Approximate Kernel Embeddings of Distributions

Dino Sejdinovic

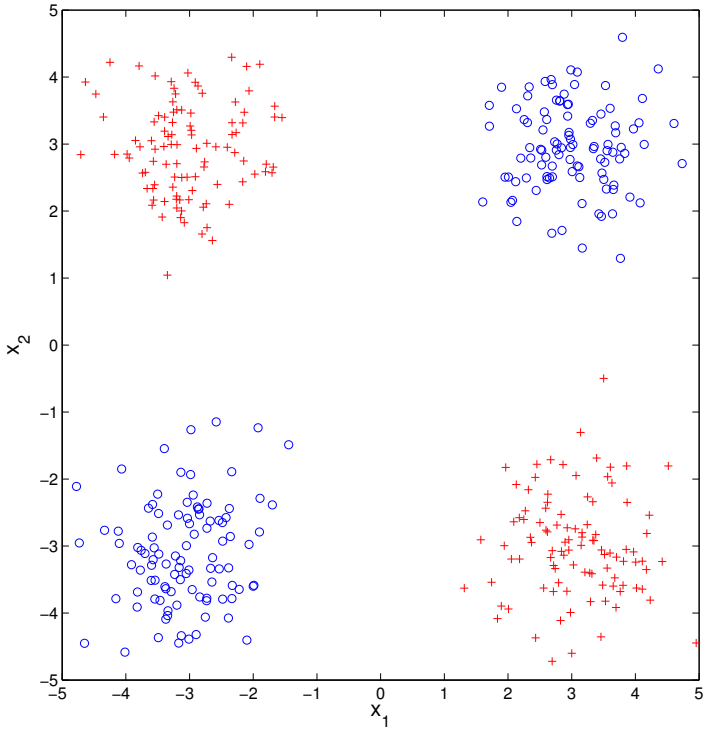
Department of Statistics
University of Oxford

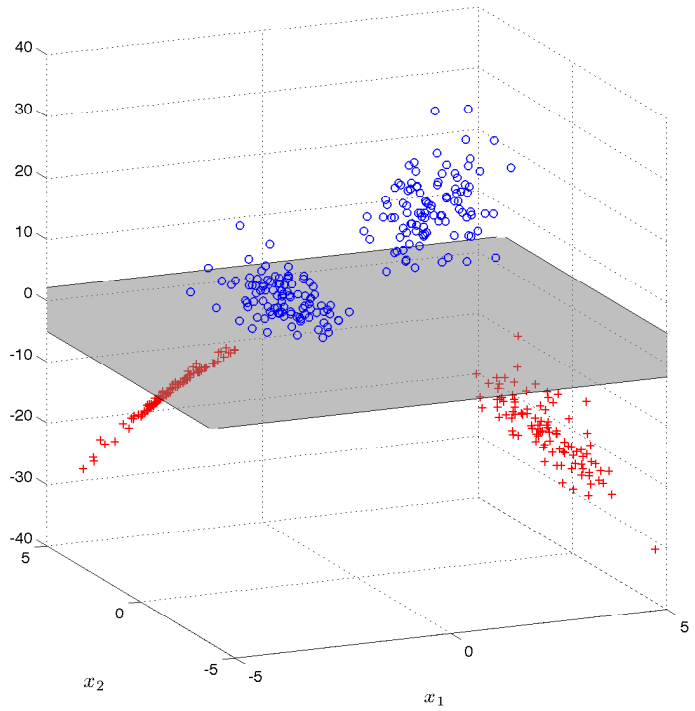
SFB Colloquium, Universität Potsdam
25/05/2018

Outline

Outline

Feature maps





- No linear classifier separates red from blue.
- Linear separation after mapping to a **higher dimensional feature space**:

$$\mathbb{R}^2 \ni \begin{pmatrix} x^{(1)} & x^{(2)} \end{pmatrix}^\top = x \quad \mapsto \quad \varphi(x) = \begin{pmatrix} x^{(1)} & x^{(2)} & x^{(1)}x^{(2)} \end{pmatrix}^\top \in \mathbb{R}^3$$

Feature maps and kernel trick

- Kernel methods on a generic domain $\mathcal{X}$ allow constructing nonlinear methods after mapping to a higher dimensional feature space:

$$\varphi : \mathcal{X} \to \mathbb{R}^D$$

- Typically need only inner products $\varphi(x_i)^\top \varphi(x_j)$ are required and the coordinates of the maps $\varphi(x_i) \in \mathbb{R}^D$ need not be computed explicitly - inner product between features can be a simple function (kernel) of x_i and x_j.
- For example, polynomial kernel $k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j) = (1 + x_i^\top x_j)^q$ on $\mathbb{R}^p$ computes q-order features - never need to compute explicit feature expansion of dimension $D = \binom{p+q}{q}$ where this inner product is defined.
- Formally, a (reproducing) kernel k is any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.

Reproducing Kernel Hilbert Space (RKHS)

Definition ([Aronszajn, 1950; Berlinet & Thomas-Agnan, 2004])

Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space*.

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$. Thus $\mathcal{H}$ servers as a canonical *feature space* with feature map $x \mapsto k(\cdot, x)$.

- Equivalently, all evaluation functionals $f \mapsto f(x)$ are continuous (norm convergence implies pointwise convergence).
- Moore-Aronszajn Theorem: every positive semidefinite $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel and has a unique RKHS $\mathcal{H}_k$.

Reproducing Kernel Hilbert Space (RKHS)

Definition ([Aronszajn, 1950; Berlinet & Thomas-Agnan, 2004])

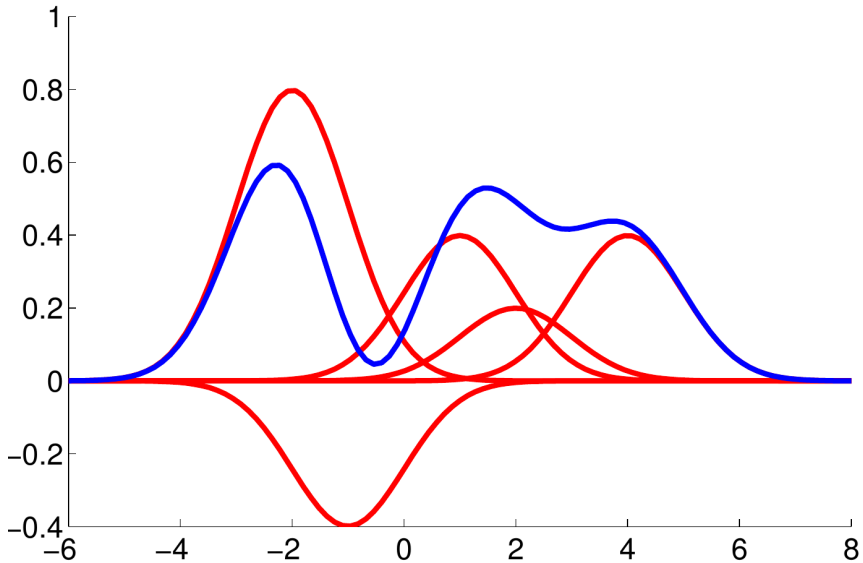
Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space*.

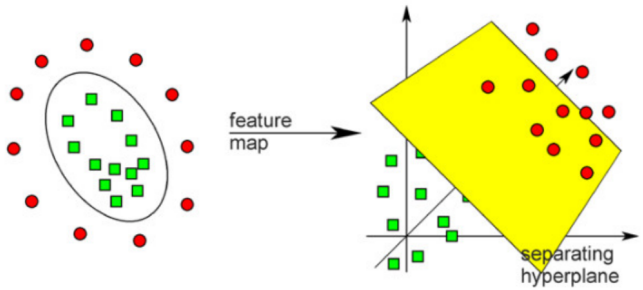
Gaussian RBF kernel $k(x, x') = \exp\left(-\frac{1}{2\gamma^2} \|x - x'\|^2\right)$ has an infinite-dimensional $\mathcal{H}$ with elements $h(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$ and their limits which give completion with respect to the inner product

$$\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(y_j, \cdot) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j).$$



Kernel Trick and Kernel Mean Trick

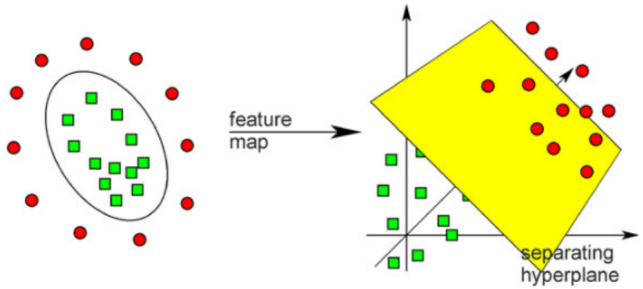
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

