Estimating Density Functionals

Barnabás Póczos
Why are we all here?
Curious
To solve these problems, our main tool is always the same
Collect data & learn from data
The world is very complicated...
We have to understand complex relationships across the data.

Basic questions about the data

- **How random is the data?**
  - How large is its entropy?

- **How large is the dependence among the instances?**
  Which variables are dependent, which ones are independent?
  - How large is their mutual information?

- **How different are the distributions of the instances?**
  - How large is the divergence between the distributions?

**Difficult & Important**

⇒ We need Entropy, Dependence, and Divergence estimators to do machine learning
Entropy, Mutual Information, Divergence

\[ H = -\int p \log p \]
\[ KL(p||q) = \int p \log \frac{p}{q} \]
\[ I = KL(p||\prod p_i) \]

**Fernandes & Gloor:** Mutual information is critically dependent on prior assumptions: **would the correct estimate of mutual information please identify itself?**

*BIOINFORMATICS* Vol. 26 no. 9 2010, pages 1135–1139

\[ D_f(p||q) = \int f \left( \frac{p(x)}{q(x)} \right) q(x) \, d\mu(x) \]

**MI** = Divergence between \( p(x_1, \ldots, x_d) \) and \( \prod_{i=1}^{d} p_i(x_i) \)
Developing efficient estimators for mutual information and related quantities is highly important in many applications.

- “Mutual information” query produces 325,000 hits on Google Scholar, and the first 10 papers have more than 30,065 citations.

- Most of these papers are application papers, e.g., in feature selection, computer vision, medical image processing, image alignment, and data fusion. As we find better estimators, such applications can simply use them.

- “Big Data” search on Google Scholar produces 181,000 hits, and the first 10 hits have 12,872 citations.

- Similarly, the “Deep Learning” search produces 106,000 hits, and the first 10 papers have 8,485 citations (as of May 28, 2017).
How should we estimate them?

Naïve plug-in approach using density estimation

- histogram
- kernel density estimation
- k-nearest neighbors [D. Loftsgaarden & C. Quesenberry. 1965.]

Density: nuisance parameter
Density estimation: difficult, curse of dimensionality!

How can we estimate them directly, without estimating the density?
Part I

Consistent estimators for

- Rényi entropy
- Rényi mutual information
- A large class of divergences that includes Rényi and $L_2$

They avoid density estimation!
Part II
Generalize ML to sets and distributions

Most machine learning algorithms operate on vectorial objects.

The world is **complicated**. Often
- hand crafted vectorial features are not good enough
- natural to work with complex inputs directly (sets or distributions...)

---

Classify galaxy clusters
- Each **galaxy** can be represented by a **feature vector**
- Each **cluster** can be represented by a **set** of these vectors
- We can’t concatenate the feature vectors into a huge vector
- do **ML on these unknown distributions** represented by sets
Part 1
Estimators

- entropy estimation
- dependence estimation
- divergence estimation

Part 2
ML on distributions

- classification, regression, clustering, anomaly detection, low-dim embedding

Applications

- computer vision
- astronomy
- other applications
ENTROPY ESTIMATION

without density estimation

Using \( X_{1:n} \doteq (X_1, \ldots, X_n) \) i.i.d. sample \( \sim f \)

Estimate Rényi entropy \( R_\alpha = \frac{1}{1 - \alpha} \log \int f^\alpha(x)dx \)
Rényi-α entropy estimators using kNN graphs

Pál, Póczos & Szepesvári. *NIPS 2010*

\[ X^1, \ldots, X^n \sim f \text{ i.i.d. samples in } \mathbb{R}^d \]

Let \( p = d - d\alpha \), \( k \) fixed.

Let \( \mathcal{N}_{k,j} \) be the set of the \( k \) nearest neighbours of \( X^j \) in \( \{X^1, \ldots, X^n\} \)

\[ k = 3 \]

**Calculate:**

\[ L_n = \sum_{j=1}^{n} \sum_{V \in \mathcal{N}_{k,j}} ||V - X^j||^p \]

\[ H_n(X^{1:n}) \triangleq \frac{1}{1-\alpha} \log \left( \frac{L_n}{\beta_{d,p,k} n^\alpha} \right) \]
Theoretical Results

Almost surely consistent

\[ H_n(X^{1:n}) \overset{\text{d}}{=} \frac{1}{1 - \alpha} \log \left( \frac{L_n}{n^{(d-p)/d\beta}} \right) \rightarrow H_\alpha(X) \]

Convergence rate

If the density \( f \) is Lipschitz, \( \alpha = 1 - p/d \), then for any \( \delta > 0 \) with probability at least \( 1 - \delta \),

\[
\left| H_n(X^{1:n}) - H_\alpha(f) \right| \leq \begin{cases} 
O \left( n^{-\frac{d-p}{d(2d-p)}} (\log(1/\delta))^{1/2-p/(2d)} \right), & \text{if } 0 < p < d - 1 \\
O \left( n^{-\frac{d-p}{d(d+1)}} (\log(1/\delta))^{1/2-p/(2d)} \right), & \text{if } d - 1 \leq p < d .
\end{cases}
\]

First high probability rate on Rényi entropy estimators.

