Kernel Mean Embeddings

George Hron, Adam Ścibior

Department of Engineering
University of Cambridge

Reading Group, March 2017

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Kernels

A kernel k is a positive definite function $\mathcal{X} \times \mathcal{X} \rightarrow$, that is for all $a_1, \ldots, a_n \in$, $x_1, \ldots, x_n \in \mathcal{X}$,

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Intuitively a kernel is a measure of similarity between two elements of $\mathcal{X}$.

Kernels

Commonly used kernels:

- polynomial

$$(\langle x_1, x_2 \rangle + 1)^d$$

- Gaussian RBF

$$e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

- Laplace

$$e^{-\frac{\|x-y\|}{\sigma}}$$

RKHS

A reproducing kernel Hilbert space (RKHS) for a kernel k is one spanned by functions $k(x, \cdot)$ for all $x \in \mathcal{X}$ with the inner product defined by

$$\langle k(x_1, \cdot), k(x_2, \cdot) \rangle = k(x_1, x_2)$$

which is well-defined on all of $\mathcal{H}$ by linearity of the inner product.

The equation above is known as the kernel trick and lets us treat $\mathcal{H}$ as an implicit feature space, where we never have to explicitly evaluate the feature map.

RKHS

$\mathcal{H}$ has the reproducing property, that is for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$,

$$\langle f, k(x, \cdot) \rangle = f(x)$$

Outline

- RKHS
- Kernel Mean Embeddings
- Characteristic kernels
- Two Sample Testing
  - MMD
  - Kernelised Stein Discrepancy
- Kernel Bayes' Rule

Expectations in RKHS

Goal: evaluating expected values of functions from an RKHS.

As with the kernel trick, it turns out that this is possible in terms of inner products, making the computation analytically tractable,

$$\mathop{\mathbb{E}}_{\mathrm{p}}[f(x)] = \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) f(x) = \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) \langle k(x,\cdot), f\rangle_{\mathcal{H}_k}$$
$$\stackrel{?}{=} \langle \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) k(x,\cdot), f\rangle_{\mathcal{H}_k} =: \langle \mu_{\mathrm{p}}, f\rangle_{\mathcal{H}_k}$$

μ_{p} is so called *kernel mean embedding* (KME) of distribution p.

Note: $\mathcal{X}$ assumed measurable throughout the whole presentation.

Expectations in RKHS

Goal: evaluating expected values of functions from an RKHS.

As with the kernel trick, it turns out that this is possible in terms of inner products, making the computation analytically tractable,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{p}}[f(x)] &= \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) f(x) = \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) \langle k(x,\cdot), f\rangle_{\mathcal{H}_k} \\ &\stackrel{?}{=} \langle \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) k(x,\cdot), f\rangle_{\mathcal{H}_k} =: \langle \mu_{\mathrm{p}}, f\rangle_{\mathcal{H}_k}\end{aligned}$$

μ_{p} is so called *kernel mean embedding* (KME) of distribution p.

Note: $\mathcal{X}$ assumed measurable throughout the whole presentation.

Expectations in RKHS

Goal: evaluating expected values of functions from an RKHS.

As with the kernel trick, it turns out that this is possible in terms of inner products, making the computation analytically tractable,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{p}}[f(x)] &= \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) f(x) = \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) \langle k(x, \cdot), f \rangle_{\mathcal{H}_k} \\ &\stackrel{?}{=} \langle \int_{\mathcal{X}} \mathrm{p}(\mathrm{d}x) k(x, \cdot), f \rangle_{\mathcal{H}_k} =: \langle \mu_{\mathrm{p}}, f \rangle_{\mathcal{H}_k}\end{aligned}$$

μ_{p} is so called *kernel mean embedding* (KME) of distribution p.

Note: $\mathcal{X}$ assumed measurable throughout the whole presentation.

