Inference with Approximate Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

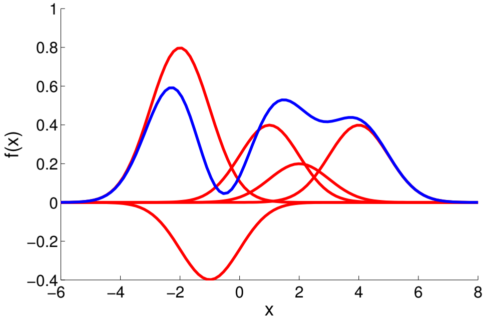
Research Collaboration Day, 31/05/2017

Outline

Outline

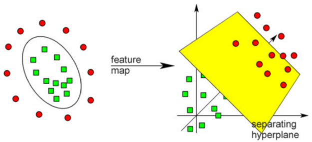
Reproducing Kernel Hilbert Spaces

- 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. $\forall x \in \mathcal{X}, \;\; k(\cdot, x) \in \mathcal{H}$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \;\; \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- RKHS can be constructed as $\mathcal{H}_k = \overline{span\,\{k(\cdot, x) \,|\, x \in \mathcal{X}\}}$ 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

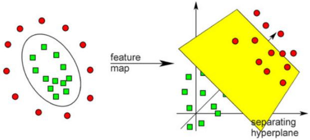
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \dots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions,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



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

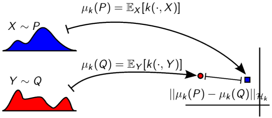
Kernel Trick and Kernel Mean Trick

- 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]

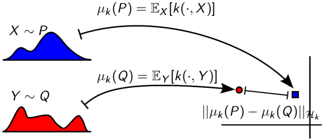
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010; Muandet et al, 2017]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P),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

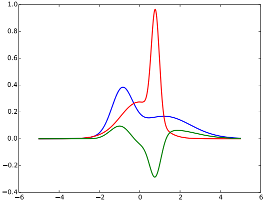


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

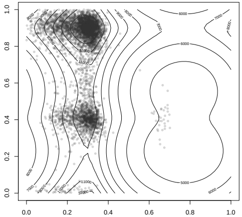
- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:





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

- **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 structured and non-Euclidean domains.



[Figure from Flaxman, Teh & DS, 2017]

Some uses of MMD

within-sample average similarity
–
between-sample average similarity

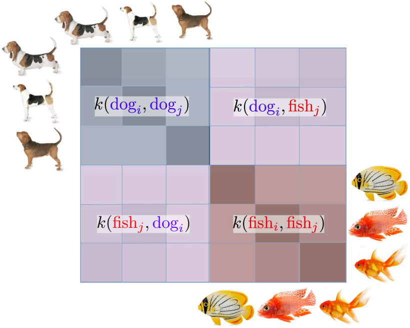


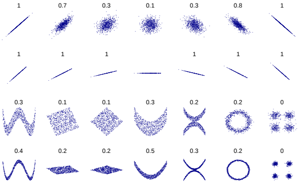
Figure by Arthur Gretton

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, 2015+]
- 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]
- 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: HSIC



cor vs. dcor

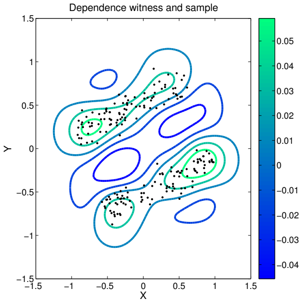
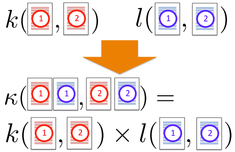


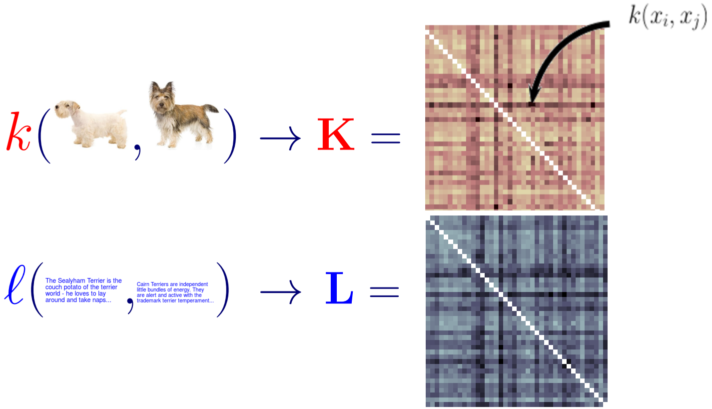
Figure by Arthur Gretton

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

Kernel dependence measures: HSIC (2)

