Distribution Regression - the Set Kernel Heuristic is Consistent

Zoltán Szabó (Gatsby Unit)

Joint work with Arthur Gretton (Gatsby Unit), Barnabás Póczos (CMU), Bharath K. Sriperumbudur (University of Cambridge)

CSML Lunch Talk Series
May 2, 2014
Outline

- Motivation, examples.
- Algorithm, consistency result.
- Numerical illustration.
Given: \( \{(x_i, y_i)\}_{i=1}^l \) samples \( \mathcal{H} \ni f =? \) such that \( f(x_i) \approx y_i \).

Typically: \( x_i \in \mathbb{R}^p, \ y_i \in \mathbb{R}^q \).

Our interest: \( x_i \)-s are distributions (\( \infty \)-dimensional objects).
In practise:

- $x_i$-s are only observable via samples: $x_i \approx \{x_{i,n}\}_{n=1}^{N} \Rightarrow$
- an $x_i$ is represented as a *bag*:
  - image = set of patches,
  - document = bag of words,
  - video = collection of images,
  - different configurations of a molecule = bag of shapes.
Given (2 bags):

\[ B_i := \{x_{i,n}\}_{n=1}^{N_i} \sim x_i, \quad (1) \]

\[ B_j := \{x_{j,m}\}_{m=1}^{N_j} \sim x_j. \quad (2) \]

Similarity of the bags (set/multi-instance/ensemble-, convolution kernel; Gärtner’02, Haussler’99):

\[ K(B_i, B_j) = \frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} k(x_{i,n}, x_{j,m}). \quad (3) \]

Many successful applications:

- classification, regression, clustering.
- fundamental question: consistency?
Example: supervised entropy learning

- Entropy of $x \sim f$: $-\int f(u) \log[f(u)] \, du$.
- Training: samples from distributions, entropy values.
- Task: estimate the entropy of a new sample set.
Example: toxic level estimation from tissues

- Toxin alters the properties/causes mutations in cells.
- Training data:
  - bag = tissue,
  - samples in the bag = cells described by some simple features,
  - output label = toxic level.
- Task: predict the toxic level given a new tissue.
Example: aerosol prediction using satellite images

- Aerosol = floating particles in the air; climate research.
- Multispectral satellite images: 1 pixel $= 200 \times 200 m^2 \in \text{bag}$.
- Bag label: ground-based (expensive) sensor.
- Task: satellite image $\rightarrow$ aerosol density.
Towards problem formulation: kernel, RKHS

\( k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R} \) kernel on \( \mathcal{D} \), if

- \( \exists \phi : \mathcal{D} \rightarrow H(\text{hilbert space}) \) feature map,
- \( k(a, b) = \langle \phi(a), \phi(b) \rangle_H \) (\( \forall a, b \in \mathcal{D} \)).

Kernel examples: \( \mathcal{D} = \mathbb{R}^d \) (\( p > 0, \theta > 0 \))

- \( k(a, b) = (\langle a, b \rangle + \theta)^p \): polynomial,
- \( k(a, b) = e^{-\|a-b\|_2^2/(2\theta^2)} \): Gaussian,
- \( k(a, b) = e^{-\theta\|a-b\|_1} \): Laplacian.

In the \( H = H(k) \) RKHS (\( \exists! \)): \( \phi(u) = k(\cdot, u) \).
Some example domains ($\mathcal{D}$), where kernels exist

- Euclidean spaces: $\mathcal{D} = \mathbb{R}^d$.
- Strings, time series, graphs, dynamical systems.
- Distributions.
Given: \((\mathcal{D}, k)\); we saw that \(u \rightarrow \varphi(u) = k(\cdot, u) \in H(k)\).