Pál, Póczos & Szepesvári, \textit{NIPS 2010}
Why is this entropy estimator consistent?

\[ \frac{1}{1 - \alpha} \log \left( \frac{L_n}{n^{(d-p)/d\beta}} \right) \rightarrow H_\alpha(X) \]

- The larger the entropy, the longer the kNN graph is.
- Quasi-subadditivity: \( L_{m+n} \leq L_m + L_n + \Delta_{m+n} \Rightarrow \exists \lim_{n \to \infty} \frac{L_n}{n^\alpha} < \infty \)

\( L_n = 10.77 \)
Part 1
Estimators

entropy estimation
dependence estimation
divergence estimation

Part 2
ML on distributions

classification, regression, clustering,
anomaly detection, low-dim embedding

Applications

computer vision
astronomy
other applications
MUTUAL INFORMATION ESTIMATION

without density estimation

Using $X_1, \ldots, X_n$ i.i.d. sample $\sim f = (f_1, \ldots, f_d)$

Estimate MI $I_{\alpha} \doteq \frac{1}{\alpha - 1} \log \int f^{\alpha}(x) \left( \prod_{i=1}^{d} f_i(x_i) \right)^{1-\alpha} \, dx$
How can we get mutual information estimators from entropy estimators?

Trick: Information is preserved under monotonic transformations.

Let \((g_1(X_1), \ldots, g_d(X_d)) = (Z_1, \ldots, Z_d) = Z\)
where \(g_j : \mathbb{R} \to \mathbb{R}, \ j = 1, \ldots, d,\) are monotone functions.

\[
I_\alpha(Z) \doteq \frac{1}{\alpha - 1} \log \int_Z \left(f_Z(z)\right)^\alpha \quad dz = I_\alpha(X)
\]

When the marginals of \(Z\) are uniform, \(\Rightarrow I_\alpha(Z) = -H_\alpha(Z)\)

\[
\Rightarrow I_\alpha(X) = I_\alpha(Z) = -H_\alpha(Z)
\]

Monotone transform \quad Uniform margins
Transformation to Get Uniform Margins

Monotone transformation leading to uniform margins?

Prob theory 101: \( X_j \sim F_j \) cont. \( \Rightarrow F_j(X_j) \sim U[0, 1] \)

The copula transformation:

Let \( X = [X_1, \ldots, X_d] \rightarrow [F_1(X_1), \ldots, F_d(X_d)] = [Z_1, \ldots, Z_d] = Z \)

A little problem: we don’t know \( F_i \) distribution functions...

Solution: Empirical distribution function (ranks are enough)
Sklar’s Theorem, 1959

Copula distribution
$$(F_X(X), F_Y(Y)) \sim C$$

The copula composition of its copula $C$ and the marginals
$$I_\alpha(X, Y) = -H_\alpha(C)$$
Copula based methods are popular in financial analysis. So popular and powerful that they led to the global financial crisis...

*WIRED MAGAZINE: 17.03*

Tech Biz : IT ⚡

Recipe for Disaster: The Formula That Killed Wall Street

By Felix Salmon ☮ 02.23.09

*It is time to make them more popular in machine learning too!*...
**REGO**: Rank-based estimation of Rényi information using Euclidean Graph Optimization

**Other Euclidean graphs**: TSP, MST, Minimal Matching, ...

1st direct, consistent Rényi mutual information estimator


**Consistency theorem:**

Let $d \geq 3$, $1/2 < \alpha < 1$. $X, X^1, \ldots, X^n$ i.i.d. with density $f = f_X$.

$\Rightarrow \hat{I}(X^1, \ldots, X^n) \rightarrow I_\alpha(X)$ almost surely as $n \rightarrow \infty$. 
Convergence Rate

If the density of the copula of $f$ is Lipschitz, $\alpha = 1 - p/d$, then for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left| \hat{I}(X^1, \ldots, X^n) - I_\alpha(X) \right| \leq$$

$$O \left( \max \left\{ n^{-\frac{d-p}{d(2d-p)}}, n^{-p/2+p/d} \right\} (\log(1/\delta))^{1/2} \right), \quad \text{if } 0 < p \leq 1;$$

$$O \left( \max \left\{ n^{-\frac{d-p}{d(2d-p)}}, n^{-1/2+p/d} \right\} (\log(1/\delta))^{1/2} \right), \quad \text{if } 1 \leq p \leq d/2$$

Pál, Póczos & Szepesvári, *NIPS 2010*
Robustness to Outliers

The amount of change caused by adding one outlier \( x \)

**Empirical mean:**  \[
\Delta_n(x) = \frac{1}{n+1} \left( x + \sum_{i=1}^{n} X_i \right) - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right) \rightarrow \infty
\]

**REGO:**

\[
\Delta_n(x) = |\hat{I}_{n+1}(X_{1:n}, x) - \hat{I}_n(X_{1:n})| = O(n^{-\alpha}) \quad 1/2 < \alpha < 1.
\]
Part 1
Estimators

- entropy estimation
- dependence estimation
- divergence estimation

Part 2
ML on distributions

- classification, regression, clustering, anomaly detection, low-dim embedding

Applications

- computer vision
- astronomy
- other applications
Using \( X_{1:n} = \{X_1, \ldots, X_n\} \sim p \quad Y_{1:m} = \{Y_1, \ldots, Y_m\} \sim q \)

