Learning on Distributions

Zoltán Szabó (Gatsby Unit)

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

Kernel methods for big data, Lille
April 2, 2014

Outline

- ITE (Information Theoretical Estimators) Toolbox.
- Distribution Regression:
  - Motivation, examples.
  - Algorithm, consistency result.
  - Numerical illustration.

Distribution based tasks: building blocks

- Entropies: uncertainty

$$H(\mathbf{x}) = -\int_{\mathbb{R}^d} f(\mathbf{u}) \log f(\mathbf{u}) \mathrm{d}\mathbf{u}.$$

Distribution based tasks: building blocks

- Entropies: uncertainty

$$H(\mathbf{x}) = -\int_{\mathbb{R}^d} f(\mathbf{u}) \log f(\mathbf{u}) \mathrm{d}\mathbf{u}.$$

- Mutual information, association indices: dependence

$$I\left(\mathbf{x}^1, \ldots, \mathbf{x}^M\right) = \int f\left(\mathbf{u}^1, \ldots, \mathbf{u}^M\right) \log \left[\frac{f\left(\mathbf{u}^1, \ldots, \mathbf{u}^M\right)}{\prod_{m=1}^M f_m(\mathbf{u}^m)}\right] \mathrm{d}\mathbf{u}.$$

- Divergences, kernels: 'distance'/inner product of probability distributions

$$D(f_1, f_2) = \int_{\mathbb{R}^d} f_1(\mathbf{u}) \log \left[\frac{f_1(\mathbf{u})}{f_2(\mathbf{u})}\right] \mathrm{d}\mathbf{u}.$$

Estimation

1. Plug-in of estimated densities:
   - Example: histogram based methods.

Estimation

1. Plug-in of estimated densities:
   - Example: histogram based methods.
   - Poor scaling!
   - Reason:
     - goal $\neq$ density estimation, but
     - *functionals* of distributions.

Estimation

1. Plug-in of estimated densities:
   - Example: histogram based methods.
   - Poor scaling!
   - Reason:
     - goal $\neq$ density estimation, but
     - *functionals* of distributions.
2. Nonparametric, non plug-in type estimators.

Existing packages

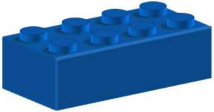
Focus on

- discrete variables, or
- quite specialized
  - applications and
  - information theoretical estimation methods.

ITE (information theoretical estimators) toolbox

Goal:

- state-of-the-art, nonparametric estimators,
- modularity:
  - high-level optimization,
  - combinations (methods/estimators).



Covered quantities

- **entropy**: Shannon entropy, Rényi entropy, Tsallis entropy (Havrda and Charvát entropy), complex entropy, Φ-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 Φ, Schweizer-Wolff's σ and κ, complex mutual information, Cauchy-Schwartz quadratic mutual information (QMI), Euclidean distance based QMI, distance covariance, distance correlation, approximate correntropy independence measure, χ^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, I 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 χ^2 divergence (χ^2 distance), Sharma-Mittal divergence,
- **association measures**: multivariate extensions of Spearman's ρ (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 β (medial correlation coefficient), multivariate conditional version of Spearman's ρ, lower/upper tail dependence via conditional Spearman's ρ,
- **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), α-divergence.

ITE: summary

- Matlab/Octave (first release).
- Multi-platform.
- GPLv3($\geq$).
- Appeared in JMLR, 2014.
- Homepage: `https://bitbucket.org/szzoli/ite/`

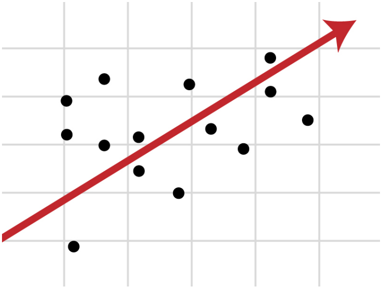
ITE: built-in tests/applications

- Consistency tests.
- Prototype: independent subspace analysis, its extensions.

- Image registration $\rightarrow$ outlier robustness.
- Distribution regression (next part).

Regression

- 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 (∞-dimensional objects).

Distribution regression: two-stage sampling difficulty

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.

Set kernels: consistency?

- Given (2 bags):

$$B_i := \{x_{i,n}\}_{n=1}^{N_i} \sim x_i, \tag{1}$$

$$B_j := \{x_{j,m}\}_{m=1}^{N_j} \sim x_j. \tag{2}$$

- Similarity of the bags (set/multi-instance/ensemble kernel):

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

- Many successful applications – no theory.
- Our results ⇒

$$\boxed{\text{statistical consistency of set kernels in regression}}.$$

Example: supervised entropy learning

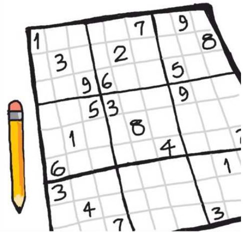
- Entropy of $x \sim f$: $-\int f(u) \log[f(u)] \mathrm{d}u$.
- Training: samples from distributions, entropy values.
- Task: estimate the entropy of a new sample set.