Let \(x\) be a distribution on \(\mathcal{D} (x \in \mathcal{M}^+_1(\mathcal{D}))\); the previous construction can be extended:

\[
\mu_x = \int_\mathcal{D} k(\cdot, u)\,dx(u) \in H(k).
\]  

(4)

If \(k\) is bounded: \(\mu_x\) is well-defined for any distribution \(x\).
Mean embedding based distribution kernel

Simple estimation of $\mu_x = \int_\mathcal{D} k(\cdot, u) d\mathcal{X}(u)$:

- **Empirical distribution**: having samples $\{x_n\}_{n=1}^N$

$$\hat{x} = \frac{1}{N} \sum_{n=1}^N \delta_{x_n}. \quad (5)$$

- **Mean embedding, inner product – empirically (set kernels!)**: 

$$\mu_{\hat{x}} = \int_\mathcal{D} k(\cdot, u) d\hat{x}(u) = \frac{1}{N} \sum_{n=1}^N k(\cdot, x_n), \quad (6)$$

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \rangle_{H(k)} = \frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} k(x_{i,n}, x_{j,m}).$$
Until now

- If we are given a domain ($\mathcal{D}$) with kernel $k$, then one can easily define/estimate the similarity of distributions on $\mathcal{D}$.
- Prototype example: $\mathcal{D} = \mathbb{R}^d$, $k = \text{Gaussian}$, $K = \text{lin. kernel}$.

The real conditions:

- $\mathcal{D}$: locally compact, Polish. $k$: $c_0$-universal.
- $K$: Hölder continuous, i.e. $\exists L > 0$, $h \in (0, 1]$

$$
\|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{\mathcal{H}(K)} \leq L \|\mu_a - \mu_b\|_{\mathcal{H}(K)}^h \quad (\forall \mu_a, \mu_b).
$$
Distribution regression problem: intuitive definition

- \( z = \{(x_i, y_i)\}_{i=1}^l: x_i \in M_1^+ (\mathcal{D}), y_i \in \mathbb{R}. \)
- \( \hat{z} = \{ (\{x_{i,n}\}_{n=1}^N, y_i) \}_{i=1}^l: x_{i,1}, \ldots, x_{i,N} \sim i.i.d. x_i. \)
- Goal: learn the relation between \( x \) and \( y \) based on \( \hat{z} \).
- Idea: embed the distributions (\( \mu \)) + apply ridge regression

\[
M_1^+ (\mathcal{D}) \xrightarrow{\mu} X(\subseteq H = H(k)) \xrightarrow{\mathcal{H} = \mathcal{H}(K)} \mathbb{R}.
\]
Objective function

- \( f_{\mathcal{H}} \in \mathcal{H} = \mathcal{H}(K) \): ideal/optimal in expected risk sense (\( \mathcal{E} \)):

\[
\mathcal{E} [f_{\mathcal{H}}] = \inf_{f \in \mathcal{H}} \mathcal{E}[f] = \inf_{f \in \mathcal{H}} \int_{X \times \mathbb{R}} [f(\mu_a) - y]^2 d\rho(\mu_a, y). \tag{7}
\]

- One-stage difficulty (\( \int \rightarrow z \)):

\[
f^{\lambda}_z = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{l} \sum_{i=1}^{l} [f(\mu_{x_i}) - y_i]^2 + \lambda \| f \|^2_{\mathcal{H}} \right). \tag{8}
\]

- Two-stage difficulty (\( z \rightarrow \hat{z} \)):

\[
f^{\lambda}_{\hat{z}} = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{l} \sum_{i=1}^{l} [f(\mu_{\hat{x}_i}) - y_i]^2 + \lambda \| f \|^2_{\mathcal{H}} \right). \tag{9}
\]
Given:
- training sample: $\hat{z}$,
- test distribution: $t$.

Prediction:

$$(f_{\hat{z}}^\lambda \circ \mu)(t) = [y_1, \ldots, y_l](K + l\lambda I_l)^{-1} \begin{bmatrix} K(\mu_{\hat{x}_1}, \mu_t) \\ \vdots \\ K(\mu_{\hat{x}_l}, \mu_t) \end{bmatrix}, \quad (10)$$

$$K = [K_{ij}] = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{l \times l}. \quad (11)$$
We studied the excess error: $E[f^\lambda_2] - E[f_H]$, i.e., the goodness compared to the best function from $\mathcal{H}$.