Estimate divergence \( R_\alpha(p\|q) \triangleq \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} \)
The Estimator

\[ \mathbb{R}^d \supseteq \mathcal{M} \]

\[ k = 2. \]

\[ k \geq 1, \text{ fixed.} \]

\[ \rho_k(i) : \text{the distance of the } k\text{-th nearest neighbor of } X_i \text{ in } X_{1:n} \]

\[ \nu_k(i) : \text{the distance of the } k\text{-th nearest neighbor of } X_i \text{ in } Y_{1:m} \]

\[ D_\alpha(p\|q) \triangleq \int p^\alpha q^{1-\alpha} \]

\[ \widehat{D}_\alpha(X_{1:n}\|Y_{1:m}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n-1)\rho_{k}^d(i)}{m \nu_{k}^d(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)} \]
Asymptotically Unbiased

The estimator
\[ \hat{D}_\alpha(X_{1:n} \| Y_{1:m}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n-1) \rho_k^d(i)}{m \nu_k^d(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1) \Gamma(k+\alpha-1)} \]

We need to prove:
\[ \lim_{n,m \to \infty} \frac{\Gamma(k-\alpha+1) \Gamma(k+\alpha-1)}{\Gamma(k)^2} \int p^\alpha q^{1-\alpha} = \lim_{n,m \to \infty} \mathbb{E} \left[ \left( \frac{(n-1) \rho_k^d(1)}{m \nu_k^d(1)} \right)^{1-\alpha} \right] \]

The r.h.s. can be rewritten as
\[ \lim_{n,m \to \infty} \mathbb{E}_{X_1 \sim p} \left[ (n-1)^{1-\alpha} \rho_k^d(1-\alpha)(1) \bigg| X_1 = x \right] \mathbb{E} \left[ m^{\alpha-1} \nu_k^d(\alpha-1)(1) \bigg| X_1 = x \right] \]

Normalized k-NN distances converge to the Erlang distribution
\[ \xi_n = (n-1) \rho_k^d(1) \to_d \xi \]

All we need is \( \{ \xi_n \to_d \xi \} \Rightarrow \{ \mathbb{E}[\xi_n^{1-\alpha}] \to \mathbb{E}[\xi^{1-\alpha}] \} \)

Póczos & Schneider, AISTATS 2011
A little problem…

\[ \{ \xi_n \to_d \xi \} \nRightarrow \{ \mathbb{E}[\xi_n] \to \mathbb{E}[\xi] \} \]

\[ f_1 \quad f_2 \quad f_3 \]

\( \mathbb{E}[\xi] < \infty, \text{ but } \mathbb{E}[\xi_1] = \infty, \mathbb{E}[\xi_2] = \infty, \ldots, \mathbb{E}[\xi_n] = \infty, \ldots \)

**Solutions:** Asymptotic uniformly integrability…

\[
\lim_{\beta \to \infty} \lim_{n \to \infty} \sup_{|u| \geq \beta} \int |u| f_n(u) \, du = 0 \text{ then } \lim_{n \to \infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi].
\]

Appendix of Póczos & Schneider, AISTATS 2011

Be careful, mistakes are easy to make!

**Need:** \( \{ \xi_n \to_d \xi \} \nRightarrow \{ \mathbb{E}[\xi_n] \to \mathbb{E}[\xi] \} \)

Strong law of large numbers [NIPS]
Be careful, some mistakes are easy to make...

**We want:** \( \{F_n(u) \to F(u) \ \forall u\} \Rightarrow \left\{ \int_0^\infty u \, dF_n(u) \to \int_0^\infty u \, dF(u) \right\} \)

**Helly–Bray theorem** \( \int_R g(u) \, dF_n(u) \to \int_R g(u) \, dF(u) \)

for each bounded, continuous function \( g : \mathbb{R} \to \mathbb{R} \)

[Annals of Statistics]

---

**Enough:** There exists an \( \varepsilon > 0 \) such that \( \limsup_{n \to \infty} \mathbb{E} \left[ \xi_n^{\gamma(1+\varepsilon)} \right] < \infty \).

**Fatou lemma:** \( \limsup_{n \to \infty} \mathbb{E} \left[ \xi_n^{\gamma(1+\varepsilon)} \right] \leq \mathbb{E} \left[ \limsup_{n \to \infty} \xi_n^{\gamma(1+\varepsilon)} \right] < \infty \)

\( \gamma(1 + \varepsilon) \) moment of an Erlang variable < \( \infty \)

**Fatou lemma:** \( \liminf_{n \to \infty} \mathbb{E} \left[ \xi_n^{\gamma(1+\varepsilon)} \right] \geq \mathbb{E} \left[ \liminf_{n \to \infty} \xi_n^{\gamma(1+\varepsilon)} \right] \)

Part 1
Estimators

- Entropy estimation
- Dependence estimation
- Divergence estimation

Part 2
ML on distributions

- Classification, regression, clustering, anomaly detection, low-dim embedding

Applications

- Computer vision
- Astronomy
- Other applications
Dealing with complex objects

- break into smaller parts, represent the input as a set of smaller parts
- treat the set elements as sample points from some **unknown distribution**
- do **ML on these unknown distributions** represented by sets
Many ML algorithms only require

- the pairwise distances between the inputs
- the inner products between the inputs

If we can estimate divergences and inner products between distributions, then we can construct ML algorithms that operate on distributions.

- Classification
- Regression
- Low-dimensional embedding
- Anomaly detection
\[ Y_1 = 1 \quad Y_2 = 0 \quad Y_3 = 1 \quad Y_m = 0 \quad ? \]

**Differences compared to standard methods on vectors**

- The inputs are distributions, density functions (not vectors)
- We don’t know these distributions, only sample sets are available (error in variables model)
Distribution Classification

We have $T$ sample sets, $(X_1, \ldots, X_T)$. [Training data]\n\[\{X_{t,1}, \ldots, X_{t,m_t}\} = X_t \sim p_t. \quad X_t \text{ has class } Y_t \in \{-1, +1\}.\]

What is the class label $Y$ of $X = \{X_1, \ldots, X_m\} \sim p$?

