Consistent Vector-valued Distribution Regression

Zoltán Szabó

Joint work with Arthur Gretton (UCL), Barnabás Póczos (CMU), Bharath K. Sriperumbudur (PSU)

UCL Workshop on the Theory of Big Data
January 8, 2015
The task

- Samples: \( \{(x_i, y_i)\}_{i=1}^{l} \). Goal: \( f(x_i) \approx y_i \), find \( f \in \mathcal{H} \).

- Distribution regression:
  - \( x_i \)-s are distributions,
  - available only through samples: \( \{x_{i,n}\}_{n=1}^{N_i} \).

  ⇒ Training examples: labelled bags.
Example: aerosol prediction from satellite images

- **Bag**: points of a multispectral satellite image over an area.
- **Label of a bag**: aerosol value.

**Engineered methods** [Wang et al., 2012]: \(100 \times \text{RMSE} = 7.5 - 8.5\).
- Using distribution regression:
  - without domain knowledge,
  - \(100 \times \text{RMSE} = 7.81\).
Context:
- machine learning: multi-instance learning,
- statistics: point estimation tasks (without analytical formula).

Applications:
- computer vision: image = collection of patch vectors,
- network analysis: group of people = bag of friendship graphs,
- natural language processing: corpus = bag of documents,
- time-series modelling: user = set of trial time-series.
Several algorithmic approaches

1. Parametric fit: Gaussian, MOG, exp. family
   [Jebara et al., 2004, Wang et al., 2009, Nielsen and Nock, 2012].

2. Kernelized Gaussian measures:
   [Jebara et al., 2004, Zhou and Chellappa, 2006].

3. (Positive definite) kernels:
   [Cuturi et al., 2005, Martins et al., 2009, Hein and Bousquet, 2005].


5. Set metrics: Hausdorff metric [Edgar, 1995]; variants
MIL dates back to [Haussler, 1999, Gärtner et al., 2002].

Sensible methods in regression: require density estimation [Póczos et al., 2013, Oliva et al., 2014] + assumptions:
1. compact Euclidean domain.
2. output = \( \mathbb{R} \).
Problem formulation

- **Given:** labelled bags
  \[
  \hat{z} = \{(\hat{x}_i, y_i)\}_{i=1}^l,
  \]
  where
  \[
  i^{th} \text{ bag: } \hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \sim x_i \in \mathcal{M}_1^+(\mathcal{D}), \ y_i \in Y.
  \]
- **Task:** find a \(\mathcal{M}_1^+(\mathcal{D}) \rightarrow Y\) mapping based on \(\hat{z}\).
- **Construction:** distribution embedding \((\mu_x) + \text{ridge regression}\)

\[
\mathcal{M}_1^+(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k) \xrightarrow{f \in \mathcal{H}=\mathcal{H}(K)} Y.
\]

- **Our goal:** risk bound compared to the regression function

\[
f_\rho(\mu_x) = \int_Y y \text{d}\rho(y|\mu_x).
\]
Goal in details

Contribution: analysis of the excess risk

\[\mathcal{E}(f_{\hat{z}}^\lambda, f_{\rho}) = \mathcal{R}[f_{\hat{z}}^\lambda] - \mathcal{R}[f_{\rho}] \leq g(l, N, \lambda) \to 0\] and rates,

\[\mathcal{R} [f] = \mathbb{E}_{(x,y)} \| f(\mu_x) - y \|^2_Y \text{ (expected risk)},\]

\[f_{\hat{z}}^\lambda = \arg \min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \| f(\mu_{\hat{x}_i}) - y_i \|^2_Y + \lambda \| f \|^2_{\mathcal{H}}, \quad (\lambda > 0)\].

