Distribution Regression – Make It Simple and Consistent*

Zoltán Szabó[1], Bharath K. Sriperumbudur[2], Barnabás Póczos[3], Arthur Gretton[1]

[1]Gatsby Unit, UCL [2]Department of Statistics, PSU [3]Machine Learning Department, CMU

Problem

- Distribution regression:
  - Input = distribution, output $\in \mathbb{R}/\mathbb{R}^d$/separable Hilbert space.
  - Challenge: sampled input distributions.
- Examples:
  - multiple instance learning (MIL),
  - point estimates of statistics (entropy/hyperparameter/...).
- Existing methods: heuristics, or require density estimation (which typically scale poorly in dimension).

Distribution Regression

- $D(\mathfrak{X})$ distributions on domain $(\mathfrak{X}, k)$.
- $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^l \overset{i.i.d.}{\sim} \mathcal{M}$: $(x_i, y_i) \in D(\mathfrak{X}) \times Y$.
- Given: $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^N, y_i)\}_{i=1}^l$, where $\{x_{i,n}\}_{n=1}^N \overset{i.i.d.}{\sim} x_i$.

- **Goal**: learn the relation between (x, y) given $\hat{\mathbf{z}}$.
- **Idea**: $D(\mathfrak{X}) \xrightarrow{\mu} X \subseteq H(k) \xrightarrow{f \in \mathcal{H}(K)} Y$.
- **Mean embedding**: $\mu_x = \int_{\mathfrak{X}} k(\cdot, u) \mathrm{d}x(u)$.

Objective Function, Algorithm

- **Cost function** (of MERR):

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

- Analytical **solution**: prediction on a new distribution t

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

- **Example**: If $Y = \mathbb{R}^d$, then $\mathcal{L}(Y) = \mathbb{R}^{d \times d}$.

Goal in Details

- **Regression function**: $f_\rho(\mu_a) = \int_Y y \mathrm{d}\rho(y|\mu_a)$.
- **Contribution**: analysis of the excess risk

$$\tilde{\mathcal{E}}(f^\lambda_{\hat{\mathbf{z}}}, f_\rho) = \mathcal{E}[f^\lambda_{\hat{\mathbf{z}}}] - \mathcal{E}[f_\rho] \le g(l, N, \lambda) \to 0 \text{ and rates,}$$
$$\mathcal{E}[f] = \mathbb{E}_{(x,y)} \|f(\mu_x) - y\|_Y^2 \text{ (expected risk).}$$

Blanket Assumptions

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

Example: If $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H \Rightarrow$ we get the set kernel

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \frac{1}{N^2} \sum_{n,m=1}^N k(x_{i,n}, x_{j,m}).$$

Performance Guarantees

- **Well-specified case** ($f_\rho \in \mathcal{H}$): f_ρ is 'c-smooth' with 'b-decaying covariance operator' and $l \ge \lambda^{-\frac{1}{b}-1}$, then

$$\tilde{\mathcal{E}}(f^\lambda_{\hat{\mathbf{z}}}, f_\rho) \le \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^{\frac{1}{b}}}.$$

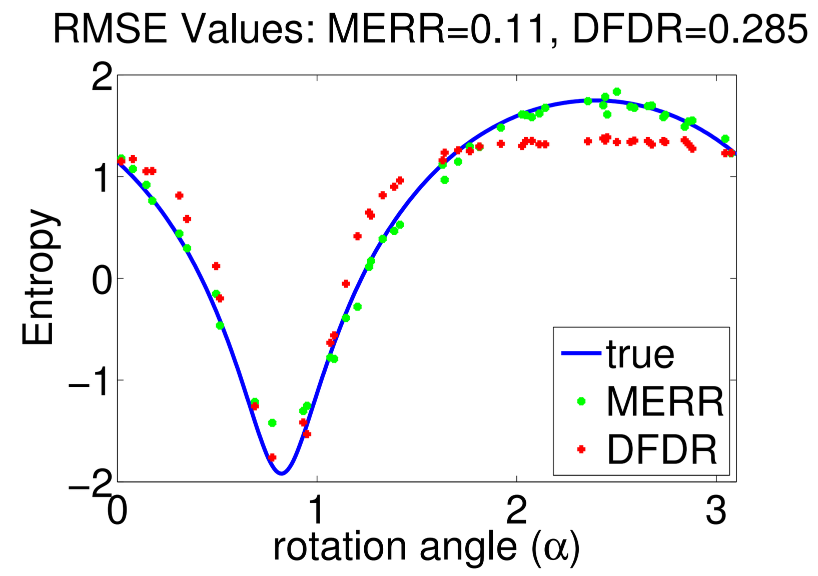
- **Misspecified case** ($f_\rho \in L^2_{\rho_X} \backslash \mathcal{H}$): f_ρ is 's-smooth', $L^2_{\rho_X}$ is separable, and $\frac{1}{\lambda^2} \le l$, then

$$\tilde{\mathcal{E}}(f^\lambda_{\hat{\mathbf{z}}}, f_\rho) \le \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}} \lambda^{\frac{3}{2}}} + \frac{1}{\sqrt{l\lambda}} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda \sqrt{l}} + \lambda^{\min(1,s)}.$$

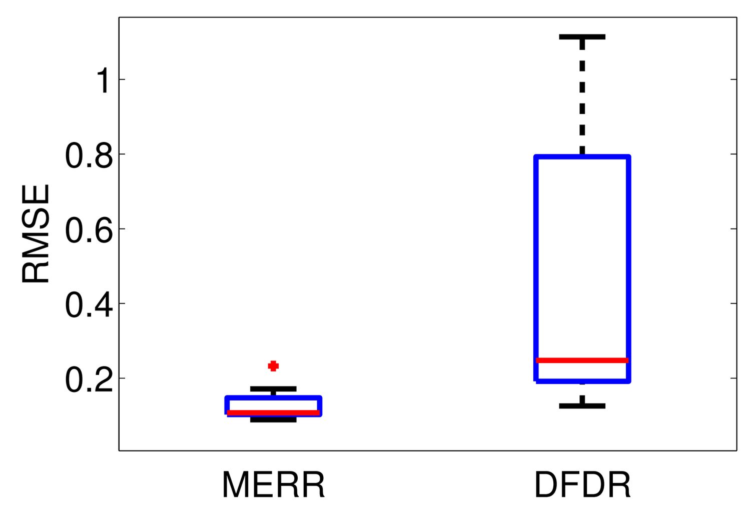
Applications

Supervised entropy learning:

- Label = entropy of the distribution represented by a bag.



(a) Entropy of Gaussian



(b) Boxplot of RMSE

Aerosol prediction:

- Bag = multispectral satellite image 'pixels' over an area.
- Label = aerosol value (highly accurate, expensive ground-based instrument).
- Performance:

Method	100×RMSE	±std
Baseline [mixture model (EM)]	7.5 − 8.5	±0.1 − 0.6
MERR: linear K, single	7.91	±1.61
MERR: linear K, ensemble	**7.86**	±1.71
MERR: nonlinear K, single	7.90	±1.63
MERR: nonlinear K, ensemble	**7.81**	±1.64

Code: in ITE (`https://bitbucket.org/szzoli/ite/`).

Acknowledgements

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.

*Data, Learning and Inference workshop (DALI), La Palma (Canaries, Spain), Apr. 10-12, 2015.