Solution: Use RKHS based SVM!

Calculate the Gram matrix \[K_{ij} \doteq \langle \phi(p_i), \phi(p_j) \rangle_K = K(p_i, p_j)\]

Dual form of SVM:
\[
\hat{\alpha} = \arg \max_{\alpha \in \mathbb{R}^T} \sum_{i=1}^{T} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K_{ij}, \quad \text{subject to } \sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C.
\]
\[Y = \text{sign} \left( \sum_{i=1}^{T} \hat{\alpha}_i y_i K(p_i, p) \right) \in \{-1, +1\}\]

Problems: We do not know $p_i$, $p$, $K(p_i, p_j)$, or $K(p_i, p)$...
Kernel Estimation

Linear kernel: $K(p, q) = \int pq$

Polynomial kernel: $K(p, q) = (\int pq + c)^s$

Gaussian kernel: $K(p, q) = \exp\left(-\frac{1}{2\sigma^2}\int (p - q)^2\right)$.

We only need to estimate $\int p^\alpha q^\beta$ terms.

We already know how!

We can also try to use other $\mu(p, q)$ divergences, e.g. Rényi ...

The $\{\hat{K}_{i,j}\}_{ij}$ Gram matrix might not be PSD!

Solution: make it symmetric, and project it to the cone of PSD matrices
OUTLINE

Part 1
Estimators

entropy estimation
dependence estimation
divergence estimation

Part 2
ML on distributions

classification, regression, clustering,
anomaly detection, low-dim embedding

Applications

computer visionastronomyother applications
Image Representation with Distributions

Dealing with complex objects
- break into smaller parts,
- represent the object as a sample set of these parts

Image patches
- Overlapping
- Non-overlapping

Patch locations
- Grid points
- Interesting points
- Random

Patch sizes
- Same
- Different,
- Hierarchy

d-dimensional sample set representation of the image
- Each image patch is represented by PCA compressed SIFT vectors.
  \[ \text{SIFT} = \text{Scale-invariant feature transform. PCA: } 128\text{dim} \Rightarrow d \text{ dim} \]
- Each image is represented as a set of these \( d \) dim feature vectors.
- Each set is considered as a sample set from some unknown distribution.
Detecting Anomalous Images

50 highway images

5 anomalies

2-dimensional sample set representation of images (128 dim SIFT $\Rightarrow$ 2 dim)

Anomaly score: divergences between the distributions of these sample sets
Detecting Anomalous Images
GMM-5 Density Approximation
Noisy USPS Dataset Classification with SDM

- Original (noiseless) USPS dataset is easy ~97%
- Each instance (image) is a set of 500 2d points
- 1000 training and 1000 test instances

Results:
- SVM on raw images: 82.1 ± .5% accuracy
- SDM on the 2D distributions, Rényi divergence: 96.0 ± .3% accuracy
Multidimensional Scaling of USPS Data

10 instances from figures 1, 2, 3, 4.
Calculate pairwise Euclidean distances.
Nonlinear embedding with MDS into 2d.

Raw images using Euclidean distance

Estimated Euclidean distance between the distributions
Local Linear Embedding of Distributions

72 rotated COIL froggies

Edge detected COIL froggy

Euclidean distance between images

Euclidean distance between distributions
Object Classification
ETH-80 [Leibe and Schiele, 2003]

8 categories, 400 images, each image is represented by 576 18 dim points

- BoW: 88.9%
- NPR: 90.1%

Póczos, Xiong, Sutherland, & Schneider, CVPR 2012
Outdoor Scenes Classification
[Oliva and Torralba, 2001]

- Best published: 91.57% (Qin and Yung, ICMV 2010)
- NPR: 92.3%

8 categories, 2688 images, each represented by 1815 53 dim points.

Póczos, Xiong, Sutherland, & Schneider, CVPR 2012
Sport Events Classification
[Li and Fei Fei, 2007]

- Best published: 86.7%
  (Zhang et al, CVPR 2011)
- NPR: 87.1%

8 categories, 1040 images, each represented by 295 to 1542 57 dim points.
**OUTLINE**

**Part 1**  
Estimators  
- entropy estimation  
- dependence estimation  
- divergence estimation

**Part 2**  
ML on distributions  
- classification, regression, clustering, anomaly detection, low-dim embedding

**Applications**  
- computer vision  
- astronomy  
- other applications
Goal: Estimate dynamical mass of galaxy clusters.

Importance: Galaxy clusters are being the largest gravitationally bound systems in the Universe. Dynamical mass measurements are important to understand the behavior of dark matter and normal matter.

Difficulty: We can only measure the velocity of galaxies not the mass of their cluster. Physicists estimate dynamical cluster mass from single velocity dispersion.

Our method: Estimate the cluster mass from the whole distribution of velocities rather than just a simple velocity distribution.
Find new scientific laws in physics

Test Catalog

PL1: $M(\sigma_v)$ power law
PL2: $M(\sigma_v)$ power law with $\kappa$
ML1: SDM with $|v_{\text{los}}|$
ML2: SDM with $|v_{\text{los}}| \& \frac{|v_{\text{los}}|}{\sigma_v}$

What are the most anomalous galaxy clusters?

