Machine Learning on Distributions

Barnabás Póczos

Carnegie Mellon University

Modern Nonparametric Methods in Machine Learning

NIPS 2012 Workshop

Dec 7, 2012

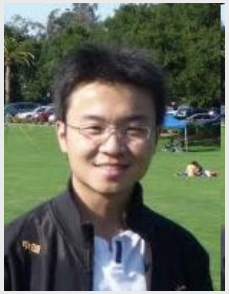
MACHINE LEARNING DEPARTMENT

AutonLab

Joint work with...

Jeff Schneider,
CMU

Larry Wasserman,
CMU



Liang Xiong,
CMU

Aarti Singh,
CMU

Dougal Sutherland,
CMU

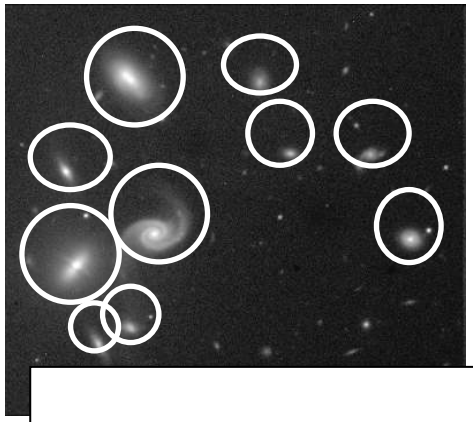
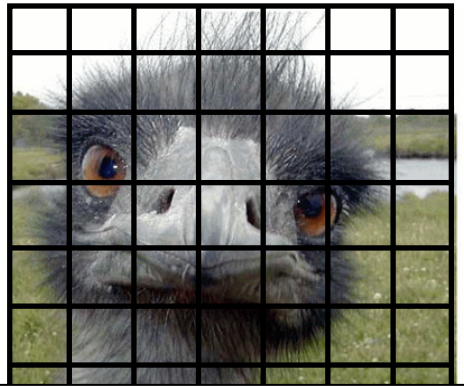
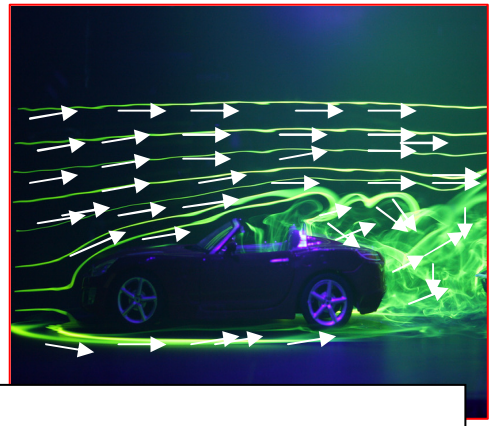
Alessandro Rinaldo,
CMU

GOAL:
Machine Learning on 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**...)

Deal

- ❑ do ***ML on these unknown distributions*** represented by sets

on

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

OUTLINE

1st Contribution Divergences	Nonparametric divergence estimation
2nd Contribution ML on distributions	classification, regression, clustering, anomaly detection, low-dim embedding, …
3rd Contribution Applications	computer vision, astronomy, turbulence data
4th Contribution Theory	Bound on prediction risk, rates

DIVERGENCE ESTIMATION

Examples: L_2 distance, L_1 distance, Kullback-Leibler divergence, f-divergence, maximum mean discrepancy, and many more…

Using $X_{1:n} = \{X_1, \ldots, X_n\} \sim p \quad Y_{1:m} = \{Y_1, \ldots, Y_m\} \sim q$

Estimate divergence $R_\alpha(p\|q) \doteq \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha}$

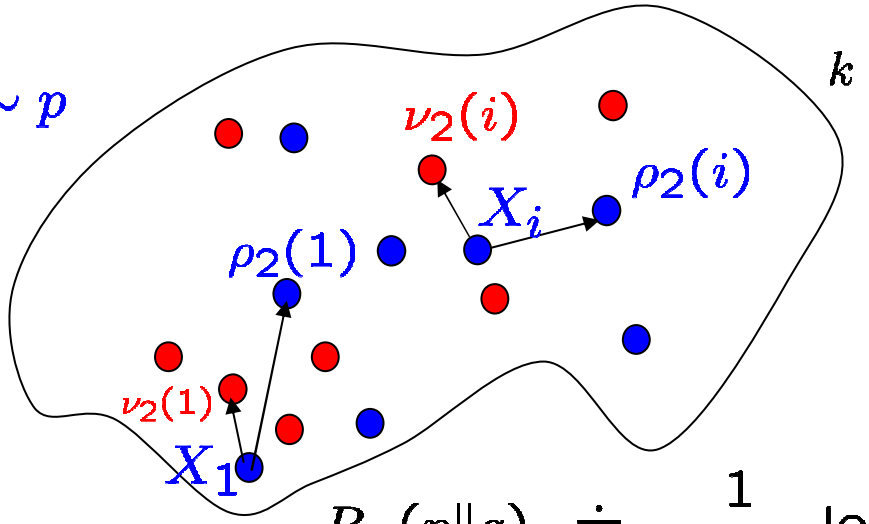
without density estimation

The Estimator

$X_{1:n} = \{X_1, \ldots, X_n\} \sim p$

$Y_{1:m} = \{Y_1, \ldots, Y_m\} \sim q$

$k \geq 1$, fixed.



$$R_\alpha(p\|q) \doteq \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha}$$

$$D_\alpha(p\|q) \doteq \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}x.$$

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

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

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

$\int p^\alpha(x) q^\beta(x) \mathrm{d}x$ can be estimated similarly.

Póczos & Schneider, *AISTATS 2011*

Does it make sense?

Theoretical results:

- Asymptotically unbiased
- L_2 consistent

Asymptotically Unbiased

The estimator

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

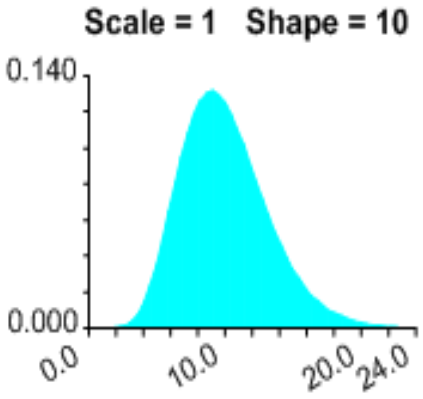
We need to prove:

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

The r.h.s. can be rewritten as

$$\lim_{n,m\to\infty}\mathbb{E}_{X_1\sim p}\left[\mathbb{E}\left[(n-1)^{1-\alpha}\rho_k^{d(1-\alpha)}(1)\middle|X_1=x\right]\mathbb{E}\left[m^{\alpha-1}\nu_k^{d(\alpha-1)}(1)\middle|X_1=x\right]\right]$$

$$\to p^{\alpha-1}(x)\frac{\Gamma(k+1-\alpha)}{\Gamma(k)}c^{\alpha-1} \to q^{1-\alpha}(x)\frac{\Gamma(k+\alpha-1)}{\Gamma(k)}c^{1-\alpha}$$



Normalized k-NN distances converge to the Erlang distribution

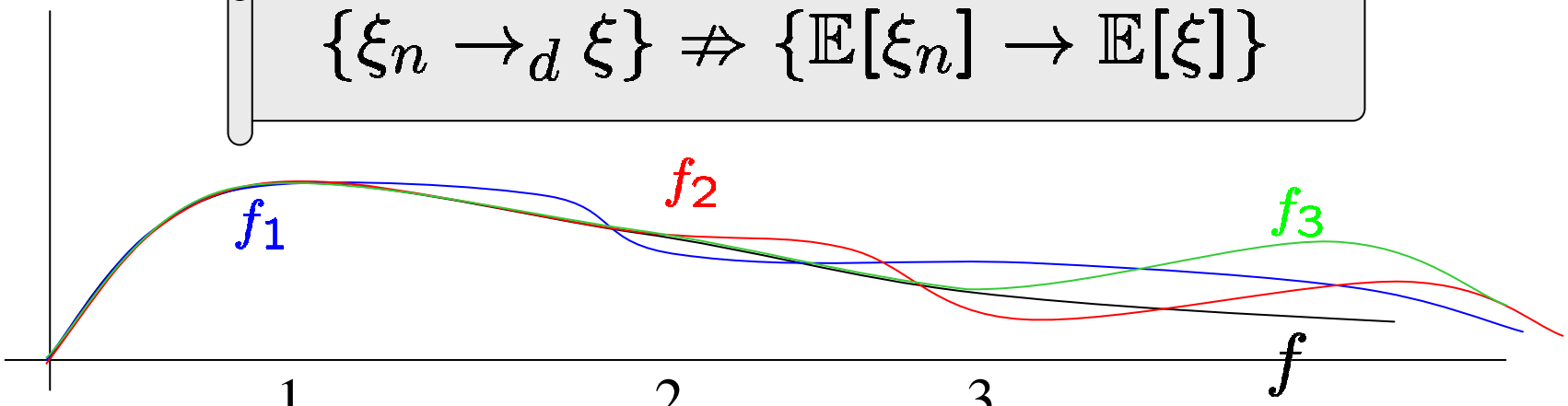
$$\xi_n = (n-1)\rho_k^d(1) \to_d \xi$$

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...

$$\{\xi_n \to_d \xi\} \not\Rightarrow \{\mathbb{E}[\xi_n] \to \mathbb{E}[\xi]\}$$



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

Solutions: *Asymptotic uniformly integrability...*

$$\lim_{\beta\to\infty} \limsup_{n\to\infty} \int_{|u|\geq\beta} |u| f_n(u) \mathrm{d}u = 0 \text{ then } \lim_{n\to\infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi].$$

Be careful, mistakes are easy to make!

Need: $\{\xi_n \to_d \xi\} \Rightarrow \{\mathbb{E}[\xi_n] \to \mathbb{E}[\xi]\}$

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 $\int_{\mathbb{R}} g(u)\, dF_n(u) \to \int_{\mathbb{R}} g(u)\, dF(u)$

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

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

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

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

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

OUTLINE

1st Contribution Divergences	Nonparametric divergence estimation
2nd Contribution ML on distributions	classification, regression, clustering, anomaly detection, low-dim embedding, …
3rd Contribution Applications	computer vision, astronomy, turbulence data
4th Contribution Theory	Bound on prediction risk, rates