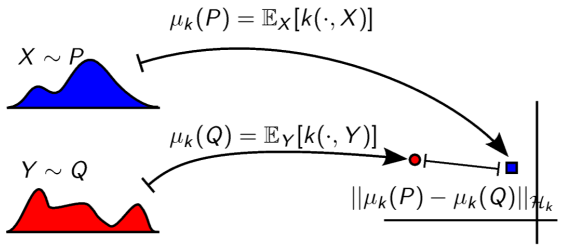
Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \dots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

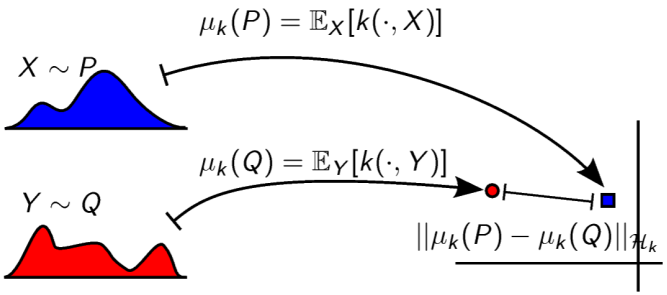
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010; Muandet et al, 2017]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \dots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions

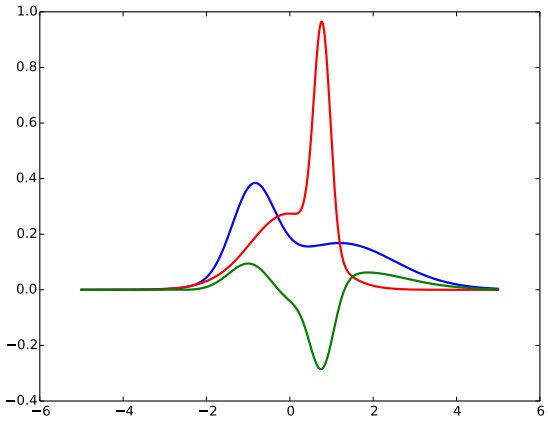


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

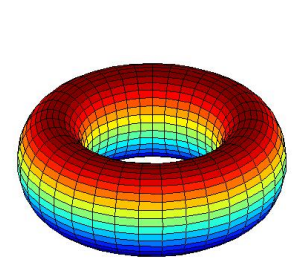
- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:

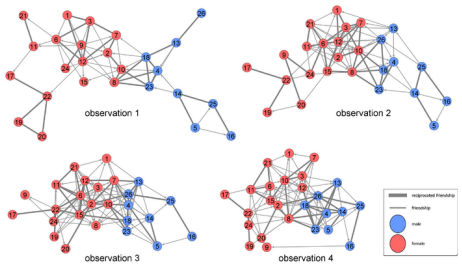




$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E} f(X) - \mathbb{E} f(Y)|$$

- Characteristic kernels: $\mathrm{MMD}_k(P, Q) = 0$ iff $P = Q$ (also metrizes weak* [Sriperumbudur,2010]).
  - Gaussian RBF $\exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$, Matérn family, inverse multiquadrics.
- Can encode structural properties in the data: kernels on non-Euclidean domains, networks, images, text...





Lorem ipsum dolor sit amet, consectetur adipiscing elit. posuere tortor vitae elit. Sed vitae metus a elit bibendum malesuada cras pulvinar. Quisque pellentesque nibh in sem. Curabitur ligula. Suspendisse potenti. Duis sit amet augue eu arcu ultrices auctor. Suspendisse elementum, nunc ut molestie elementum, neque augue vulputate elit, eu blandit enim velit vitae nulla. Duis sed.

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

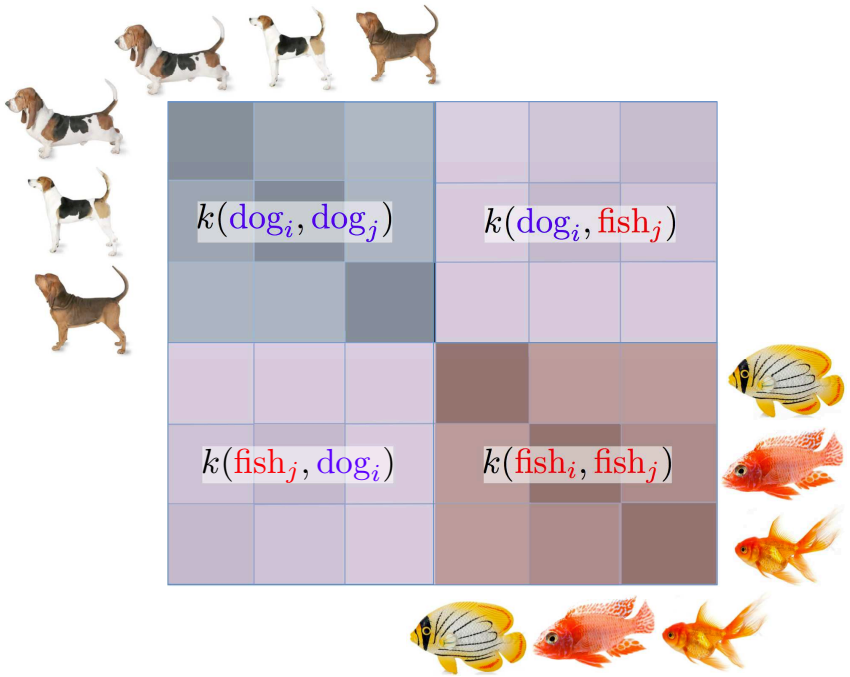


Figure by Arthur Gretton

MMD has been applied to:

- two-sample tests and independence tests (on graphs, text, audio...) [Gretton et al, 2009, Gretton et al, 2012]
- model criticism and interpretability [Lloyd & Ghahramani, 2015; Kim, Khanna & Koyejo, 2016]
- analysis of Bayesian quadrature [Briol et al, 2018]
- ABC summary statistics [Park, Jitkrittum & DS, 2015; Mitrovic, DS & Teh, 2016]
- summarising streaming data [Paige, DS & Wood, 2016]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- MMD-GAN: training deep generative models [Dziugaite, Roy & Ghahramani, 2015; Sutherland et al, 2017; Li et al, 2017]

$$\text{MMD}_k^2(P, Q) = \mathbb{E}_{X,X' \overset{i.i.d.}{\sim} P} k(X, X') + \mathbb{E}_{Y,Y' \overset{i.i.d.}{\sim} Q} k(Y, Y') - 2\mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).$$

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

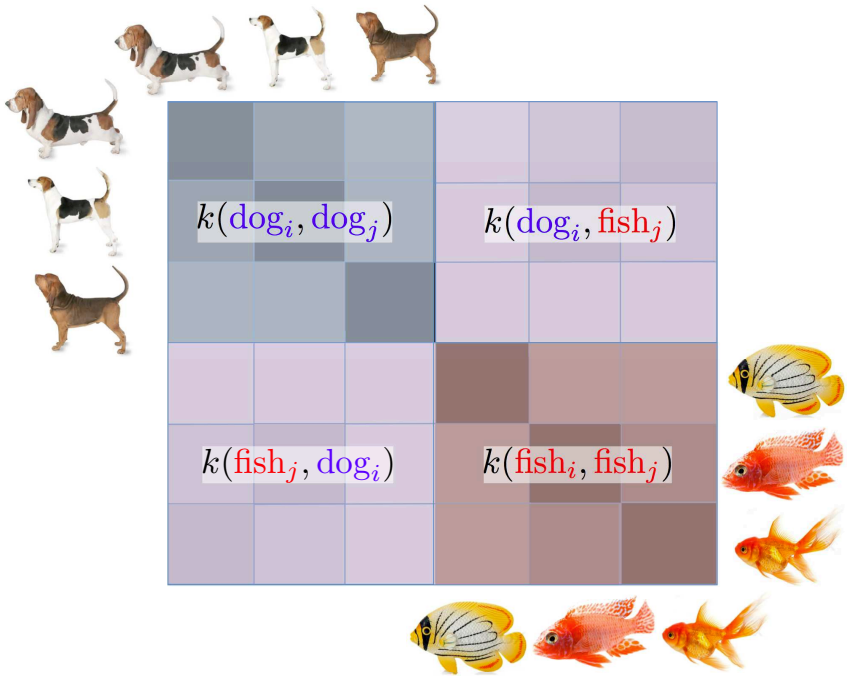


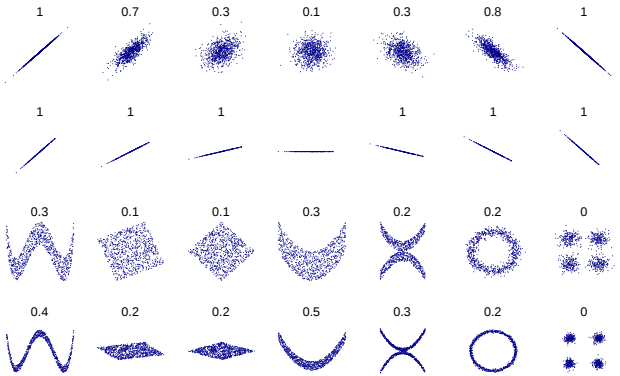
Figure by Arthur Gretton

MMD has been applied to:

