# Learning with Approximate Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

Statistical Learning Workshop, Lancaster, 23/03/2017

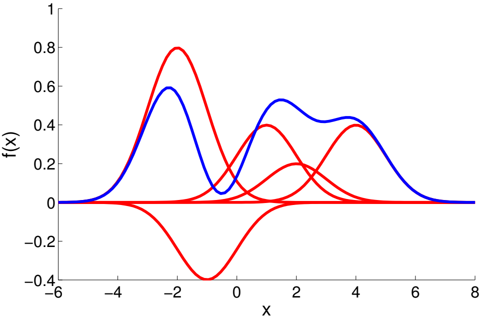
# Outline

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

# Outline

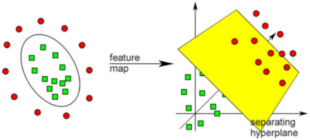
# Reproducing Kernel Hilbert Spaces

- 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. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, 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 $\mathcal{H}_k = \overline{span \{k(\cdot, x) \mid x \in \mathcal{X}\}}$ 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

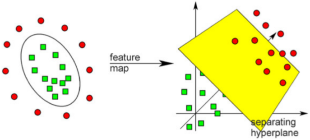
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  *inner products readily available*
  - nonlinear decision boundaries, nonlinear regression functions,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




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

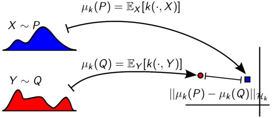
# Kernel Trick and Kernel Mean Trick

- 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]

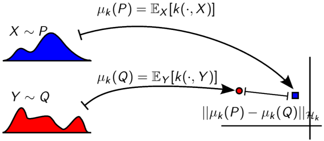
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X\sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q)\rangle_{\mathcal{H}_k}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

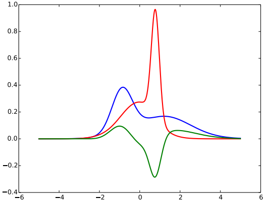



[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

# Maximum Mean Discrepancy

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

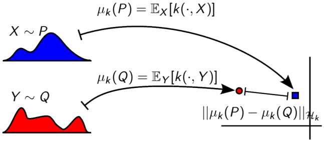


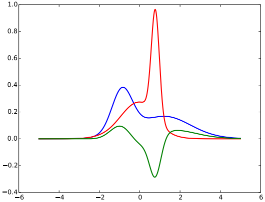




$$\text{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k :\, \|f\|_{\mathcal{H}_k} \le 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

# 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 $P = Q$.
  - 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

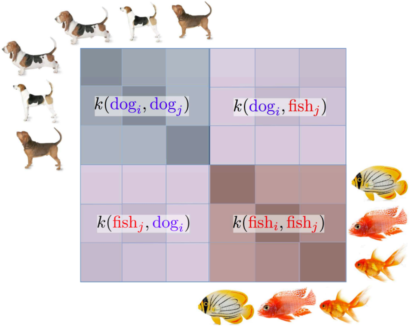



Figure by Arthur Gretton

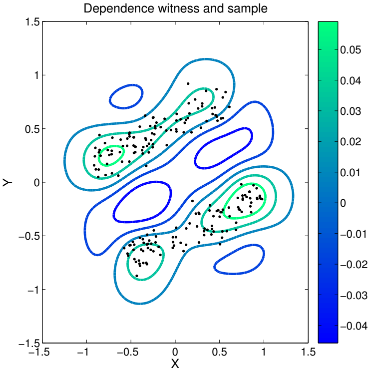
MMD has been applied to:

- 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]

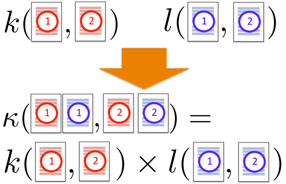
$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}_{X,X' \overset{i.i.d.}{\sim} P} k(X,X') + \mathbb{E}_{Y,Y' \overset{i.i.d.}{\sim} Q} k(Y,Y') - 2\mathbb{E}_{X\sim P, Y\sim Q} k(X,Y).$$

# Kernel dependence measures

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




- 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 Covariance (dCov) of [Szekely et al, 2007]: HSIC with Brownian motion covariance kernels [DS et al, 2013]

# Outline

# 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) $\Lambda$.

- 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].