Result: with high probability

$$E[f^\lambda_2] - E[f_H] \to 0,$$

(12)

if we appropriately choose the $(l, N, \lambda)$ triplet.
Let the $T : \mathcal{H} \to \mathcal{H}$ covariance operator be

$$T = \int_X K(\cdot, \mu_a)K^*(\cdot, \mu_a)d\rho_X(\mu_a) = \int_X K(\cdot, \mu_a)\delta_{\mu_a}d\rho_X(\mu_a)$$

with eigenvalues $t_n$ ($n = 1, 2, \ldots$).

Let $\rho \in \mathcal{P}(b, c)$ be the set of distributions on $X \times \mathbb{R}$:

- $\alpha \leq n^b t_n \leq \beta$ ($\forall n \geq 1; \alpha > 0, \beta > 0$),
- $\exists g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ ($R > 0$),

where $b \in (1, \infty)$, $c \in [1, 2]$. 
Consistency result: convergence rates

High-level idea:

- The excess error can be upper bounded on $\mathcal{P}(b, c)$ as:

$$g(l, N, \lambda) = \mathcal{E} \left[ f_2^\lambda \right] - \mathcal{E} \left[ f_{\mathcal{H}} \right] \leq \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^{\frac{1}{b}}}.$$ 

- We choose ($h = 1$, i.e., $K$ is Lipschitz)
  
  - $\lambda = \lambda_{l,N} \to 0$:
    
    - by matching two terms,
    - $g(l, N, \lambda) \to 0$; moreover, make the 2 equal terms dominant.
  
- $l = N^a$ ($a > 0$).
Convergence rate: results

1 = 2: If $\lambda = \left[\frac{\log(N)}{N}\right]^{\frac{1}{c+3}}, \quad \frac{1}{b+c} \leq a$, then

$$g(N) = O\left(\left[\frac{\log(N)}{N}\right]^{\frac{c}{c+3}}\right) \rightarrow 0.$$  (13)
Convergence rate: results

1 = 2: If \( \lambda = \left[ \frac{\log(N)}{N} \right]^{\frac{1}{c+3}}, \frac{1}{b+c} \leq a, \) then

\[ g(N) = \mathcal{O} \left( \left[ \frac{\log(N)}{N} \right]^{\frac{c}{c+3}} \right) \rightarrow 0. \]  \hspace{1cm} (13)

1 = 3: If \( \lambda = N^{a-\frac{1}{2}} \log^{\frac{1}{2}}(N), \frac{1}{6} \leq a < \min \left( \frac{1}{2} - \frac{1}{c+3}, \frac{1}{2} \left( \frac{b-1}{b-2} \right) \right), \)

\[ g(N) = \mathcal{O} \left( \frac{1}{N^{3a-\frac{1}{2}} \log^{\frac{1}{2}}(N)} \right) \rightarrow 0. \]  \hspace{1cm} (14)

1 = 4: If \( \lambda = \left[ N^{a-1} \log(N) \right]^{\frac{b}{3b-1}}, \max(\frac{b-1}{4b-2}, \frac{1}{3b}) \leq a < \frac{bc+1}{3b+bc}, \)

\[ g(N) = \mathcal{O} \left( \frac{1}{N^{a+\frac{a}{3b-1}} - \frac{1}{3b-1} \log^{\frac{1}{3b-1}}(N)} \right) \rightarrow 0. \]  \hspace{1cm} (15)
Convergence rate: results

- $2 = 3$: $\emptyset$ (the matched terms can not be made dominant).
- $2 = 4$: If $\lambda = \frac{1}{N^{bc+1}}$, $a < \frac{bc+1}{3b+bc}$, then

$$g(N) = \Theta \left( \frac{1}{N^{abc}} \right) \to 0.$$  

(16)

- $3 = 4$: If $\lambda = \frac{1}{N^{b-1}}$, $2 < b$, $a < \frac{b-1}{2(2b-1)}$, then

$$g(N) = \Theta \left( \frac{1}{N^{2a-\frac{ab}{b-1}}} \right) \to 0.$$  

(17)
Problem: learn the entropy of Gaussians in a supervised manner.