- two-sample tests and independence tests (on graphs, text, audio...) [Gretton et al, 2009, Gretton et al, 2012]
- model criticism and interpretability [Lloyd & Ghahramani, 2015; Kim, Khanna & Koyejo, 2016]
- analysis of Bayesian quadrature [Briol et al, 2018]
- ABC summary statistics [Park, Jitkrittum & DS, 2015; Mitrovic, DS & Teh, 2016]
- summarising streaming data [Paige, DS & Wood, 2016]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- MMD-GAN: training deep generative models [Dziugaite, Roy & Ghahramani, 2015; Sutherland et al, 2017; Li et al, 2017]

$$\widehat{\text{MMD}}_k^2(P, Q) = \frac{1}{n_x(n_x-1)} \sum_{i \neq j} k(X_i, X_j) + \frac{1}{n_y(n_y-1)} \sum_{i \neq j} k(Y_i, Y_j) - \frac{2}{n_x n_y} \sum_{i,j} k(X_i, Y_j).$$

Kernel dependence measures: HSIC



cor vs. dcor

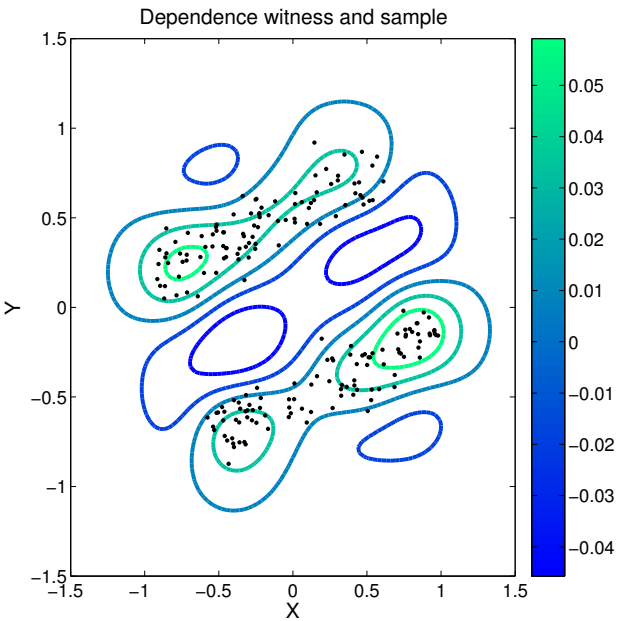
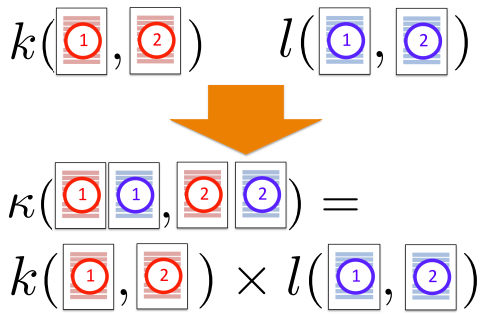


Figure by Arthur Gretton

- $HSIC^2(X, Y; \kappa) = \|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|^2_{\mathcal{H}_\kappa}$
- Hilbert-Schmidt norm of the feature-space cross-covariance [Gretton et al, 2009]
- dependence witness is a smooth function in the RKHS $\mathcal{H}_\kappa$ of functions on $\mathcal{X} \times \mathcal{Y}$



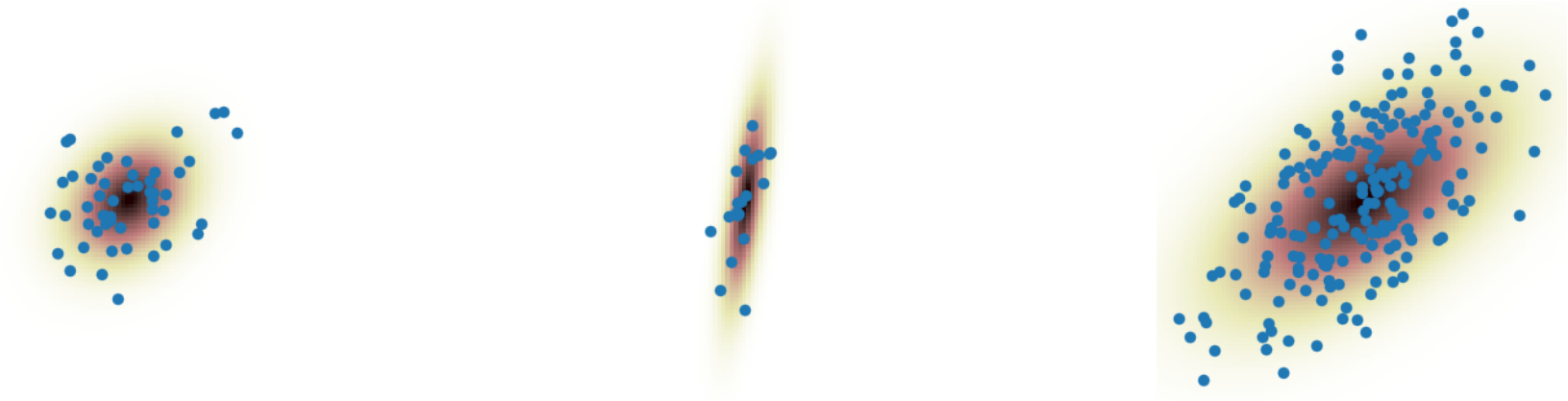
- Independence testing framework that generalises Distance Correlation (dcor) of [Szekely et al, 2007]: HSIC with Brownian motion kernels [DS et al, 2013]
- Extends to multivariate interaction and joint dependence measures [DS et al, 2013; Pfister et al, 2017]

Distribution Regression

-0.856 0.562 1.39

- Labels $y_i = f(P_i)$ but observe only $\{x_i^j\}_{j=1}^{N_i} \sim P_i$.
- The goal: build a predictive model $\hat{y}_\star = f(\{x_\star^j\}_{j=1}^{N_\star})$ for a new sample $\{x_\star^j\}_{j=1}^{N_\star} \sim P_\star$.
- Represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Now can use the induced inner product structure on empirical measures to build a regression model:
  - Linear kernel on the RKHS: $K(\hat{\mu}_i, \hat{\mu}_j) = \langle \hat{\mu}_i, \hat{\mu}_j \rangle_{\mathcal{H}_k} = \frac{1}{N_i N_j} \sum_{r,s} k(x_i^r, x_j^s)$
  - Gaussian kernel on the RKHS: $K(\hat{\mu}_i, \hat{\mu}_j) = \exp(-\gamma \|\hat{\mu}_i - \hat{\mu}_j\|_{\mathcal{H}_k}^2)$

Distribution Regression



-0.856 0.562 1.39

- Labels $y_i = f(P_i)$ but observe only $\{x_i^j\}_{j=1}^{N_i} \sim P_i$.
- The goal: build a predictive model $\hat{y}_\star = f(\{x_\star^j\}_{j=1}^{N_\star})$ for a new sample $\{x_\star^j\}_{j=1}^{N_\star} \sim P_\star$.
- Represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Now can use the induced inner product structure on empirical measures to build a regression model:
  - Linear kernel on the RKHS: $K(\hat{\mu}_i, \hat{\mu}_j) = \langle \hat{\mu}_i, \hat{\mu}_j \rangle_{\mathcal{H}_k} = \frac{1}{N_i N_j} \sum_{r,s} k(x_i^r, x_j^s)$
  - Gaussian kernel on the RKHS: $K(\hat{\mu}_i, \hat{\mu}_j) = \exp(-\gamma \|\hat{\mu}_i - \hat{\mu}_j\|_{\mathcal{H}_k}^2)$

Distribution Regression

- supervised learning where labels are available at the group, rather than at the individual level.

