Learning with Approximate Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

RegML Workshop, Simula, Oslo, 06/05/2017

Outline

1. Preliminaries on Kernel Embeddings

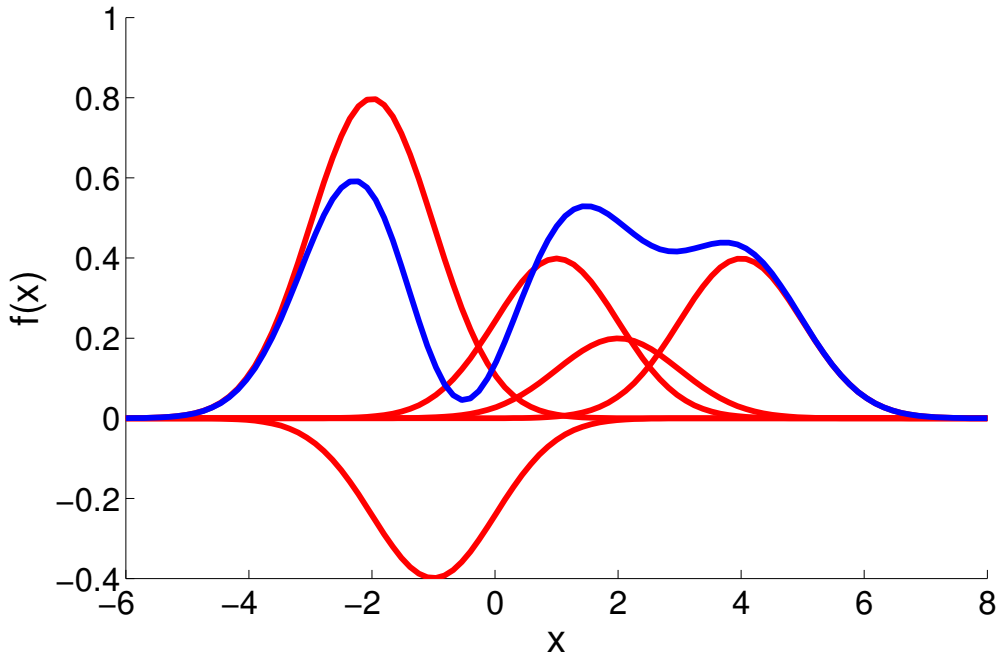
2. Testing and Learning on Distributions with Symmetric Noise Invariance

Outline

1. Preliminaries on Kernel Embeddings
2. Testing and Learning on Distributions with Symmetric Noise Invariance

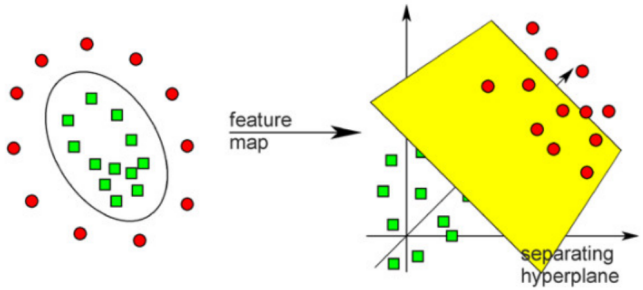
Reproducing Kernel Hilbert Spaces

- RKHS: a Hilbert space of functions on $\mathcal{X}$ with continuous evaluation $f \mapsto f(x)$, $\forall x \in \mathcal{X}$ (norm convergence implies pointwise convergence).
- Each RKHS corresponds to a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, s.t.
  1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.
- RKHS can be constructed as $\mathcal{H}_k = \overline{span\left\{k(\cdot, x) \,|\, x \in \mathcal{X}\right\}}$ and includes functions $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ and their pointwise limits.