Example: age prediction from images

- Training: (image, age) pairs; image = bag of features.
- Goal: estimate the age of a person being on a new image.

Example: Sudoku difficulty estimation

- Sudoku: special constraint satisfaction problem.
- Spiking neural networks (SNN)
  - can be used to solve such problems,
  - have stationary distribution under mild conditions.
- Sudoku $\leftrightarrow$ stationary distribution of the SNN.



Example: aerosol prediction using satellite images

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

Towards problem formulation: kernel, RKHS

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel on $\mathcal{D}$, if
  - $\exists \varphi : \mathcal{D} \to H$(ilbert space) feature map,
  - $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{D})$.
- 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 $(\exists!)$: $\varphi(u) = k(\cdot, u)$.

Some example domains ($\mathcal{D}$), where kernels exist

- Euclidean spaces: $\mathcal{D} = \mathbb{R}^d$.
- Strings, time series, graphs, dynamical systems.

- Distributions.

Distribution kernel: example (used in our work)

- Given: $(\mathcal{D}, k)$; we saw that $u \to \varphi(u) = k(\cdot, u) \in H(k)$.
- Let x be a distribution on $\mathcal{D}$ ($x \in \mathcal{M}_1^+(\mathcal{D})$); the previous construction can be extended:

$$\mu_x = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}x(u) \in H(k). \tag{4}$$

- If k is bounded: μ_x is well-defined for *any* distribution x.

Mean embedding based distribution kernel

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

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

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

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

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

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

Mini summary

- 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: $\mathcal{D} = \mathbb{R}^d$, $k =$ Gaussian, $K =$ lin. kernel.

- The *real* conditions:
  - $\mathcal{D}$: locally compact, Polish. k: c_0-universal.
  - K: Hölder continuous.

Distribution regression problem: intuitive definition

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

$$M_1^+(\mathcal{D}) \xrightarrow{\mu} X(\subseteq H = H(k)) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} \mathbb{R}.$$

Objective function

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

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

- One-stage difficulty ($\int \to \mathbf{z}$):

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

- Two-stage difficulty ($\mathbf{z} \to \hat{\mathbf{z}}$):

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \left(\frac{1}{l} \sum_{i=1}^{l} [f(\mu_{\hat{x}_i}) - y_i]^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (9)$$

Algorithmically: ridge regression $\Rightarrow$ simple solution

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

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

$$\mathbf{K} = [K_{ij}] = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{l \times l}. \quad (11)$$

Consistency result

- We studied
  - the excess error: $\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right]$, i.e,
  - the goodness compared to the best function from $\mathcal{H}$.
- Result: with high probability

$$\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right] \to 0, \tag{12}$$

if we appropriately choose the (l, N, λ) triplet.

Consistency result: $\mathcal{P}(b,c)$ class

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

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

  with eigenvalues t_n $(n = 1, 2, \ldots)$.
- Let $\rho \in \mathcal{P}(b,c)$ be the set of distributions on $X \times \mathbb{R}$:
  - $\alpha \le n^b t_n \le \beta \quad (\forall n \ge 1; \alpha > 0, \beta > 0)$,
  - $\exists g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \le R$ $(R > 0)$,

  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:

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

- We choose
  - $\lambda=\lambda_{l,N}\rightarrow 0$:
    - by matching two terms,
    - $g(l,N,\lambda)\rightarrow 0$; moreover, make the 2 equal terms dominant.
  - $l=N^{a}$ $(a>0)$.

Convergence rate: results

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

$$g(N) = \mathcal{O}\left(\left[\frac{\log(N)}{N}\right]^{\frac{c}{c+3}}\right) \to 0. \tag{13}$$