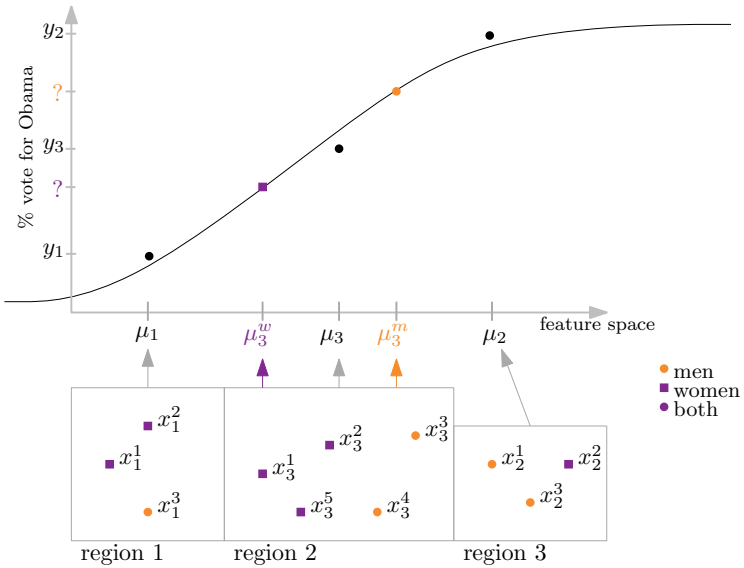


Figure from Flaxman et al, 2015

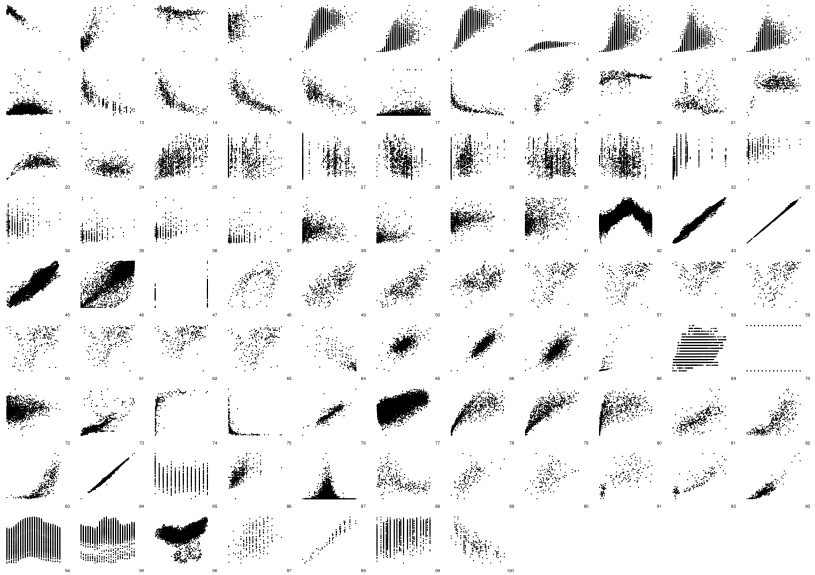


Figure from Mooij et al, 2014

- classifying text based on word features [Yoshikawa et al, 2014; Kusner et al, 2015]
- aggregate voting behaviour of demographic groups [Flaxman et al, 2015; 2016]
- image labels based on a distribution of small patches [Szabo et al, 2016]
- "traditional" parametric statistical inference by learning a function from sets of samples to parameters: ABC [Mitrovic et al, 2016], EP [Jitkrittum et al, 2015]
- identify the cause-effect direction between a pair of variables from a joint sample [Lopez-Paz et al,2015]

Kernel methods at scale

- Expressivity of kernel methods comes at a price of $O(n^2)$ or $O(n^3)$ in the number of observations n (due to having to compute, store and often invert the Gram matrix)
- Problematic when we have a lot of observations (and this is exactly when we want to use a rich expressive model with a high-dimensional hypothesis class!)
- Scaling up kernel methods is a very active research area

  [Sonnenburg et al, 2006; Rahimi & Recht, 2007; Le, Sarlos & Smola, 2013; Wilson et al, 2014; Dai et al, 2014; Sriperumbudur & Szabo, 2015; Bach, 2015; Avron et al, 2017].
- Main idea: study the RKHS and construct a (random) low-dimensional space with similar inner product structure for a given data - then undo the kernel trick(!?)

explicit basis functions

$\downarrow$

implicit basis functions

$\downarrow$

explicit random basis functions

Random Fourier features: Inverse Kernel Trick

Bochner's representation: Assume that k is a positive definite translation-invariant kernel on $\mathbb{R}^p$. Then k can be written as

$$\begin{aligned} k(x,y) &= \int_{\mathbb{R}^p} \exp\left(i\omega^\top (x-y)\right) d\Lambda(\omega) \\ &= 2\int_{\mathbb{R}^p} \left\{\cos\left(\omega^\top x\right)\cos\left(\omega^\top y\right) + \sin\left(\omega^\top x\right)\sin\left(\omega^\top y\right)\right\} d\Lambda(\omega) \end{aligned}$$

for some positive measure (w.l.o.g. a probability distribution) Λ.

- Sample m frequencies $\Omega = \{\omega_j\}_{j=1}^m \sim \Lambda$ and use a Monte Carlo estimator of the kernel function instead [Rahimi & Recht, 2007]:

$$\begin{aligned} \hat{k}(x,y) &= \frac{2}{m}\sum_{j=1}^{m} \left\{\cos\left(\omega_j^\top x\right)\cos\left(\omega_j^\top y\right) + \sin\left(\omega_j^\top x\right)\sin\left(\omega_j^\top y\right)\right\} \\ &= \langle \xi_\Omega(x), \xi_\Omega(y)\rangle_{\mathbb{R}^{2m}}, \end{aligned}$$

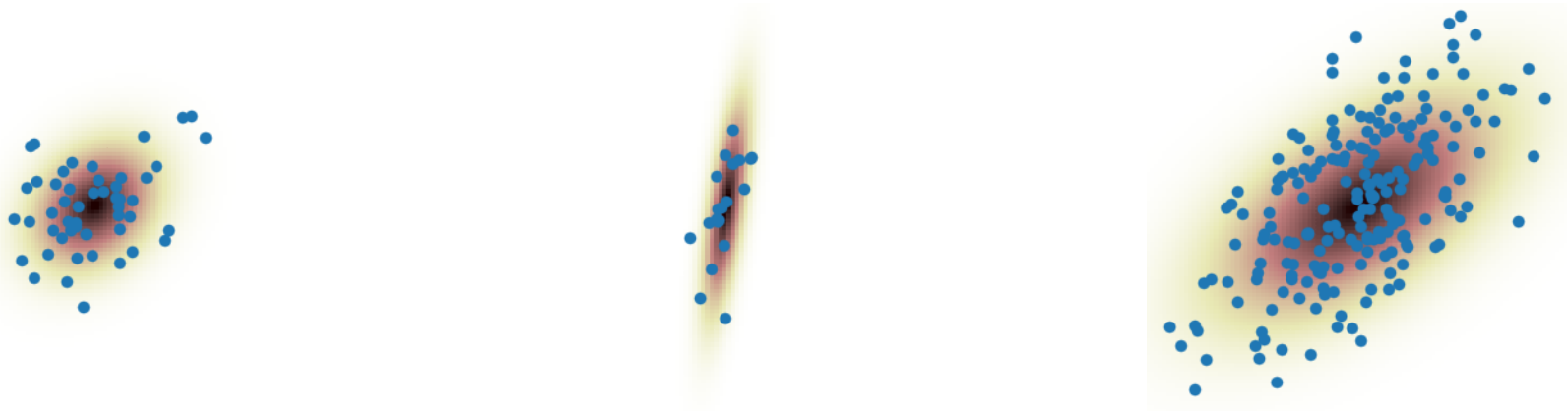
  with an explicit set of features $\xi_\Omega : x \mapsto \sqrt{\frac{2}{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots\right]^\top$.
- The cost drops: $O(n^3) \to O(m^2 n + m^3)$, $O(n^2) \to O(mn + m^2)$. How fast does m need to grow with n? Often sublinear and can be as low as $\log n$ without sacrificing convergence rates [Bach, 2015; Rudi et al, 2017, Avron et al, 2017].

This talk:

- How to model uncertainty of kernel embeddings in distribution regression?
  - A simple Bayesian model for kernel mean embeddings leads to shrinkage estimators with better predictive performance in high noise regimes.
- When measuring nonparametric distances between distributions, can we disentangle the differences in the noise from the differences in the signal?
  - Weighted distance between the empirical phase functions can lead to distribution regression which is more robust to changes in measurement noise.

Outline

Uncertainty in Bag Sizes



-0.856 0.562 1.39

- Recall: we represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Empirical mean in infinite-dimensional space? Stein's phenomenon? Shrinkage estimators can be better behaved [Muandet et al, 2013]
- These inputs (with or without shrinkage) are *noisy* - we do not observe the true embedding μ_i. Moreover, bags with small N_i are noisier - can this uncertainty be included in the predictive model?

Bayesian Approaches to Distribution Regression
Ho Chung Leon Law, Dougal Sutherland, DS, and Seth Flaxman
AISTATS 2018
http://proceedings.mlr.press/v84/law18a.html

Uncertainty in Mean Embeddings

- The empirical mean embedding is $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$
- Bayesian model for kernel mean embeddings [Flaxman,DS,Cunningham & Filippi, UAI 2016]:
  - Place prior on the RKHS $\mu_i \sim GP\left(m_0(\cdot), r(\cdot,\cdot)\right)$ (requires care due to 0/1 laws [Kallianpur, 1970; Wahba, 1990; Steinwart, 2014+])
  - Posit normal likelihood for the *evaluations of the embedding* at a set of points $\mathbf{u}$:

$$\hat{\mu}_i(\mathbf{u}) | \mu_i(\mathbf{u}) \sim \mathcal{N}(\mu_i(\mathbf{u}), \Sigma_i / N_i)$$

  - Leads to a closed-form GP posterior $\mu_i | \{x_i^j\}$:

$$\mu_i(\mathbf{z}) | \{x_i^j\} \sim \mathcal{N}\Big(R_{\mathbf{zu}}(R_{\mathbf{uu}} + \Sigma_i/N_i)^{-1}(\hat{\mu}_i - m_0) + m_0, \\ R_{\mathbf{zz}} - R_{\mathbf{zu}}(R_{\mathbf{uu}} + \Sigma_i/N_i)^{-1} R_{\mathbf{uz}} \Big)$$

  - Recovers frequentist shrinkage estimator of mean embeddings [Muandet et al, 2013] (but with r instead of k), similar to James-Stein estimator.