Kernel Trick and Kernel Mean Trick

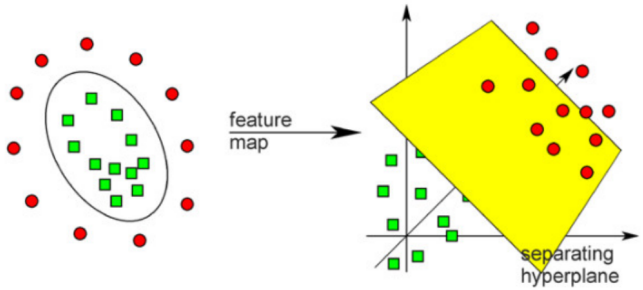
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions,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



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

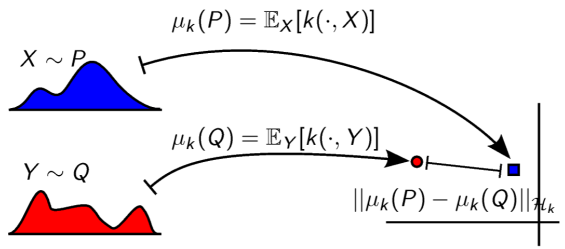
Kernel Trick and Kernel Mean Trick

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



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

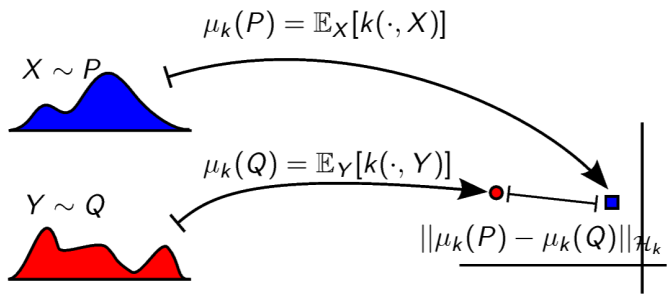
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k}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

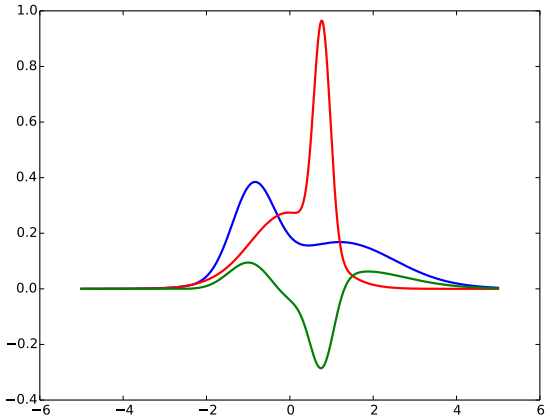


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

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

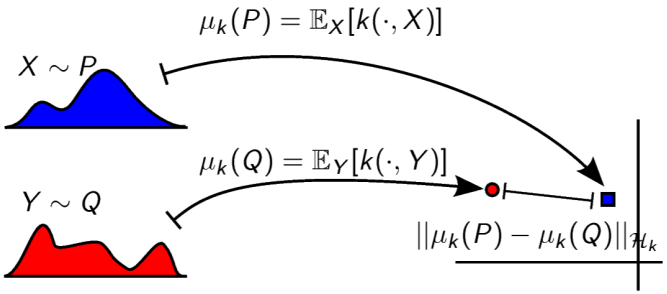


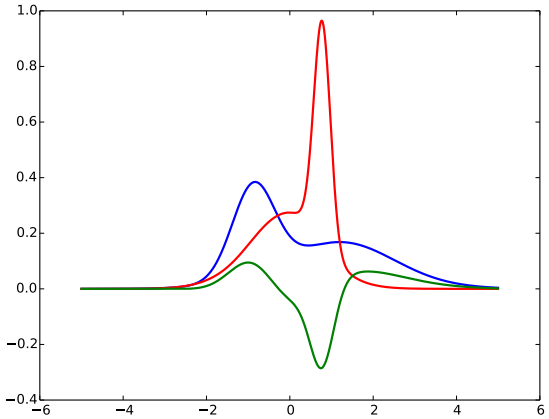


$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

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 $P = Q$.
  - Gaussian RBF $\exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$, Matérn family, inverse multiquadrics.
- For characteristic kernels on LCH $\mathcal{X}$, MMD metrizes weak* topology on probability measures [Sriperumbudur,2010],

$$\mathrm{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \rightsquigarrow P.$$