Formally:

- \( A = [A_{i,j}] \in \mathbb{R}^{2 \times 2}, A_{ij} \sim U[0, 1]. \)
- 100 sample sets: \( \{N(0, \Sigma_u)\}_{u=1}^{100} \), where
  - 100 = 25(training) + 25(validation) + 50(testing).
  - one set = 500 i.i.d. 2D points,
  - \( \Sigma_u = R(\beta_u)AA^T R(\beta_u)^T \),
  - \( R(\beta_u) \): 2d rotation,
  - angle \( \beta_u \sim U[0, \pi] \).
Goal: learn the entropy of the first marginal

\[ H = \frac{1}{2} \ln (2\pi e \sigma^2), \quad \sigma^2 = M_{1,1}, \quad M = \sum_{u} \in \mathbb{R}^{2 \times 2}. \quad (18) \]

Baseline: kernel smoothing based distribution regression (applying density estimation) \( \Rightarrow \) DFDR.

Performance: RMSE boxplot over 25 random experiments.
Supervised entropy learning: results

RMSE: MERR=0.75, DFDR=2.02

![Graph showing entropy vs. rotation angle (β)]

- True
- MERR
- DFDR

![Box plots comparing MERR and DFDR RMSE]
Numerical illustration: aerosol prediction

- **Bags:**
  - randomly selected pixels,
  - within a 20km radius around an AOD sensor.
- 800 bags, 100 instances/bag.
- **Instances:** \( x_{i,n} \in \mathbb{R}^{16} \).

Zoltán Szabó  Distribution Regression
Baseline: state-of-the-art mixture model
- EM optimization,
- $800 = 4 \times 160(\text{training}) + 160(\text{test})$; 5-fold CV, 10 times.
- Accuracy: $100 \times RMSE(\pm \text{std}) = 7.5 - 8.5 (\pm 0.1 - 0.6)$.

Ridge regression:
- $800 = 3 \times 160(\text{training}) + 160(\text{validation}) + 160(\text{test})$,
- 5-fold CV, 10 times,
- validation: $\lambda$ regularization, $\theta$ kernel parameter.
We picked 10 kernels \((k)\): Gaussian, exponential, Cauchy, generalized t-student, polynomial kernel of order 2 and 3 \((p = 2\) and 3\), rational quadratic, inverse multiquadratic kernel, Matérn kernel (with \(\frac{3}{2}\) and \(\frac{5}{2}\) smoothness parameters).

We also studied their ensembles.

Explored parameter domain:

\[
(\lambda, \theta) \in \left\{ 2^{-65}, 2^{-64}, \ldots, 2^{-3} \right\} \times \left\{ 2^{-15}, 2^{-14}, \ldots, 2^{10} \right\}.
\]

First, \(K\) was linear.
Kernel definitions \((p = 2, 3)\):

\[
k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \quad k_e(a, b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \quad (19)
\]

\[
k_C(a, b) = \frac{1}{1 + \frac{\|a-b\|_2^2}{\theta^2}}, \quad k_t(a, b) = \frac{1}{1 + \|a - b\|_2}, \quad (20)
\]

\[
k_p(a, b) = (\langle a, b \rangle + \theta)^p, \quad k_r(a, b) = 1 - \frac{\|a-b\|_2^2}{\|a-b\|_2^2 + \theta}, \quad (21)
\]

\[
k_i(a, b) = \frac{1}{\sqrt{\|a-b\|_2^2 + \theta^2}}, \quad (22)
\]

\[
k_{M, \frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3} \|a-b\|_2}{\theta}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\theta}}, \quad (23)
\]

\[
k_{M, \frac{5}{2}}(a, b) = \left(1 + \frac{\sqrt{5} \|a-b\|_2}{\theta} + \frac{5 \|a-b\|_2^2}{3\theta^2}\right) e^{-\frac{\sqrt{5}\|a-b\|_2}{\theta}}. \quad (24)
\]
Aerosol prediction: results ($K$: linear)