Distribution Regression Model

- Model label as a function of the "true" kernel mean embedding:

$$y_i = f(\mu_i) + \epsilon, \qquad \mu_i = \mathbb{E}_{X \sim P_i} k(\cdot, X)$$

- Linear model on the evaluation of kernel mean embedding at a set of "landmark points" $\mathbf{z}$:

$$f(\mu_i) = \beta^\top \mu_i(\mathbf{z})$$

- Can model uncertainty in β (BLR) or in μ_i (shrinkage) or in both (BDR, which requires MCMC due to non-conjugacy).
- **Shrinkage**: Integrate likelihood $y_i \sim \mathcal{N}(f(\mu_i), \sigma^2)$ through the posterior $\mu_i | \{x_i^j\}$ to obtain

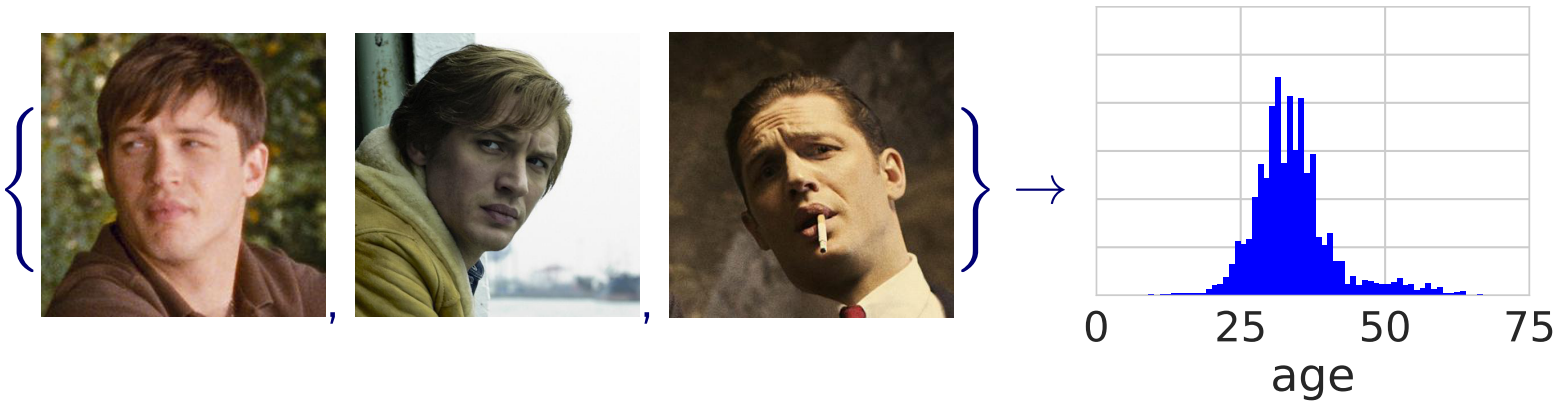
$$y_i \mid \{x_i^j\}, \beta \sim \mathcal{N}(\xi_i^\beta, \nu_i^\beta)$$

$$\xi_i^\beta = \beta^\top R_{\mathbf{z}\mathbf{x}_i} \left(R_{\mathbf{x}_i\mathbf{x}_i} + \frac{\Sigma_i}{N_i} \right)^{-1} (\hat{\mu}_i - m_0) + \beta^\top m_0$$

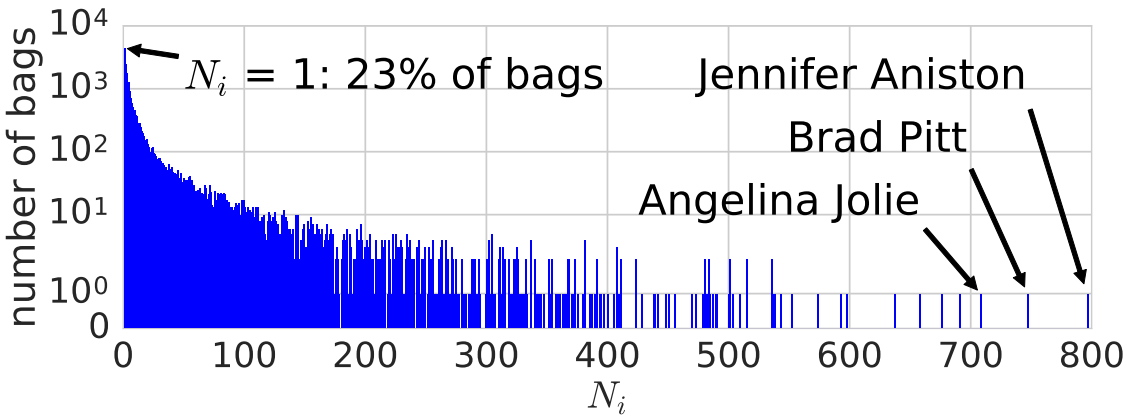
$$\nu_i^\beta = \beta^\top \left(R_{\mathbf{z}\mathbf{z}} - R_{\mathbf{z}\mathbf{x}_i} \left(R_{\mathbf{x}_i\mathbf{x}_i} + \frac{\Sigma_i}{N_i} \right)^{-1} R_{\mathbf{x}_i\mathbf{z}}^\top \right) \beta + \sigma^2.$$

- Can be optimized to find MAP of β, σ^2, kernel parameters, locations of landmark points, ...

Age prediction from images



- IMDb-Wiki database of images with age labels
  - *Very* noisy labels in the dataset
- Distribution regression: group pictures of actors, predict *mean age*
- Image features: last hidden layer from a convolutional neural network by [Rothe et al, IJCV 2016]
- *Lots* of variation in N_i:



Age prediction from images

Propagating uncertainty using shrinkage helps!

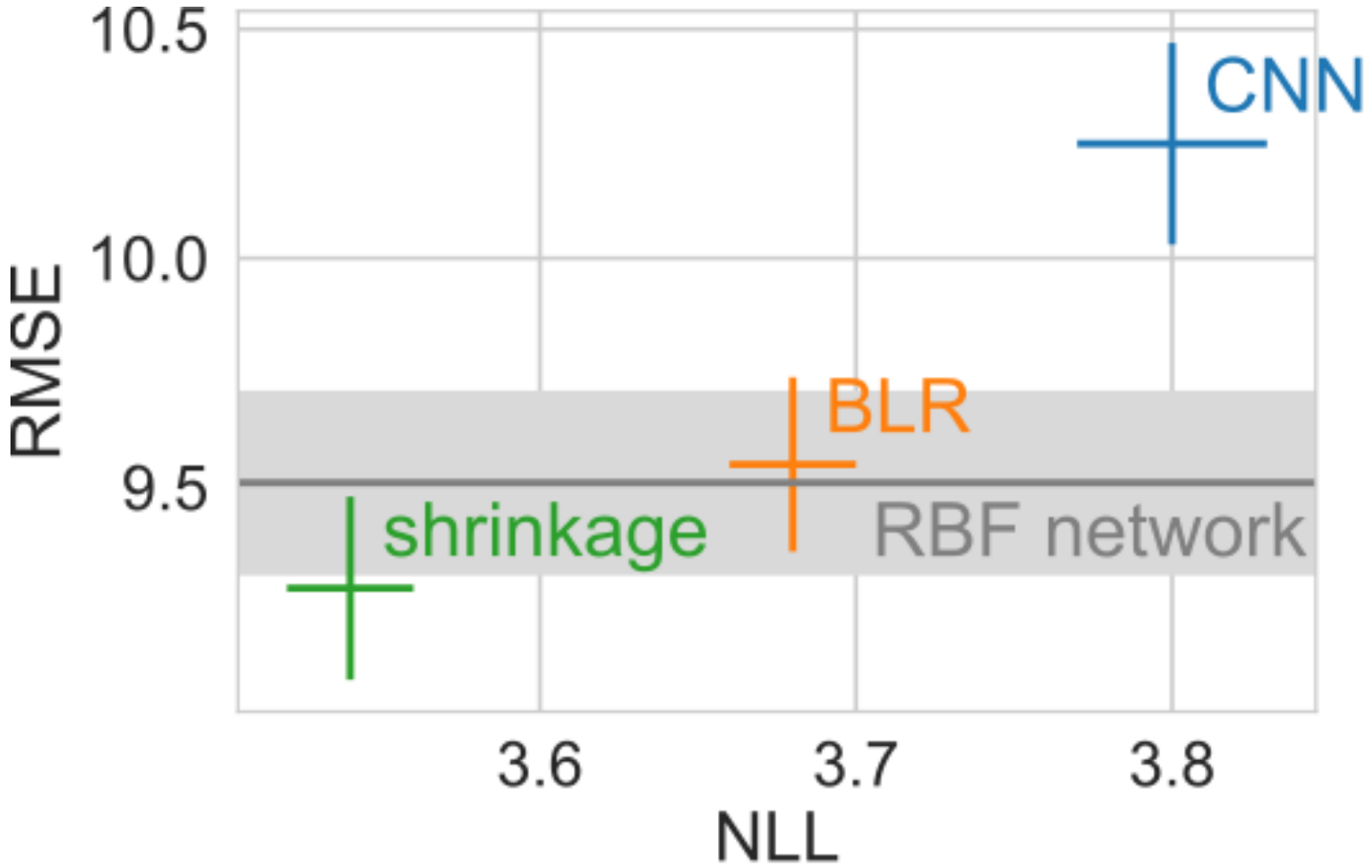


Figure: Results across ten data splits (means and standard deviations). RBF net is tuned for RMSE, other methods for NLL. CNN takes the mean of the predictive distributions of [Rothe-IJCV-2016] for each point in the bag.