The most anomalous galaxy cluster contains mostly
- star forming blue galaxies
- irregular galaxies

B. Póczos, L. Xiong & J. Schneider, UAI, 2011.

Credits: ESA, NASA
Find the parameters of Universe

Given a distribution of particles, our goal is to predict the parameters of the simulated universe.
OUTLINE

Part 1
Estimators
- entropy estimation
- dependence estimation
- divergence estimation

Part 2
ML on distributions
- classification, regression, clustering, anomaly detection, low-dim embedding

Applications
- computer vision
- astronomy
- other applications
Understanding Turbulences

Credits: ESA, NASA, PPPL, Wikipedia
Turbulence Data Classification

Simulated fluid flow through time
(JHU Turbulence Research Group)

Goal: find interesting events, patterns, phenomena

• 11 positive, 20 negative examples
• Results: Leave one out cross-val: 97%

What happened? Something interesting happened?

Positive (vortex)  Negative  Negative
Finding Vortices

Classification probabilities
Find Interesting Phenomena in Turbulence Data

Anomaly detection with 1-class SDM

Anomaly scores
Find Interesting Phenomena in Turbulence Data

Vorticity Scores

Xiong, Póczos, and Schneider, *NIPS 2011*.
Agriculture
Surrogate robotic system in the field
The surrogate system collecting data at the TAMU field site. The carriage supports two boom assemblies each one of which carries a sensor pod. The carriage slides up and down on the column allowing full scanning of a plant.
Surrogate robotic system in the field

The carriage/dual-boom assembly moves up and down the column at a constant scanning speed. At its highest travel point the assembly clears the canopy (right).
Data collection with sensor pods

A sensor pod is deployed into a row and scans a plant
<table>
<thead>
<tr>
<th>Name</th>
<th>Range</th>
<th>RMSE error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf angle*</td>
<td>75.94</td>
<td>3.30 (4.35%)</td>
</tr>
<tr>
<td>Leaf radiation angle*</td>
<td>120.66</td>
<td>4.34 (3.60%)</td>
</tr>
<tr>
<td>Leaf length*</td>
<td>35.00</td>
<td>0.87 (2.49%)</td>
</tr>
<tr>
<td>Leaf width [max]</td>
<td>3.61</td>
<td>0.27 (7.48%)</td>
</tr>
<tr>
<td>Leaf width [average]</td>
<td>2.99</td>
<td>0.21 (7.02%)</td>
</tr>
<tr>
<td>Leaf area*</td>
<td>133.45</td>
<td>8.11 (6.08%)</td>
</tr>
</tbody>
</table>
Extensions

**L₂ divergence:**

\[ L(p \| q) \doteq \left( \int_M (p(x) - q(x))^2 \, dx \right)^{1/2} \]

\[ \hat{L}^2(X_1:N \| Y_1:M) \doteq \]

\[
\frac{1}{N} \sum_{n=1}^{N} \left[ \frac{k - 1}{(N - 1)c\rho_k^d(X_n)} - \frac{2(k - 1)}{Mc\nu_k^d(X_n)} + \frac{(N - 1)c\rho_k^d(X_n)(k - 2)(k - 1)}{(Mc\nu_k^d(X_n))^2 k} \right],
\]

where \( k - 2 > 0 \).

**Conditional Rényi Mutual Information:**

\[ I_\alpha(X, Y \| Z) \doteq \int p_Z(z) D_\alpha(p(X, Y \| Z = z) \| p(X \| Z = z)p(Y \| Z = z)) \ | Z = z) \]

\[ \hat{I}_\alpha = \frac{1}{\alpha - 1} \log \frac{1}{N} \sum_{n=1}^{N} \frac{(c_{xyz}(1-\alpha))^{d_{xyz}(1-\alpha)}(X_n, Y_n, Z_n)}{(c_{xz}(1-\alpha))^{d_{xz}(1-\alpha)}(X_n, Z_n)} \frac{(c_{yz}(1-\alpha))^{d_{yz}(1-\alpha)}(Y_n, Z_n)}{(c_{y}(1-\alpha))^{d_{y}(1-\alpha)}(Z_n)} B^2, \]

where \( B^2 = \frac{\Gamma^4(k)}{\Gamma^2(k-\alpha+1)\Gamma^2(k+\alpha-1)} \).

Rates for kNN Estimators
Let $\mathcal{X} := [0, 1]^D$ denote the unit cube in $\mathbb{R}^D$, and let $\mu$ denote the Lebesgue measure.

Suppose $P$ is an unknown $\mu$-absolutely continuous Borel probability measure supported on $\mathcal{X}$, and let $p : \mathcal{X} \to [0, \infty)$ denote the density of $P$.

Consider a (known) differentiable function $f : (0, \infty) \to \mathbb{R}$.

Given $n$ samples $X_1, \ldots, X_n$ drawn IID from $P$, we are interested in estimating the functional

$$F(P) := \mathbb{E}_{X \sim P} [f(p(X))].$$
We will work with distances induced by the $r$-norm

$$\|x\|_r := \left( \sum_{i=1}^{D} x_i^r \right)^{1/r}$$

and define

$$c_{D,r} := \frac{(2\Gamma(1 + 1/r))^D}{\Gamma(1 + D/r)} = \mu(B(0, 1)),$$

where $B(x, \varepsilon) := \{y \in \mathbb{R}^D : \|x - y\|_r < \varepsilon\}$ denotes the open radius-$\varepsilon$ ball centered at $x$. 
K-NN Distances

Suppose we have \( n \) samples \( X_1, \ldots, X_n \) drawn IID from \( P \).