100 × $\text{RMSE}(\pm \text{std})$ [baseline: 7.5 − 8.5 (±0.1 − 0.6)]:

<table>
<thead>
<tr>
<th></th>
<th>$k_G$</th>
<th>$k_e$</th>
<th>$k_C$</th>
<th>$k_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.97 (±1.81)</td>
<td>8.25 (±1.92)</td>
<td>7.92 (±1.69)</td>
<td>8.73 (±2.18)</td>
</tr>
<tr>
<td>$k_p(p = 2)$</td>
<td>12.5 (±2.63)</td>
<td>171.24 (±56.66)</td>
<td>9.66 (±2.68)</td>
<td>7.91 (±1.61)</td>
</tr>
<tr>
<td>$k_M,\frac{3}{2}$</td>
<td>8.05 (±1.83)</td>
<td>7.98 (±1.75)</td>
<td>ensemble</td>
<td>7.86 (±1.71)</td>
</tr>
</tbody>
</table>
We fed the mean embedding distance \( \| \mu_x - \mu_y \|_{H(k)} \) to the previous kernels.

Example (RBF on mean embeddings – valid kernel):

\[
K(\mu_a, \mu_b) = e^{-\frac{\|\mu_a - \mu_b\|^2_{H(k)}}{2\theta_k^2}} \quad (\mu_a, \mu_b \in X). \tag{25}
\]

We studied the efficiency of (i) single, (ii) ensembles of kernels \([(k, K) \text{ pairs}]\).
Aerosol prediction: nonlinear $K$, results

- **Baseline:**
  - Mixture model (EM): $7.5 - 8.5$ (±0.1 – 0.6),
  - Linear $K$ (single): $7.91$ (±1.61).
  - Linear $K$ (ensemble): $7.86$ (±1.71).

- **Nonlinear $K$:**
  - Single: $7.90$ (±1.63),
  - Ensemble: $7.81$ (±1.64).
Summary

- **Problem:**
  - consistency of set kernels in regression,
  - open for 15 years.

- Examined solution: ridge regression; simple alg.!

- Contribution (on arXiv: 1402.1754):
  - consistency; convergence rate.

- Code: in ITE (https://bitbucket.org/szzoli/ite/).
Thank you for the attention!

Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350.
Given: \( X \neq \emptyset \) set.

\( \tau \subseteq 2^X \) is called a \textit{topology} on \( X \) if:

1. \( \emptyset \in \tau, \ X \in \tau \).
2. Finite intersection: \( O_1 \in \tau, \ O_2 \in \tau \Rightarrow O_1 \cap O_2 \in \tau \).
3. Arbitrary union: \( O_i \in \tau \ (i \in I) \Rightarrow \bigcup_{i \in I} O_i \in \tau \).

Then, \((X, \tau)\) is called a \textit{topological space}; \( O \in \tau \): \textit{open sets}. 

Zoltán Szabó

Distribution Regression
Topology: examples

- $\tau = \{\emptyset, \mathcal{X}\}$: indiscrete topology.
- $\tau = 2^\mathcal{X}$: discrete topology.
- $(\mathcal{X}, d)$ metric space:
  - Open ball: $B_\epsilon(x) = \{y \in \mathcal{X} : d(x, y) < \epsilon\}$.
  - $O \subseteq \mathcal{X}$ is open if for $\forall x \in O \exists \epsilon > 0$ such that $B_\epsilon(x) \subseteq O$.
  - $\tau := \{O \subseteq \mathcal{X} : O$ is an open subset of $\mathcal{X}\}$.
Given: \((X, \tau)\). \(A \subseteq X\) is

- **closed** if \(X \setminus A \in \tau\) (i.e., its complement is open),
- **compact** if for any family \((O_i)_{i \in I}\) of open sets with \(A \subseteq \bigcup_{i \in I} O_i, \exists i_1, \ldots, i_n \in I\) with \(A \subseteq \bigcup_{j=1}^n O_{i_j}\).