We consider two settings:

1. well-specified case: \( f_{\rho} \in \mathcal{H}, \)
2. misspecified case: \( f_{\rho} \in L^2_{\rho_X} \setminus \mathcal{H}. \)
Kernel \((k, K)\), RKHS

- \(k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}\) kernel on \(\mathcal{D}\), if \(\exists \varphi : \mathcal{D} \rightarrow H(\text{hilbert})\)

\[
k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H.
\]

- \(\exists! \text{ RKHS}: H(k) = \{\mathcal{D} \rightarrow \mathbb{R} \text{ functions}\}, \varphi(u) = k(\cdot, u)\).
- 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,
  - Graphs, texts, time series, distributions.
Kernel \((k, K)\), RKHS

- \(k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}\) kernel on \(\mathcal{D}\), if \(\exists \varphi : \mathcal{D} \to H(\text{ilbert})\)

\[
k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H.
\]

- \(\exists!\) RKHS: \(H(k) = \{ \mathcal{D} \to \mathbb{R} \text{ functions} \}\), \(\varphi(u) = k(\cdot, u)\).
- 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,
  - Graphs, texts, time series, distributions.
- Note: \(\mathcal{H}(k) = \{ X \to Y \text{ functions} \}\), \(K(\mu_x, \mu_{x'}) \in \mathcal{L}(Y)\).
Step-1 (distribution embedding): $\mathcal{M}^{+}_{1}(\mathcal{D}) \xrightarrow{\mu^i} X \subseteq H(k)$

- Given: kernel $k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$.
- Mean embedding of a distribution $x, \hat{x}_i \in \mathcal{M}^{+}_{1}(\mathcal{D})$:
  \[
  \mu_x = \int_{\mathcal{D}} k(\cdot, u) dx(u) \in H(k), \\
  \mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u) d\hat{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_i, n).
  \]
- $Y = \mathbb{R}$, linear $K \Rightarrow$ set kernel:
  \[
  K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \left\langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \right\rangle_H = \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_i, n, x_j, m).
  \]
Step-2 (ridge regression): analytical solution

- **Given:**
  - training sample: $\hat{z}$,
  - test distribution: $t$.

- **Prediction:**

  $$(f_\hat{z}^\lambda \circ \mu)(t) = k(K + l\lambda I_l)^{-1}[y_1; \ldots; y_l],$$  
  $$K = [K_{ij}] = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l \times l},$$  
  $$k = [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_l}, \mu_t)] \in \mathcal{L}(Y)^{1 \times l}.$$  

- Specially: $Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}$; $Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^{d \times d}$. 

Blanket assumptions

- \( \mathcal{D} \): separable, topological domain.
- \( k \): bounded, continuous.
- \( K \): bounded, Hölder continuous (\( h \in (0, 1] \): exponent).
- \( X = \mu (\mathcal{M}_1^+(\mathcal{D})) \in \mathcal{B}(H) \).
- \( Y \): separable Hilbert.
If in addition

1. well-specified case: $f_\rho$ is 'c-smooth' with 'b-decaying covariance operator' and $l \geq \lambda^{-\frac{1}{b}}-1$, then

$$
\mathcal{E}(f_\lambda^z, f_\rho) \leq \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^{\frac{1}{b}}}.
$$

(4)

2. misspecified case: $f_\rho$ is 's-smooth', $L^2_{\rho X}$ is separable, and $\frac{1}{\lambda^2} \leq l$, then

$$
\mathcal{E}(f_\lambda^z, f_\rho) \leq \frac{\log^h(l)}{N^{\frac{h}{2}} \lambda^{\frac{3}{2}}} + \frac{1}{\sqrt{l} \lambda} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda \sqrt{l}} + \lambda^{\min(1,s)}.
$$

(5)
Misspecified case: assume

- $s \geq 1$, $h = 1$ ($K$: Lipschitz),
- $[1] = [3]$ in (5) $\Rightarrow \lambda$; $l = N^a$ ($a > 0$)
- $t = lN^a$: total number of samples processed.

Then

1. $s = 1$ ('most difficult' task): $\mathcal{E}(\hat{f}_\lambda, f_\rho) \approx t^{-0.25}$,
2. $s \to \infty$ ('simplest' problem): $\mathcal{E}(\hat{f}_\lambda, f_\rho) \approx t^{-0.5}$. 

