots \partial x_d^{r_d}}$. The Hölder class $\Sigma(s, K)$ of functions on $L_2(\mathcal{X})$ consists of the functions f that satisfy

$$|D^r f(x) - D^r f(y)| \leq K \|x - y\|^{s-r},$$

for all $x, y \in \mathcal{X}$ and for all r s.t. $|r| \leq \lfloor s \rfloor$.

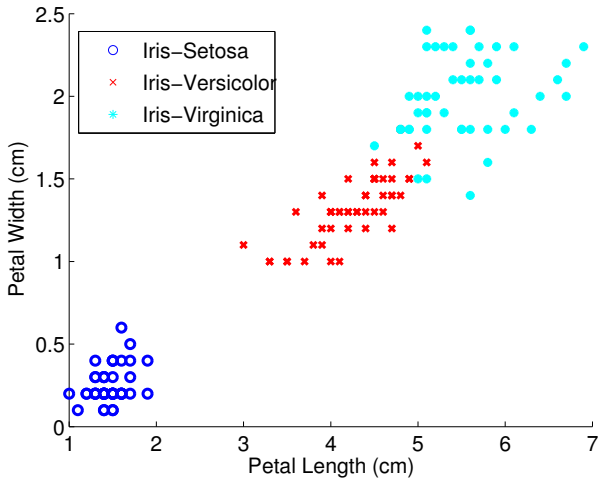
- I.e., f is continuously differentiable up to order $\lfloor s \rfloor$.

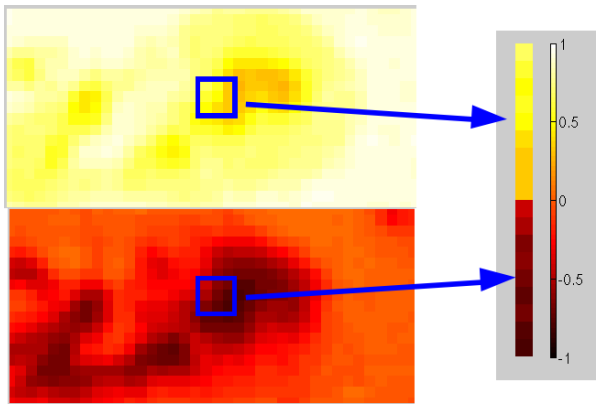
Parametric Estimation of Distributional Functionals

- Assume a parametric model on the densities (e.g. Gaussian)
- Estimate the parameters of the model (e.g. mean and variance)
- Use the values in the densities and calculate the distributional functional from the formula

Parametric Estimation of Distributional Functionals

- Assume a parametric model on the densities (e.g. Gaussian)
- Estimate the parameters of the model (e.g. mean and variance)
- Use the values in the densities and calculate the distributional functional from the formula
- May require numerical integration (computationally intensive)
- The parametric model may be a poor fit





Problem Setup

- N i.i.d. samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ from f_1
- N i.i.d. samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ from f_2.
- Can we accurately estimate $G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) dx$?
  - Approaches can be generalized to 1 or more distributions

Problem Setup

- N i.i.d. samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ from f_1
- N i.i.d. samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ from f_2.
- Can we accurately estimate $G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) dx$?
  - Approaches can be generalized to 1 or more distributions

My Contributions:

1. Derive MSE convergence rates for k-nn plug-in estimators **without** boundary correction
2. Apply ensemble estimation theory to obtain estimators that achieve the parametric MSE convergence rate $O(1/N)$
3. Derive central limit theorems for the estimators