**Closure of** \(A \subseteq X\):

\[
\bar{A} := \bigcap_{A \subseteq C \text{ closed in } X} C.
\] (26)

For \(A \subseteq X\) the **subspace topology** on \(A\): \(\tau_A = \{O \cap A : O \in \tau\}\).
(\mathcal{X}, \tau) is a *Hausdorff space*, if

- for any \( x \neq y \in \mathcal{X} \) \( \exists U, V \in \tau \) such that \( x \in U, \; y \in V, \; U \cap V = \emptyset \).

- In other words, disjunct points have disjunct open environments.

- Example: metric spaces.
• $A \subseteq X$ is dense if $\overline{A} = X$.
• $(X, \tau)$ is separable if $\exists$ countable, dense subset of $X$.
  Counterexample: $l^\infty / L^\infty$.
• $\tau_1 \subseteq \tau$ is a basis of $\tau$ if every open set is the union of sets in $\tau_1$. Example: open balls in a metric space.
• $(X, \tau)$ is Polish if $\tau$ has a countable basis and $\exists$ metric defining $\tau$. Example: complete separable metric spaces.
$\mathcal{X}$, $\tau$:

- $V \subseteq \mathcal{X}$ is a *neighborhood* of $x \in \mathcal{X}$ if $\exists O \in \tau$ such that $x \in O \subseteq V$.

- is called *locally compact* if for $\forall x \in \mathcal{X} \exists$ compact neighborhood of $x$. Example: $\mathbb{R}^d$; not compact.
Examples: LCH, but not (necessarily) compact

- Euclidean spaces: $\mathbb{R}^d$, not compact.
- Discrete spaces: LCH. Compact $\iff |X| < \infty$.
- Open/closed subsets of an LCH: LC in subspace topology. Example: unit ball (open/closed).
Examples: Hausdorff, but not locally compact

- \(\mathbb{Q}\), topology inherited from \(\mathbb{R}\).
  - In other words, not every subset of an LCH is LC.
- Infinite dimensional Hilbert spaces.
  - Example: complex \(L^2([0, 1])\); \(\{f_n(x) = e^{2\pi inx}, n \in \mathbb{Z}\}\): ONB.
\( (\mathcal{X}, 2^{\mathcal{X}}) \): complete metric space.

Discrete metric (inducing the discrete topology):

\[
d(x, y) = \begin{cases} 
0, & \text{if } x = y \\
1, & \text{if } x \neq y
\end{cases}.
\] (27)

Discrete space: separable \( \iff \) \( |\mathcal{X}| \) is countable.
Let $C_0(\mathcal{D}) = \mathcal{D} \to \mathbb{R}$ continuous functions vanishing at infinity, i.e.,

$$\{ u \in \mathcal{D} : |g(u)| \geq \epsilon \}$$

is compact for $g \in C_0(\mathcal{D})$, $\forall \epsilon > 0$. $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is $c_0$-universal if

- $\| k \|_{\infty} := \sup_{u \in \mathcal{D}} \sqrt{k(u, u)} < \infty$,
- $k(\cdot, u) \in C_0(\mathcal{D})$ ($\forall u \in \mathcal{D}$),
- $H = H(k)$ is dense in $C_0(\mathcal{D})$ w.r.t. the uniform norm.
ITE: covered quantities

- **entropy**: Shannon entropy, Rényi entropy, Tsallis entropy (Havrda and Charvát entropy), complex entropy, $\Phi$-entropy ($f$-entropy), Sharma-Mittal entropy,

- **mutual information**: generalized variance, kernel canonical correlation analysis, kernel generalized variance, Hilbert-Schmidt independence criterion, Shannon mutual information (total correlation, multi-information), $L_2$ mutual information, Rényi mutual information, Tsallis mutual information, copula-based kernel dependency, multivariate version of Hoeffding’s $\Phi$, Schweizer-Wolff’s $\sigma$ and $\kappa$, complex mutual information, Cauchy-Schwartz quadratic mutual information (QMI), Euclidean distance based QMI, distance covariance, distance correlation, approximate correntropy independence measure, $\chi^2$ mutual information (Hilbert-Schmidt norm of the normalized cross-covariance operator, squared-loss mutual information, mean square contingency),