Outline

The problem with test bags

- supervised learning where labels are available at the group, rather than at the individual level.

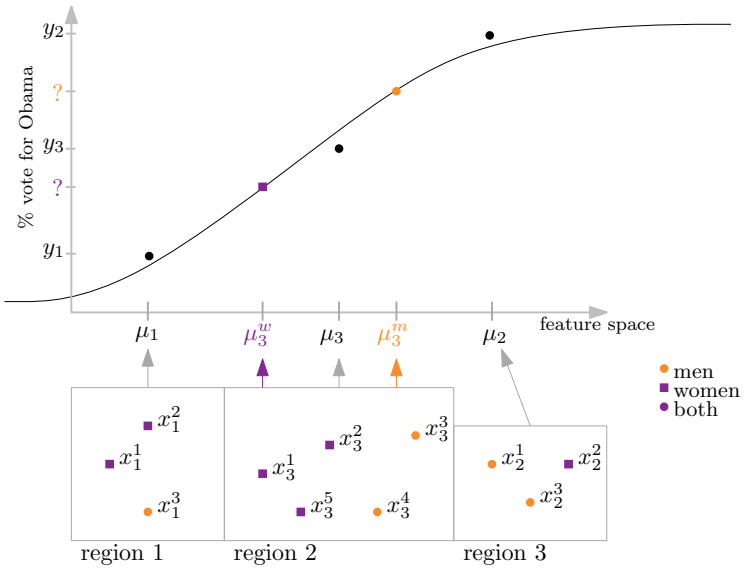


Figure from Flaxman et al, 2015

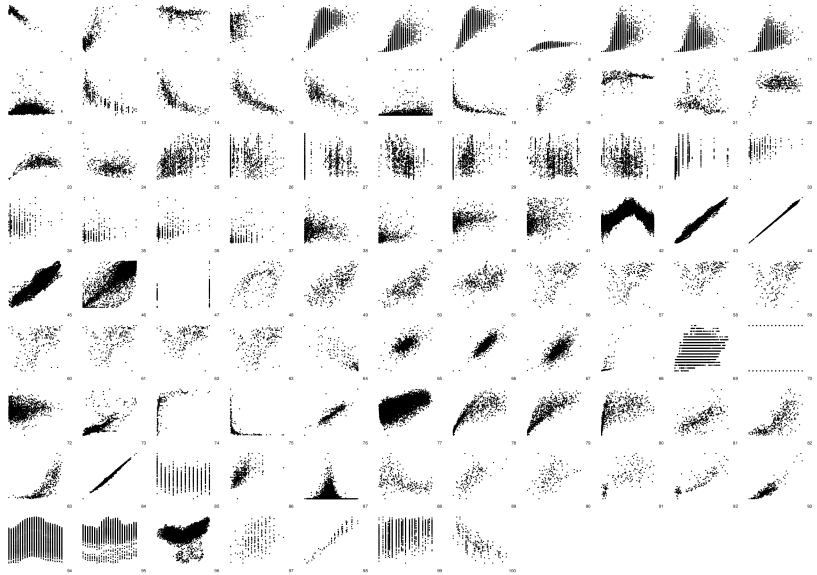
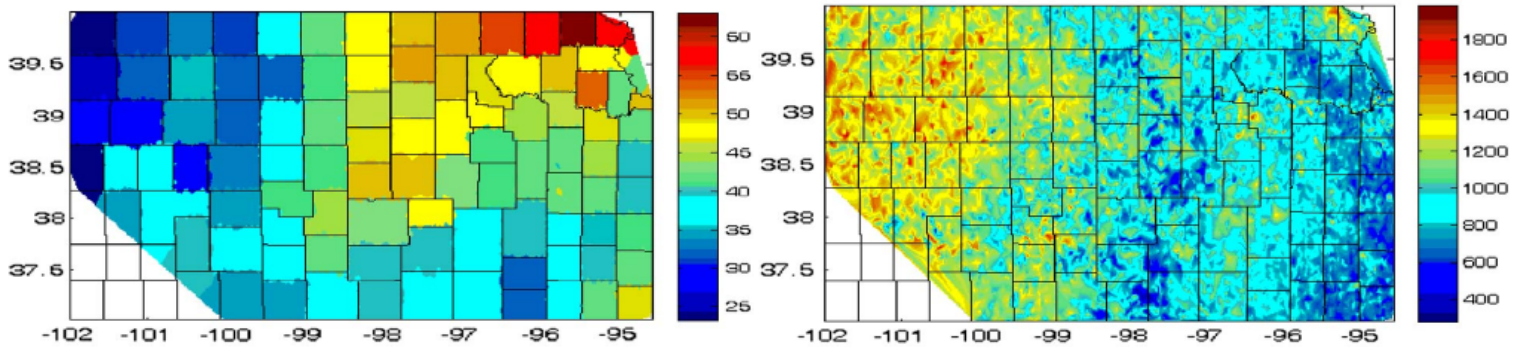


Figure from Mooij et al, 2014

  - aggregate voting behaviour of demographic groups [Flaxman et al, 2015; 2016]
  - identify the cause-effect direction between a pair of variables from a joint sample [Lopez-Paz et al,2015]
- Possible (distributional) covariate shift?

Bag-specific noises in Distribution Regression



figures from Wang et al, 2012

Aerosol / Crop Yield prediction from multispectral data [Wang et al, 2012]:

- The labels y_i provided by the ground sensors (aerosol), government records (crop yield)
- "Multiple-instance regression": randomly subsample multispectral (16-dim) pixels (satellite imaging data) within 20km radius of ground sensor (aerosol) / within a county (crop yield)
- *Large image variability* due to surface properties.
- Arguably different noise distribution ("cloudy pixels") in different images.

All possible differences between generating processes?

- differences between embeddings can be due to different types of measurement noise or data collection artefacts
  - With a large sample-size, uncovers potentially irrelevant sources of variability
- Covariate shift in distribution regression?
  - Each bag of observations could be impaired by a *different measurement noise process*. Also, test bags could have different measurement noise than train bags.
- Both problems require learning a representation *invariant to some form of a noise model* (here we will assume that the noise is symmetric and the signal is asymmetric).

Testing and Learning on Distributions with Symmetric Noise Invariance
Ho Chung Leon Law, Chris Yau, and DS
NIPS 2017
https://arxiv.org/abs/1703.07596

Random Fourier features: Inverse Kernel Trick

Bochner's representation: Assume that k is a positive definite **translation-invariant** kernel on $\mathbb{R}^p$. Then k can be written as

$$\begin{aligned} k(x,y) &= \int_{\mathbb{R}^p} \exp\left(i\omega^\top (x-y)\right) d\Lambda(\omega) \\ &= 2\int_{\mathbb{R}^p} \left\{\cos\left(\omega^\top x\right)\cos\left(\omega^\top y\right) + \sin\left(\omega^\top x\right)\sin\left(\omega^\top y\right)\right\} d\Lambda(\omega) \end{aligned}$$

for some positive measure (w.l.o.g. a probability distribution) Λ.

- Sample m frequencies $\Omega = \{\omega_j\}_{j=1}^m \sim \Lambda$ and use a Monte Carlo estimator of the kernel function instead [Rahimi & Recht, 2007]:

$$\begin{aligned} \hat{k}(x,y) &= \frac{2}{m}\sum_{j=1}^{m} \left\{\cos\left(\omega_j^\top x\right)\cos\left(\omega_j^\top y\right) + \sin\left(\omega_j^\top x\right)\sin\left(\omega_j^\top y\right)\right\} \\ &= \langle \xi_\Omega(x), \xi_\Omega(y)\rangle_{\mathbb{R}^{2m}}, \end{aligned}$$

with an explicit set of features $\xi_\Omega : x \mapsto \sqrt{\frac{2}{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots\right]^\top$.