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

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

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

# 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 $\Lambda$.

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

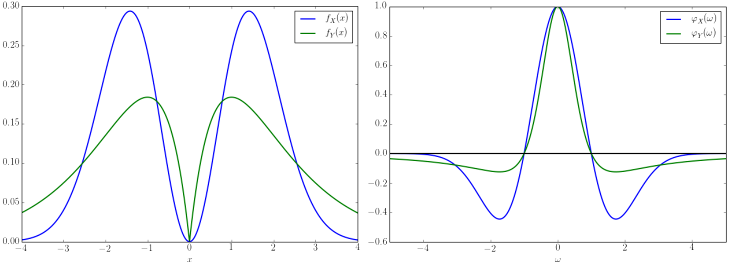
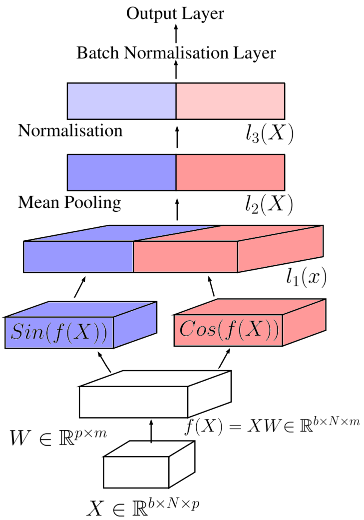



Figure: Example of two indecomposable distributions which have the same phase function. **Left**: densities. **Right**: characteristic functions.

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

# Learning Phase Features




- 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 $\Lambda$), 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

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

- 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].

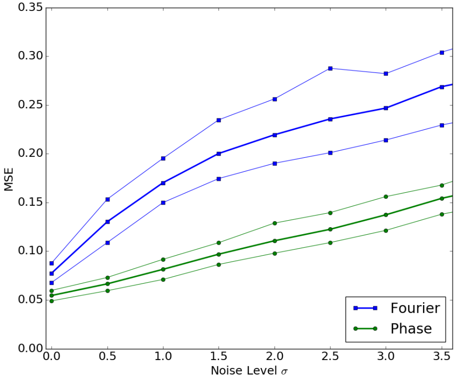



Figure: MSE of $\theta$, using the Fourier and phase neural network averaged over $100$ runs. Here noise $\sigma$ is varied between $0$ and $3.5$, and the $5^{th}$ and the $95^{th}$ percentile is shown.

# Aerosol Dataset with Covariate Shift

- Aerosol MISR1 dataset [Wang et al, 2012; Szabo et al, 2015]
- Aerosol Optical Depth (AOD) multiple-instance learning problem with $800$ bags, each containing $100$ randomly selected 16-dim multispectral pixels (satellite imaging) within $20$km radius of AOD sensor.
- The label $y_i$ provided by the ground AOD sensors.
- The test data is impaired by additive SPD noise components.

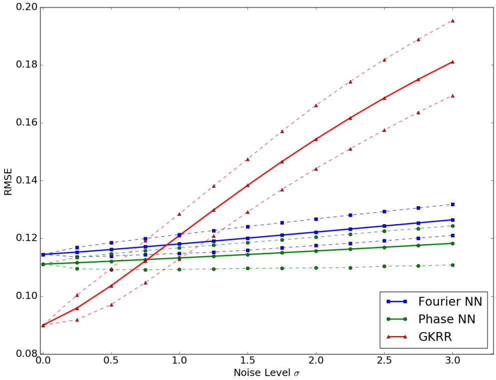



Figure: RMSE on the test set, corrupted by various levels of noise, using the Fourier and phase neural network and GKKR averaged over $100$ runs. Here noise-to-signal ratio $\sigma$ is varied between $0$ and $3.0$, and the $5^{th}$ and the $95^{th}$ percentile is shown.

# 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.

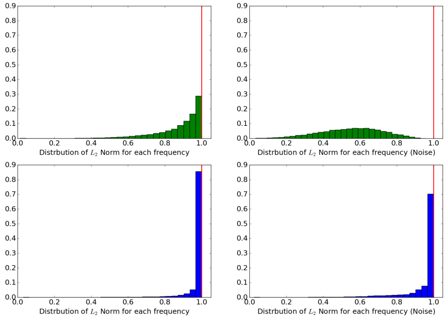



Figure: Histograms for the distribution of the modulus of Fourier features over each frequency $w$ for the Aerosol data (test set). **Top Green:** Random Fourier Features $w$ (with the optimised kernel bandwidth) **Bottom Blue:** Learned Fourier features $w$ from the Fourier Neural Network

# 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.

# Outline

# Bayesian Model for Embeddings

- In MMD, HSIC and other applications of embeddings, we estimate $\mu = \int k(\cdot, x)\mathsf{P}(dx)$ with its empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i)$.
- Empirical mean over an infinite-dimensional case? Due to Stein's phenomenon, shrinkage estimators are better behaved [Muandet et al, 2013] and are reported to improve performance in kernel PCA and in testing power [Ramdas & Wehbe, 2015].
- Can we formulate a Bayesian inference procedure for kernel embeddings?
- Two challenges:
  - How to construct a valid prior over the RKHS?
  - What is the likelihood of our observations given the kernel embedding?