For any \( x \in \mathbb{R}^D \), we define the \( k \)-nearest neighbor distance \( \varepsilon_k(x) \) by \( \varepsilon_k(x) = \| x - X_i \|_r \), where \( X_i \) is the \( k \)th-nearest element (in \( \| \cdot \|_r \)) of the set \( \{ X_1, \ldots, X_n \} \) to \( x \).

For divergence estimation, if we also have \( n \) samples \( Y_1, \ldots, Y_n \) drawn IID from \( Q \), then we similarly define \( \delta_k(x) \) by \( \delta_k(x) = \| x - Y_i \|_r \), where \( Y_i \) is the \( k \)th-nearest element of \( \{ Y_1, \ldots, Y_n \} \) to \( x \).
The $k$-NN density estimator

$$
\hat{p}_k(x) = \frac{k/n}{\mu(B(x, \varepsilon_k(x)))} = \frac{k/n}{c_D \varepsilon_k^D(x)}
$$

is well-studied nonparametric density estimator (Loftsgaarden, 1965), motivated by the observations that, for small $\varepsilon > 0$,

$$
p(x) \approx \frac{P(B(x, \varepsilon))}{\mu(B(x, \varepsilon))},
$$

and that, $P(B(x, \varepsilon_k(x))) \approx k/n$. 
k-NN density Estimator Properties

One can show that, for $x \in \mathbb{R}^D$ at which $p$ is continuous, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then $\hat{p}_k(x) \to p(x)$ in probability (Loftsgaarden, 1965, Theorem 3.1).

The Plug-In estimator:

Thus, a natural approach for estimating $F(P)$ is the plug-in estimator of $F(P) := \mathbb{E}_{X \sim P} [f(p(X))]$ is:

$$\hat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f(\hat{p}_k(X_i)).$$
Plug-In Estimator Properties

Since $\hat{p}_k \to p$ in probability pointwise as $k, n \to \infty$ and $f$ is smooth, one can show $\hat{F}_{PI}$ is consistent.

Sricharan,(2010) showed a convergence rate of $O\left(n^{-\min\left\{\frac{2\beta}{\beta+D},1\right\}}\right)$ for $\beta$-Hölder continuous densities by setting $k \asymp n^{\frac{\beta}{\beta+d}}$. 
Issues with the Plug-In Estimator

Unfortunately, while necessary to ensure \( \nabla [\hat{p}_k(x)] \to 0 \), the requirement \( k \to \infty \) is computationally burdensome.

Furthermore, increasing \( k \) can increase the bias of \( \hat{p}_k \) due to over-smoothing (see later), suggesting that this may be sub-optimal for estimating \( F(P) \).
Fixed-k functional estimators

\[
F(P) := \mathbb{E}_{X \sim P} [f(p(X))]. \quad \hat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f(\hat{p}_k(X_i)).
\]

An alternative approach is to fix \( k \) as \( n \to \infty \).

Since \( \hat{F}_{PI} \) is itself an empirical mean, unlike \( \mathbb{V} [\hat{p}_k(x)] \),

\[ \mathbb{V} [\hat{F}_{PI}] \to 0 \text{ as } n \to \infty. \]
Bias Correction

\[ F(P) := \mathbb{E}_{X \sim P} [f(p(X))] \quad \text{and} \quad \widehat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f(\hat{p}_k(X_i)) \]

A more critical complication of fixing \( k \) is the bias of the plug-in estimator.

Since \( f \) is typically non-linear, the non-vanishing variance of \( \hat{p}_k \) translates into asymptotic bias.
Bias Correction

\[ F(P) := \mathbb{E}_{X \sim P} [f(p(X))] \quad \text{and} \quad \hat{F}_{PI} := \frac{1}{n} \sum_{i=1}^{n} f(\hat{p}_k(X_i)) \]

A solution is to derive a **bias correction** function \( \mathcal{B} \) such that

\[ \mathbb{E}_{X, X_1, \ldots, X_n} [\mathcal{B}(f(\hat{p}_k(X)))] = \mathbb{E}_{X, X_1, \ldots, X_n} [f(p_{\varepsilon_k}(X)(X))]. \]

That is,

\[ \mathbb{E}_{X, X_1, \ldots, X_n} \left[ \mathcal{B} \left( f \left( \frac{k/n}{\mu(B(x, \varepsilon_k(X)))} \right) \right) \right] = \mathbb{E}_{X, X_1, \ldots, X_n} \left[ f \left( \frac{P(B(X, \varepsilon_k(X)))}{\mu(B(X, \varepsilon_k(X)))} \right) \right]. \]

That is we correct the bias due to that \( P(B(x, \varepsilon_k(x))) \) is unknown and estimated with \( k/n \).

\[ p_{\varepsilon_k}(x)(x) := \frac{P(B(x, \varepsilon_k(x)))}{\mu(B(x, \varepsilon_k(x)))} \] is a consistent estimate of \( p(x) \) with \( k \) fixed, but it is not computable, since \( P \) is unknown.
The bias correction $\mathcal{B}$ gives us an asymptotically unbiased estimator

$$
\hat{F}_B(P) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{B} \left( f \left( \hat{p}_k(X_i) \right) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{B} \left( f \left( \frac{k/n}{\mu(B(X_i, \varepsilon_k(X_i)))} \right) \right).
$$

that uses $k/n$ in place of $P(B(x, \varepsilon_k(x)))$.