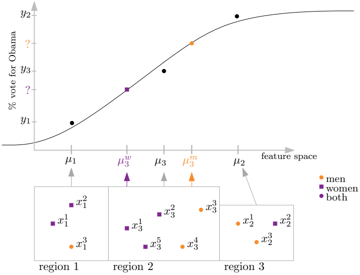


Hilbert-Schmidt Independence Criterion (**HSIC**): similarity between the kernel matrices $\left\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \right\rangle = \boxed{\mathrm{Tr}\left(\tilde{\mathbf{K}}\tilde{\mathbf{L}}\right)}$, where $\tilde{\mathbf{K}} = \mathbf{HKH}$, and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

[Gretton et al, 2008; Fukumizu et al, 2008; Song et al, 2012]

Distribution Regression

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



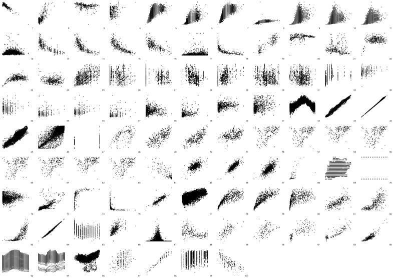


Figure from Flaxman et al, 2015   Figure from Mooij et al, 2014

- 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?

Bag-specific noises in Distribution Regression

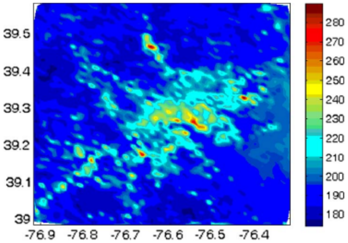


figure from Wang et al, 2012

Aerosol MISR1 Dataset [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.
- Large image variability due to surface properties, but small spatial variability of AOD – can be treated as distribution regression.
- The label y_i provided by the ground AOD sensors.
- Different noise ("cloudy pixels") distribution in different images.

This talk:

- Kernel embeddings as *nonparametric modules* which "automate" difficult choices in *parametric (Bayesian) inference*.
    - This talk considered summary statistics for ABC, but there are several other examples (proposal distributions in MCMC, passing messages in Expectation Propagation...)
- 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 be used for learning algorithms on distribution inputs which are robust to measurement noise and covariate shift.

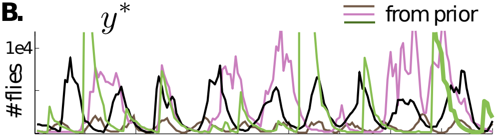
Outline

Motivating example: ABC for modelling ecological dynamics

- Given: a time series $\mathbf{Y} = (Y_1, \ldots, Y_T)$ of population sizes of a blowfly.
- Model: A dynamical system for blowfly population (a discretised ODE) [Nicholson, 1954; Gurney et al, 1980; Wood, 2010; Meeds & Welling, 2014]

$$Y_{t+1} = P Y_{t-\tau} \exp\left(-\frac{Y_{t-\tau}}{Y_0}\right) e_t + Y_t \exp(-\delta \epsilon_t),$$

where $e_t \sim$ Gamma $\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$, $\epsilon_t \sim$ Gamma $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
Parameter vector: $\theta = \{P, Y_0, \sigma_d, \sigma_p, \tau, \delta\}$.



- Goal: For a prior $p(\theta)$, sample from $p(\theta|\mathbf{Y})$.
  - Cannot evaluate $p(\mathbf{Y}|\theta)$. But, can sample from $p(\cdot|\theta)$.
  - For $\mathbf{X} = (X_1, \ldots, X_T) \sim p(\cdot|\theta)$, how to measure distance $\rho(\mathbf{X}, \mathbf{Y})$?