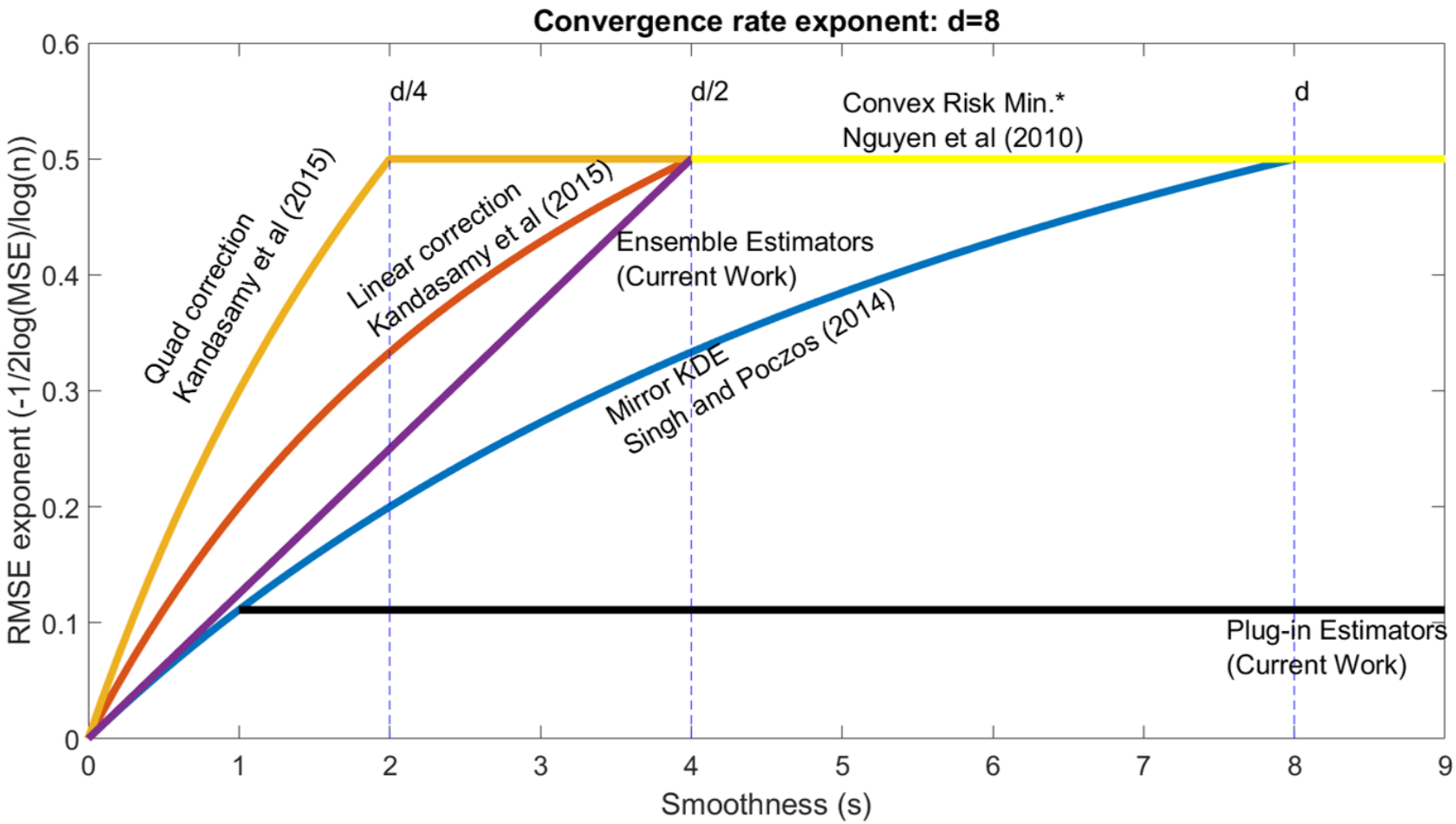
Related Work: Entropy Estimation

- Discrete entropy estimation
  - Shannon entropy: Good (1953); Miller (1955); Zahl (1977); etc.
  - Minimax estimates of various entropies and divergences recently derived by Prof. Weissman's group (Han et al, 2015, 2016; Jiao et al, 2015)
- Continuous (differential) entropy estimation
  - Nonlinear density functional estimation (Ibragimov and Khas'minskii, 1978; Levit, 1978; Hall and Marron, 1987; Bickel and Ritov, 1988; Ritov and Bickel, 1990)
  - Optimal kernel density estimator approaches (Birge and Massart, 1995; Gine and Mason, 2008; Laurent et al, 1996)
  - **Optimally weighted ensemble estimation** with KDE and boundary correction (Sricharan et al, 2013)

Related Work on Nonparametric Divergence Estimation

	Current Work	Moon et al. (2016)	Moon et al. (2014)	Noshad et al. (2017)	Singh & Póczos (2014)	Nguyen et al (2010)	Krishnamurthy et al (2014)	Kandasamy et al (2015)
Parametric MSE rate?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CLT?	Yes	Yes	Yes	?	?	?	?	Yes
General functionals?	Yes	Yes	X	X	Yes	X	X	Yes
Boundary complexity	Low	Low	High	Med.	High	Low	High	High
Computational burden	Low	Med.	Low	Low	Med.	High	Med.	Med.

Recent Work on Divergence Estimation (Continuous)



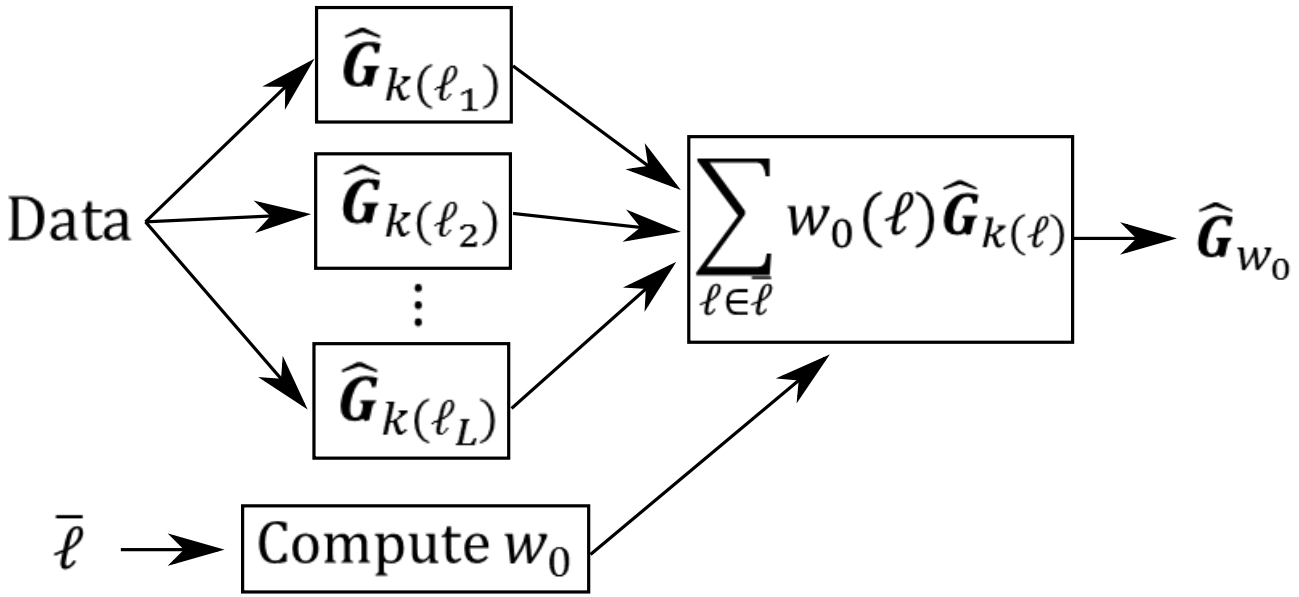
k-nn vs. KDE

- Advantages of k-nn approaches over KDE (Moon et al., 2016)
  1. Computationally faster
  2. Locally adaptive
  3. Easier to tune