Convergence rate: results

- $\boxed{1} = \boxed{2}$: If $\lambda = \left[\frac{\log(N)}{N}\right]^{\frac{1}{c+3}}$, $\frac{\frac{1}{b}+c}{c+3} \le a$, then
$$g(N) = \mathcal{O}\left(\left[\frac{\log(N)}{N}\right]^{\frac{c}{c+3}}\right) \to 0. \tag{13}$$
- $\boxed{1} = \boxed{3}$: If $\lambda = N^{a-\frac{1}{2}} \log^{\frac{1}{2}}(N)$, $\frac{1}{6} \le a < \min\left(\frac{1}{2} - \frac{1}{c+3}, \frac{\frac{1}{2}\left(\frac{1}{b}-1\right)}{\frac{1}{b}-2}\right)$,
$$g(N) = \mathcal{O}\left(\frac{1}{N^{3a-\frac{1}{2}} \log^{\frac{1}{2}}(N)}\right) \to 0. \tag{14}$$
- $\boxed{1} = \boxed{4}$: If $\lambda = \left[N^{a-1}\log(N)\right]^{\frac{b}{3b-1}}$, $\max\left(\frac{b-1}{4b-2}, \frac{1}{3b}\right) \le a < \frac{bc+1}{3b+bc}$,
$$g(N) = \mathcal{O}\left(\frac{1}{N^{a+\frac{a}{3b-1}-\frac{1}{3b-1}} \log^{\frac{1}{3b-1}}(N)}\right) \to 0. \tag{15}$$

Convergence rate: results

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

$$g(N) = \mathcal{O}\left(\frac{1}{N^{\frac{abc}{bc+1}}}\right) \to 0. \tag{16}$$

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

$$g(N) = \mathcal{O}\left(\frac{1}{N^{2a-\frac{ab}{b-1}}}\right) \to 0. \tag{17}$$

Numerical illustration: supervised entropy learning

- 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
    - $100 = 25(\text{training}) + 25(\text{validation}) + 50(\text{testing})$.
    - 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 $\beta_u \sim U[0,\pi]$.

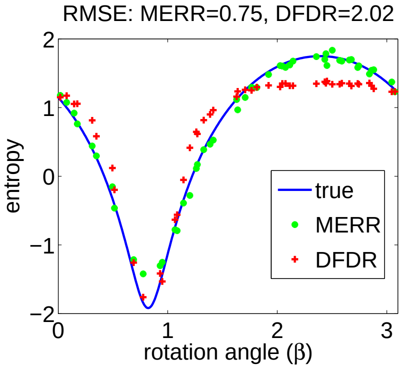
Supervised entropy learning: goal, performance measure

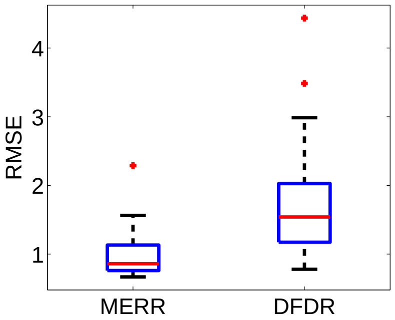
- Goal: learn the entropy of the first marginal

$$H = \frac{1}{2}\ln\left(2\pi e\sigma^2\right), \quad \sigma^2 = M_{1,1}, \quad M = \Sigma_u \in \mathbb{R}^{2\times 2}. \quad (18)$$

- Baseline: kernel smoothing based distribution regression (applying density estimation) =: DFDR.
- Performance: RMSE boxplot over 25 random experiments.

Supervised entropy learning: results





Numerical illustration: aerosol prediction

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

Aerosol prediction - baseline

- Baseline: state-of-the-art mixture model
  - EM optimization,
  - $800 = 4 \times 160$(training) $+ 160$(test); 5-fold CV, 10 times.
  - Accuracy: $100 \times RMSE(\pm \text{ std}) = 7.5 - 8.5$ $(\pm 0.1 - 0.6)$.
- Ridge regression:
  - $800 = 3 \times 160$(training) $+ 160$(validation) $+ 160$(test),
  - 5-fold CV, 10 times,
  - validation: λ regularization, θ kernel parameter.

Aerosol prediction: kernel k

- 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.

Aerosol prediction: kernel definitions

Kernel definitions ($p = 2, 3$):

$$k_G(a,b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \qquad k_e(a,b) = e^{-\frac{\|a-b\|_2}{2\theta^2}}, \tag{19}$$

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

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

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

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

$$k_{M,\frac{5}{2}}(a,b) = \left(1+\frac{\sqrt{5}\,\|a-b\|_2}{\theta}+\frac{5\,\|a-b\|_2^2}{3\theta^2}\right) e^{-\frac{\sqrt{5}\|a-b\|_2}{\theta}}. \tag{24}$$

Aerosol prediction: results (K: linear)

$100 \times RMSE(\pm std)$ [baseline: $7.5 - 8.5$ ($\pm 0.1 - 0.6$)]:

k_G 7.97 ($\pm$1.81)	k_e 8.25 ($\pm$1.92)	k_C 7.92 ($\pm$1.69)	k_t 8.73 ($\pm$2.18)
$k_p(p=2)$ 12.5 ($\pm$2.63)	$k_p(p=3)$ 171.24 ($\pm$56.66)	k_r 9.66 ($\pm$2.68)	k_i **7.91** ($\pm$**1.61**)
$k_{M,\frac{3}{2}}$ 8.05 ($\pm$1.83)	$k_{M,\frac{5}{2}}$ 7.98 ($\pm$1.75)	ensemble **7.86** ($\pm$**1.71**)	

Best combination in the ensemble: $k = k_G, k_C, k_i$.