Machine Learning on Distributions

Many ML algorithms only require

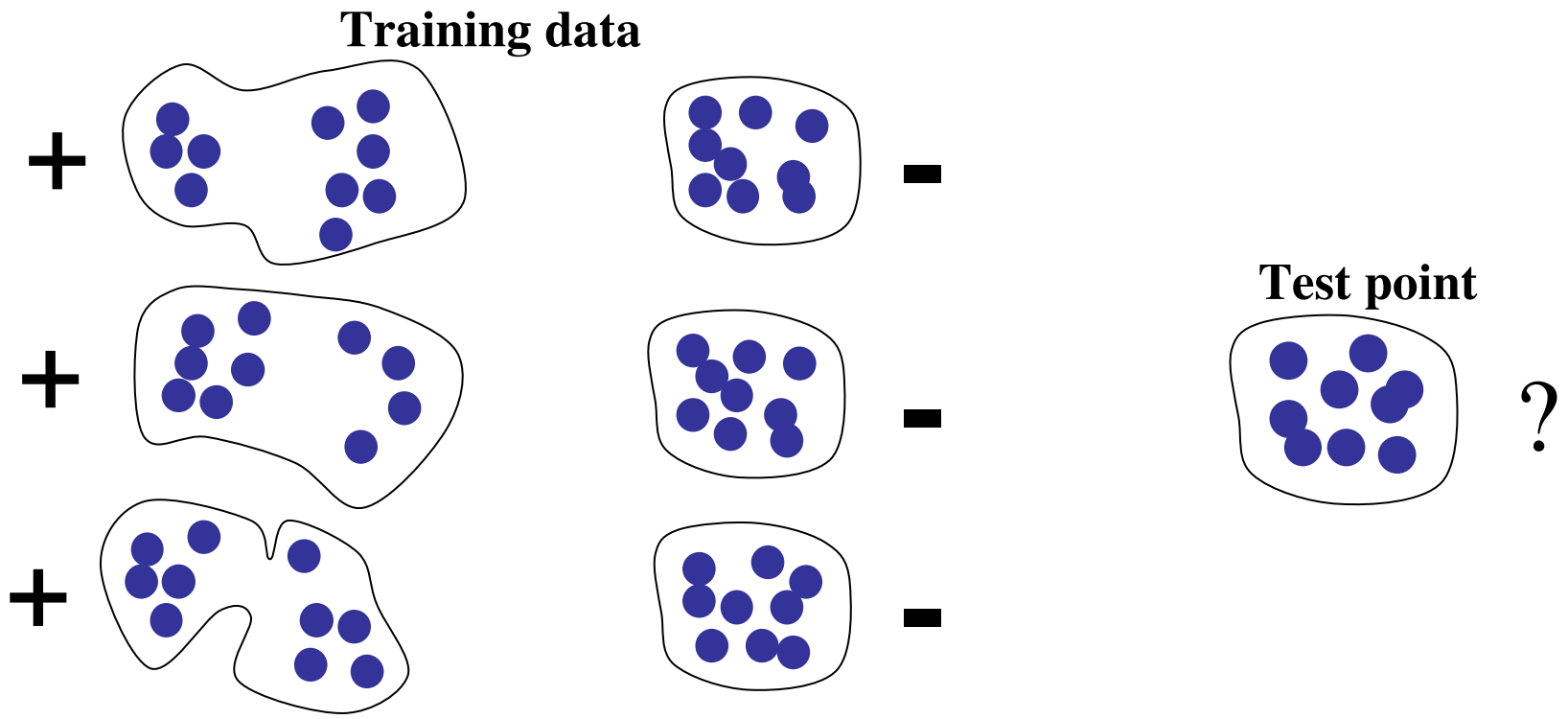
- the pairwise **distances** between the inputs
- the **inner products** between the inputs

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

Distribution Classification

Goal: Classify the following distributions



Differences compared to standard methods on vectors

- The inputs are distributions (not vectors)
- We don't even know these distributions, only sample sets are available

Support Distribution Machine

We have T sample sets, $(\mathbf{X}_1, \ldots, \mathbf{X}_T)$. [**Training data**]
$\{X_{t,1}, \ldots, X_{t,m_t}\} = \mathbf{X_t} \sim p_t$. $\mathbf{X_t}$ has class $Y_t \in \{-1, +1\}$.

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

Solution: Use RKHS based SVM! $\Rightarrow$ SDM

Calculate the Gram matrix $K_{ij} \doteq \langle \phi(p_i), \phi(p_j) \rangle_{\mathcal{K}} = K(p_i, p_j)$

Dual form of SVM:

$$\hat{\alpha} = \arg\max_{\alpha \in \mathbb{R}^T} \sum_{i=1}^{T} \alpha_i - \frac{1}{2} \sum_{i,j}^{T} \alpha_i \alpha_j y_i y_j K_{ij}, \quad \text{subject to } \textstyle\sum_i \alpha_i y_i = 0, \; 0 \leq \alpha_i \leq C.$$

$$Y = \text{sign}(\sum_{i=1}^{T} \hat{\alpha}_i y_i K(p_i, p)) \in \{-1, +1\}$$

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

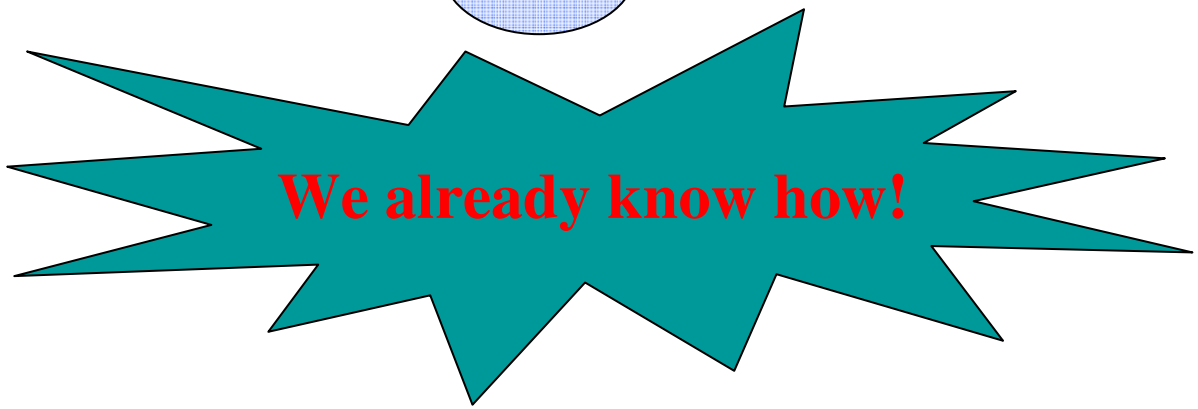
Kernel Estimation

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

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

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

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



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

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

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

Other approaches: set kernels, MMD, Mean map kernel [Smola et al 2007], representer theorem on distributions, support measure machines [Muandet et al 2012],…

OUTLINE

1st Contribution Divergences	Nonparametric divergence estimation
2nd Contribution ML on distributions	classification, regression, clustering, anomaly detection, low-dim embedding, …
3rd Contribution Applications	computer vision, astronomy, turbulence data
4th Contribution Theory	Bound on prediction risk, rates

Image Representation with Distributions

Dealing with complex objects

- ❑ break into smaller parts,
- ❑ represent the object as a sample set of these parts

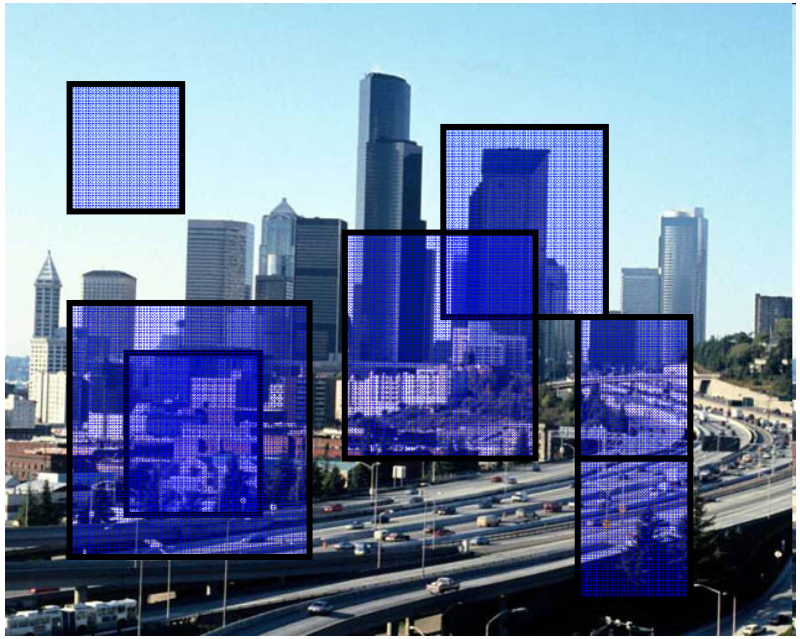


Image patches
- •Overlapping
- •Non-overlapping

Patch locations
- •Grid points
- •Interesting points
- •Random

Patch sizes
- •Same
- •Different,
- •Hierarchy

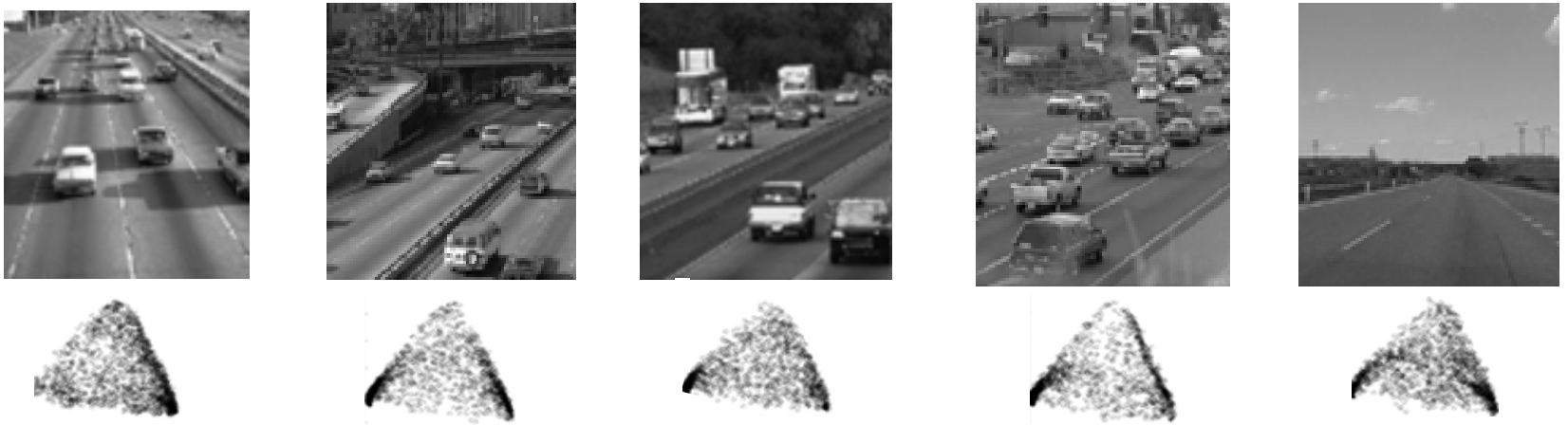
d-dimensional sample set representation of the image

- ❑ Each image ***patch*** is represented by PCA compressed SIFT vectors.
  SIFT = *Scale-invariant feature transform.* ***PCA:*** *128dim⇒ d 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***.