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

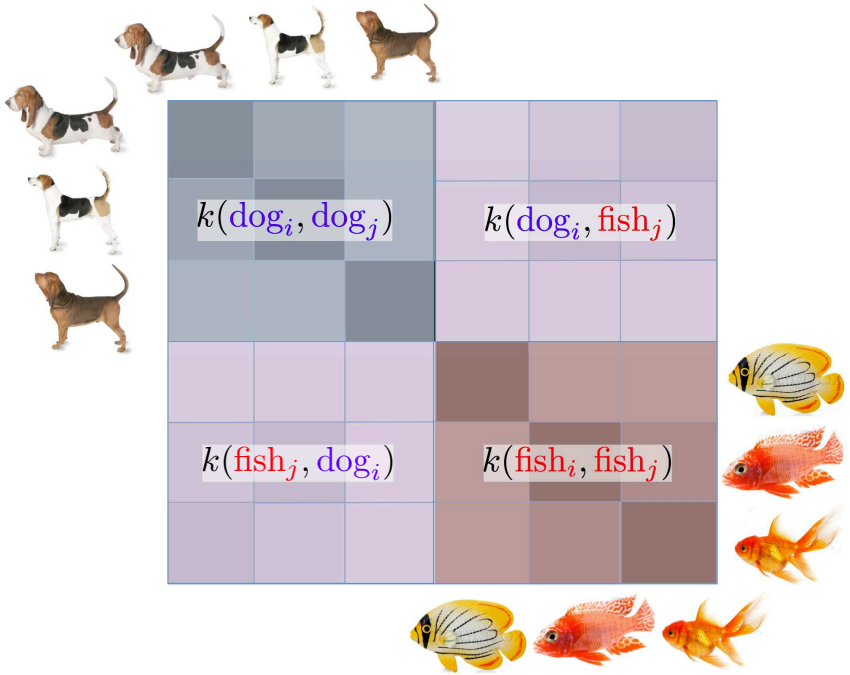


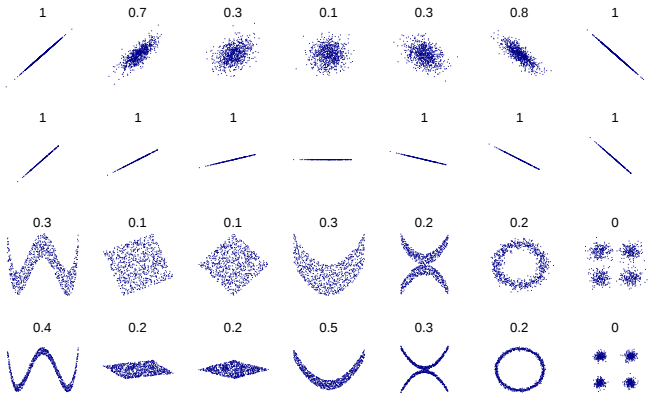
Figure by Arthur Gretton

MMD has been applied to:

- two-sample tests and independence tests [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, 2015+]
- ABC summary statistics [Park, Jitkrittum & DS, 2015]
- summarising streaming data [Paige, DS & Wood, 2016]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- training deep generative models [Dziugaite, Roy & Ghahramani, 2015; Sutherland et al, 2017]

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

Kernel dependence measures



cor vs. dcor

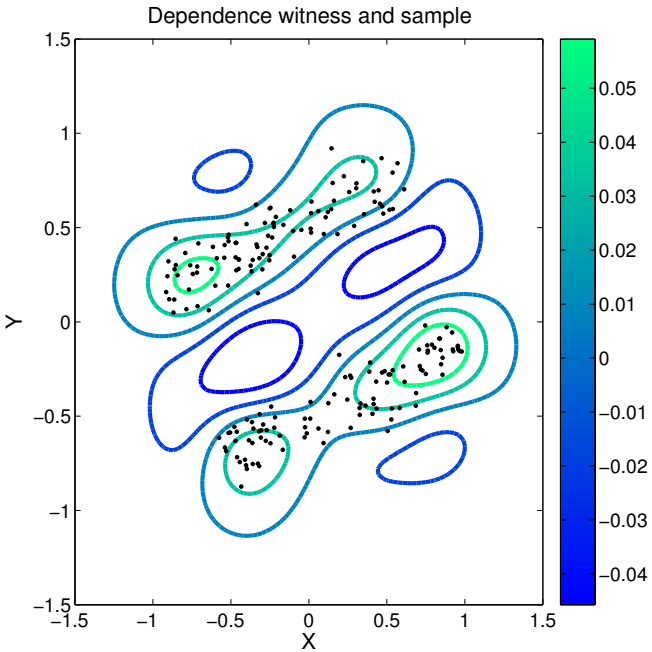


Figure by Arthur Gretton

$$HSIC^2(X, Y; \kappa) = \|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|^2_{\mathcal{H}_\kappa}$$