Existence of KME

Riesz representation theorem: For every *bounded linear functional* $\mathcal{T} : \mathcal{H} \to \mathbb{R}$ (resp. $\mathcal{T} : \mathcal{H} \to \mathbb{C}$), there exists a unique $g \in \mathcal{H}$ such that $\mathcal{T}(f) = \langle g, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Expectation is a linear functional, and $\forall f \in \mathcal{H}$ we have,

$$\mathop{\mathbb{E}}_{\mathrm{p}} f(x) \leq \left| \mathop{\mathbb{E}}_{\mathrm{p}} f(x) \right| \leq \mathop{\mathbb{E}}_{\mathrm{p}} \left| f(x) \right| = \mathop{\mathbb{E}}_{\mathrm{p}} \left| \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \right| \leq \|f\|_{\mathcal{H}} \mathop{\mathbb{E}}_{\mathrm{p}} \left\| k(x, \cdot) \right\|_{\mathcal{H}}$$

Thus if $\mathbb{E}_{\mathrm{p}} \sqrt{k(x, x)} < \infty$, then $\mu_{\mathrm{p}} \in \mathcal{H}$ exists, and $\mathbb{E}_{\mathrm{p}} f(x) = \langle \mu_{\mathrm{p}}, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Existence of KME

Riesz representation theorem: For every *bounded linear functional* $\mathcal{T} : \mathcal{H} \to \mathbb{R}$ (resp. $\mathcal{T} : \mathcal{H} \to \mathbb{C}$), there exists a unique $g \in \mathcal{H}$ such that $\mathcal{T}(f) = \langle g, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Expectation is a linear functional, and $\forall f \in \mathcal{H}$ we have,

$$\mathop{\mathbb{E}}_{\mathrm{p}} f(x) \leq \left| \mathop{\mathbb{E}}_{\mathrm{p}} f(x) \right| \leq \mathop{\mathbb{E}}_{\mathrm{p}} \left| f(x) \right| = \mathop{\mathbb{E}}_{\mathrm{p}} \left| \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \right| \leq \| f \|_{\mathcal{H}} \mathop{\mathbb{E}}_{\mathrm{p}} \left\| k(x, \cdot) \right\|_{\mathcal{H}}$$

Thus if $\mathbb{E}_{\mathrm{p}} \sqrt{k(x, x)} < \infty$, then $\mu_{\mathrm{p}} \in \mathcal{H}$ exists, and $\mathbb{E}_{\mathrm{p}} f(x) = \langle \mu_{\mathrm{p}}, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Existence of KME

Riesz representation theorem: For every *bounded linear functional* $\mathcal{T}: \mathcal{H} \to \mathbb{R}$ (resp. $\mathcal{T}: \mathcal{H} \to \mathbb{C}$), there exists a unique $g \in \mathcal{H}$ such that $\mathcal{T}(f) = \langle g, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Expectation is a linear functional, and $\forall f \in \mathcal{H}$ we have,

$$\mathop{\mathbb{E}}_{\mathrm{p}} f(x) \leq \left| \mathop{\mathbb{E}}_{\mathrm{p}} f(x) \right| \leq \mathop{\mathbb{E}}_{\mathrm{p}} \left| f(x) \right| = \mathop{\mathbb{E}}_{\mathrm{p}} \left| \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \right| \leq \|f\|_{\mathcal{H}} \mathop{\mathbb{E}}_{\mathrm{p}} \left\| k(x, \cdot) \right\|_{\mathcal{H}}$$

Thus if $\mathbb{E}_{\mathrm{p}} \sqrt{k(x, x)} < \infty$, then $\mu_{\mathrm{p}} \in \mathcal{H}$ exists, and $\mathbb{E}_{\mathrm{p}} f(x) = \langle \mu_{\mathrm{p}}, f \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$.

Estimating KME

Because μ_{p} is generally unknown, it has to be estimated.