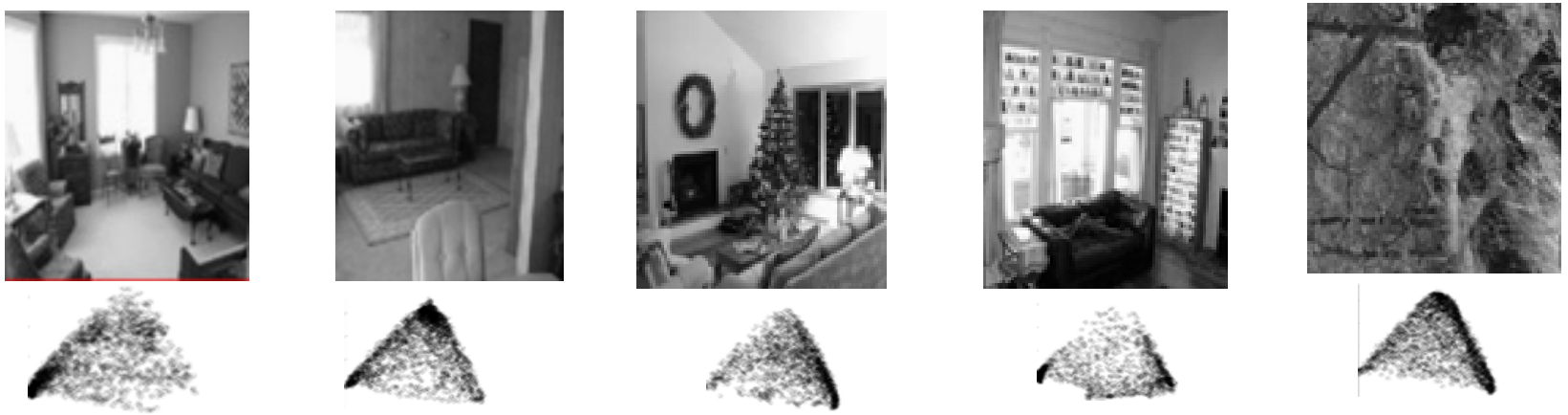
Detecting Anomalous Images

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

50 highway images



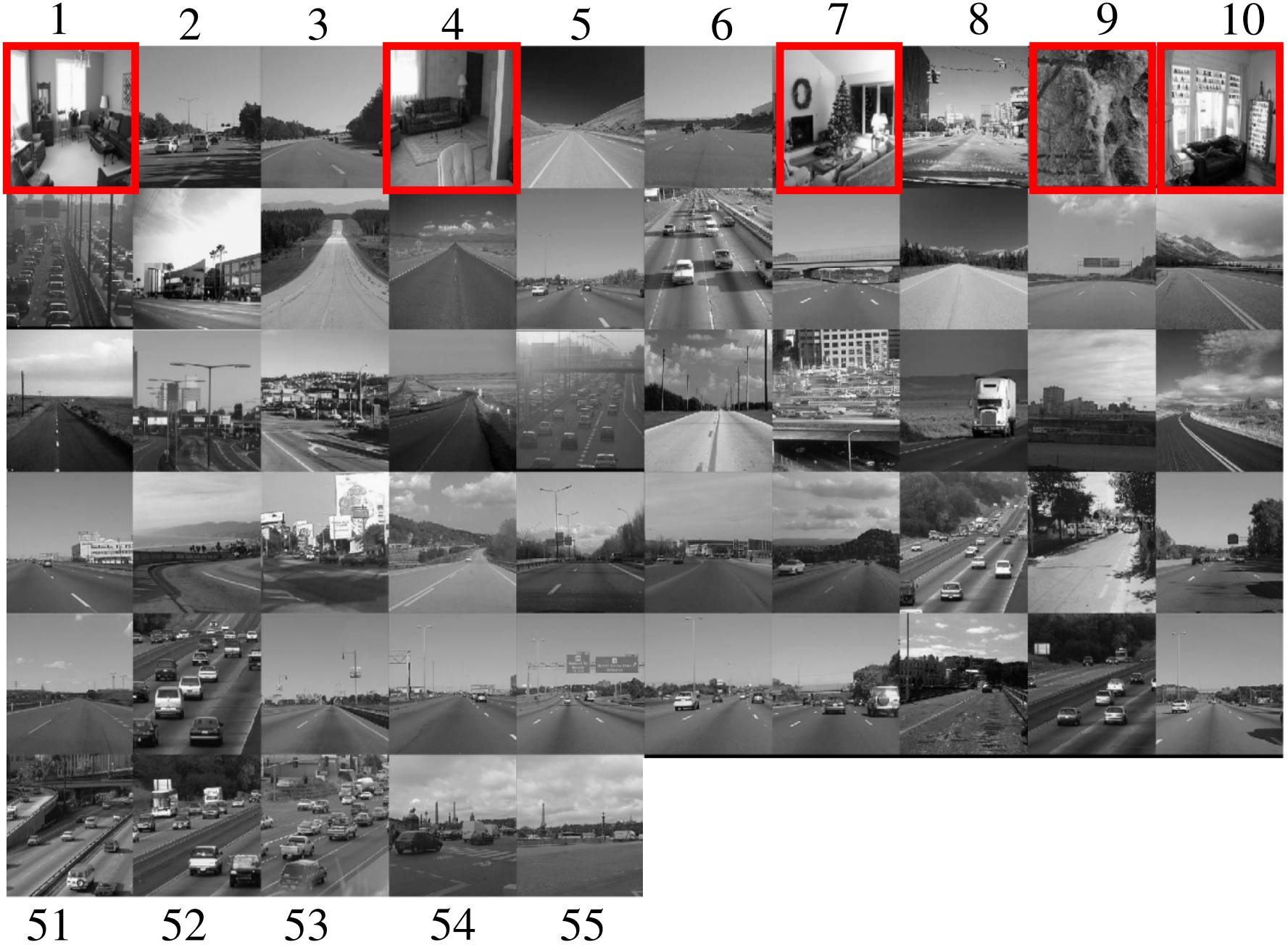
5 anomalies



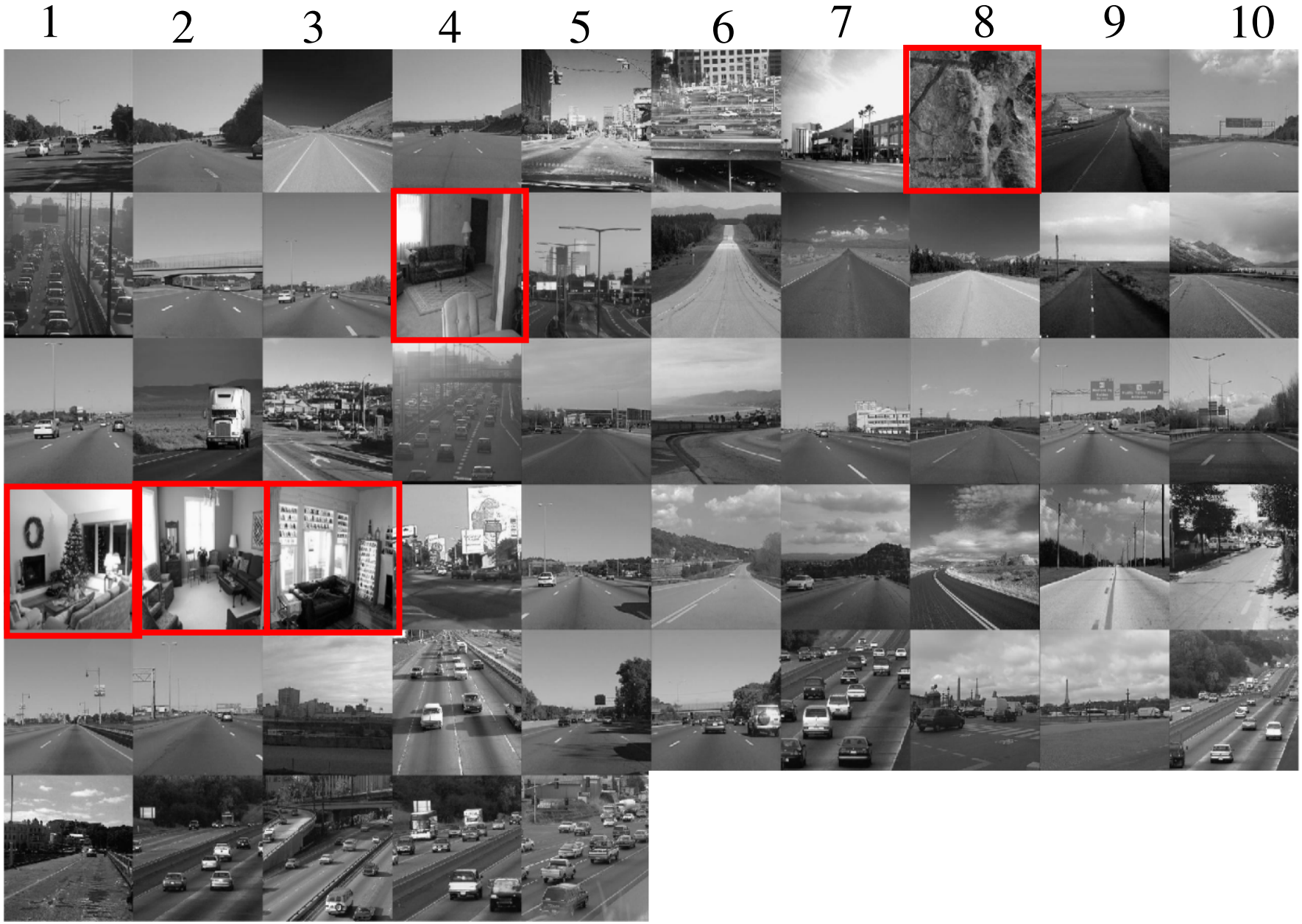
2-dimensional sample set representation of images (128 dim SIFT $\Rightarrow$ 2 dim)

Anomaly score: divergences between the distributions of these sample sets

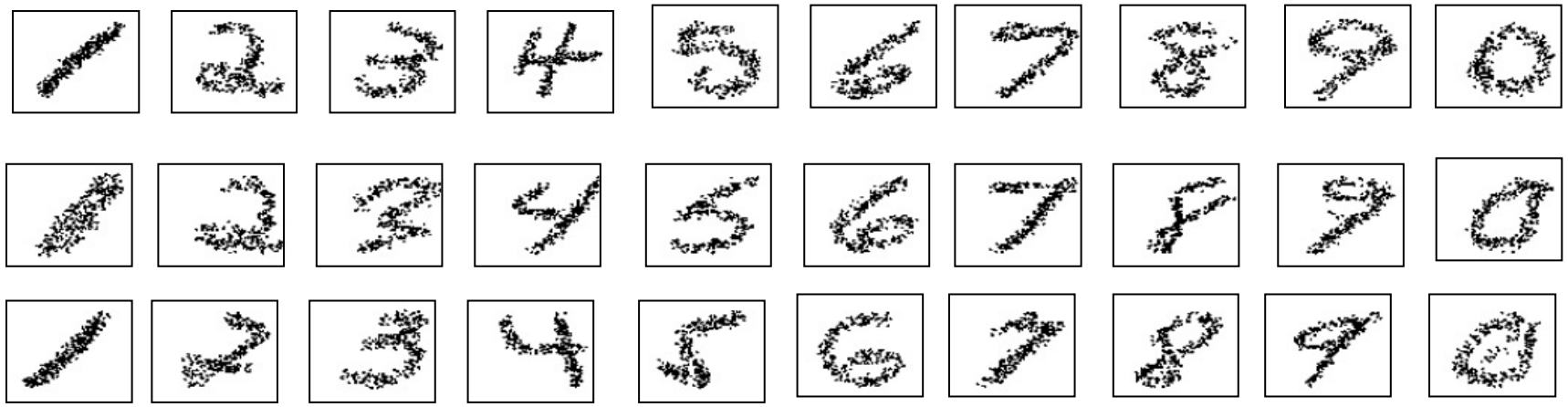
Detecting Anomalous Images



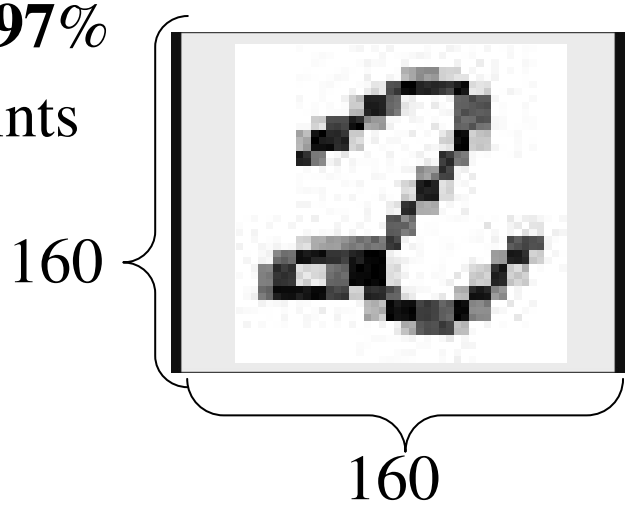
GMM-5 Density Approximation



Noisy USPS Dataset Classification



- 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

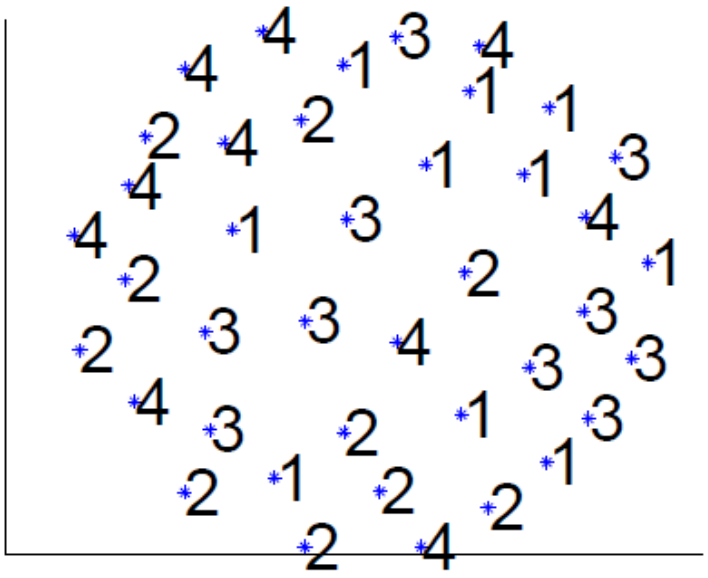
SDM on the 2D distributions, Rényi divergence: **96.0 ±.3%** accuracy