- **divergence**: Kullback-Leibler divergence (relative entropy, directed divergence), $L_2$ divergence, Rényi divergence, Tsallis divergence, Hellinger distance, Bhattacharyya distance, maximum mean discrepancy (kernel distance), $J$-distance (symmetrised Kullback-Leibler divergence, $J$ divergence), Cauchy-Schwartz divergence, Euclidean distance based divergence, energy distance (specially the Cramer-Von Mises distance), Jensen-Shannon divergence, Jensen-Rényi divergence, K divergence, L divergence, $f$-divergence (Csiszár-Morimoto divergence, Ali-Silvey distance), non-symmetric Bregman distance (Bregman divergence), Jensen-Tsallis divergence, symmetric Bregman distance, Pearson $\chi^2$ divergence ($\chi^2$ distance), Sharma-Mittal divergence,

- **association measures**: multivariate extensions of Spearman’s $\rho$ (Spearman’s rank correlation coefficient, grade correlation coefficient), correntropy, centered correntropy, correntropy coefficient, correntropy induced metric, centered correntropy induced metric, multivariate extension of Blomqvist’s $\beta$ (medial correlation coefficient), multivariate conditional version of Spearman’s $\rho$, lower/upper tail dependence via conditional Spearman’s $\rho$,

- **cross quantities**: cross-entropy,

- **kernels on distributions**: expected kernel (summation kernel, mean map kernel), Bhattacharyya kernel, probability product kernel, Jensen-Shannon kernel, exponentiated Jensen-Shannon kernel, Jensen-Tsallis kernel, exponentiated Jensen-Rényi kernel(s), exponentiated Jensen-Tsallis kernel(s),

- **+some auxiliary quantities**: Bhattacharyya coefficient (Hellinger affinity), $\alpha$-divergence.
ITE: summary

- Matlab/Octave (first release).
- Multi-platform.
- GPLv3(≥).
- Appeared in JMLR, 2014.
- Homepage: https://bitbucket.org/szzoli/ite/
• Consistency tests.
• Prototype: independent subspace analysis, its extensions.

• Image registration $\rightarrow$ outlier robustness.
• Distribution regression.
$K = \text{RBF: Lipschitz on compact } \mathcal{D} \text{ domains}$

Let

$$K(\mu_a, \mu_b) = e^{-\frac{\|\mu_a - \mu_b\|^2_H}{2\sigma^2}}. \quad (29)$$

Needed: $\exists \ L > 0$ such that

$$\|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{\mathcal{H}} \leq L \|\mu_a - \mu_b\|_H \quad (\forall \mu_a, \mu_b \in X). \quad (30)$$

L.h.s.:

$$\|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{\mathcal{H}}^2 = K(\mu_a, \mu_a) + K(\mu_b, \mu_b) - 2K(\mu_a, \mu_b) \quad (31)$$

$$= 2 \left[1 - e^{-\frac{\|\mu_a - \mu_b\|^2_H}{2\sigma^2}}\right]. \quad (32)$$
The statement is equivalent to $\exists \, L' > 0$:

$$
u(\mu_a, \mu_v) := \frac{1 - e^{-\frac{\|\mu_a - \mu_b\|^2}{2\sigma^2}}}{\|\mu_a - \mu_b\|^2_H} \leq L'. \quad (33)$$

Idea:

- $\mathcal{D}$ compact $\Rightarrow$ $X$ compact; $\Rightarrow$ $X \times X$ compact (Tychonoff T.).
- $u = u_2 \circ u_1$ continuous (continuity of $u_i$-s):

$$
u_1 : X \times X \to \mathbb{R}^{\geq 0}, \quad \nu_1(\mu_a, \mu_b) = \|\mu_a - \mu_b\|^2_H, \quad (34)$$

$$
u_2 : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}, \quad \nu_2(\nu) = \frac{1 - e^{-\frac{\nu}{2\sigma^2}}}{\nu}. \quad (35)$$

- Continuous image ($u$) of a compact set ($X \times X$) is compact.