[ Instead of $\hat{F}(P) := \frac{1}{n} \sum_{i=1}^{n} f \left( \hat{p}_k(X_i) \right)$ ]

**Bias Correction for Divergences:**

This bias correction idea extends naturally to divergences:

$$
\hat{F}_B(P, Q) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{B} \left( f \left( \hat{p}_k(X_i), \hat{q}_k(X_i) \right) \right).
$$
### Known Bias Correction Functions

<table>
<thead>
<tr>
<th>Functional Name</th>
<th>Functional Form</th>
<th>Correction</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shannon Entropy</td>
<td>$\mathbb{E} \left[ \log p(X) \right]$</td>
<td>Additive constant: $\psi(n) - \psi(k) + \log(k/n)$</td>
<td>Kozachenko and Leonenko [1987], Goria et al. [2005]</td>
</tr>
<tr>
<td>Rényi-$\alpha$ Entropy</td>
<td>$\mathbb{E} \left[ p^{\alpha-1}(X) \right]$</td>
<td>Multiplicative constant: $\frac{\Gamma(k)}{\Gamma(k+1-\alpha)}$</td>
<td>Leonenko et al. [2008], Leonenko and Pronzato [2010]</td>
</tr>
<tr>
<td>KL Divergence</td>
<td>$\mathbb{E} \left[ \log \frac{p(X)}{q(X)} \right]$</td>
<td>None*</td>
<td>Wang et al. [2009]</td>
</tr>
<tr>
<td>$\alpha$-Divergence</td>
<td>$\mathbb{E} \left[ \left( \frac{p(X)}{q(X)} \right)^{\alpha-1} \right]$</td>
<td>Multiplicative constant: $\frac{\Gamma^2(k)}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$</td>
<td>Poczos and Schneider [2011]</td>
</tr>
</tbody>
</table>

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x} \, dx$$ is the gamma function, and $$\psi(x) = \frac{d}{dx} \log \left( \Gamma(x) \right)$$ is the digamma function.

$\alpha$ is a parameter in $\mathbb{R}\setminus\{1\}$.

For the KL divergence, the bias corrections for $p$ and $q$ exactly cancel.
Bias Correction: Entropy Special Case

As an example, if \( f = \log \) (as in Shannon entropy), then it can be shown that, for any continuous \( p \),

\[
\mathbb{E}_{X,X_1,\ldots,X_n} \left[ \log P(B(X, \varepsilon_k(X))) \right] = \psi(k) - \psi(n).
\]

Hence, for \( B_{n,k} := \psi(k) - \psi(n) + \log(n) - \log(k) \),

\[
\mathbb{E}_{X_1,\ldots,X_n} \left[ f \left( \frac{k/n}{\mu(B(x, \varepsilon_k(x)))} \right) \right] + B_{n,k} = \mathbb{E}_{X_1,\ldots,X_n} \left[ f \left( \frac{P(B(x, \varepsilon_k(x)))}{\mu(B(x, \varepsilon_k(x)))} \right) \right].
\]

giving the estimator of Kozachenko, 1987.

\[
\hat{H} = \psi(k) - \psi(n) + \log(n) - \log(k) + \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{k/n}{c_D \varepsilon_k^D(x)} \right)
\]
Here, we discuss some of these challenges, motivating the assumptions we make to overcome them.

First, these estimators are sensitive to regions of low probability (i.e., $p(x)$ small), for two reasons:

1. Many functions $f$ of interest (e.g., $f = \log$ or $f(z) = z^\alpha$, $\alpha < 0$) have singularities at 0.

2. The $k$-NN estimate $\hat{p}_k(x)$ of $p(x)$ is highly biased when $p(x)$ is small. For example, for $p$ $\beta$-Hölder continuous ($\beta \in (0, 2]$), one has (Mack, 1979, Theorem 2)

$$
\text{Bias}(\hat{p}_k(x)) \approx \left( \frac{k}{np(x)} \right)^{\beta/D}.
$$
For these reasons, it has been common in the analysis of k-NN estimators to make the following assumption:

**(A1)** $p$ is bounded away from zero on its support. That is, $p_* := \inf_{x \in \mathcal{X}} p(x) > 0$. 

Discussion of Assumptions
Discussion of Assumptions

Second, unlike many functional estimators, the fixed-\(k\) estimators we consider do not attempt correct for **boundary bias** (i.e., bias incurred due to discontinuity of \(p\) on the boundary \(\partial \mathcal{X}\) of \(\mathcal{X}\)).

Either of the following assumptions would suffice to obtain finite-sample rates:

**(A2)** \(p\) is continuous not only on the interior of \(\mathcal{X}\) but also on \(\partial \mathcal{X}\) (i.e., \(p(x) \to 0\) as \(\text{dist}(x, \partial \mathcal{X}) \to 0\)).