Multidimensional Scaling of USPS Data

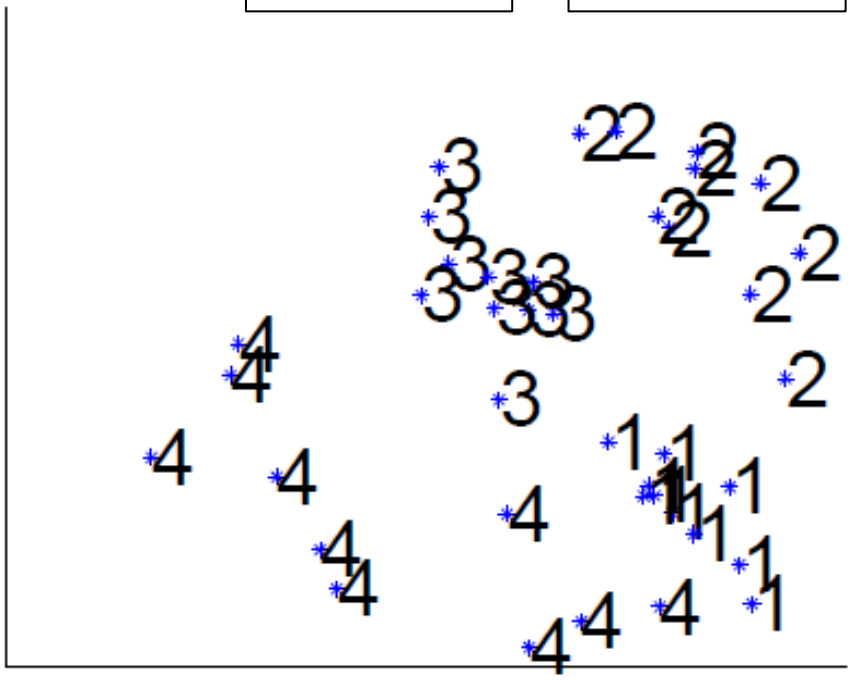
10 instances from figures 1,2,3,4.

Calculate pairwise Euclidean distances.

Nonlinear embedding with MDS into 2d.



Raw **images**
using Euclidean distance

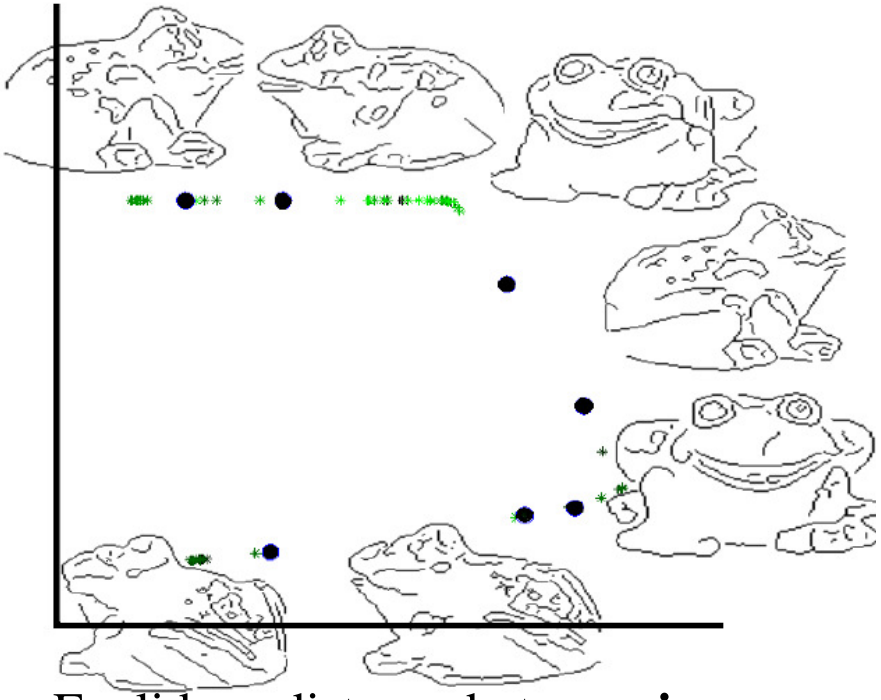


Estimated Euclidean distance
between the **distributions**

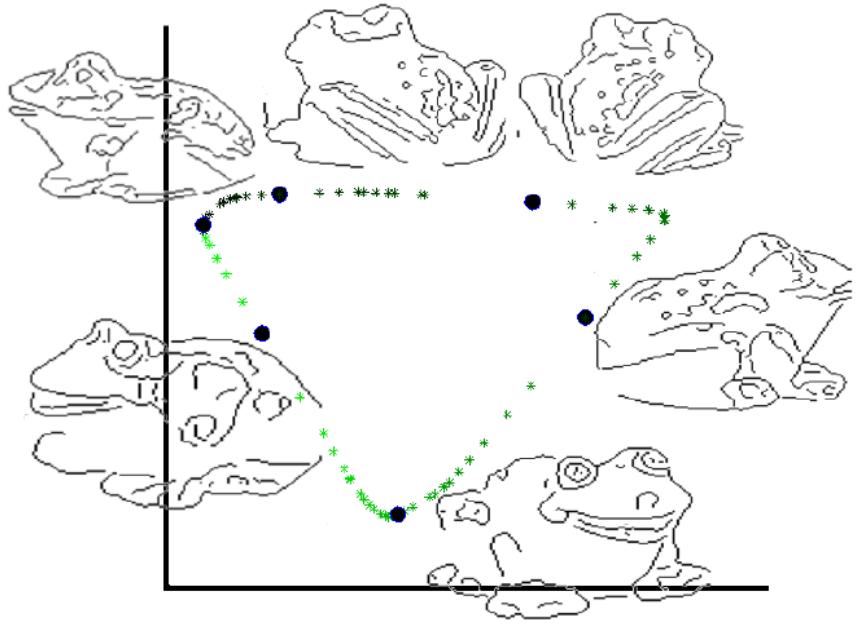
Local Linear Embedding of Distributions

72 rotated COIL froggies

Edge detected COIL froggy



Euclidean distance between **images**

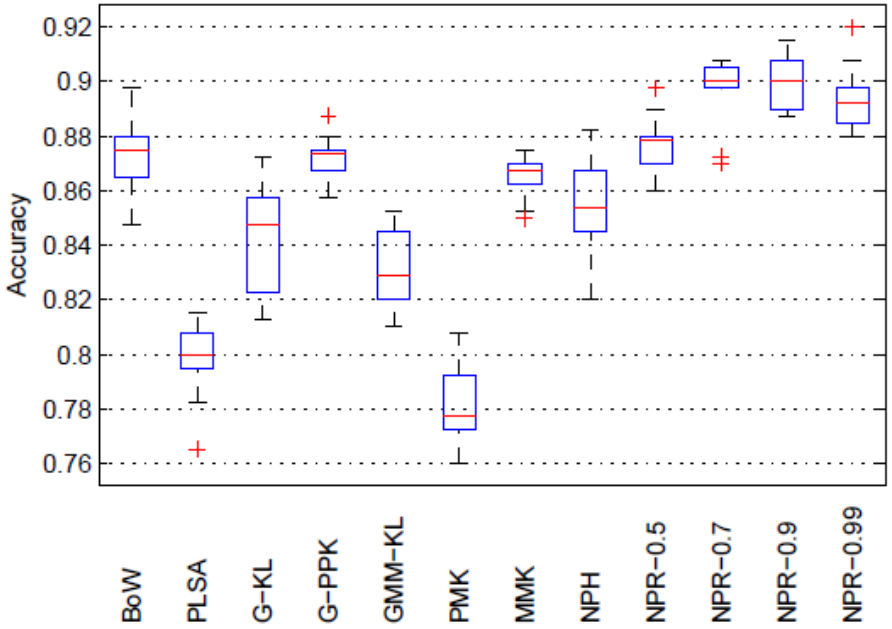


Euclidean distance between **distributions**

Object Classification

ETH-80 [Leibe and Schiele, 2003]

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



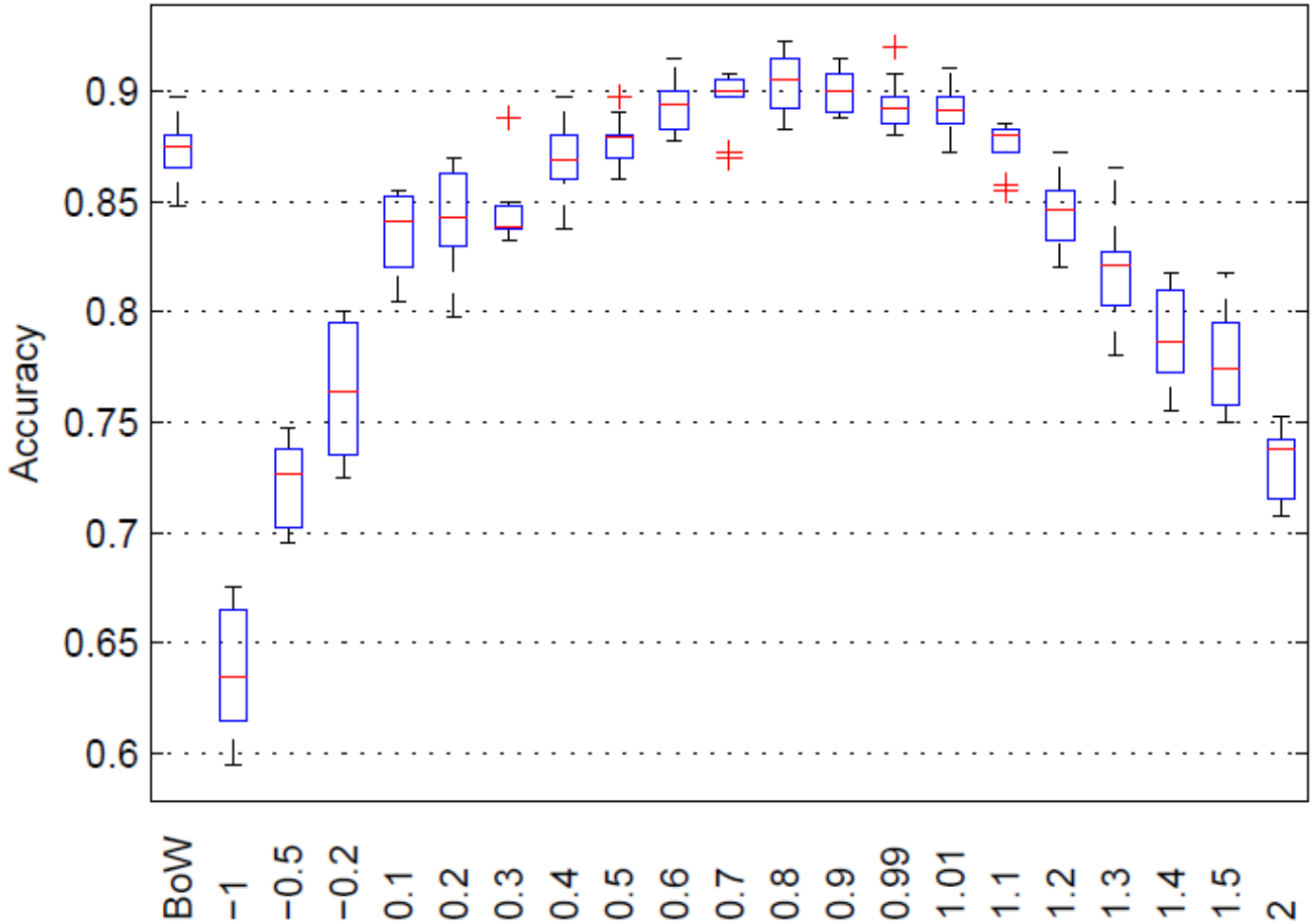
- BoW: **87.33%**
- NPR: **89.93%**

2-fold CV,10 runs

Póczos, Xiong, Sutherland, & Schneider, *CVPR 2012*

Object Classification

ETH-80 [Leibe and Schiele, 2003]



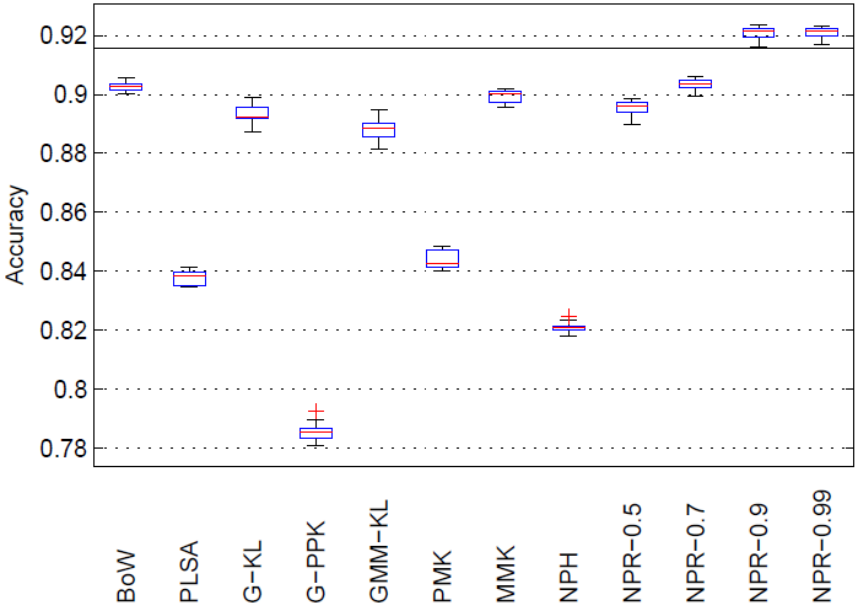
Classification accuracies on ETH-80 with Renyi-αs for twenty αs, as well as BoW.