Aerosol prediction: nonlinear K

- 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) pairs].

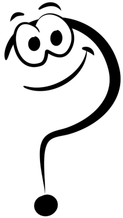
Aerosol prediction: nonlinear K, results

- Baseline:
  - Mixture model (EM): $7.5 - 8.5$ ($\pm 0.1 - 0.6$),
  - Linear K (single): 7.91 (± 1.61).
  - Linear K (ensemble): **7.86** ($\pm$**1.71**).
- Nonlinear K:
  - Single: 7.90 (± 1.63),
  - Ensemble:
    - Accuracy: **7.81** ($\pm$**1.64**),
    - $(k, K) = (k_i, k_t), \left(k_{M,\frac{3}{2}}, k_{M,\frac{3}{2}}\right), (k_C, k_G)$.

Summary

- Problem: distribution regression.
- Difficulty: two-stage sampling.
- Examined solution: ridge regression; simple alg.!
- Contribution (on arXiv):
  - consistency; convergence rate.
  - specially: consistency of set kernels in regression.
- ITE toolbox (Bitbucket, $\ni$MERR).

Thank you for the attention!



Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350.

Topological space, open sets

- Given: $\mathcal{X} \neq \emptyset$ set.
- $\tau \subseteq 2^{\mathcal{X}}$ is called a *topology* on $\mathcal{X}$ if:
    1. $\emptyset \in \tau$, $\mathcal{X} \in \tau$.
    2. Finite intersection: $O_1 \in \tau$, $O_2 \in \tau \Rightarrow O_1 \cap O_2 \in \tau$.
    3. Arbitrary union: $O_i \in \tau$ $(i \in I) \Rightarrow \cup_{i \in I} O_i \in \tau$.

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

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 $\mathcal{X}\}$.

Closed set, compact set, closure, subspace topology

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

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

Closure of $A \subseteq \mathcal{X}$:

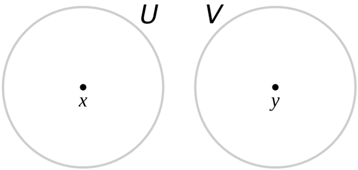
$$\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{X}} C. \tag{26}$$

For $A \subseteq \mathcal{X}$ the *subspace topology* on A: $\tau_A = \{O \cap A : O \in \tau\}$.

Hausdorff space

$(\mathcal{X}, \tau)$ 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.



Dense subset, separability, basis of a topology, Polish

- $A \subseteq \mathcal{X}$ is *dense* if $\bar{A} = \mathcal{X}$.
- $(\mathcal{X}, \tau)$ is *separable* if $\exists$ countable, dense subset of $\mathcal{X}$. Counterexample: l^∞ / L^∞.
- $\tau_1 \subseteq \tau$ is a *basis* of τ if every open is union of sets in τ_1. Example: open balls in a metric space.
- $(\mathcal{X}, \tau)$ is *Polish* if τ has a countable basis and $\exists$ metric defining τ. Example: complete separable metric spaces.

Environment, locally compact 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 $\Leftrightarrow |\mathcal{X}| < \infty$.
- 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.

The discrete space

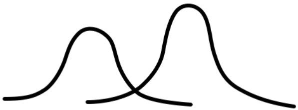
- $(\mathcal{X}, 2^{\mathcal{X}})$: complete metric space.
- Discrete metric (inducing the discrete topology):

$$d(x, y) = \begin{Bmatrix} 0, \text{ if } x = y \\ 1, \text{ if } x \neq y \end{Bmatrix}. \tag{27}$$

- Discrete space: separable $\Leftrightarrow$ $|\mathcal{X}|$ is countable.

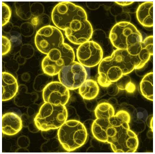
Example: hyperparameter selection

- Training: samples from MOGs with component number labels.
- Task:
  - given: samples from a new MOG distribution,
  - predict: the number of components.



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.



Kernel k: c_0-universal

Let $C_0(\mathcal{D}) = \mathcal{D} \to \mathbb{R}$ continuous functions vanishing at infinity, i.e.,

$$\{u \in \mathcal{D} : |g(u)| \geq \epsilon\} \tag{28}$$

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.