Performance guarantee: example
Nonlinear $K$ examples

$Y = \mathbb{R}$; $\mathcal{D}$: compact, metric; $k$: universal $\Rightarrow$ Hölder $K$-s:

<table>
<thead>
<tr>
<th>$K_G$</th>
<th>$K_e$</th>
<th>$K_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e^{-\frac{|\mu_a - \mu_b|^2_H}{2\theta^2}}$</td>
<td>$e^{-\frac{|\mu_a - \mu_b|_H}{2\theta^2}}$</td>
<td>$\left(1 + |\mu_a - \mu_b|^2_H / \theta^2\right)^{-1}$</td>
</tr>
<tr>
<td>$h = 1$</td>
<td>$h = \frac{1}{2}$</td>
<td>$h = 1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$K_t$</th>
<th>$K_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\left(1 + |\mu_a - \mu_b|_H^\theta\right)^{-1}$</td>
<td>$\left(|\mu_a - \mu_b|^2_H + \theta^2\right)^{-\frac{1}{2}}$</td>
</tr>
<tr>
<td>$h = \frac{\theta}{2}$ ($\theta \leq 2$)</td>
<td>$h = 1$</td>
</tr>
</tbody>
</table>

They are functions of $\|\mu_a - \mu_b\|_H$ $\Rightarrow$ computation: similar to set kernel.
Problem: distribution regression.

Literature: large number of heuristics.

Contribution:
- a simple ridge solution is consistent,
- specially, the set kernel is so (15-year-old open question).

Code ∈ ITE toolbox:
- https://bitbucket.org/szzoli/ite/

Details (submitted to JMLR):
Thank you for the attention!

Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.
Well/misspecified assumptions.

- Topological definitions, separability.
- Vector-valued RKHS.
- Weak topology on $\mathcal{M}_1^+(\mathcal{D})$.
- Measurability of $\mu$.
- Universal kernel examples.
Well-specified case: $\rho \in \mathcal{P}(b, c)$

- Let the $T : \mathcal{H} \to \mathcal{H}$ covariance operator be

$$T = \int_{\mathcal{X}} K(\cdot, \mu_a)K^*(\cdot, \mu_a)d\rho_X(\mu_a)$$

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

- Assumption: $\rho \in \mathcal{P}(b, c) =$ set of distributions on $\mathcal{X} \times \mathcal{Y}$
  - $\alpha \leq n^b t_n \leq \beta$ ($\forall n \geq 1; \alpha > 0, \beta > 0$),
  - $\exists g \in \mathcal{H}$ such that $f_\rho = T^{c-1}g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ ($R > 0$),

  where $b \in (1, \infty)$, $c \in [1, 2]$.

- Intuition: $b$ – effective input dimension, $c$ – smoothness of $f_\rho$. 
Let $\tilde{T}$ be the extension of $T$ from $\mathcal{H}$ to $L^2_{\rho_X}$:

$$S^*_K : \mathcal{H} \hookrightarrow L^2_{\rho_X},$$

$$S_K : L^2_{\rho_X} \rightarrow \mathcal{H}, \quad (S_K g)(\mu_u) = \int_X K(\mu_u, \mu_t)g(\mu_t)d\rho_X(\mu_t),$$

$$\tilde{T} = S^*_K S_K : L^2_{\rho_X} \rightarrow L^2_{\rho_X}.$$

Our range space assumption on $\rho$: $f_\rho \in \text{Im} \left( \tilde{T}^s \right)$ for some $s \geq 0$. 
Misspecified case: note on the separability of $L^2_{\rho_X}$

$L^2_{\rho_X}$: separable $\iff$ measure space with $d(A, B) = \rho_X(A \triangle B)$ is so [Thomson et al., 2008].
Given: $\mathcal{D} \neq \emptyset$ set.