Outdoor Scenes Classification
[Oliva and Torralba, 2001]

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

- Best published: **91.57%** (Qin and Yung, ICMV 2010)
- NPR: **92.11%**

10 fold CV, 10 runs



Accuracies on the OT dataset; the horizontal line shows the best previously reported result

Póczos, Xiong, Sutherland, & Schneider, *CVPR 2012*

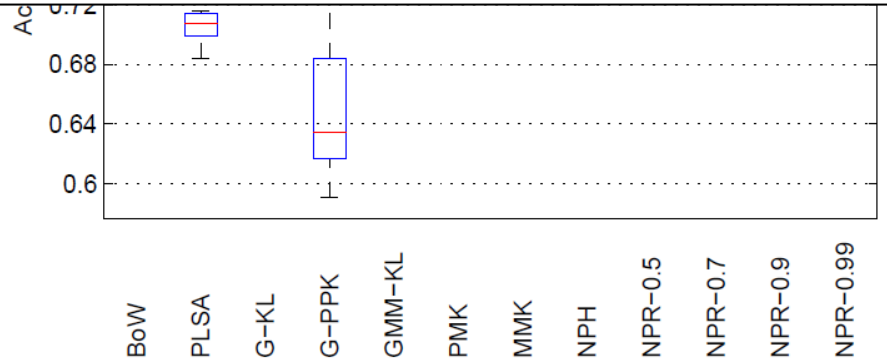
Sport Events Classification
[Li and Fei Fei, 2007]

Computer Vision:

- **Goal**: best results
- Sophisticated feature construction, complex algorithms, heuristics

Distribution based approach:

- simple, easy to implement
- standard SIFT features
- 3 datasets, 3 best performances



10 fold CV, 10 runs

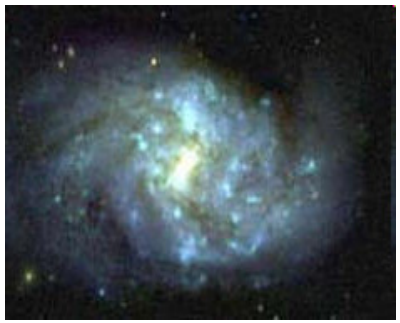
(Zhang et al, CVPR 2011)

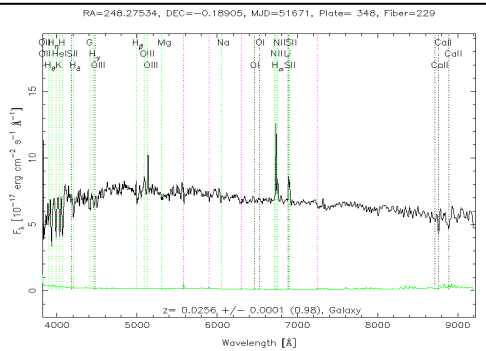
- NPR: **87.18%**

Finding Unusual Galaxy Clusters

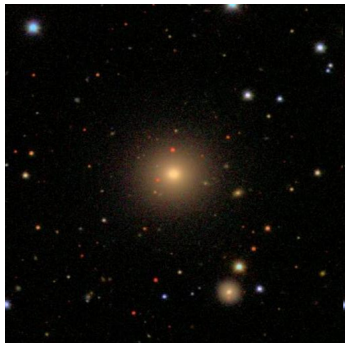
Sloan Digital Sky Survey (SDSS)

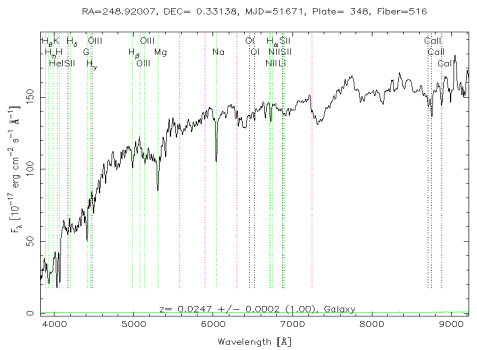
- ❑ continuum spectrum⇒ 2d
- ❑505 galaxy clusters
  (10-50 galaxies in each)
- ❑7530 galaxies





Blue galaxy





Red galaxy

What are the most anomalous galaxy clusters?

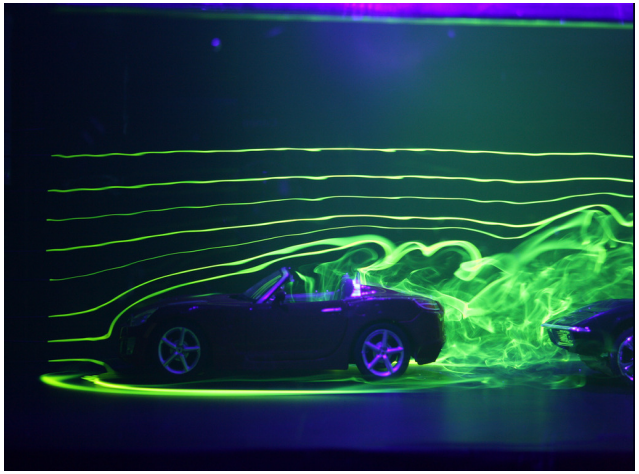
The most anomalous galaxy cluster contains mostly

- ❑ star forming blue galaxies
- ❑ irregular galaxies

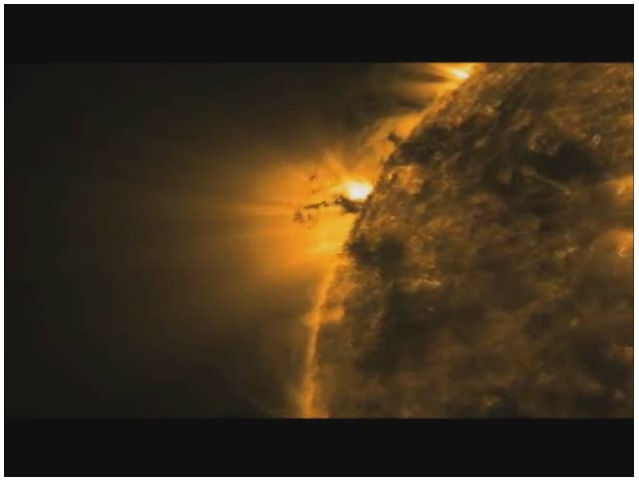
B. Póczos, L. Xiong & J. Schneider, UAI, 2011. Credits: ESA, NASA

Understanding Turbulences

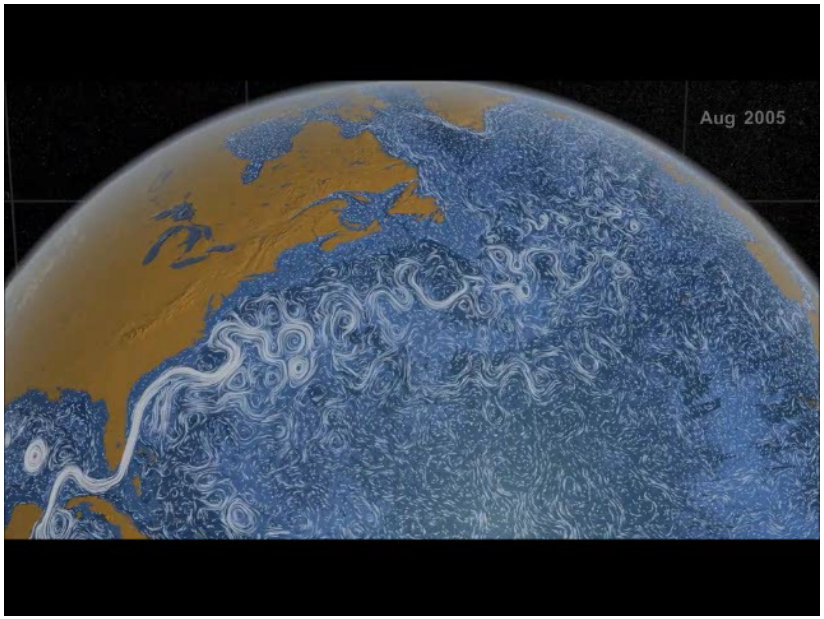
More economical cars



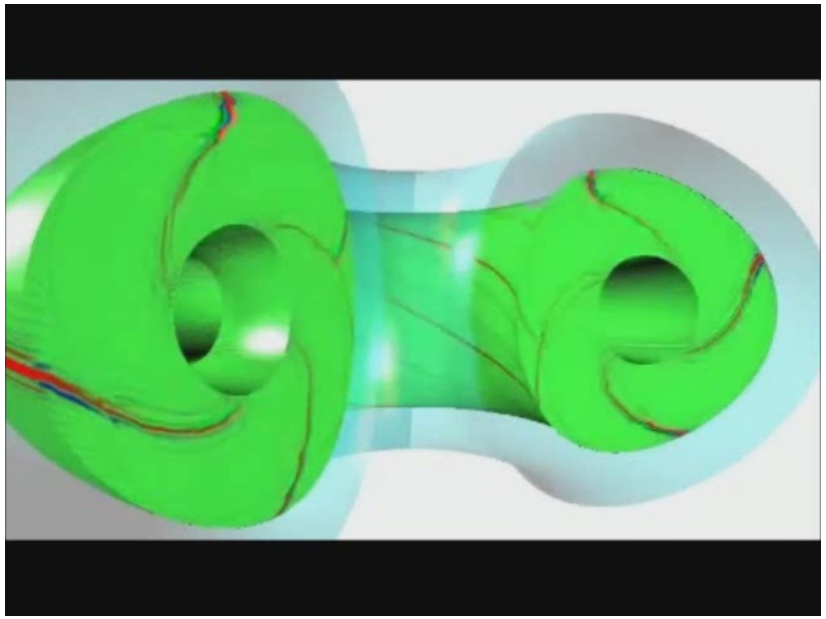
Magnetic storms, solar winds



Safer, faster airplanes



Ocean currents



Fusion power plants:
keep the plasma togerther

Credits: ESA, NASA, PPPL, Wikipedia

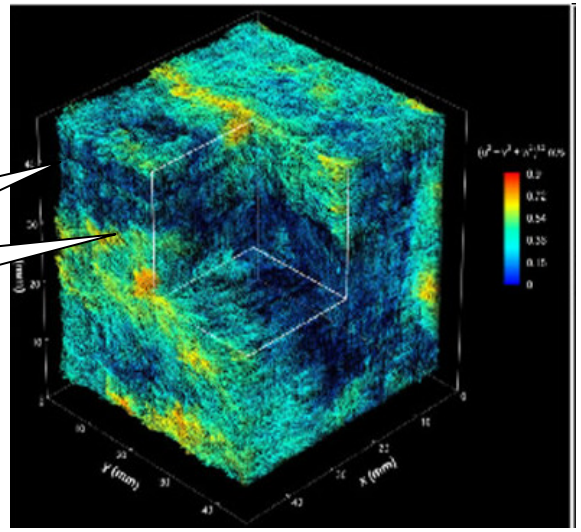
Turbulence Data Classification

Simulated fluid flow through time

(JHU Turbulence Research Group)