Bayesian Learning of Kernel Embeddings.
**UAI 2016.**
Seth Flaxman, DS, John Cunningham, and Sarah Filippi.
`http://arxiv.org/abs/1603.02160`

# Priors on RKHS

Since sample paths of a GP with kernel $k$ lie outside RKHS $\mathcal{H}_k$ with probability 1 Kallianpur's 0-1 law, [Kallianpur, 1970; Wahba, 1990], use

$$r(x, x') = \int k(x, u)k(u, x')\nu(du)$$

in which case $f \in \mathcal{H}_k$ with probability 1 by nuclear dominance theory [Lukic and Beder, 2001; Pillai et al, 2007].

For some simple cases, kernel $r$ analytically available, e.g. for a Gaussian kernel $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\theta^2}\right)$ and $\nu(du) \propto \exp\left(-\frac{\|u\|^2}{2\eta^2}\right) du$:

$$r(x, x') \propto \exp\left(-\frac{\|x-x'\|^2}{4\theta^2} - \frac{\|(x+x')/2\|^2}{4\theta^2+\eta^2}\right).$$

- Has a nonstationary component, but similar to another (smoother) Gaussian kernel with bandwidth $\theta\sqrt{2}$ when $\eta$ is large.

# Likelihood

We need a likelihood linking the kernel mean embedding $\mu$ to the observations $\{x_i\}_{i=1}^n$ Consider evaluating $\widehat{\mu}$ induced by $\{x_i\}_{i=1}^n$ at some $x \in \mathcal{X}$ - we link $\widehat{\mu}(x)$ to $\mu(x)$ using a Gaussian distribution with variance $\tau^2/n$:

$$p(\widehat{\mu}(x)|\mu(x)) = \mathcal{N}(\widehat{\mu}(x); \mu(x), \tau^2/n), \quad x \in \mathcal{X}.$$

Obviously wrong - both $\mu$ and $\hat{\mu}$ are smooth functions. In general covariance will depend both on $k$ and $P$.

# Posterior of the embedding

Standard conjugacy results give:

$$\mu(\mathbf{x}) \mid \widehat{\mu}(\mathbf{x}) \sim \mathcal{N}(R(R + (\tau^2/n)I_n)^{-1}\widehat{\mu}(\mathbf{x}), R - R(R + (\tau^2/n)I_n)^{-1}R),$$

where $R$ is the $n \times n$ matrix such that its $(i,j)$-th element is $r(x_i, x_j)$.

- Recovers the frequentist shrinkage estimator of [Muandet et al, 2013] as the posterior mean (with $R$ instead of $K$).
- Allows to account for uncertainty in kernel embeddings in the inference procedures, e.g. when estimating a witness function for the two-sample test.

# Learning hyperparameters

Kernel $k = k_\theta$ typically has hyperparameters $\theta$, e.g., bandwidth of the Gaussian (SE) kernel.

**Idea**: Integrate out the kernel mean embedding $\mu_\theta$ and consider the probability of our observations $\{x_i\}_{i=1}^n$ given the hyperparameters $\theta$.

Fix a set of points $z_1, \ldots, z_m$ in $\mathcal{X} \subset \mathbb{R}^D$, with $m \geq D$.

$$\widehat{\mu_\theta}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \phi_{\mathbf{z}}(X_i)|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \frac{\tau^2}{n} I_m\right),$$

with the mapping $\phi_{\mathbf{z}} : \mathbb{R}^D \mapsto \mathbb{R}^m$, given by

$$\phi_{\mathbf{z}}(x) := [k_\theta(x, z_1), \ldots, k_\theta(x, z_m)] \in \mathbb{R}^m.$$

How good this model is depends on how far $\phi_{\mathbf{z}}(X_i)|\mu_\theta$ is from $\mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right)$. Similarly to e.g. KPCA, this is essentially a "Gaussian in the feature space" assumption. Testable using a kernel two-sample test on the RKHS [Kellner & Celisse, 2014].