k-nn vs. KDE

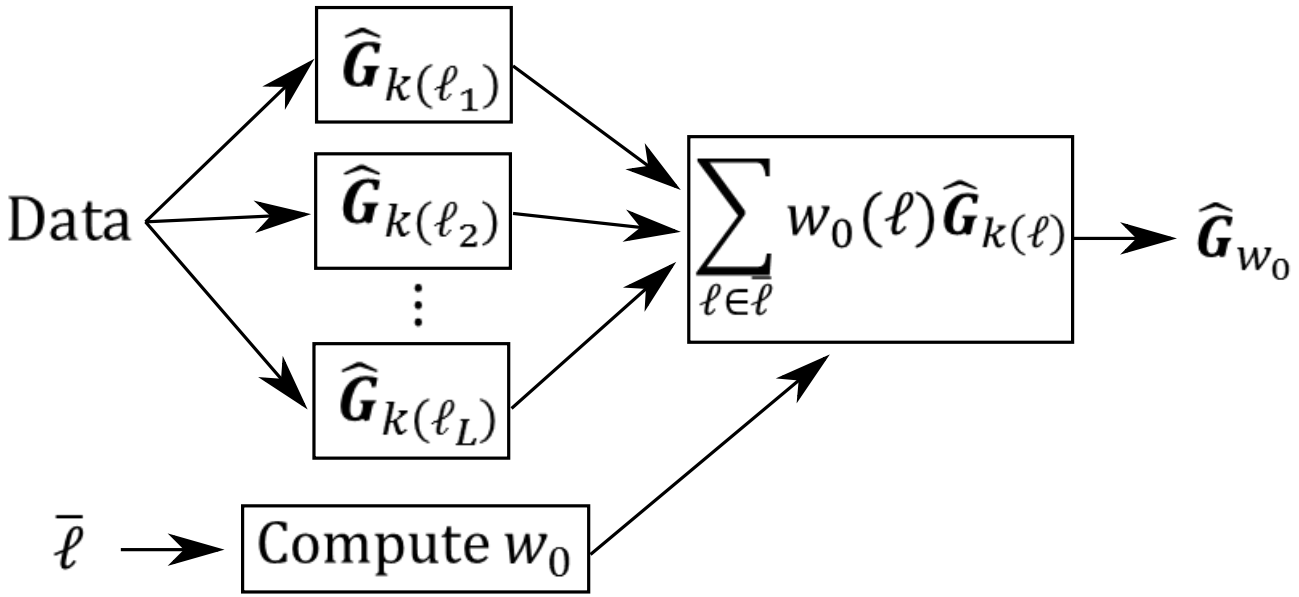
- Advantages of k-nn approaches over KDE (Moon et al., 2016)
  1. Computationally faster
  2. Locally adaptive
  3. Easier to tune
- Disadvantage of k-nn approaches
  - Much more difficult to analyze

My Approach: Optimally Weighted Ensemble Estimation



If bias and variance of $\hat{\mathbf{G}}_{k(\ell)}$ as a function of ℓ is known, we can choose w_0 to decrease the bias with little increase in the variance.

My Approach: Optimally Weighted Ensemble Estimation

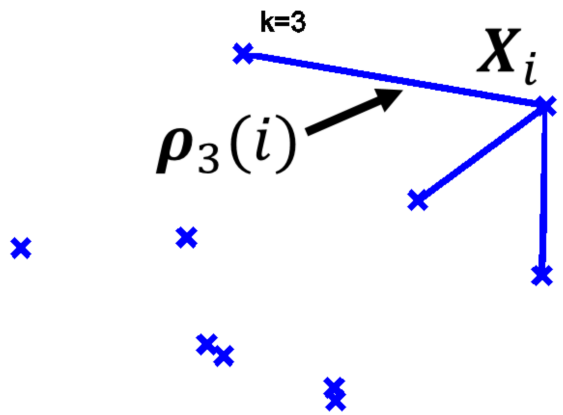


If bias and variance of $\hat{\mathbf{G}}_{k(\ell)}$ as a function of ℓ is known, we can choose w_0 to decrease the bias with little increase in the variance.

1. Define $\hat{\mathbf{G}}_{k(\ell)}$
2. Derive MSE of $\hat{\mathbf{G}}_{k(\ell)}$
3. Compute w_0

k-nn density estimator

- Standard k-nn density estimator $\hat{\mathbf{f}}_k(X_i) \propto \frac{1}{\rho_k^d(i)}$ (Loftsgaarden and Quesenberry, 1965)



- Denote the estimators as $\hat{\mathbf{f}}_{1,k}$ and $\hat{\mathbf{f}}_{2,k}$

Divergence Plug-in Estimator

- The plug-in estimator:

$$\hat{\mathsf{G}}_k = \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{\mathsf{f}}_{1,k}(\mathbf{X}_i), \hat{\mathsf{f}}_{2,k}(\mathbf{X}_i)\right).$$

Divergence Plug-in Estimator

- The plug-in estimator:

$$\hat{\mathsf{G}}_k = \frac{1}{N}\sum_{i=1}^{N} g\left(\hat{\mathsf{f}}_{1,k}(\mathbf{X}_i), \hat{\mathsf{f}}_{2,k}(\mathbf{X}_i)\right).$$

- Does the estimator $\hat{\mathsf{G}}_k$ converge to $G(f_1, f_2)$? If so, at what rate?

Convergence Rates: Variance Result

Theorem (Efron-Stein Inequality)

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$ *are independent random variables (vectors) with* $\boldsymbol{X}_i$ *and* $\boldsymbol{X}'_i$ *having the same distribution for all* i*. Let* $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *and let* $\boldsymbol{X}^{(i)} = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}'_i, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_n\right)$*. Then for any function* f

$$\mathbb{V}\left[f(\boldsymbol{X})\right] \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{X}) - f\left(\boldsymbol{X}^{(i)}\right)\right)^2\right]$$

Convergence Rates: Variance Result

Theorem (Efron-Stein Inequality)

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$ *are independent random variables (vectors) with* $\boldsymbol{X}_i$ *and* $\boldsymbol{X}'_i$ *having the same distribution for all* i*. Let* $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ *and let* $\boldsymbol{X}^{(i)} = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}'_i, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_n\right)$*. Then for any function* f

$$\mathbb{V}\left[f(\boldsymbol{X})\right] \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{X}) - f\left(\boldsymbol{X}^{(i)}\right)\right)^2\right]$$

- Bounds the variance of a general function (e.g. the plug-in estimator)
- We can use it to bound the variance of the plug-in estimator

Convergence Rates: Variance Result

Theorem (Moon et al, 2017a)

If the functional g is Lipschitz continuous in both of its arguments with Lipschitz constant C_g, then the ***variance*** *of $\hat{\mathbf{G}}_k$ is $O(1/N)$.*

- K. Moon, K. Sricharan, A. Hero, "Ensemble Estimation of Distributional Functionals via k-Nearest Neighbors," at arXiv:1707.03083.