Goal: find interesting events, patterns, phenomena

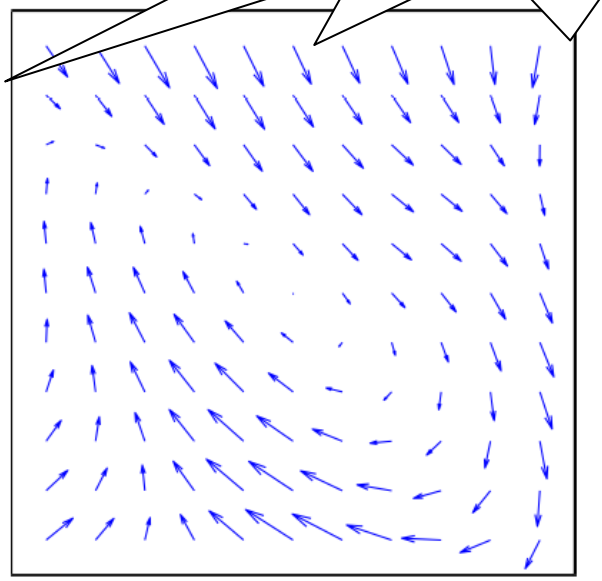
find vortices



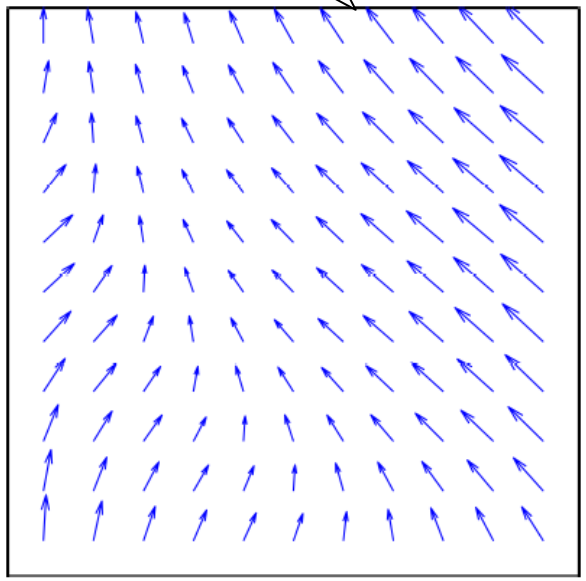
Something interesting happened?

- 11 positive, 2[illegible]
- **Results:** Leave one out cross-val : 97%

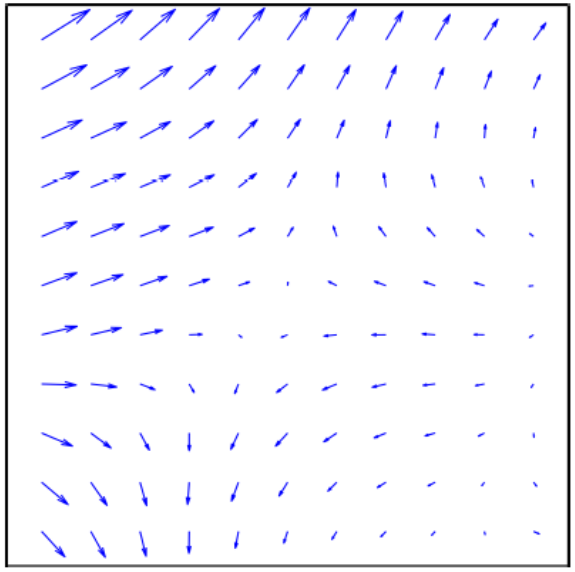
[illegible]



Positive (vortex)

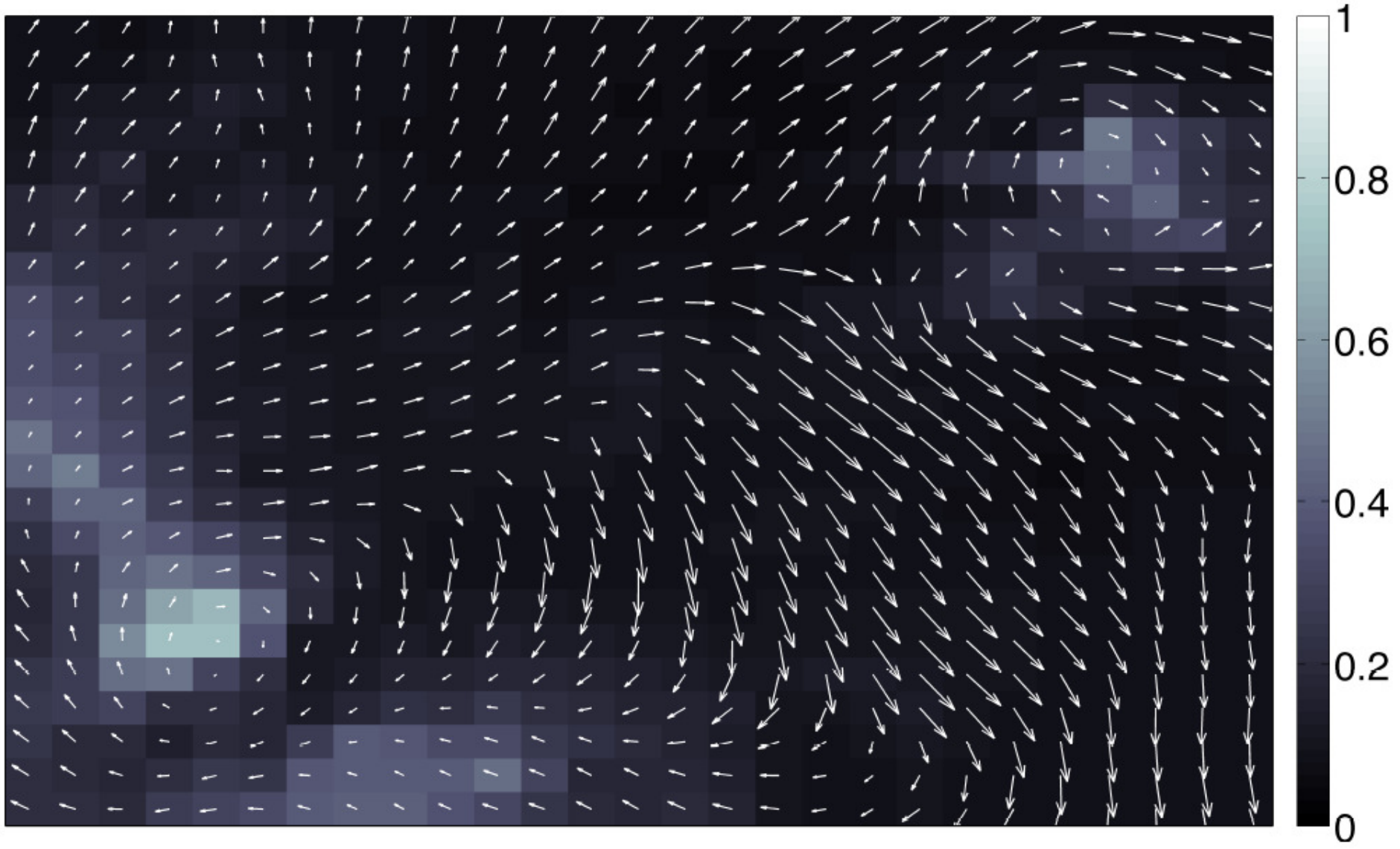


Negative



Negative

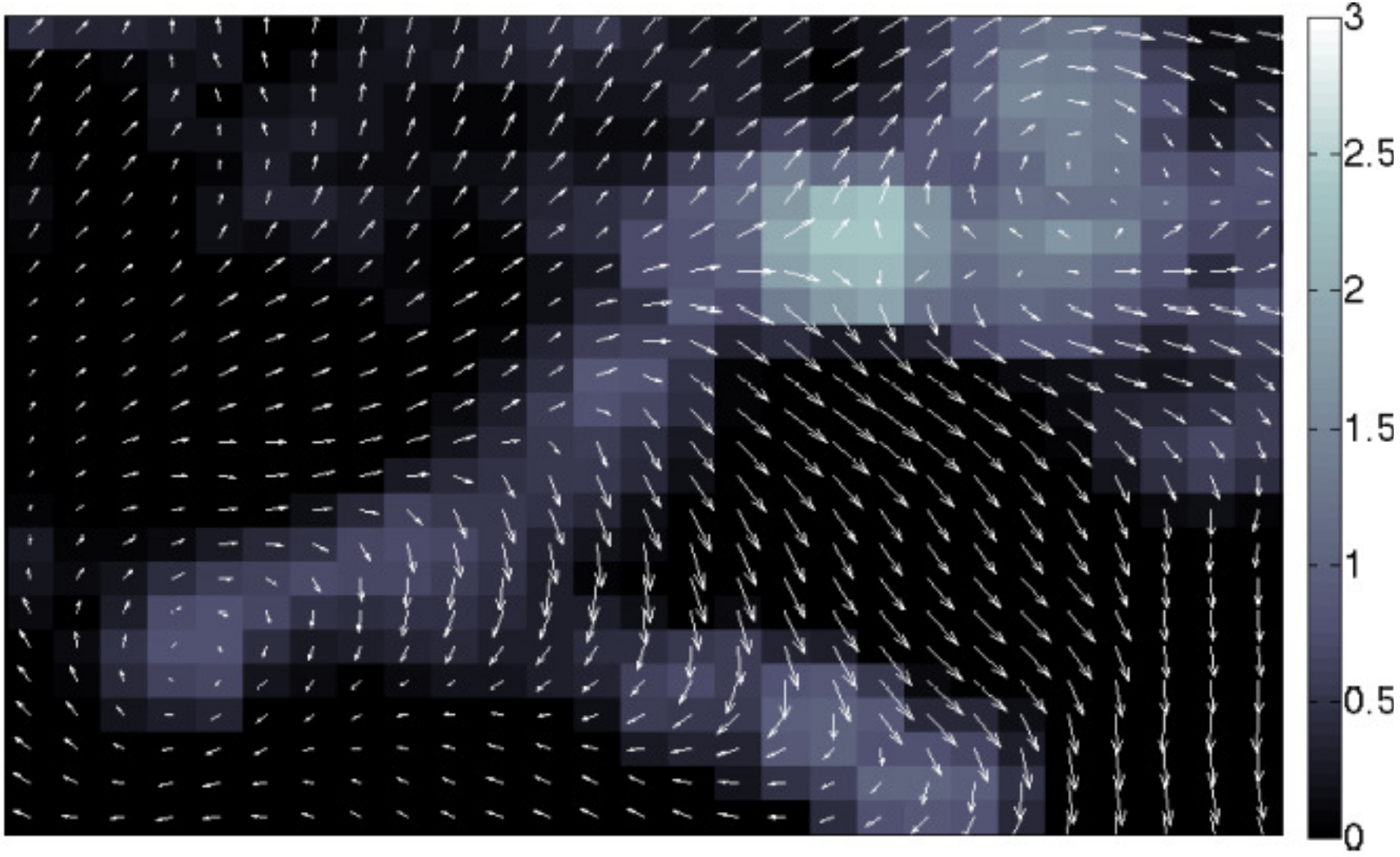
Finding Vortices



Classification probabilities

Find Interesting Phenomena in Turbulence Data

Anomaly detection with 1-class SDM



Anomaly scores

OUTLINE

1st Contribution Divergences	Nonparametric divergence estimation
2nd Contribution ML on distributions	classification, regression, clustering, anomaly detection, low-dim embedding, …
3rd Contribution Applications	computer vision, astronomy, turbulence data
4th Contribution Theory	Bound on prediction risk, rates

So far, we got good experimental results in applications

Theory???