# Marginal (pseudo)likelihood

Assume

$$\phi_{\mathbf{z}}(X_i)|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right).$$

and apply change of variable to the mapping $x \mapsto \phi_{\mathbf{z}}(x)$, $\phi_{\mathbf{z}} : \mathbb{R}^D \mapsto \mathbb{R}^m$: what model does this imply on the original space?

- $X|\mu_\theta, \theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right) \times \gamma_\theta(x)$, with the Jacobian term $\gamma_\theta(x) = \left(\det\left[\sum_{l=1}^m \frac{\partial k_\theta(x,z_l)}{\partial x^{(i)}} \frac{\partial k_\theta(x,z_l)}{\partial x^{(j)}}\right]_{ij}\right)^{1/2}$
- Integrate out the embedding $\mu_\theta$:

$$\begin{aligned} p(x_1,\ldots,x_n|\theta) &= \int p(x_1,\ldots,x_n|\mu_\theta,\theta)p(\mu_\theta|\theta)d\mu_\theta \\ &= \mathcal{N}\left(\phi_{\mathbf{z}}(\mathbf{x}); \mathbf{0}, \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right)\prod_{i=1}^n \gamma_\theta(x_i). \end{aligned}$$

- Computational complexity: using Kronecker structure $\mathcal{O}(m^3 + mn)$ for the Gaussian log-likelihood and $\mathcal{O}(nD^3 + nmD^2)$ for the Jacobian term (Gaussian kernel).

# Marginal (pseudo)likelihood for a challenging two-sample test

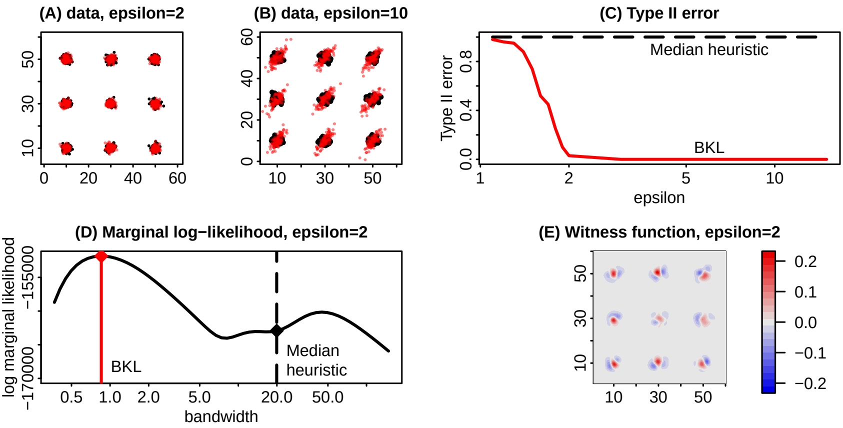
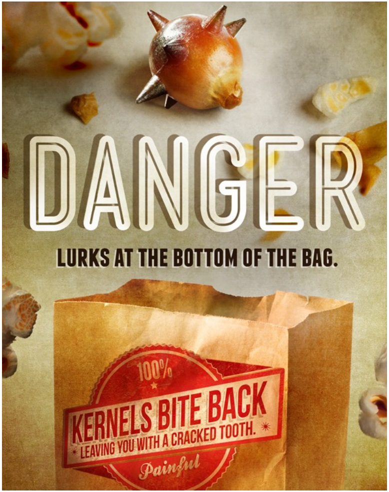



Figure: Comparing samples from a grid of isotropic Gaussians (black dots) to samples from a grid of non-isotropic Gaussians (red dots) with a ratio $\epsilon$ of largest to smallest covariance eigenvalues. BKL marginal log-likelihood is maximised for a lengthscale of 0.85 whereas the median heuristic suggests a value of 20.

# Summary

- A simple Bayesian model on kernel embeddings recovers shrinkage estimators.
- Marginal (pseudo)likelihood of observations given the kernel hyperparameters allows optimization or sampling of hyperparameters as well.
- Can discover multiscale properties in the data – where there is a mismatch between the global scale of the distribution and the scale at which differences or dependencies are present.
- Potentially a drop-in replacement for median heuristic in unsupervised settings?


DANGER
LURKS AT THE BOTTOM OF THE BAG.
100%
KERNELS BITE BACK
LEAVING YOU WITH A CRACKED TOOTH.
Painful