Characteristic Functions and (Approximate) Kernel Embeddings

Approximate mean embedding using random Fourier features [Rahimi & Recht, 2007] is simply the evaluation (real and complex part stacked together) of the characteristic function at the frequencies $\{\omega_j\}_{j=1}^m \sim \Lambda$:

$$\begin{aligned}\Phi(P) &= \mathbb{E}_{X\sim P}\xi_\Omega(X) \\ &= \sqrt{\frac{2}{m}}\mathbb{E}_{X\sim P}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots, \cos\left(\omega_m^\top x\right), \sin\left(\omega_m^\top x\right)\right]^\top\end{aligned}$$

If k is translation-invariant, MMD becomes the weighted L_2-distance between the characteristic functions of P and Q [Sriperumbudur, 2010].

$$\|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 = \int_{\mathbb{R}^d} |\varphi_P(\omega) - \varphi_Q(\omega)|^2 \, d\Lambda(\omega),$$

Used for distribution regression [Sutherland et al, 2015] and for sketching / compressive learning [Keriven et al, 2016].

The Noise and the Signal

The key idea comes from nonparametric deconvolution of [Delaigle and Hall, 2016].

- define a *symmetric positive definite (SPD) noise component* to be any random vector E on $\mathbb{R}^d$ with a positive characteristic function, $\varphi_E(\omega) = \mathbb{E}_{X \sim E}\left[\exp(i\omega^\top E)\right] > 0$, $\forall \omega \in \mathbb{R}^d$ (but E is not a.s. 0)
  - symmetric about zero, i.e. E and $-E$ have the same distribution
  - if E has a density, it must be a positive definite function
  - spherical zero-mean Gaussian distribution, as well as multivariate Laplace, Cauchy or Student's t (but not uniform).
- define an (SPD-)*decomposable* random vector X if its characteristic function can be written as $\varphi_X = \varphi_{X_0}\varphi_E$, with E SPD noise component.
- Big modelling assumption: only the indecomposable components of distributions are of interest.

Phase Discrepancy and Phase Features

[Delaigle and Hall, 2016] construct density estimators for nonparametric deconvolution, i.e. estimate density f_0 of X_0 with observations $X_i \sim X_0 + E$. E has unknown SPD distribution. Matching *phase functions*:

$$\rho_X(\omega) = \frac{\varphi_X(\omega)}{|\varphi_X(\omega)|} = \exp(i\tau_X(\omega))$$

Phase function is *invariant to SPD noise* as it only changes the amplitude of the characteristic function.

We are not interested in density estimation but in *measuring differences up to SPD noise*. In analogy to MMD, define **phase discrepancy**:

$$\mathrm{PhD}(X, Y) = \int_{\mathbb{R}^d} |\rho_X(\omega) - \rho_Y(\omega)|^2 \, d\Lambda(\omega)$$

for some spectral measure Λ.

Trivial to construct *phase distribution embeddings* by simply normalising standard approximate mean embeddings to unit norm:

$$\Psi(P_X) = \sqrt{\frac{1}{m}} \left[\frac{\mathbb{E}\xi_{\omega_1}(X)}{\|\mathbb{E}\xi_{\omega_1}(X)\|}, \ldots, \frac{\mathbb{E}\xi_{\omega_m}(X)}{\|\mathbb{E}\xi_{\omega_m}(X)\|} \right]^\top$$

where $\xi_{\omega_j}(x) = \left[\cos\left(\omega_j^\top x\right), \sin\left(\omega_j^\top x\right)\right]$.

Synthetic Example

$$\theta_i \overset{i.i.d.}{\sim} \Gamma(\alpha, \beta), \quad Z_i \overset{i.i.d.}{\sim} U[0, \sigma],$$

$$\{X_i^j\}_j | \theta_i, Z_i \overset{i.i.d.}{\sim} \frac{\Gamma(\theta_i/2, 1/2)}{\sqrt{2\theta_i}} + \mathcal{N}(0, Z_i),$$

- Goal: Learn a mapping $\{X_i^j\} \mapsto \theta_i$ for Semi-Automatic ABC [Fearnhead and Prangle, 2010; Mitrovic, DS, and Teh, 2016].

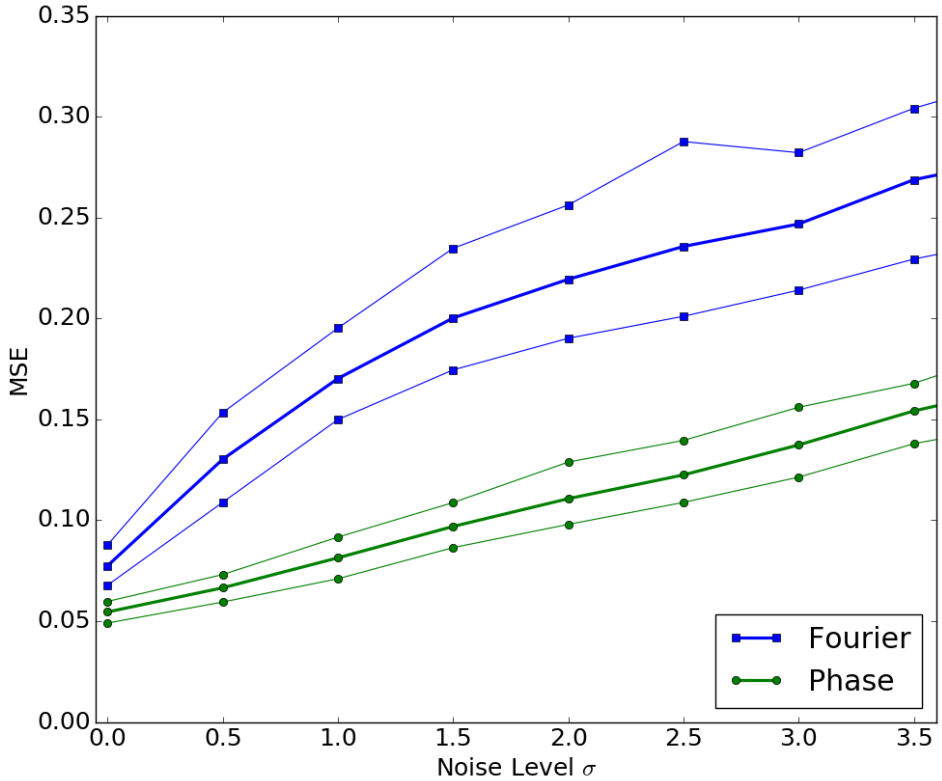
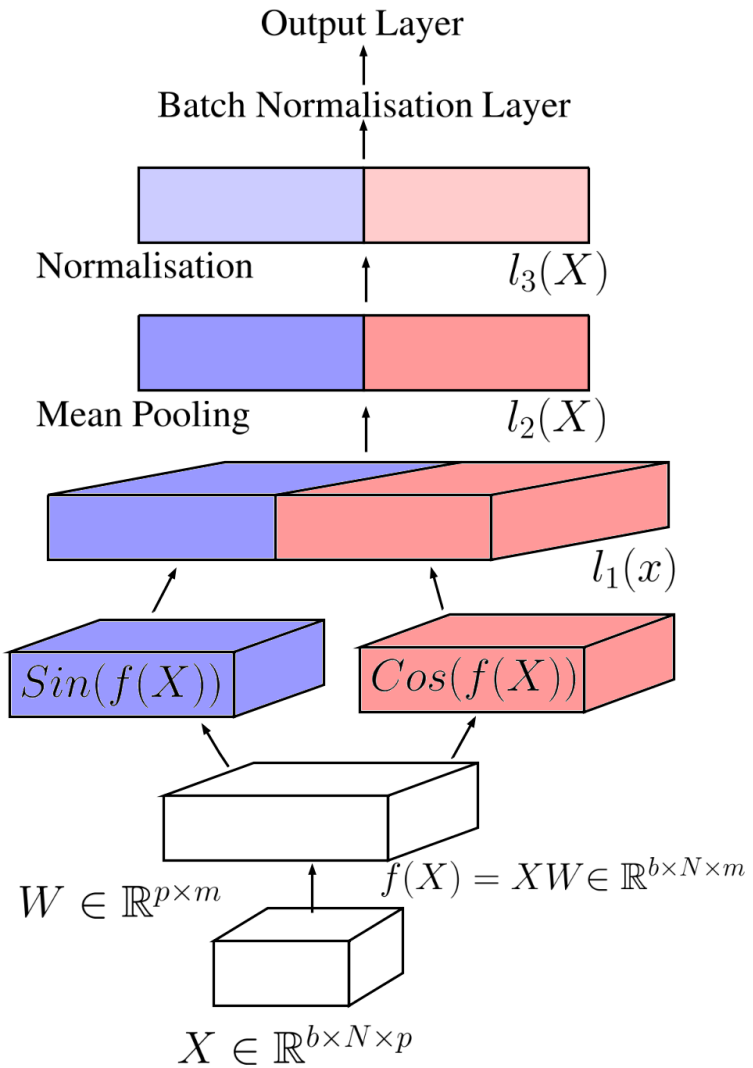


Figure: MSE of θ, using the Fourier and phase neural network based SA-ABC averaged over 100 runs. Here noise σ is varied between 0 and 3.5, and the 5^{th} and the 95^{th} percentile is shown.