Convergence Rates: Variance Result

Theorem (Moon et al, 2017a)

If the functional g is Lipschitz continuous in both of its arguments with Lipschitz constant C_g, then the ***variance*** *of $\hat{\mathbf{G}}_k$ is $O(1/N)$.*

- K. Moon, K. Sricharan, A. Hero, "Ensemble Estimation of Distributional Functionals via k-Nearest Neighbors," at arXiv:1707.03083.
- Proof involves the Efron-Stein inequality
  - Complicated by the dependencies between different k-nn neighborhoods
  - Analyze the effects on the k-nn graph when one sample differs

Bias Assumptions

1. The densities f_1 and f_2 are $s \geq 2$ times differentiable and are bounded away from zero and infinity.
2. The functional g has an infinite number of mixed derivatives.
   - Most functionals of interest fulfill this criterion.
3. The boundary of the densities' support set $\mathcal{S} = [-1, 1]^d$
   - Easily generalized to support sets with smooth boundaries

Bias Results

Theorem (Moon et al, 2017a)

The bias of $\hat{\mathbf{G}}_k$ *is*

$$\mathbb{B}\left[\hat{\mathbf{G}}_k\right] = c_1\left(\frac{k}{N}\right)^{\frac{1}{d}} + \frac{b_1}{k} + O\left(\frac{1}{k} + \left(\frac{k}{N}\right)^{\frac{2}{d}}\right).$$

- Constants are independent of k and N
- K. Moon, K. Sricharan, A. Hero, "Ensemble Estimation of Distributional Functionals via k-Nearest Neighbors," arXiv.

Bias Results

Theorem (Moon et al, 2017a)

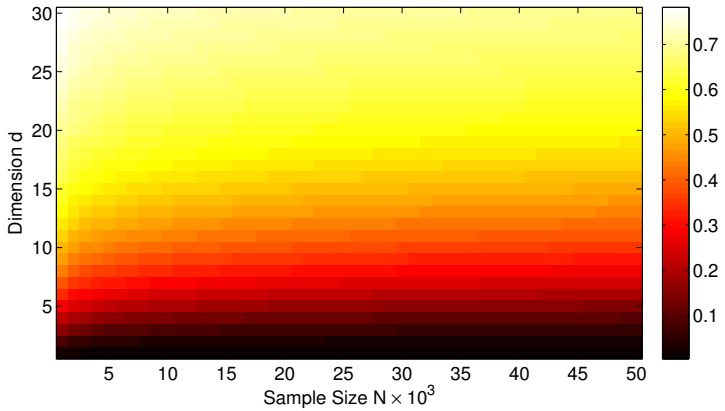
The bias of $\hat{\mathbf{G}}_k$ *is*

$$\mathbb{B}\left[\hat{\mathbf{G}}_k\right] = c_1 \left(\frac{k}{N}\right)^{\frac{1}{d}} + \frac{b_1}{k} + O\left(\frac{1}{k} + \left(\frac{k}{N}\right)^{\frac{2}{d}}\right).$$

- Constants are independent of k and N
- K. Moon, K. Sricharan, A. Hero, "Ensemble Estimation of Distributional Functionals via k-Nearest Neighbors," arXiv.
- Proof involves conditioning on the k-nn distances to use KDE proof techniques and to handle the support set boundary.

Bias Results

- Plug-in methods are **highly biased** for large dimension d



Heat map of predicted bias of divergence functional plug-in estimator

- Use **weighted ensemble** estimation to improve bias

$$\hat{\mathbf{G}}_w := \sum_{\ell \in \bar{\ell}} w(\ell) \hat{\mathbf{G}}_{k(\ell)}$$

Bias Results

Theorem (Moon et al, 2017a)

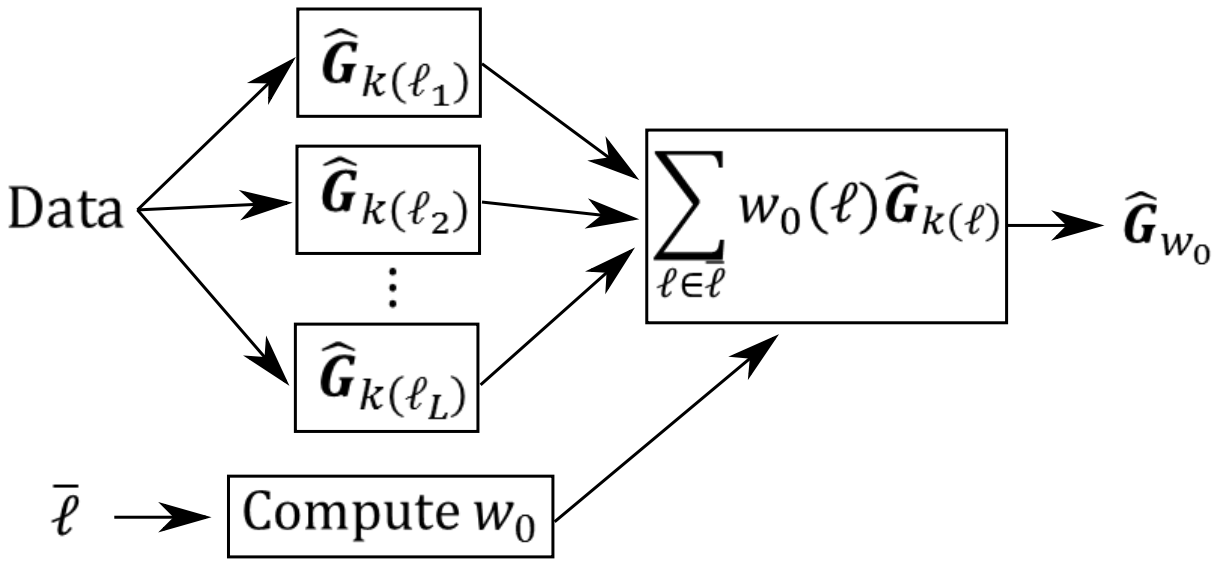
The bias of $\hat{\mathbf{G}}_k$ *is*

$$\mathbb{B}\left[\hat{\mathbf{G}}_k\right] = \sum_{j=1}^{s} c_{3,j} \left(\frac{k}{N}\right)^{\frac{j}{d}} + \frac{b_1}{k} + O\left(\frac{1}{k} + \left(\frac{k}{N}\right)^{\frac{s}{d}}\right).$$

- Constants are independent of k and N
- K. Moon, K. Sricharan, A. Hero, "Ensemble Estimation of Distributional Functionals via k-Nearest Neighbors," at arXiv:1707.03083.