- 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 $\mathcal{X} \times \mathcal{Y}$

$$k(\boxed{1}, \boxed{2}) \quad l(\boxed{1}, \boxed{2})$$

$$\kappa(\boxed{1}\boxed{1}, \boxed{2}\boxed{2}) = k(\boxed{1}, \boxed{2}) \times l(\boxed{1}, \boxed{2})$$

- Independence testing framework that generalises Distance Correlation (dcor) of [Szekely et al, 2007]: HSIC with Brownian motion covariance kernels [DS et al, 2013]

Distribution Regression

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

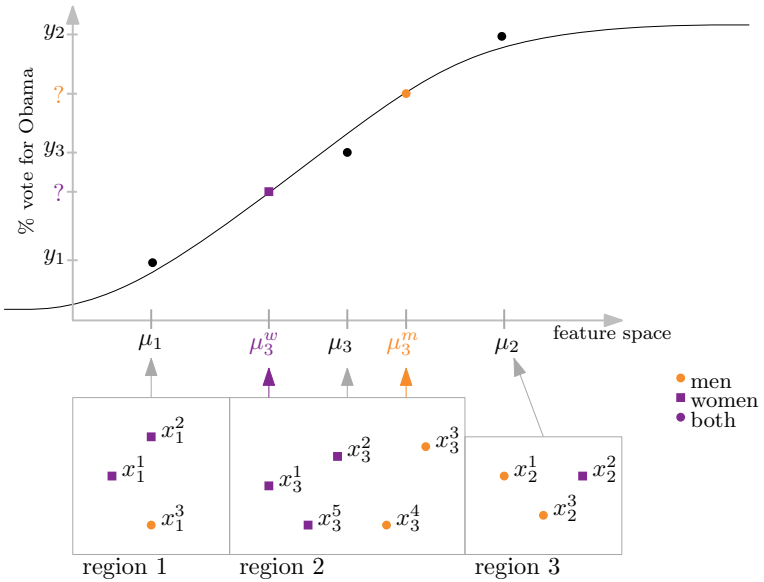


Figure from Flaxman et al, 2015

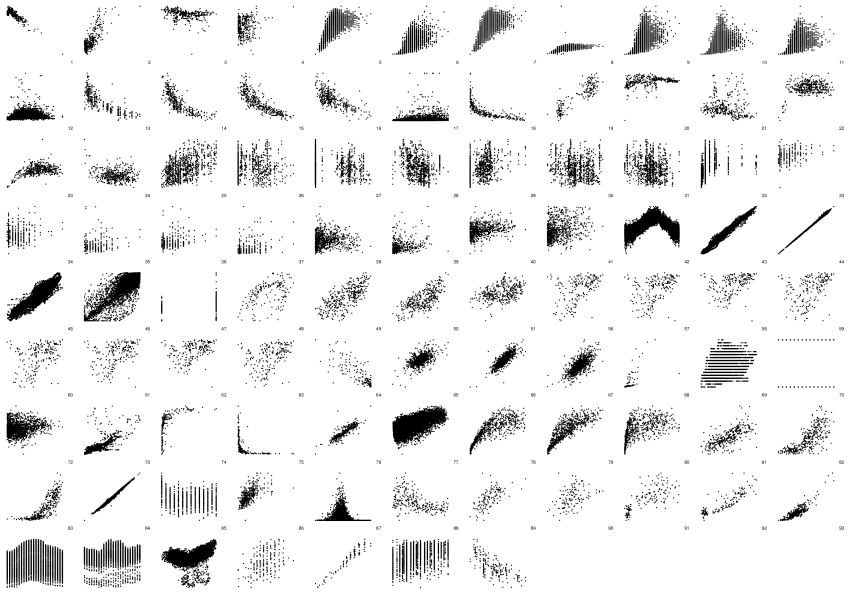


Figure from Mooij et al, 2014

- 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]

- Possible (distributional) covariate shift?

