Inference with Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

OxWaSP Workshop, Warwick, 07/10/2016
Outline

1. Preliminaries on Kernel Embeddings
2. Using Kernel MMD as a criterion in ABC
3. Bayesian Learning of Embeddings
Outline

1 Preliminaries on Kernel Embeddings

2 Using Kernel MMD as a criterion in ABC

3 Bayesian Learning of Embeddings
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 $\mathcal{H}_k = \text{span} \{ k(\cdot, x) | x \in \mathcal{X} \}$ 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 $x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s$

- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  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 \( x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s \)

- \( \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y) \)
inner products readily available

- nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data

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 \( P \mapsto [\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)] \in \mathbb{R}^s \)

- \( \langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y) \)
inner products easy to estimate

- nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions

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

Maximum Mean Discrepancy

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

\[
\text{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|
\]
Maximum Mean Discrepancy

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

  \[
  \text{MMD}_k(P, Q) = \| \mu_k(P) - \mu_k(Q) \|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k: \| f \|_{\mathcal{H}_k} \leq 1} \left| \mathbb{E} f(X) - \mathbb{E} f(Y) \right|
  \]

- **Characteristic** kernels: $\text{MMD}_k(P, Q) = 0$ iff $P = Q$.
  - Gaussian RBF $\exp\left(-\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],

  \[
  \text{MMD}_k(P_n, P) \to 0 \iff P_n \Rightarrow P.
  \]
Some uses of MMD

within-sample average similarity

between-sample average similarity

MMD has been applied to:

- independence tests [Gretton et al, 2009]
- two-sample tests [Gretton et al, 2012]
- training generative neural networks for image data [Dziugaite, Roy & Ghahramani, 2015]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- model criticism in Automatic Statistician [Lloyd & Ghahramani, 2015]
- similarity measure between observed and simulated data in ABC [Park, Jitkrittum & DS, 2015]

\[
\text{MMD}^2_k(P, Q) = \mathbb{E}_{X, X' \sim_i \sim d. P} k(X, X') + \mathbb{E}_{Y, Y' \sim_i \sim d. Q} k(Y, Y') - 2 \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).
\]
Kernel dependence measures

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

- Dependence witness is a smooth function in the RKHS \( \mathcal{H}_\kappa \) of functions on \( X \times Y \)
  \[ \kappa(\boxed{1}, \boxed{2}) = k(\boxed{1}, \boxed{1}) \times l(\boxed{1}, \boxed{2}) \]

- Independence testing framework that generalises Distance Covariance (dCov): HSIC with Brownian motion covariance kernels
  
  [Szekely et al, 2007; DS et al, 2013]
Kernel dependence measures (2)

The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps...

Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament...

Hilbert-Schmidt Independence Criterion (HSIC): similarity between the kernel matrices

\[
\langle \tilde{K}, \tilde{L} \rangle = \text{Tr} \left( \tilde{K} \tilde{L} \right), \text{ where } \tilde{K} = HKH, \text{ and }
\]

\[
H = I - \frac{1}{n} 11^\top
\]

is the centering matrix.

[Gretton et al, 2008; Fukumizu et al, 2008; Song et al, 2012]
1 Preliminaries on Kernel Embeddings

2 Using Kernel MMD as a criterion in ABC

3 Bayesian Learning of Embeddings
K2-ABC: Approximate Bayesian Computation with Kernel Embeddings.
AISTATS 2016
Mijung Park, Wittawat Jitkrittum, and DS.
Code: https://github.com/wittawatj/k2abc
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]

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

where $e_t \sim \text{Gamma} \left( \frac{1}{\sigma_p^2}, \sigma_p^2 \right)$, $\epsilon_t \sim \text{Gamma} \left( \frac{1}{\sigma_d^2}, \sigma_d^2 \right)$.