**(A3)** \(p\) is supported on all of \(\mathbb{R}^D\). That is, the support of \(p\) has no boundary.
Unfortunately, both assumptions (A2) and (A3) are inconsistent with (A1). Our approach is to assume (A2) and replace assumption (A1) with a much milder assumption that $p$ is \textit{locally lower bounded} on its support in the following sense:

\[(A4) \text{ There exist } \rho > 0 \text{ and a function } p_* : \mathcal{X} \rightarrow (0, \infty) \text{ such that, for all } x \in \mathcal{X}, r \in (0, \rho], \quad p_*(x) \leq \frac{P(B(x,r))}{\mu(B(x,r))}.\]

We can show that assumption (A4) is in fact very mild; as long as $p$ is continuous on $\mathcal{X}$, such a $p_*$ exists for any desired $\rho > 0$. For simplicity, we will use $\rho = \sqrt{D} = \text{diam}(\mathcal{X})$. 
(A4) is a Mild Assumption

(Lemma: Existence of Local Bounds) If $p$ is continuous on $\mathcal{X}$ and strictly positive on the interior $\mathcal{X}^\circ$ of $\mathcal{X}$, then, for $\rho := \sqrt{D} = \text{diam}(\mathcal{X})$, there exists a continuous function $p_* : \mathcal{X} \to (0, \infty)$ and a constant $p^* \in (0, \infty)$ such that

$$0 < p_*(x) \leq \frac{P(B(x, r))}{\mu(B(x, r))} \leq p^* < \infty, \quad \forall x \in \mathcal{X}^\circ, r \in (0, \rho].$$
Concentration of k-NN Distances

We now show that the existence of local lower and upper bounds implies concentration of the $k$-NN distance of around a term of order $(\frac{k}{np(x)})^{1/D}$.

Lemma: [Concentration of $k$-NN Distances] Suppose $p$ is continuous on $\mathcal{X}$ and strictly positive on $\mathcal{X}^\circ$. Then, for any $x \in \mathcal{X}^\circ$,

1. if $r > (\frac{k}{p_*(x)n})^{1/D}$, then $\mathbb{P}[\varepsilon_k(x) > r] \leq e^{-p_*(x)r^Dn} \left( e^{p_*(x)r^Dn} \frac{k}{k} \right)^k$.

2. if $r \in \left[0, (\frac{k}{p^*_n})^{1/D}\right)$, then $\mathbb{P}[\varepsilon_k(x) < r] \leq e^{-p_*(x)r^Dn} \left( ep^*_r^Dn \frac{k}{k} \right)^{kp_*(x)/p^*}$.
Concentration of k-NN Distances

Concentration Corollary:

For any $\alpha > 0,$

$$\mathbb{E} [\varepsilon_k^\alpha(x)] \leq \left(1 + \frac{\alpha}{D}\right) \left(\frac{k}{c_{D,r}np_*(x)}\right)^{\frac{\alpha}{D}}.$$
Main results: Bias Bound:

Suppose that, for some $\beta \in (0, 2]$, $p$ is $\beta$-Hölder continuous with constant $L > 0$ on $\mathcal{X}$, and $p$ is strictly positive on $\mathcal{X}^\circ$. Let $p_*$ and $p^*$ be as before. Let $f : (0, \infty) \to \mathbb{R}$ be differentiable, and define $M_{f,p} : \mathcal{X} \to [0, \infty)$ by

$$M_{f,p}(x) := \sup_{z \in [p_*(x), p^*]} \left| \frac{d}{dz} f(z) \right|$$

Assume

$$C_f := \mathbb{E}_{X \sim p} \left[ \frac{M_{f,p}(X)}{(p_*(X))^{\beta/D}} \right] < \infty.$$

Then,

$$\left| \hat{F}_\beta(P) - F(P) \right| \leq C_f L \left( \frac{k}{n} \right)^{\frac{\beta}{D}}.$$
Main results: Variance Bound

Suppose that $\mathcal{B} \circ f$ is continuously differentiable and strictly monotone. Assume that $C_{f,p} := \mathbb{E}_{X \sim P}[\mathcal{B}^2(f(p_*(X)))] < \infty$, and that $C_f := \int_0^\infty e^{-y}y^k f(y) < \infty$.

Then, for

$$C_V := 2 \left(1 + N_{k,D}\right)(3 + 4k)\left(C_{f,p} + C_f\right), \quad \text{we have} \quad \forall \left[\hat{F}_B(P)\right] \leq \frac{C_V}{n}.$$
Main results: MSE Bound

Under the previous conditions we have that

\[ E \left[ \left( \hat{H}_k(X) - H(X) \right)^2 \right] \leq C_f^2 L^2 \left( \frac{k}{n} \right)^{2\beta/D} + \frac{C_V}{n}. \]
Benefits

Compared to plug-in estimators, fixed-k estimators:

- are faster to compute,

- can also exhibit superior rates of convergence
Main result: Under some conditions

for $\beta$-Hölder continuous ($\beta \in (0, 2]$) densities on $[0, 1]^D$,

- the bias of fixed-$k$ estimators decays as $O\left(n^{-\beta/D}\right)$,
- the variance decays as $O\left(n^{-1}\right)$,
- giving a mean squared error of $O\left(n^{-2\beta/D} + n^{-1}\right)$.

- Hence, the estimators converge at the parametric $O(n^{-1})$ rate when $\beta \geq D/2$, and at the slower rate $O(n^{-2\beta/D})$ otherwise.
Dependence estimation

Entropy estimation

Dependence estimation

Divergence estimation

Part 1

Applications

classification, regression, clustering, anomaly detection, low-dim embedding

Part 2

Applications

computer vision, astronomy, other applications

• Outperforms state-of-the-art results in CV benchmarks
• Solves new problems in Astronomy, Turbulence data analysis, Agriculture

Take Me Home!

• direct, consistent estimators, rates
• 1st Rényi MI estimator: robust, rank statistics only
• 1st divergence estimators

• Support Distribution Machines