Outline

1 Preliminaries on Kernel Embeddings

2 Testing and Learning on Distributions with Symmetric Noise Invariance

All possible differences between generating processes?

- differences discovered by an MMD two-sample test can be due to different types of measurement noise or data collection artefacts
  - With a large sample-size, uncovers potentially irrelevant sources of variability: slightly different calibration of the data collecting equipment, different numerical precision, different conventions of dealing with edge-cases
- Learning on distributions: each label y_i in supervised learning is associated to a whole bag of observations $B_i = \{X_{ij}\}_{j=1}^{N_i}$ – assumed to come from a probability distribution P_i
  - Each bag of observations could be impaired by a different measurement noise process. Distributional covariate shift: different measurement noise on test bags?
- Both problems require encoding the distribution with a representation invariant to symmetric noise.

Testing and Learning on Distributions with Symmetric Noise Invariance.
Ho Chung Leon Law, Christopher Yau, DS.
http://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 $\xi_\Omega : x \mapsto \sqrt{\frac{2}{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots\right]^\top$.
- How fast does m need to grow with n? Can be sublinear for regression [Bach, 2015].

Approximate Mean Embeddings and Characteristic Functions

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

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

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

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

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

The Noise and the Signal

Adopting similar ides 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.
- Assume that 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.

Construct distribution features by simply normalising approximate mean embeddings:

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

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

Phase and Indecomposability

Is phase discrepancy a metric on indecomposable random variables?