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 $\rho(\mathbf{X}, \mathbf{Y})$?
Observe a dataset $Y$,

\[ p(\theta|Y) \propto p(\theta)p(Y|\theta) \]

\[ = p(\theta) \int p(X|\theta) \, d\delta_Y(X) \]

\[ \approx p(\theta) \int p(X|\theta)\kappa\epsilon(X, Y) \, dX, \]

where $\kappa\epsilon(X, Y)$ defines similarity of $X$ and $Y$.

\[(\text{ABC likelihood}) \quad p_\epsilon(Y|\theta) := \int p(X|\theta)\kappa\epsilon(X, Y) \, dX.\]

Simplest choices for $\kappa\epsilon$: $1(\rho(X, Y) < \epsilon)$ or $\exp(-\rho^2(X, Y)/\epsilon)$

- $\rho$: a distance function between observed and simulated data
Data Similarity via Summary Statistics

- Distance $\rho$ is typically defined via summary statistics
  \[ \rho(X, Y) = \| s(X) - s(Y) \|_2. \]

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

- Distance $\rho$ is typically defined via summary statistics
  
  \[ \rho(X, Y) = \| s(X) - s(Y) \|_2. \]

- How to select the summary statistics $s(\cdot)$? Unless $s(\cdot)$ is sufficient, targets the incorrect (partial) posterior $p(\theta|s(Y))$ rather than $p(\theta|Y)$.

- Hard to quantify additional bias.
  - Adding more summary statistics decreases "information loss":
    \[ p(\theta|s(Y)) \approx p(\theta|Y) \]
  - $\rho$ computed on a higher dimensional space - without appropriate calibration of distances therein, leads to a higher rejection rate so need to increase $\epsilon$: $p_\epsilon(\theta|s(Y)) \not\approx p(\theta|s(Y))$

- Contribution: Use a nonparametric distance (MMD) between the empirical measures of datasets $X$ and $Y$.
  - No need to design $s(\cdot)$.
  - Rejection rate does not blow up since MMD penalises the higher order moments via Mercer expansion.
Embeddings via Mercer Expansion

Mercer Expansion

For a compact metric space $\mathcal{X}$, and a continuous kernel $k$,

$$k(x, y) = \sum_{r=1}^{\infty} \lambda_r e_r(x) e_r(y),$$

with $\{\lambda_r, e_r\}_{r \geq 1}$ eigenvalue, eigenfunction pairs of $f \mapsto \int f(x) k(\cdot, x) dP(x)$ on $L_2(P)$, with $\lambda_r \to 0$, as $r \to \infty$. $e_r$ are typically functions of increasing “complexity”, i.e., Hermite polynomials of increasing degree.

$$\mathcal{H}_k \ni k(\cdot, x) \iff \left\{ \sqrt{\lambda_r} e_r(x) \right\} \in \ell_2$$

$$\mathcal{H}_k \ni \mu_k(P) \iff \left\{ \sqrt{\lambda_r} \mathbb{E} e_r(X) \right\} \in \ell_2$$

$$\left\| \mu_k(\hat{P}) - \mu_k(\hat{Q}) \right\|_{\mathcal{H}_k}^2 = \sum_{r=1}^{\infty} \lambda_r \left( \frac{1}{n_x} \sum_{t=1}^{n_x} e_r(X_t) - \frac{1}{n_y} \sum_{t=1}^{n_y} e_r(Y_t) \right)^2$$
**K2-ABC (proposed method)**

- **Input:** observed data $Y$, threshold $\epsilon$
- **Output:** Empirical posterior $\sum_{i=1}^{M} w_i \delta_{\theta_i}$

1. for $i = 1, \ldots, M$ do
2. Sample $\theta_i \sim p(\theta)$
3. Sample pseudo dataset $X_i \sim p(\cdot | \theta_i)$
4. $\tilde{w}_i = \kappa_\epsilon(X_i, Y) = \exp\left(-\frac{\text{MMD}^2(X_i, Y)}{\epsilon}\right)$
5. end for
6. $w_i = \tilde{w}_i / \sum_{j=1}^{M} \tilde{w}_j$ for $i = 1, \ldots, M$

- Two kernels: $k$ (in MMD) and $\kappa_\epsilon$, hence “K2”
Blow Fly Population Modelling

Number of blow flies over time

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

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

Simulated trajectories with inferred posterior mean of \( \theta \)

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

Computation of \( \widehat{\text{MMD}}^2(X,Y) \) costs \( O(n^2) \).
- Linear-time unbiased estimators of \( \text{MMD}^2 \) or random feature expansions reduce the cost to \( O(n) \).
Summary: K2-ABC

- A dissimilarity criterion for ABC based on MMD between empirical distributions of observed and simulated data.
- No “information loss” due to insufficient statistics.
- Simple and effective when parameters model marginal distribution of observations (variants for conditional distributions readily available).
Outline