$\tau \subseteq 2^{\mathcal{D}}$ is called a topology on $\mathcal{D}$ if:

1. $\emptyset \in \tau$, $\mathcal{D} \in \tau$.
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, $(\mathcal{D}, \tau)$ is called a topological space; $O \in \tau$: open sets.
Closed-, compact set, closure, dense subset, separability

Given: \((\mathcal{D}, \tau)\). \(A \subseteq \mathcal{D}\) is

- **closed** if \(\mathcal{D} \setminus A \in \tau\) (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 \(A \subseteq \mathcal{D}\):

\[
\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{D}} C.
\]  

(A \subseteq \mathcal{D}\) is **dense** if \(\bar{A} = \mathcal{D}\).

- \((\mathcal{D}, \tau)\) is **separable** if \(\exists\) countable, dense subset of \(\mathcal{D}\).
  Counterexample: \(L^\infty/L^\infty\).
Vector-valued RKHS

Definition:

- A $\mathcal{H} \subseteq Y^X$ Hilbert space of functions is RKHS if
  
  $A_{\mu_x,y} : f \mapsto \langle y, f(\mu_x) \rangle_Y$ 

  is continuous for $\forall \mu_x \in X, y \in Y$.

- The evaluation functional is continuous in every direction.

Riesz representation theorem $\Rightarrow$

- $\exists K_{\mu_t} \in \mathcal{L}(Y, \mathcal{H})$:

  $K(\mu_x, \mu_t)(y) = (K_{\mu_t}y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X)$, or shortly

  $K(\cdot, \mu_t)(y) = K_{\mu_t}y,$ 

  $\mathcal{H}(K) = \overline{\text{span}}\{K_{\mu_t}y : \mu_t \in X, y \in Y\}$. 

Zoltán Szabó
Consistent Vector-valued Distribution Regression
Examples \((Y = \mathbb{R}^d)\):

1. \(K_i : X \times X \to \mathbb{R}\) kernels \((i = 1, \ldots, d)\). Diagonal kernel:

\[
K(\mu_a, \mu_b) = \text{diag}(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \tag{10}
\]

2. Combination of \(D_j\) diagonal kernels \([D_j(\mu_a, \mu_b) \in \mathbb{R}^{r \times r}, A_j \in \mathbb{R}^{r \times d}]\):

\[
K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \tag{11}
\]
Def.: It is the weakest topology such that the mapping is continuous for all $h \in C_b(\mathcal{D})$, where

$$C_b(\mathcal{D}) = \{(\mathcal{D}, \tau) \to \mathbb{R} \text{ bounded, continuous functions}\}.$$
Measurability of $\mu$

- $k$: bounded, continuous $\Rightarrow$
  - $\mu : (\mathcal{M}_1^+(\mathcal{D}), \mathcal{B}(\tau_w)) \rightarrow (H, \mathcal{B}(H))$ measurable.
  - $\mu$ measurable, $X \in \mathcal{B}(H) \Rightarrow \rho$ on $X \times Y$: well-defined.

- If $\mathcal{D}$ is compact metric, $k$ is universal, then $\mu$ is continuous and $X \in \mathcal{B}(H)$. 

Zoltán Szabó
Consistent Vector-valued Distribution Regression
On every compact subset of $\mathbb{R}^d$:

\[
k(a, b) = e^{-\frac{\|a-b\|^2}{2\sigma^2}}, \quad (\sigma > 0)
\]

\[
k(a, b) = e^{\beta \langle a, b \rangle}, \quad (\beta > 0), \text{ or more generally}
\]

\[
k(a, b) = f(\langle a, b \rangle), \quad f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (\forall a_n > 0)
\]

\[
k(a, b) = (1 - \langle a, b \rangle)^\alpha, \quad (\alpha > 0).
\]

In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 870–873.


Nonparametric divergence estimation with applications to machine learning on distributions.

In *Uncertainty in Artificial Intelligence (UAI)*, pages 599–608.