Ensemble Estimation Setup

- $\bar{\ell} = \{\ell_1, \ell_2, \ldots, \ell_L\}$ a set of real, positive numbers
- An ensemble of estimators $\left\{\hat{\mathbf{G}}_{k(\ell)}\right\}_{\ell \in \bar{\ell}}$ of parameter $G(f_1, f_2)$ and weights w with $\sum_{\ell \in \bar{\ell}} w(\ell) = 1$



Ensemble Estimation Procedure

- Choose $k(\ell) = \ell\sqrt{N}$

$$\mathbb{B}\left[\hat{\mathbf{G}}_w\right] = \sum_{\ell\in\bar{\ell}}\sum_{j=1}^{s} c_{4,i} w(\ell)\ell^{j/d} N^{\frac{-j}{2d}} + O\left(\sqrt{L}||w||_2\left(N^{\frac{-s}{2d}} + N^{\frac{-1}{2}}\right)\right)$$

- Calculate (offline) **optimal weight** w_0 to zero out lower order bias terms:

$$\begin{array}{rl} \min_w & ||w||_2 \\ \textit{subject to} & \sum_{\ell\in\bar{\ell}} w(\ell) = 1, \\ & \sum_{\ell\in\bar{\ell}} w(\ell)\ell^{j/d} = 0,\ j \in \{1,\ldots,s\}. \end{array}$$

- **MSE** of $\hat{\mathbf{G}}_{w_0}$ is $O\left(\frac{1}{N}\right)$ if $s \geq d$

Ensemble Estimation Procedure

- Choose $k(\ell) = \ell\sqrt{N}$

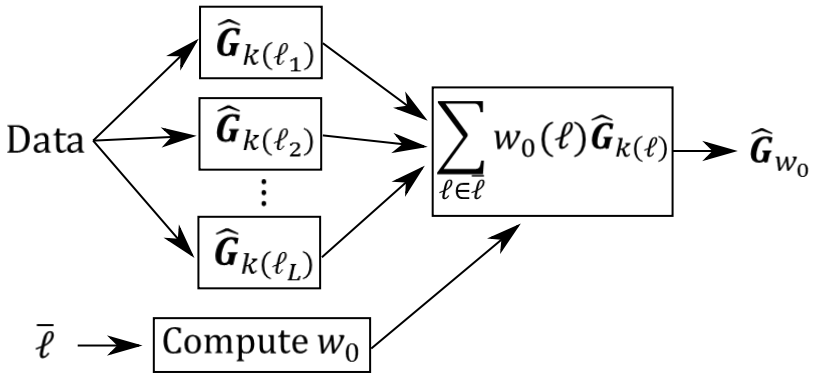
$$\mathbb{B}\left[\hat{\mathbf{G}}_w\right] = \sum_{\ell\in\bar{\ell}}\sum_{j=1}^{s} c_{4,i} w(\ell)\ell^{j/d} N^{\frac{-j}{2d}} + O\left(\sqrt{L}||w||_2\left(N^{\frac{-s}{2d}} + N^{\frac{-1}{2}}\right)\right)$$

- Calculate (offline) **optimal weight** w_0 to zero out lower order bias terms:

$$\begin{array}{rl} \min_w & ||w||_2 \\ \textit{subject to} & \sum_{\ell\in\bar{\ell}} w(\ell) = 1, \\ & \sum_{\ell\in\bar{\ell}} w(\ell)\ell^{j/d} = 0,\ j \in \{1,\ldots,s\}. \end{array}$$

- **MSE** of $\hat{\mathbf{G}}_{w_0}$ is $O\left(\frac{1}{N}\right)$ if $s \geq d$
- Under extra assumptions on g (fulfilled for KL and Renyi divergence), we can define an ensemble estimator with MSE $O\left(\frac{1}{N}\right)$ if $s > \frac{d}{2}$.

Summary



1. Derive the bias and variance of the base estimator (if unknown)
2. Calculate the optimal weight vector w_0 according to the bias and variance expressions
3. Collect data if not already available
4. Calculate the k-nn density estimators for the corresponding values of $k(\ell)$
5. Plug into the formula for $\hat{\mathbf{G}}_{k(\ell)}\ \forall \ell \in \bar{\ell}$
6. Take the weighted average of the estimates: $\sum_{\ell\in\bar{\ell}} w_0(\ell)\hat{\mathbf{G}}_{k(\ell)}$

Central Limit Theorem

Theorem

Assume that g is twice differentiable. Further assume that $k(\ell) \to \infty$ as $N \to \infty$ for each $\ell \in \bar{\ell}$. The asymptotic distribution of the centered and scaled weighted ensemble estimator $\hat{\mathbf{G}}_w$ is given by a standard normal random variable.

- Enables inference and hypothesis testing

Extensions

- KDE plug-in divergence estimation (ISIT 2016; *Entropy*, 2018)
  - Ensemble of bandwidth parameters
  - CLT
  - Uniform convergence over class of bounded Holder smooth densities

Extensions

- KDE plug-in divergence estimation (ISIT 2016; *Entropy*, 2018)
  - Ensemble of bandwidth parameters
  - CLT
  - Uniform convergence over class of bounded Holder smooth densities
- KDE plug-in mutual information estimation (ICASSP 2017; ISIT 2017)
  - Continuous case
  - Mixture of discrete and continuous random variables (first estimators that achieve parametric rate)
  - CLT

Extensions

- KDE plug-in divergence estimation (ISIT 2016; *Entropy*, 2018)
  - Ensemble of bandwidth parameters
  - CLT
  - Uniform convergence over class of bounded Holder smooth densities
- KDE plug-in mutual information estimation (ICASSP 2017; ISIT 2017)
  - Continuous case
  - Mixture of discrete and continuous random variables (first estimators that achieve parametric rate)
  - CLT
- Other k-nn ensemble approaches (ISIT 2017; ICASSP 2018)