Data Similarity via Summary Statistics

- Distance ρ is typically defined via summary statistics

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

- How to select the summary statistics $s(\cdot)$? Unless $s(\cdot)$ is sufficient, even as $\epsilon \to 0$, targets an incorrect (partial) posterior $p(\theta|s(\mathbf{Y}))$ rather than $p(\theta|\mathbf{Y})$.
- Hard to quantify additional bias.
  - Adding more summary statistics decreases "information loss": $p(\theta|s(\mathbf{Y})) \approx p(\theta|\mathbf{Y})$
  - ρ computed on a higher dimensional space - without appropriate calibration of distances therein, leads to a higher rejection rate so need to increase ϵ: $p_\epsilon(\theta|s(\mathbf{Y})) \not\approx p(\theta|s(\mathbf{Y}))$

Data Similarity via Summary Statistics

- Distance ρ is typically defined via summary statistics

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

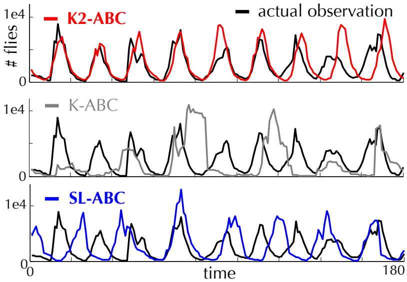
- How to select the summary statistics $s(\cdot)$? Unless $s(\cdot)$ is sufficient, even as $\epsilon \to 0$, targets an incorrect (partial) posterior $p(\theta|s(\mathbf{Y}))$ rather than $p(\theta|\mathbf{Y})$.
- Hard to quantify additional bias.
  - Adding more summary statistics decreases "information loss": $p(\theta|s(\mathbf{Y})) \approx p(\theta|\mathbf{Y})$
  - ρ computed on a higher dimensional space - without appropriate calibration of distances therein, leads to a higher rejection rate so need to increase ϵ: $p_\epsilon(\theta|s(\mathbf{Y})) \not\approx p(\theta|s(\mathbf{Y}))$
- A very simple idea: Use a nonparametric distance (MMD) between the empirical measures of datasets $\mathbf{X}$ and $\mathbf{Y}$).
  - No need to design $s(\cdot)$.
  - Rejection rate does not blow up since MMD penalises the higher order moments (Mercer expansion).

Blowfly example

Number of blow flies over time

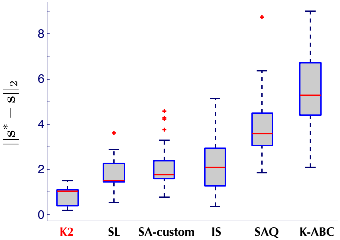
$$Y_{t+1} = PY_{t-\tau} \exp\left(-\frac{Y_{t-\tau}}{Y_0}\right) e_t + Y_t \exp(-\delta \epsilon_t)$$

- $e_t \sim$ Gam $\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$ and $\epsilon_t \sim$ Gam $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
- Want $\theta := \{P, Y_0, \sigma_d, \sigma_p, \tau, \delta\}$.

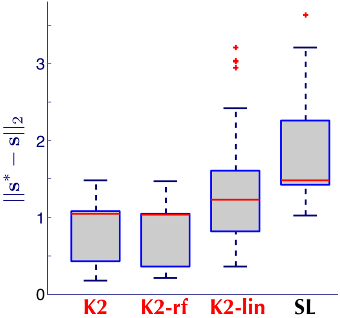


- Simulated trajectories with inferred posterior mean of θ
  - Observed sample of size 180.
  - Other methods use handcrafted 10-dimensional summary statistics $s(\cdot)$ from [Meeds & Welling, 2014]: quantiles of marginals, first-order differences, maximal peaks, etc.

Blowfly example: comparisons



- Let $\tilde{\theta}$ be the posterior mean.
- Simulate $\mathbf{X} \sim p(\cdot|\tilde{\theta})$.
- $\mathbf{s} = s(\mathbf{X})$ and $\mathbf{s}^* = s(\mathbf{Y})$.
- Improved mean squared error on $\mathbf{s}$, even though SL-ABC, SA-custom explicitly operate on $\mathbf{s}$ while K2-ABC does not.