Phase and Indecomposability

Is phase discrepancy a metric on indecomposable random variables? No

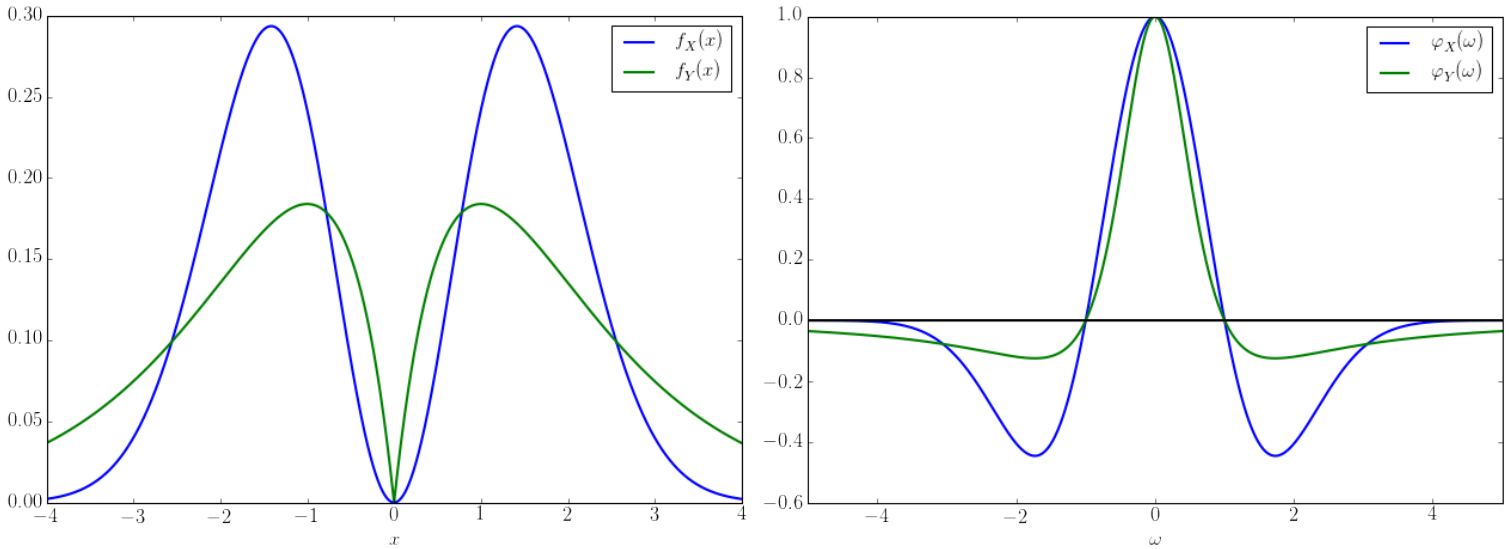
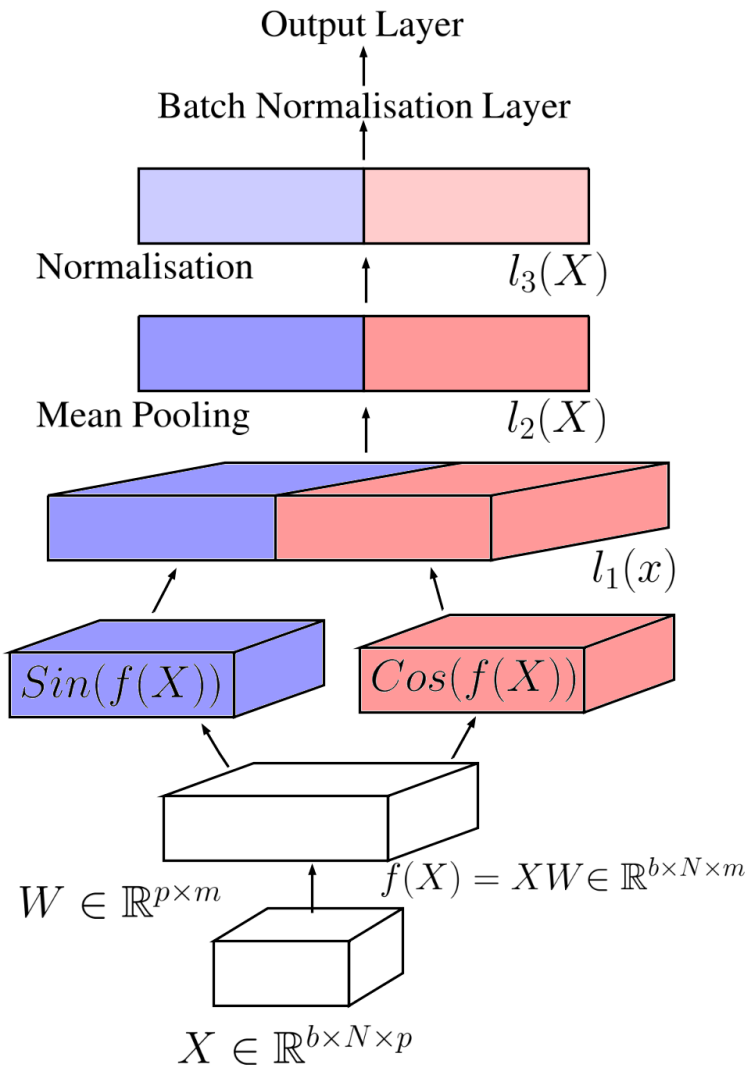


Figure: Example of two indecomposable distributions which have the same phase function. **Left**: densities. **Right**: characteristic functions.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^2 \exp(-x^2/2), \quad f_Y(x) = \frac{1}{2}|x| \exp(-|x|).$$

Learning Phase Features



- Random Phase features have a large variance, due to empirical normalisation.
- Given a supervised signal, we can 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 coupled cos/sin activation functions, mean pooling and normalisation.
- Straightforward implementation in `Tensorflow` (code: `https://github.com/hcllaw/Fourier-Phase-Neural-Network`)

Synthetic Example

$$
\begin{aligned}
\theta &\sim \Gamma(\alpha, \beta), \\
Z &\sim U[0, \sigma], \\
\epsilon|Z &\sim \mathcal{N}(0, Z), \\
\{X_i\}|\theta, \epsilon &\overset{i.i.d.}{\sim} \frac{\Gamma\left(\theta/2, 1/2\right)}{\sqrt{2\theta}} + \epsilon,
\end{aligned}
$$

- Goal: Learn a mapping $\{X_i\} \mapsto \theta$
- Can be used for semi-automatic ABC [Fearnhead & Prangle, 2012] with kernel distribution regression for summary statistics [Mitrovic, DS & Teh, 2016].