Advantages of Ensemble Estimators

They

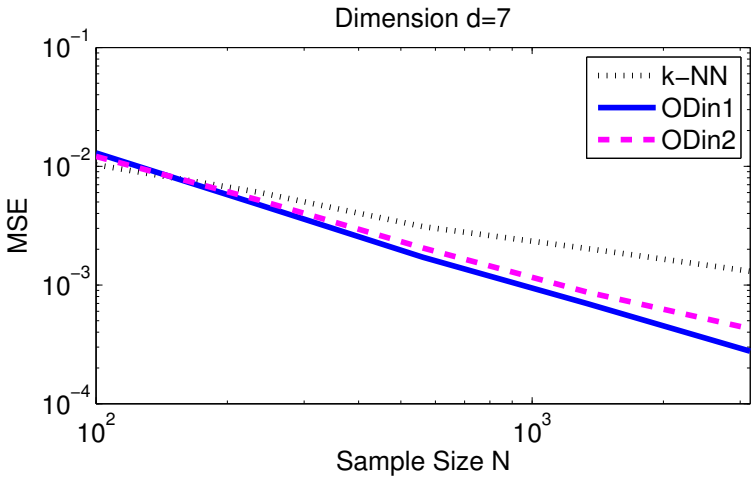
- Apply to general distributional functionals
- Are simpler to implement than competing estimators
  - No knowledge of the densities' support set required
- Achieve the parametric MSE rate ($O(1/N)$) if densities are at least $d/2$ times differentiable
  - Competitive with other estimators
- Are computationally tractable
- Have a central limit theorem for performing hypothesis testing

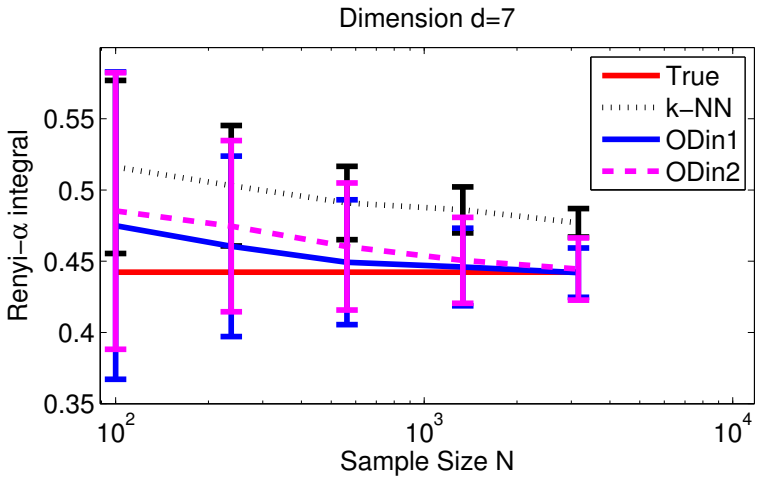
Experiments: Renyi-α divergence integral

$$G_\alpha(f_1, f_2) = \int f_1(x)^\alpha f_2(x)^{1-\alpha} dx$$

- $\alpha = 0.5$
- f_1 and f_2 are truncated Gaussian distributions
- The covariance matrices are all identical and diagonal
- The means differ
- Varied sample size
- Compared ensemble estimator to k-nn plug-in
- $\bar{\ell}$ =50 linearly spaced values between 0.3 and 3

Numerical Validation





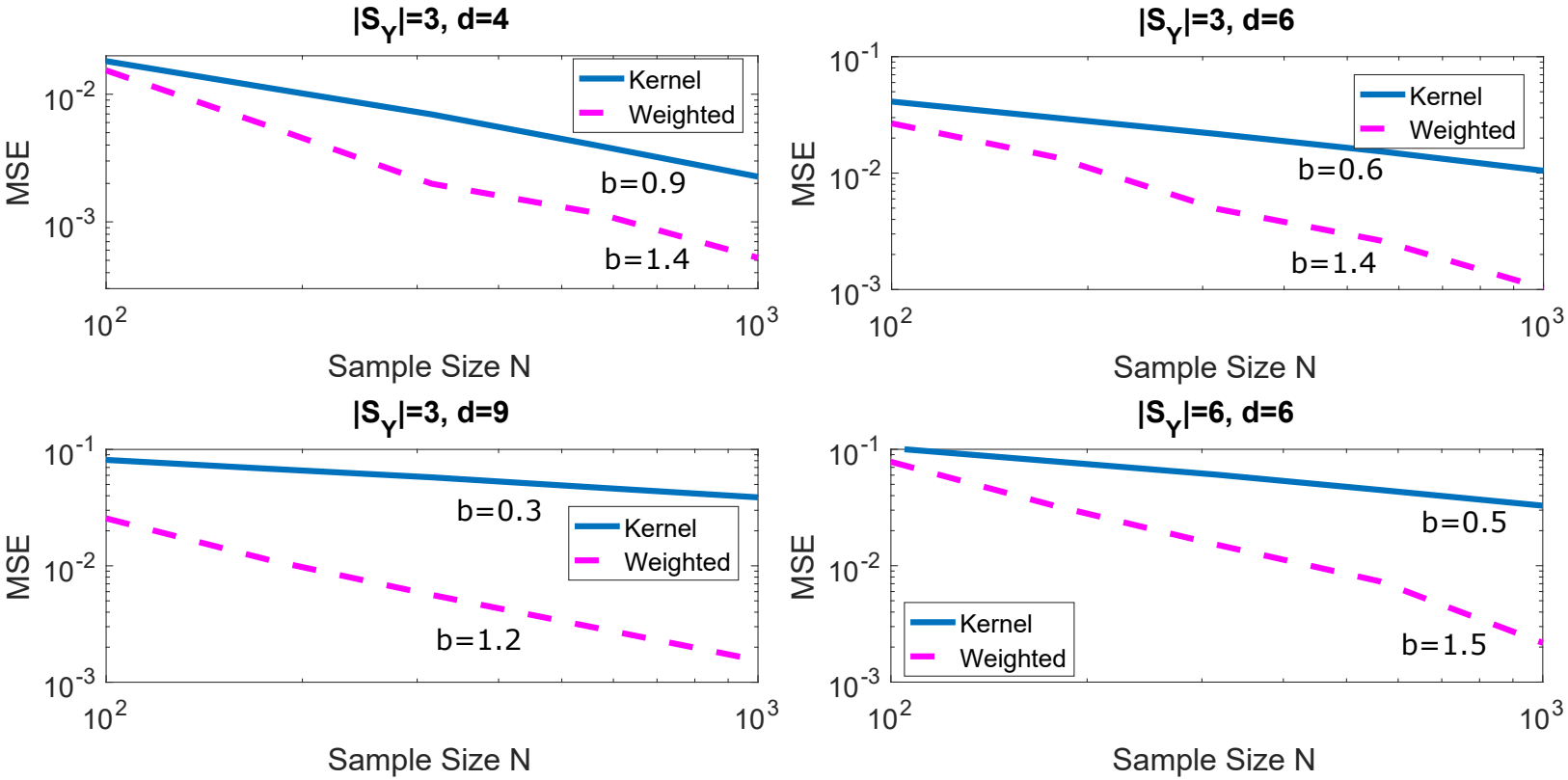
- ODin1 and 2 are different weighting schemes
- Weighted estimators are much less biased