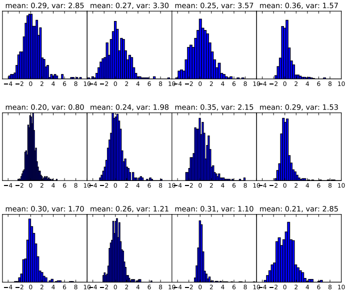


- Computation of $\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y})$ costs $O(n^2)$.
- Linear-time unbiased estimators of MMD^2 or random feature expansions reduce the cost to $O(n)$.

[M. Park, W. Jitkrittum, and DS. **K2-ABC: Approximate Bayesian Computation with Kernel Embeddings**, AISTATS 2016. code: https://github.com/wittawatj/k2abc]

ABC and Modelling Invariance

$$\theta \sim \Gamma(\alpha, \beta), \quad Z \sim U[0, \sigma],$$
$$\{\epsilon_i\}|Z \overset{i.i.d.}{\sim} \mathcal{N}(0, Z),$$
$$X_i|\theta, \epsilon_i \sim \frac{\Gamma(\theta/2, 1/2)}{\sqrt{2\theta}} + \epsilon_i,$$



- MMD is simple and effective when $\{X_i\} \overset{i.i.d.}{\sim} p(\cdot|\theta)$. However, in the model above there is an additional variability in $\{X_i\}$ due to the noise distribution which differs for every bag of observations.
- Semi-Automatic ABC [Fearnhead & Prangle, 2012] uses posterior mean estimates $\hat{\mathbb{E}}[\theta|\{X_i\}]$ as summary statistics, which requires learning a map $\{X_i\} \mapsto \theta$, using e.g. distribution regression from (conditional) kernel embeddings [Mitrovic, DS and Teh, 2016].
    - If $\{X_i\}$, Z are both observed can build a regression from the joint distribution $p(\mathbf{X}, Z)$ or from the conditional $p(\mathbf{X}|Z)$ (note that θ parametrizes $\{X_i\}|Z$)
    - But Z is generally not observed on the real data – a different idea: build a regression function invariant to Z?

Outline

All possible differences between generating processes?

- 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?
- 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
- 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

Characteristic Functions and (Approximate) Kernel Embeddings

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

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

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 $\{\omega_j\}_{j=1}^m \sim \Lambda$:

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

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**:

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

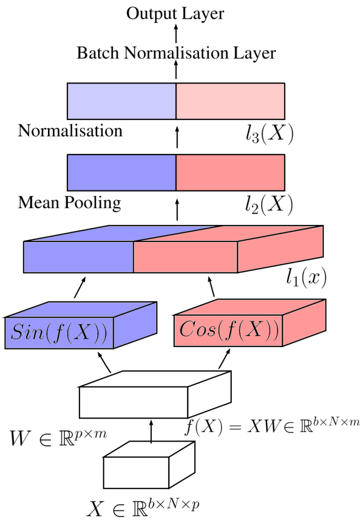
for some spectral measure Λ.

Construct distribution features by simply normalising approximate mean embeddings 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 $\xi_{\omega_j}(x) = \left[\cos\left(\omega_j^\top x\right), \sin\left(\omega_j^\top x\right)\right]$.

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 coupled cos/sin activation functions, mean pooling and normalisation.
- Straightforward implementation in `Tensorflow` (code: `https://github.com/hcllaw/Fourier-Phase-Neural-Network`)

Synthetic Example

$$\theta \sim \Gamma(\alpha, \beta), \quad Z \sim U[0, \sigma],$$
$$\{\epsilon_i\}|Z \stackrel{i.i.d.}{\sim} \mathcal{N}(0, Z),$$
$$X_i|\theta, \epsilon_i \sim \frac{\Gamma(\theta/2, 1/2)}{\sqrt{2\theta}} + \epsilon_i,$$

- Goal: Learn a mapping $\{X_i\} \mapsto \theta$ for Semi-Automatic ABC.