Distribution regression

Standard, finite dimensional regression
Nadaraya-Watson, 1964

Model: $Y = f(X) + \mu,\ X \in \mathbb{R}^d,\ Y \in \mathbb{R}$

$\mathbb{E}[\mu] = 0,\ X, \mu$ are independent.

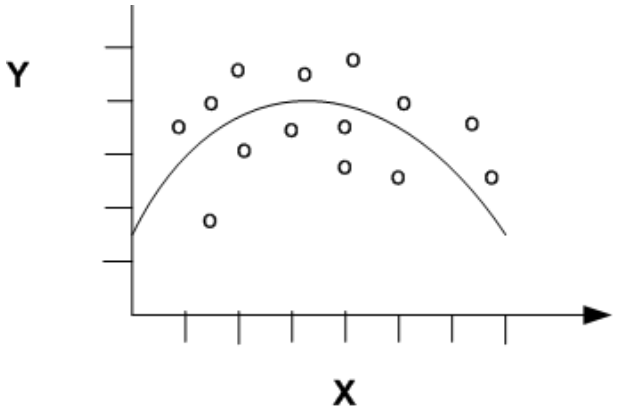
Training points:

$(X_1, Y_1),\ (X_2, Y_2),\ \ldots (X_m, Y_m)$

$X_i \sim P$, i.i.d., $Y_i \sim Q(\cdot|X_i)$, where $X_i \in \mathbb{R}^d,\ Y \in \mathbb{R}$

Test point: $x \in \mathbb{R}^d$

Goal: Estimate the $f(x) = \mathbb{E}[Y|X = x]$ function.



Data

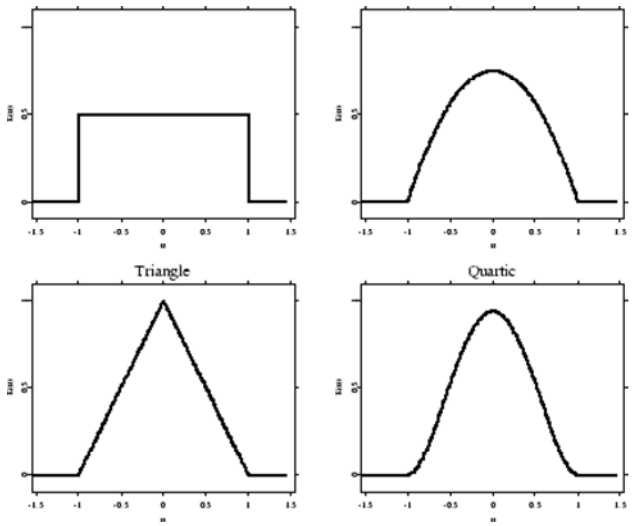
Kernel regression:

$$\widehat{Y}(x) = \frac{\sum_{i=1}^{m} Y_i K(\frac{x - X_i}{h_m})}{\sum_{j=1}^{m} K(\frac{x - X_j}{h_m})}$$

Consistency theorem of the estimator:

When $h_m \to 0,\ m h_m^d \to \infty,$
+ some other conditions

$\Rightarrow$ consistent



Kernel functions: *K(x)*

Distribution Regression Problem Definition

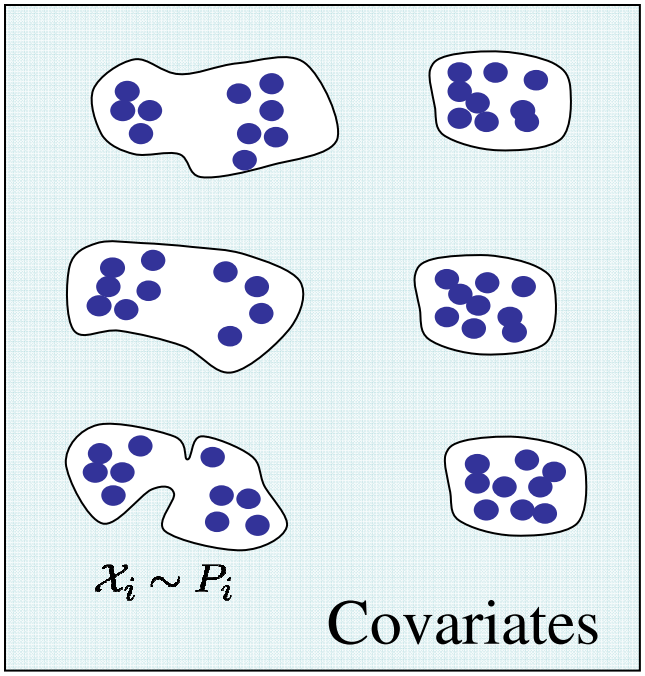
Model:

(P_1, Y_1), (P_2, Y_2), $\ldots$ (P_m, Y_m), $P_i \sim \mathcal{P}$, $Y_i \sim Q(\cdot|P_i)$ i.i.d.,
where P_i is a distribution on $\mathcal{K} \subset \mathbb{R}^k$, $Y \in \mathbb{R}$

We do not observe P_i directly!
We observe a sample from P_i.

$$\mathcal{X}_i = X_{i1}, \ldots, X_{in_i} \sim P_i$$

$$Y_i = f(P_i) + \mu_i \qquad \mathcal{X}_i \sim P_i \sim \mathcal{P}$$



Goal: Estimate $\mathbb{E}[Y|P]$ using the samples $\mathcal{X}, (\mathcal{X}_1, Y_1), \ldots, (\mathcal{X}_m, Y_m)$.

Difficulties

Error in variables $P_1, \ldots, P_m$.
Dimension of a distribution is ∞, but we need $h_m \to 0$, $mh_m^d \to \infty$

The kernel-kernel estimator

Regression function estimator

$$\widehat{f}(\mathcal{X}) = \widehat{f}(\mathcal{X}; \mathcal{X}_1, \ldots, \mathcal{X}_m) = \begin{cases} \dfrac{\sum_i Y_i K\left(\frac{D(\mathcal{X}_i,\mathcal{X})}{h}\right)}{\sum_i K\left(\frac{D(\mathcal{X}_i,\mathcal{X})}{h}\right)} & \text{if } \sum_i K\left(\frac{D(\mathcal{X}_i,\mathcal{X})}{h}\right) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

where $D(\mathcal{X}_i, \mathcal{X}_j) = D(\widehat{P}_i, \widehat{P}_j) = \int |\widehat{p}_i(x) - \widehat{p}_j(x)|\, dx.$

Kernel density estimator

$$\widehat{p}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{b_i^k} B\left(\frac{\|x - X_{ij}\|}{b_i}\right)$$

with kernel B and bandwidth b_i
$\|\cdot\|$ is the Euclidean distance, $k = dim(x)$

Goal: Bound $R(m, n_1, \ldots, n_m) = \mathbb{E}|\widehat{f}(\widehat{P}) - f(P)|.$

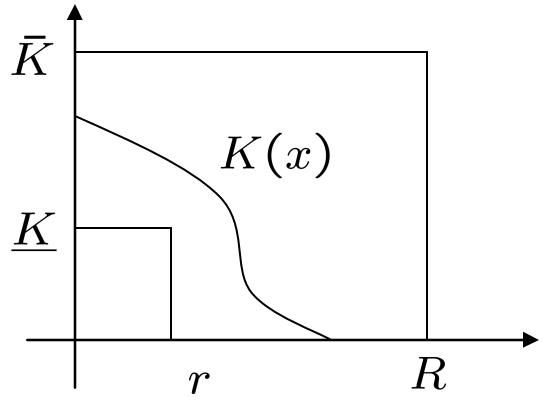
Assumptions

(A1) **Hölder continuous functional.**

$$f \in \left\{ f : \ |f(P_i) - f(P_j)| \leq L\, D(P_i, P_j)^{\beta} \right\} \qquad L > 0 \text{ and } 0 < \beta \leq 1$$

(A2) **Asymmetric boxed and Lipschitz kernel.**

$$\underline{K} I_{\{x \in \mathcal{B}(0,r)\}} \leq K(x) \leq \bar{K} I_{\{x \in \mathcal{B}(0,R)\}} \ \forall x > 0$$



(A3) **Lipschitz class of distributions** P, P_i

(A4) **Bounded regression.**

$\sup_{P \in \mathcal{P}} |f(P)| < f_{\max} < \infty \quad \mathbb{P}(|Y_i| \leq B_Y) = 1$ for some $B_Y < \infty$.

(A5) **Lower bound on** $\min_{1 \leq i \leq m+1} n_i$

$n = \min_{1 \leq i \leq m+1} n_i$. We assume that $e^{n^{\frac{k}{2+k}}}/m \to \infty$ as $m \to \infty$.

(A6) **Relationship between *n* and *h***

Assume that $C_* n^{-\frac{1}{2+k}} \leq rh/4$ where C_* is a constant.

Bounding the risk

$$R(m,n) = \mathbb{E}\left[|\hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - f(P)|\right]$$

Let $\mathcal{B}(P,h) = \{\tilde{P} \in \mathbb{D} : D(\tilde{P}, P) \leq h\}$.

The L_1 ball of distributions around P with radius h.

Small ball probabilities $\Phi_P(h) \doteq \mathbb{P}(P_1 \in \mathcal{B}(P,h)|P)$ is a function of P.

Theorem

Suppose that the assumptions stated above hold.

Let $b = n^{-\frac{1}{2+k}}$ be the bandwidth in the density estimators $\widehat{p}_i$.

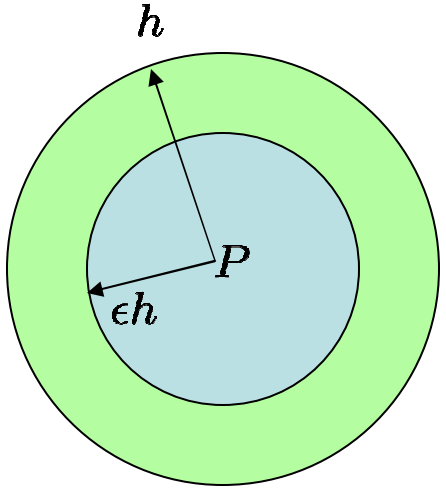
$$R(m,n) \leq \frac{1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] C_1 n^{-\frac{1}{2+k}} + C_2 h^{\beta} + C_3\sqrt{\frac{1}{m}}\sqrt{\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]}$$
$$+ \frac{C_4}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] + (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}}.$$

Doubling Measure

Definition [Kpotufe, 2012]

$\mathcal{P}$ is a doubling measure with effective dimension d if, for every P, $h > 0$, and $0 < \epsilon < 1$,

$$\frac{\phi_P(h)}{\phi_P(\epsilon h)} = \frac{\mathcal{P}(\mathcal{B}(P,h))}{\mathcal{P}(\mathcal{B}(P,\epsilon h))} \leq \left(\frac{c}{\epsilon}\right)^d.$$



Remark

If d denotes the doubling dimension of measure $\mathcal{P}$, then $\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]$ can be upper bounded as follows:

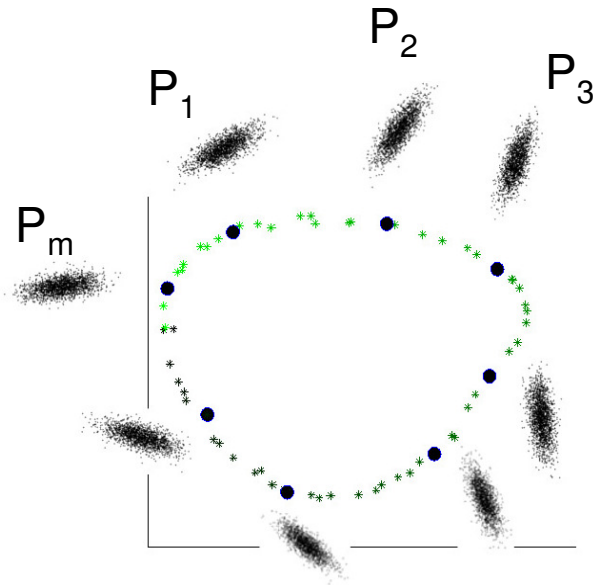
$$\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] = \mathbb{E}\left[\frac{\Phi_P(1)}{\Phi_P(rh/2)}\frac{1}{\Phi_P(1)}\right] \leq C(rh/2)^{-d}\mathbb{E}\left[\frac{1}{\Phi_P(1)}\right] \leq \frac{C}{h^d}.$$