Experiments: Renyi-α MI integral

$$G_\alpha(\mathbf{X};\mathbf{Y}) = \sum_{y\in\mathcal{S}_Y} f_Y(y) \int \left(\frac{f_X(x)}{f_{X|Y}(x|y)} \right)^\alpha f_{X|Y}(x|y)dx$$

- $\alpha = 0.5$
- $(\mathbf{X}, \mathbf{Y})$ are drawn from a mixture of truncated Gaussian distributions where $\mathbf{Y}$ is discrete (indicates which Gaussian)
- The conditional covariance matrices of $\mathbf{X}|\mathbf{Y}$ are all identical and diagonal
- The conditional means of $\mathbf{X}|\mathbf{Y}$ all differ
- Varied d_X and $|\mathcal{S}_Y|$
- Compared ensemble estimator to KDE plug-in

Experiments



- MSE$= O\left(N^{-b}\right)$
- As dimension increases, the performance gap between the estimators increases

Single-Cell RNA-sequencing Data

- Single-cell mouse bone marrow data described earlier
  - 19 identified cell types, 2730 cells
  - Data imputation via MAGIC (van Dijk, Moon et al., 2017)
  - Estimated Renyi-α MI with $\alpha = 0.5$ between gene expression ($\mathbf{X}$) and cell type ($\mathbf{Y}$)
  - KEGG pathways and key genes from Paul et al. (2015)

Single-Cell RNA-sequencing Data

- Single-cell mouse bone marrow data described earlier
  - 19 identified cell types, 2730 cells
  - Data imputation via MAGIC (van Dijk, Moon et al., 2017)
  - Estimated Renyi-α MI with $\alpha = 0.5$ between gene expression ($\mathbf{X}$) and cell type ($\mathbf{Y}$)
  - KEGG pathways and key genes from Paul et al. (2015)

	Platelets	Erythrocytes	Neutrophils	Macrophages	Combined	Key Genes	Random
Estimated MI (mean$\pm$std)	0.24 ± 0.11	0.66 ± 0.11	0.27 ± 0.10	0.15 ± 0.09	1.65 ± 0.36	2.50 ± 0.64	0.007 ± 0.07
Prob. Error	43%	42.9%	43.1%	47.4%	35.5%	27.8%	55.9%

Estimating the Bayes error of the Iris data set

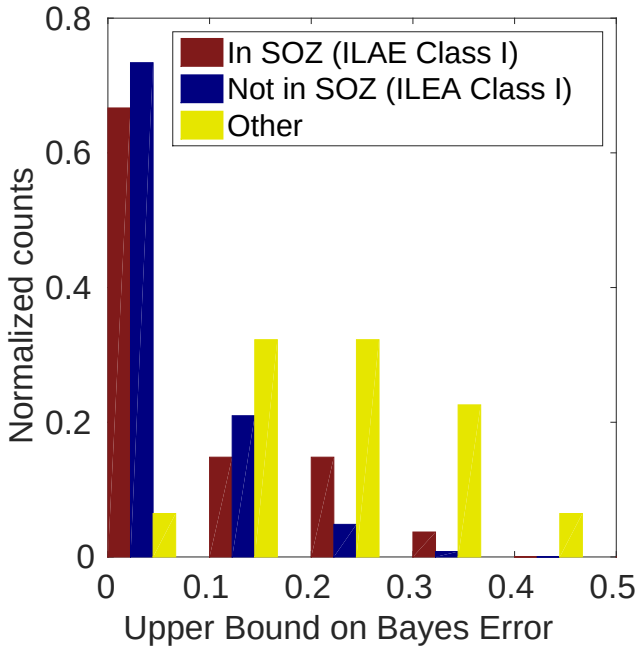
- 50 samples available from each class
- Setosa class is linearly separable from the others
- Estimated a 95% confidence interval on the Chernoff bound of the Bayes error
- Compared the results to the performance of a standard Quadratic Discriminant Analysis classifier

	Setosa-Versicolor	Setosa-Virginica	Versicolor-Virginica
CI	(0,0.0013)	(0,0.0002)	(0,0.0726)
QDA	0	0	0.04