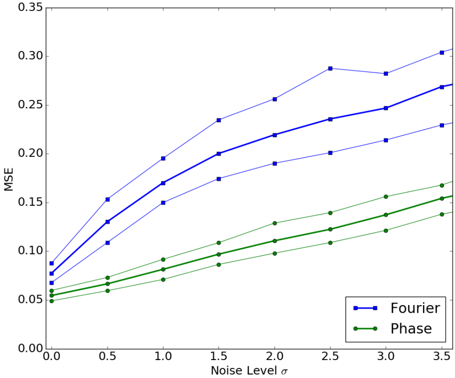


Figure: MSE of θ, using the Fourier and phase neural network based SA-ABC averaged over 100 runs. Here noise σ is varied between 0 and 3.5, and the 5^{th} and the 95^{th} percentile is shown.

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

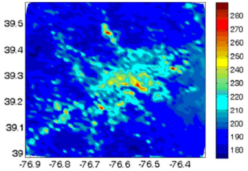


figure from Wang et al, 2012

- 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.

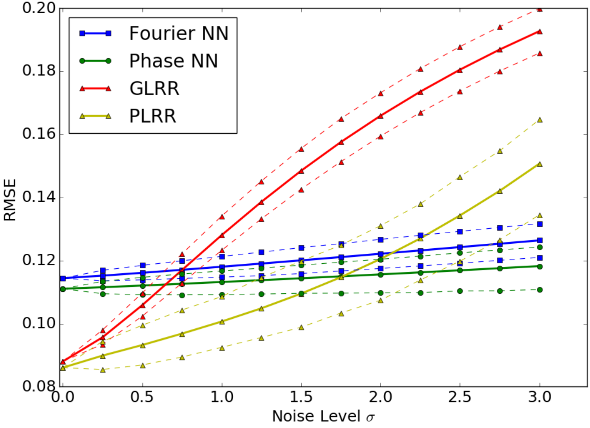
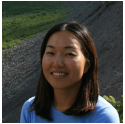


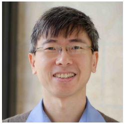
Figure: RMSE on the test set, corrupted by various levels of noise on the test set. 5^{th} and the 95^{th} percentile is shown.

References

- Mijung Park, Wittawat Jitkrittum, and DS, K2-ABC: Approximate Bayesian Computation with Kernel Embeddings, in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016, PMLR 51:398-407.



- Jovana Mitrovic, DS, and Yee Whye Teh, DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression, in *International Conference on Machine Learning (ICML)*, 2016, PMLR 48:1482-1491.



- Ho Chung Leon Law, Christopher Yau, and DS, Testing and Learning on Distributions with Symmetric Noise Invariance, *ArXiv e-prints:1703.07596*, 2017.

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

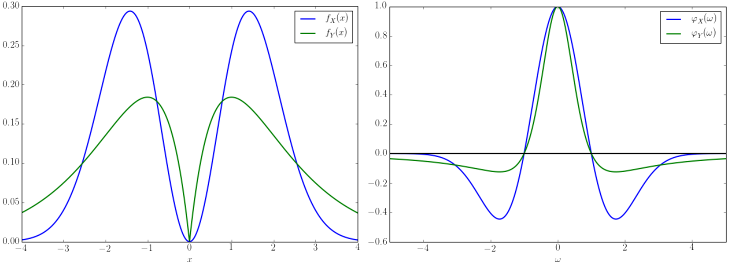


Figure: Example of two indecomposable distributions which have the same phase function. **Left**: densities. **Right**: characteristic functions.

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

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”)

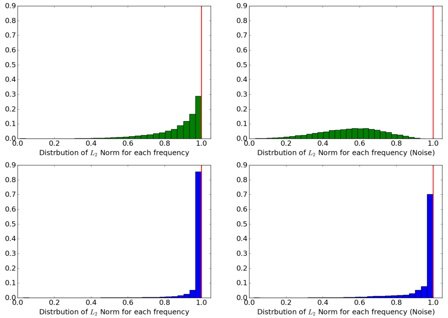


Figure: Histograms for the distribution of the modulus of Fourier features over each frequency w for the Aerosol data (test set); **Green:** Random Fourier Features (with the kernel bandwidth optimised on training data)
Bottom Blue: Learned Fourier features; **Left:** Original test set; **Right:** Test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