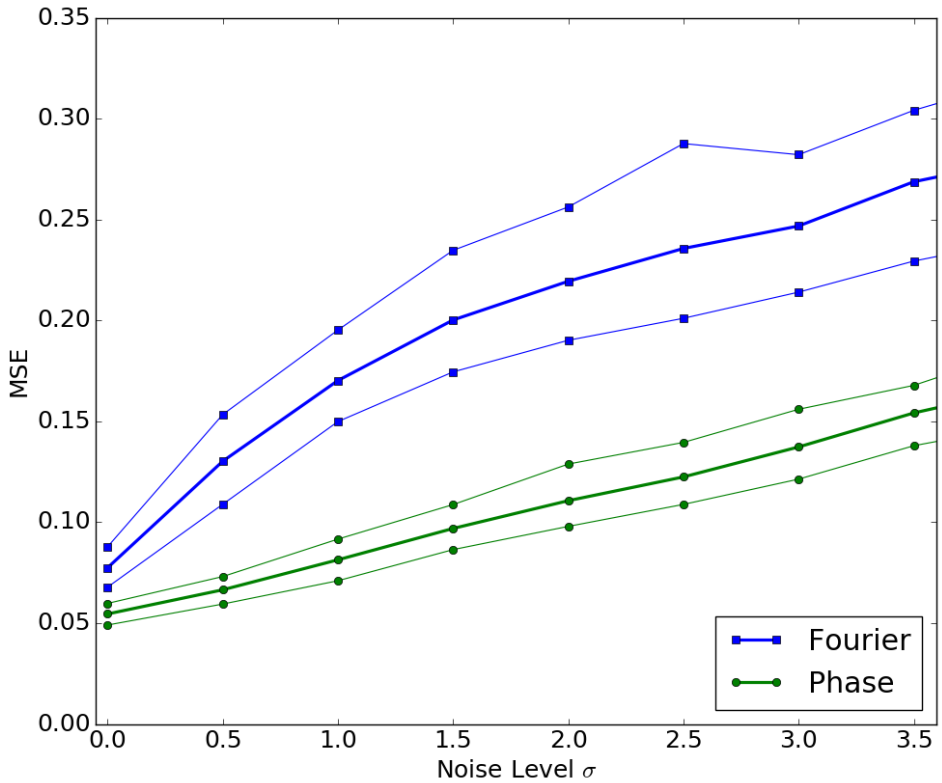


Figure: MSE of θ, using the Fourier and phase neural network averaged over 100 runs. Here noise σ is varied between 0 and 3.5, and the 5^{th} and the 95^{th} percentile is shown.

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

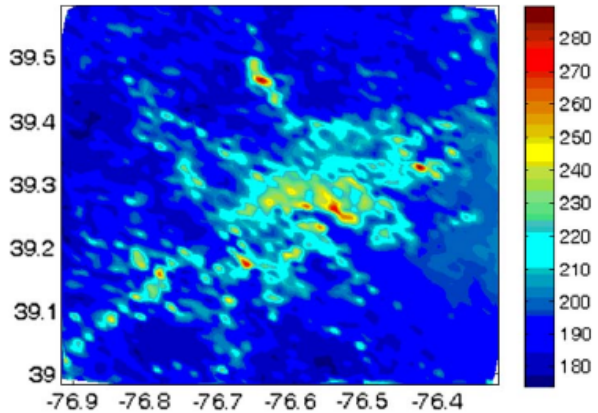


figure from Wang et al, 2012

- 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.
- Image variability due to surface properties – small spatial variability of AOD.
- The label y_i provided by the ground AOD sensors.

The test data is impaired by additive SPD noise components.