Learning Phase Features



- Given a supervised signal, we can also optimise a set of frequencies $\{w_i\}_{i=1}^m$ that will give us a useful discriminative representation. In other words, we are no longer focusing on a specific translation-invariant kernel k (specific Λ), but are learning Fourier/phase features.
- A neural network with one hidden layer, coupled cos/sin activation functions, mean pooling and normalisation.
- Straightforward implementation in `Tensorflow` (code: `https://github.com/hcllaw/Fourier-Phase-Neural-Network`)

Aerosol MISR1 Dataset [Wang et al, 2012] with Covariate Shift

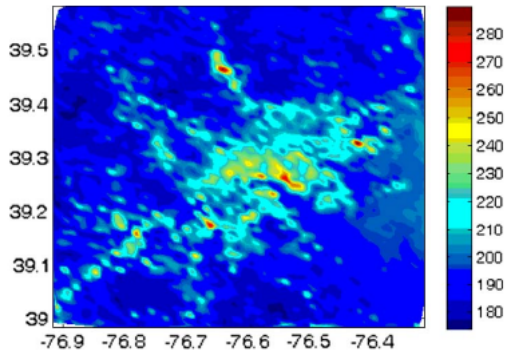


figure from Wang et al, 2012

- Aerosol Optical Depth (AOD) multiple-instance learning problem with 800 bags, each containing 100 randomly selected 16-dim multispectral pixels (satellite imaging) within 20km radius of AOD sensor.

The test data is impaired by additive SPD noise components.

Learning frequencies is key for robustness to noise.

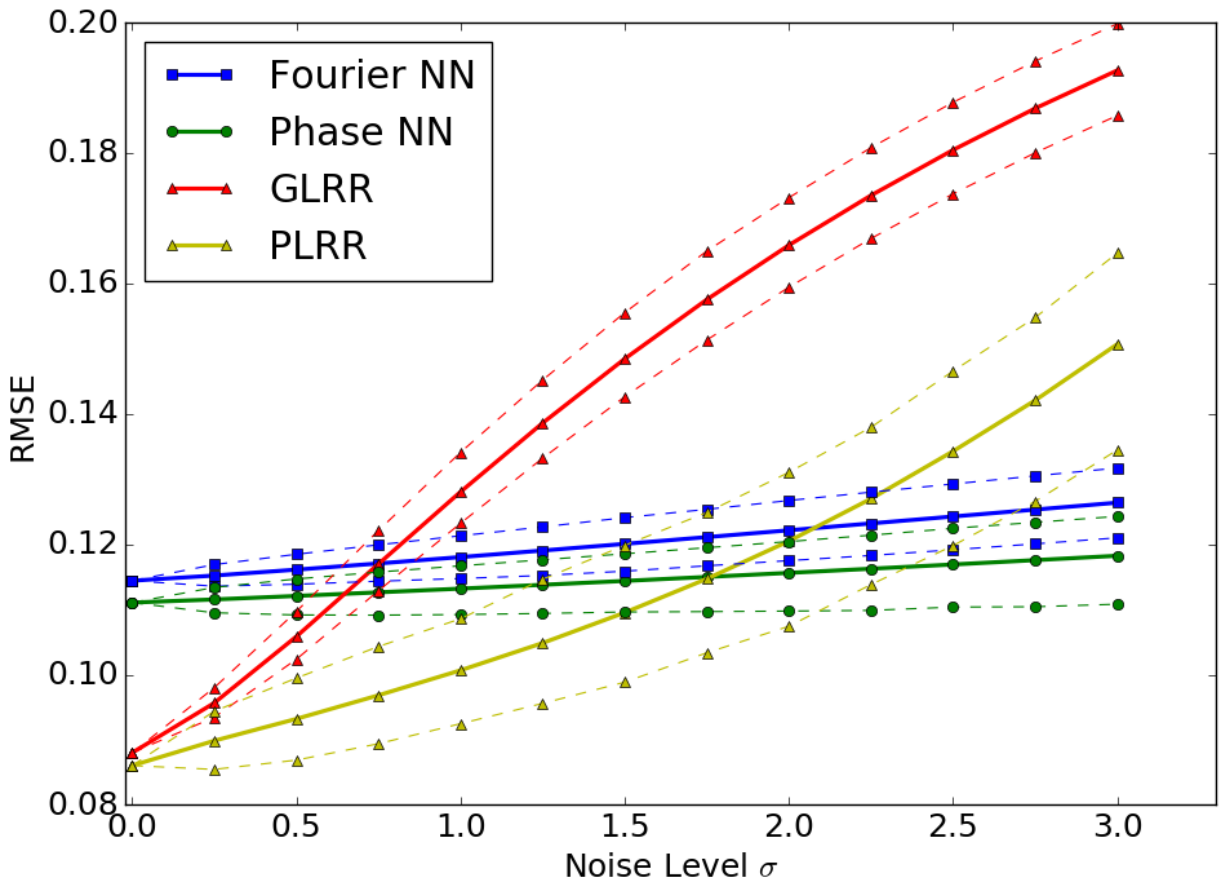


Figure: RMSE on the test set, corrupted by various levels of noise on the test set. 5^{th} and the 95^{th} percentile is shown.

Can Fourier features learn invariance?

- Discriminative frequencies learned on the "noiseless" training data correspond to *Fourier features* that are nearly normalised (i.e. they are close to unit norm).
- This means that the Fourier NN has *learned to be approximately invariant* based on training data, indicating that Aerosol data potentially has irrelevant SPD noise components ("cloudy pixels")
- In practice, use both types of features (characteristic + phase) and let data speak for itself.

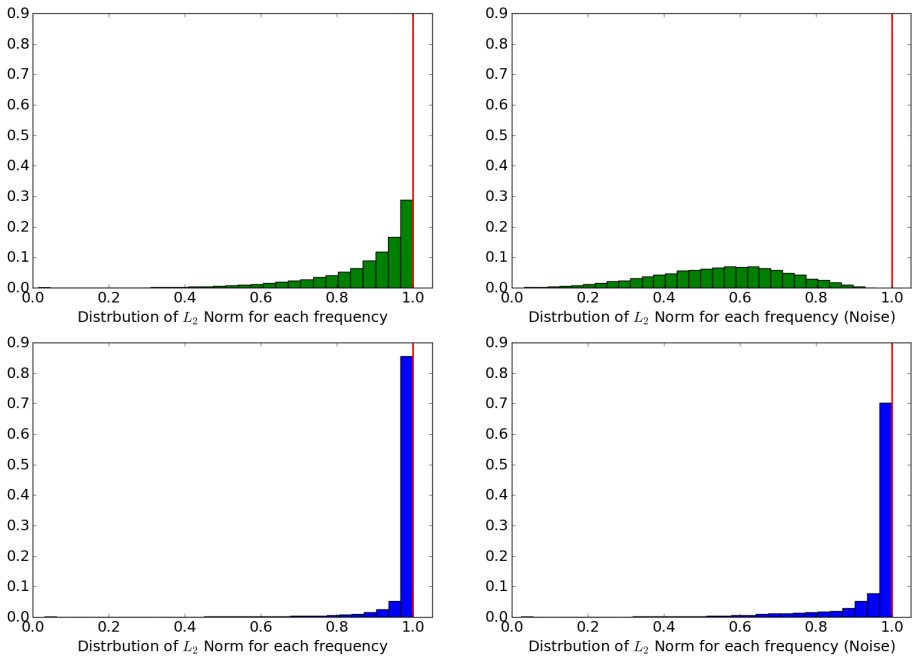


Figure: Histograms for the distribution of the modulus of Fourier features over each frequency w for the Aerosol data (test set); **Green:** Random Fourier Features (with the kernel bandwidth optimised on training data)
Bottom Blue: Learned Fourier features; **Left:** Original test set; **Right:** Test set with (additional) noise.

Summary

- Both contributions study distribution regression probems, where the responses are available at the group level, and demonstrate how statistical modelling can be brought to bear to address questions of uncertainty and invariance.
  - Modelling uncertainty can be vital for predictive performance on noisy datasets
  - Encoding invariance can make models more robust to irrelevant variation in the data
- Increasing confluence between statistical modelling and machine learning – making use of the well engineered deep learning (black-box) infrastructure, while carefully considering appropriate statistical models.
- Flexibility of the RKHS framework and kernel mean embeddings as a common ground between deep learning and statistical inference.

References

- Ho Chung Leon Law, Dougal J. Sutherland, DS, and Seth Flaxman, Bayesian Approaches to Distribution Regression, in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018, PMLR 84:1167-1176.
- Ho Chung Leon Law, Christopher Yau, and DS, Testing and Learning on Distributions with Symmetric Noise Invariance, in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1343-1353.