- Results are relatively consistent

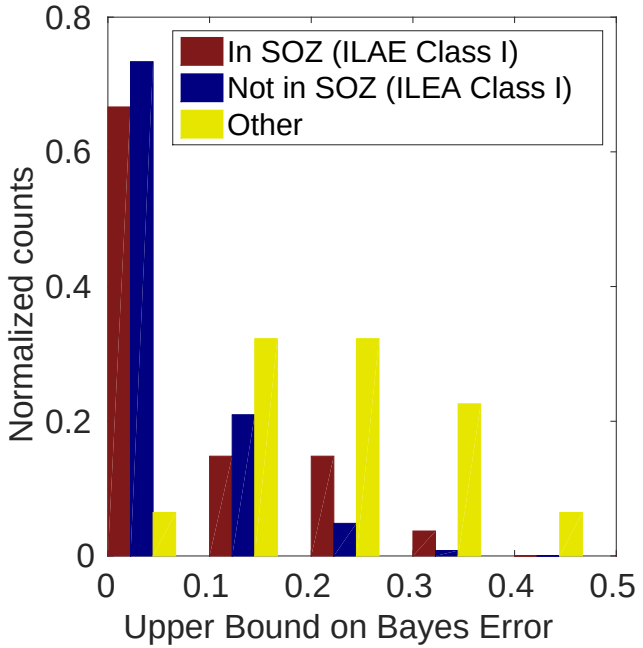
HFO Classification Problem

- HFOs typically "look different" during a seizure (ictal) than when no seizure is occurring (interictal)
- Classification problem: classify interictal HFOs as "ictal-like" or "interictal-like"
- Should be easier in healthy tissue (i.e. outside of the SOZ)



HFO Classification Problem

- HFOs typically "look different" during a seizure (ictal) than when no seizure is occurring (interictal)
- Classification problem: classify interictal HFOs as "ictal-like" or "interictal-like"
- Should be easier in healthy tissue (i.e. outside of the SOZ)



- Ictal and interictal HFOs are fairly distinguishable in ILAE Class I patients (good surgery outcomes)
- Poor distinguishability may be a biomarker for poor surgery outcome

Conclusion

- Distributional functional estimation is very useful
  - Provides a framework for machine learning problems

Conclusion

- Distributional functional estimation is very useful
  - Provides a framework for machine learning problems
- I derived convergence rates for a k-nn plug-in estimator of distributional functionals

Conclusion

- Distributional functional estimation is very useful
  - Provides a framework for machine learning problems
- I derived convergence rates for a k-nn plug-in estimator of distributional functionals
- I obtained an estimator with MSE convergence of $O\left(\frac{1}{N}\right)$ by applying the theory of optimally weighted ensemble estimation
  - General g
  - Simple to implement
  - Converges rapidly
  - Performs well for higher dimensions
  - Central limit theorem

Conclusion: My Applications

- Estimating intrinsic dimension for dimensionality reduction of sunspot images
  - Moon et al, 2014 (ICIP); Moon et al, 2016a (SWSC)
- Testing whether samples come from the same distribution
  - Moon et al, 2016a (SWSC)
- Clustering sunspot images using distributional features
  - Moon et al, 2016b (SWSC)
- Estimating bounds on the Bayes error
  - Moon and Hero, 2014 (NIPS); Moon et al, 2015 (SPW)
- Analysis of Neural data
  - Gliske, Moon et al, 2016 (ICASSP)
- Structure Learning
  - Moon et al, 2017 (ICASSP)

Future Work

- Extension to manifolds
  - Estimate intrinsic dimension simultaneously?

Future Work

- Extension to manifolds
  - Estimate intrinsic dimension simultaneously?
- Scalable methods of mutual information estimation
  - Structure learning on large datasets
  - Multivariate mutual information

Future Work

- Extension to manifolds
  - Estimate intrinsic dimension simultaneously?
- Scalable methods of mutual information estimation
  - Structure learning on large datasets
  - Multivariate mutual information
- Extension to non-iid samples
  - E.g. time series

Future Work

- Extension to manifolds
  - Estimate intrinsic dimension simultaneously?
- Scalable methods of mutual information estimation
  - Structure learning on large datasets
  - Multivariate mutual information
- Extension to non-iid samples
  - E.g. time series
- Can we relax smoothness assumptions further?
  - Minimax rates suggest so (Kandasamy et al, 2015)

Future Work

- Extension to manifolds
  - Estimate intrinsic dimension simultaneously?
- Scalable methods of mutual information estimation
  - Structure learning on large datasets
  - Multivariate mutual information
- Extension to non-iid samples
  - E.g. time series
- Can we relax smoothness assumptions further?
  - Minimax rates suggest so (Kandasamy et al, 2015)
- Density estimation
- k-nn classification

Questions?

References

- Extended ISIT paper available on arXiv (same name)
- KR Moon, M Noshad, S Yasaei Sekeh, AO Hero III, "Information Theoretic Structure Learning with Confidence," in *2017 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017b.
- KR Moon, K Sricharan, K Greenewald, and AO Hero III, "Improving convergence of divergence functional ensemble estimators," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.
- K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 356–360.
- K. R. Moon and A. O. Hero III, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.

References

- D. van Dijk, et al., "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion," *Cell,* 2018.
- F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, et al., "Transcriptional heterogeneity and lineage commitment in myeloid progenitors," *Cell*, vol. 163, no. 7, pp. 1663–1677, 2015.
- C Chow and C Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.
- D. Pál, B. Póczos, and C. Szepesvári, "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, pp. 1849–1857, 2010.

References

- J. Lewi, R. Butera, and L. Paninski, "Real-time adaptive informationtheoretic optimization of neurophysiology experiments," in *Advances in Neural Information Processing Systems*, pp. 857–864, 2006.
- E. Schneidman, W. Bialek, and M. J. B. II, "An information theoretic approach to the functional classification of neurons," *Advances in Neural Information Processing Systems*, vol. 15, pp. 197–204, 2003.
- S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 2116–2124, 2015.
- C. Salge, C. Glackin, and D. Polani, "Changing the environment based on empowerment as intrinsic motivation," *Entropy*, vol. 16, no. 5, pp. 2789–2819, 2014.

References

- K Kandasamy, A Krishnamurthy, B Poczos, L Wasserman, and J Robins, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, 2015, pp. 397–405.
- A Kraskov, H Stögbauer, and P Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138, 2004.
- S Khan, S Bandyopadhyay, A R Ganguly, S Saigal, D J Erickson III, V Protopopescu, and G Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, pp. 026209, 2007.
- Nguyen, X. and Wainwright, M. and Jordan, M., "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory,* vol.

References

- H. Liu, L. Wasserman, and J. D. Lafferty, "Exponential concentration for mutual information estimation with application to forests," in *Advances in Neural Information Processing Systems*, pp. 2537–2545, 2012.