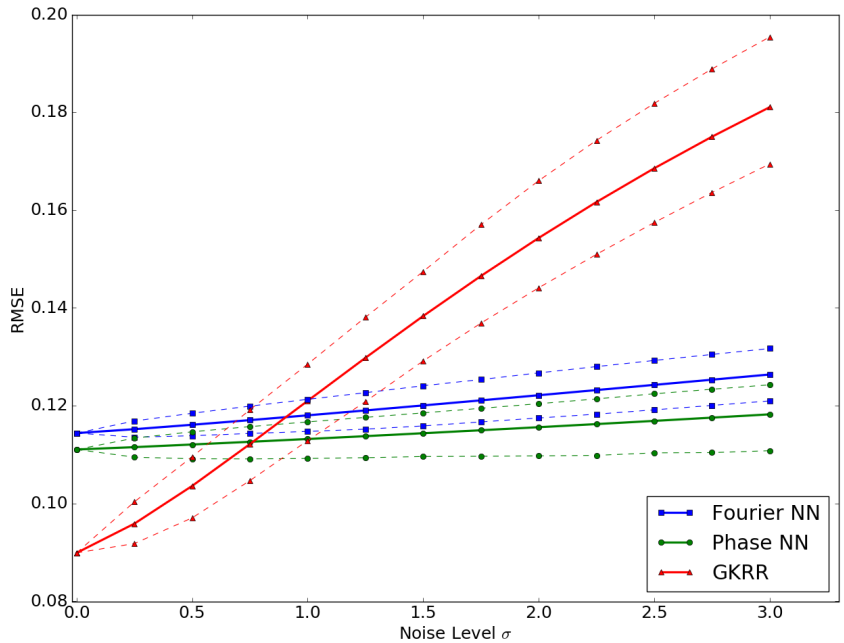


Figure: RMSE on the test set, corrupted by various levels of noise, using the Fourier and phase neural network and GKKR averaged over 100 runs. Here noise-to-signal ratio σ is varied between 0 and 3.0, and the 5^{th} and the 95^{th} percentile is shown.

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")

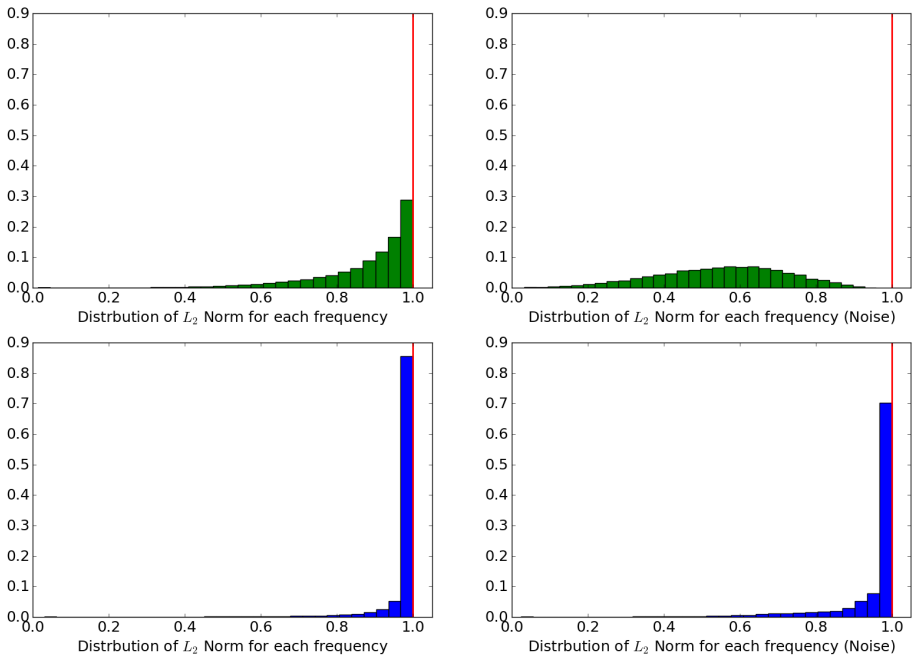


Figure: Histograms for the distribution of the modulus of Fourier features over each frequency w for the Aerosol data (test set); **Green:** Random Fourier Features (with the kernel bandwidth optimised on training data)
Bottom Blue: Learned Fourier features; **Left:** Original test set; **Right:** Test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

Summary

- When measuring nonparametric distances between distributions, can we disentangle the differences in noise from the differences in the signal?
- We considered two different ways to encode invariances to symmetric noise:
  - MMD for asymmetry (not discussed in the talk) in paired sample differences, $MMD(X-Y, Y-X)$, which can be used to construct a two-sample test up to symmetric noise.
  - weighted distance between the empirical phase functions for learning algorithms on distribution inputs which are robust to measurement noise and covariate shift.