Theorem

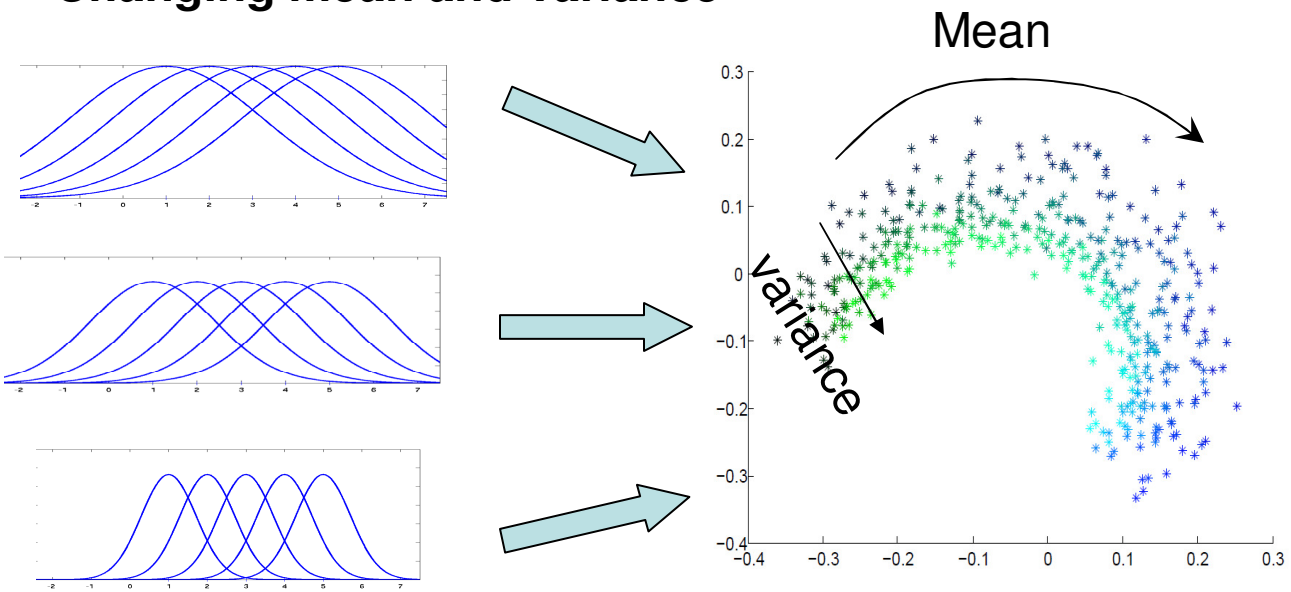
$$R(m,n) \leq \frac{C_1}{h^{d+1}n^{1/(k+2)}} + C_2h^\beta + C_3\sqrt{\frac{1}{mh^d}}$$

Examples for finite doubling dimension

Rotated 2D Gaussian distributions



1D Gaussian distributions;
Changing mean and variance



Rates

$$R(m,n) \le \frac{C_1}{h^{d+1} n^{1/(k+2)}} + C_2 h^{\beta} + C_3 \sqrt{\frac{1}{mh^d}}$$

First case

$$\sqrt{\frac{1}{mh^d}} = \Omega\left(\frac{C_1}{h^{d+1} n^{1/(k+2)}}\right) \Rightarrow \begin{cases} R(m,n) = O\left(m^{-\beta/(2\beta+d)}\right) \\ h = \Theta\left(m^{-\frac{1}{2\beta+d}}\right) = \Omega\left(n^{-\frac{1}{(k+2)(\beta+d+1)}}\right) = \Omega\left(n^{-\frac{1}{k+2}}\right) \\ n = \Omega\left(m^{\frac{\beta+d+1}{2\beta+d}(k+2)}\right) \\ \text{n is large compared to m} \end{cases}$$

Second case

$$\sqrt{\frac{1}{mh^d}} = O\left(\frac{1}{h^{d+1} n^{1/(k+2)}}\right) \Rightarrow \begin{cases} R(m,n) = O\left(\frac{1}{h^{d+1} n^{1/(k+2)}} + h^{\beta}\right) \\ h = \Theta\left(n^{-\frac{1}{(k+2)(\beta+d+1)}}\right) = \Omega\left(n^{-\frac{1}{k+2}}\right) \\ m = \Omega\left(n^{\frac{2\beta+d}{(k+2)(\beta+d+1)}}\right) \\ \text{m is large compared to n} \end{cases}$$

Proofs

1st Step

$$R(m,n) = \mathbb{E}|\hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - f(P)|$$
$$\leq \mathbb{E}|\hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) - \hat{f}(P; P_1, \ldots, P_m)| + \mathbb{E}|\hat{f}(P; P_1, \ldots, P_m) - f(P)|$$

2nd Step

Finish the proof! ☺

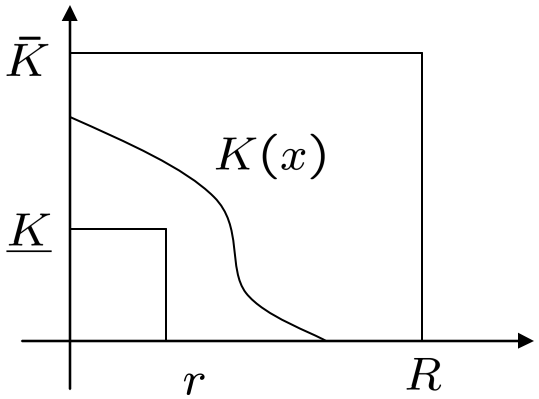
Similar rates using knn regression when d>2.

(…curse of low-dimensionality…)

Why small ball probabilities?

$$K_i = K\left(\frac{D(\hat{P}_i, \hat{P})}{h}\right)$$

$$\hat{f}(\hat{P}) = \hat{f}(\hat{P}; \hat{P}_1, \ldots, \hat{P}_m) = \begin{cases} \frac{\sum_i Y_i K_i}{\sum_i K_i} & \text{if } \sum_i K_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$



Lemma

$$\mathbb{P}(\sum_{i=1}^{m} K_i = 0) \leq \mathbb{P}(\sum_{i=1}^{m} K_i < \underline{K}) \leq = \frac{1}{em}\mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]$$

Proof [based on Györfi et al, 2002]

$$\mathbb{P}\Big(\sum_{i=1}^{m} K_i < \underline{K}\Big) \leq \mathbb{P}\left(\sum_{i=1}^{m} I_{\{D(P_i,P)\geq rh\}} = 0\right)$$

$$= \mathbb{E}[\mathbb{P}(\sum_{i=1}^{m} I_{\{D(P_i,P)\geq rh\}} = 0 \,|P)]$$

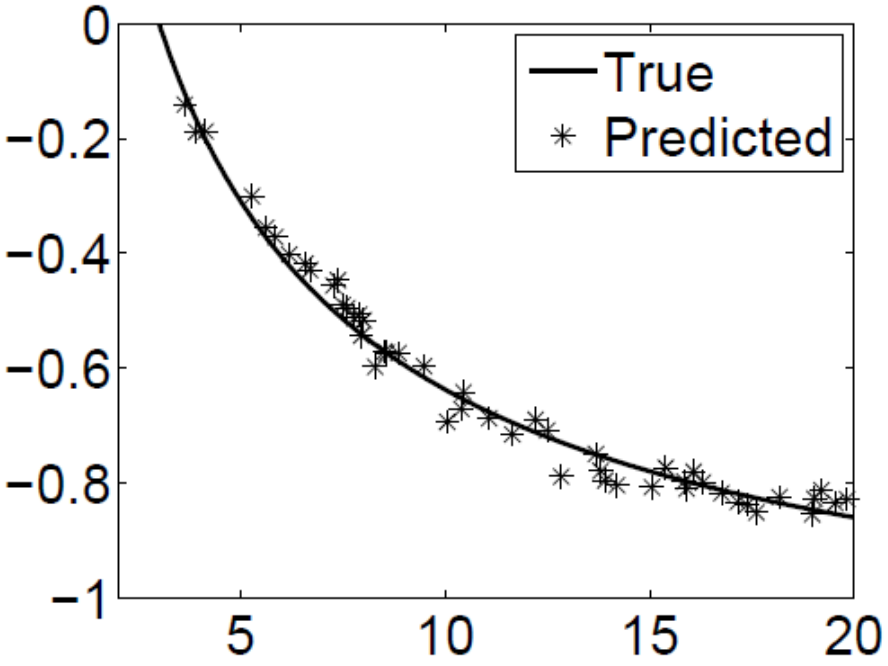
$$= \int \mathbb{P}\left(\sum_{i=1}^{m} I_{\{D(P_i,P)\geq rh\}} = 0 \,\middle|\, P\right) d\mathcal{P}(P)$$

$$= \int \prod_{i=1}^{m} [1 - \mathcal{P}(P_i \in \mathcal{B}(P, rh)|P)] d\mathcal{P}(P)$$

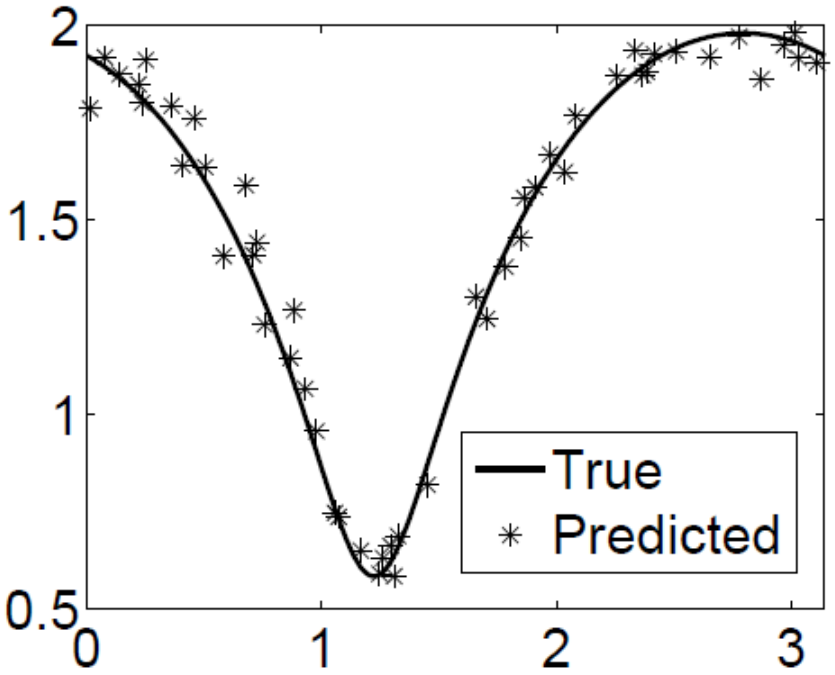
Proof continued...

$$
\begin{aligned}
&= \int \prod_{i=1}^{m} [1 - \mathcal{P}(P_i \in \mathcal{B}(P, rh)|P)] d\mathcal{P}(P) \\
&= \int [1 - \mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)]^m d\mathcal{P}(P) \\
&\leq \int \exp[-m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)] d\mathcal{P}(P) \\
&= \int \exp[-m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)] \frac{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} d\mathcal{P}(P) \\
&\leq \max_{u>0} u \exp(-u) \int \frac{d\mathcal{P}(P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} \\
&\leq \frac{1}{e} \int \frac{d\mathcal{P}(P)}{m\mathcal{P}(P_1 \in \mathcal{B}(P, rh)|P)} \\
&= \frac{1}{em} \mathbb{E}\left[\frac{1}{\Phi_P(rh)}\right]
\end{aligned}
$$

Distribution Regression



Skewness of Beta



Entropy of Gaussian

Take Me Home!

1st Contribution Divergences	Nonparametric divergence estimation
Applications	
4th Contribution Theory	Bound on prediction risk, rates

Thanks for your attention! ☺