1 Preliminaries on Kernel Embeddings

2 Using Kernel MMD as a criterion in ABC

3 Bayesian Learning of Embeddings
Right... But how do you choose your kernel?

- Frequentists cross-validate, Bayesians optimize marginal likelihood...
- But with kernel embeddings, neither is typically available (e.g. hypothesis testing or ABC).
- **Median heuristic**: bandwidth parameter
  \[ \theta = \text{median}(\|x_i - x_j\|_2) \] for e.g. Gaussian kernel
  \[ k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\theta^2}\right) \]

---

**Bayesian Learning of Kernel Embeddings.**
**UAI 2016.**
Seth Flaxman, DS, John Cunningham, and Sarah Filippi.
Bayesian Model for Embeddings

- In MMD and HSIC, we estimate embedding $\mu = \int k(\cdot, x)P(dx)$ 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?
A classical result, Kallianpur’s 0-1 law, [Kallianpur, 1970; Wahba, 1990]: sample paths of a GP with kernel \( k \) lie outside RKHS \( \mathcal{H}_k \) with probability 1.

Recall Mercer’s expansion \( k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x)e_i(x') \), for the eigenvalue-eigenfunction pairs \( \{(\lambda_i, e_i)\}_{i=1}^{n} \), which gives representation

\[
f \sim \mathcal{GP}(0,k) : \quad f = \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i e_i, \quad \{Z_i\}_{i=1}^{\infty} \text{i.i.d. } \sim \mathcal{N}(0,1).
\]

But then \( \|f\|^2_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{\lambda_i Z_i^2}{\lambda_i} = \sum_{i=1}^{\infty} Z_i^2 = \infty \) so \( f \notin \mathcal{H}_k \) a.s.

However, one can use a prior \( f \sim \mathcal{GP}(0,r) \) with

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

for any finite measure \( \nu \) in which case \( f \in \mathcal{H}_k \) with probability 1: nuclear dominance theory established by [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 $\hat{\mu}$ induced by $\{x_i\}_{i=1}^n$ at some $x \in X$ - we link $\hat{\mu}(x)$ to $\mu(x)$ using a Gaussian distribution with variance $\tau^2/n$:

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

Motivation by the Central Limit Theorem:

$$\sqrt{n}(\hat{\mu}(x) - \mu(x)) \xrightarrow{D} \mathcal{N}(0, \text{var}_{X \sim \mathcal{P}[k(X, x)]}).$$

A heteroscedastic noise model is certainly more appropriate, but let’s keep this (obviously wrong) model for now.
Standard conjugacy results give:

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

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.
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 $m \geq D$.

$$
\hat{\mu}_\theta(z) = \frac{1}{n} \sum_{i=1}^n \phi_z(X_i)|\mu_\theta \sim \mathcal{N} \left( \mu_\theta(z), \frac{\tau^2}{n} I_m \right),
$$

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

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

How good this model is depends on how far $\phi_z(X_i)|\mu_\theta$ is from $\mathcal{N} \left( \mu_\theta(z), \tau^2 I_m \right)$. 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

\[ \phi_{z}(X_i | \mu_\theta) \sim \mathcal{N}(\mu_\theta(z), \tau^2 I_m). \]

and apply change of variable to the mapping \( x \mapsto \phi_z(x), \phi_z : \mathbb{R}^D \rightarrow \mathbb{R}^m \):

what model does this imply on the original space?

\[
p(x_1, \ldots, x_n | \theta) = \int p(x_1, \ldots, x_n | \mu_\theta, \theta)p(\mu_\theta | \theta) d\mu_\theta \\
= \int \mathcal{N}(\phi_z(x); [\mu_\theta(z)^\top \cdots \mu_\theta(z)^\top]^\top, \tau^2 I_{mn}) \left[ \prod_{i=1}^{n} \gamma_\theta(x_i) \right] p(\mu_\theta | \theta) d\mu_\theta \\
= \mathcal{N}(\phi_z(x); 0, 1_n 1_n^\top \otimes R_{\theta,zz} + \tau^2 I_{mn}) \prod_{i=1}^{n} \gamma_\theta(x_i).
\]

- Jacobian term: \( \gamma_\theta(x) = \left( \det \left[ \sum_{l=1}^{m} \frac{\partial k_\theta(x,z_l)}{\partial x^{(i)}} \frac{\partial k_\theta(x,z_l)}{\partial x^{(j)}} \right]_{ij} \right)^{1/2} \).

- 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).
**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.
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