A natural (and minimax optimal) estimator is the sample mean,

$$\hat{\mu}_{\mathrm{p}} := \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k(x_n, \cdot)$$

and in particular,

$$\hat{\mu}_{\mathrm{p}}(x) = \langle \hat{\mu}_{\mathrm{p}}, k(x, \cdot) \rangle_{\mathcal{H}} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k(x_n, x) \xrightarrow{\mathrm{N} \to \infty} \mathop{\mathbb{E}}_{\mathrm{p}(x')} k(x', x)$$

Estimating KME

Because μ_{p} is generally unknown, it has to be estimated.

A natural (and minimax optimal) estimator is the sample mean,

$$\hat{\mu}_{\mathrm{p}} := \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k(x_n, \cdot)$$

and in particular,

$$\hat{\mu}_{\mathrm{p}}(x) = \langle \hat{\mu}_{\mathrm{p}}, k(x, \cdot) \rangle_{\mathcal{H}} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k(x_n, x) \xrightarrow{\mathrm{N} \to \infty} \mathop{\mathbb{E}}_{\mathrm{p}(x')} k(x', x)$$

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Representing Distributions via KME

A positive definite kernel is called *characteristic* if,

$$\begin{aligned} \mathcal{T} : \mathcal{M}_1^+(\mathcal{X}) &\to \mathcal{H} \\ \mathrm{p} &\mapsto \mu_{\mathrm{p}} \end{aligned}$$

is injective; $\mathcal{M}_1^+(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$. [5, 6]

Characteristic kernels

Proving a kernel is characteristic is non-trivial in general, but sufficient conditions exist. Three well known examples:

- *Universality*: If k is continuous, $\mathcal{X}$ compact, and $\mathcal{H}_k$ dense in $\mathcal{C}(\mathcal{X})$ wrt L_∞, then k is characteristic. [6, 8]

- *Integral strict positive definiteness*: A bounded measurable kernel k is called *integrally strictly positive definite* if $\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, \mu(\mathrm{d}x)\mu(\mathrm{d}y) > 0$ for all non–zero finite signed Borel measures μ on $\mathcal{X}$. [22]

- *Some stationary kernels*: For $\mathcal{X} = \mathbb{R}^d$, a stationary kernel k is characteristic **iff** $\operatorname{supp} \Lambda(\omega) = \mathbb{R}^d$, where $\Lambda(\omega)$ is the spectral density of k (cf. Bochner's theorem). [22]

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

t-test

Have $\{x_1, \ldots, x_{\mathrm{N}}\} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$, and $\{y_1, \ldots, y_{\mathrm{M}}\} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$, where all parameters are unknown.

Declare $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$. Then for,

$$\hat{\mu}_1 := \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} x_n \qquad \hat{\mu}_2 := \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} y_m$$

$\hat{\mu}_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i^2}{\mathrm{N}}), \forall i \in \{1, 2\}$, and $\hat{\mu}_1 - \hat{\mu}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{\mathrm{N}} + \frac{\sigma_2^2}{\mathrm{M}})$.

Hence $t \overset{H_0}{\sim} \mathrm{t}_{\,\nu}$, $\nu := \frac{(s_1^2/\mathrm{N} + s_2^2/\mathrm{M})^2}{\frac{(s_1^2/\mathrm{N})^2}{\mathrm{N}-1} + \frac{(s_2^2/\mathrm{M})^2}{\mathrm{M}-1}}$, where,

$$t := \frac{(\hat{\mu}_1 - \hat{\mu}_2) - 0}{\sqrt{\frac{s_1^2}{\mathrm{N}} + \frac{s_2^2}{\mathrm{M}}}}, \quad s_1^2 := \frac{\sum_{i=1}^{\mathrm{N}} (x_i - \hat{\mu}_1)^2}{\mathrm{N} - 1}, \quad s_2^2 := \frac{\sum_{i=1}^{\mathrm{M}} (y_i - \hat{\mu}_2)^2}{\mathrm{M} - 1}.$$

t-test

Have $\{x_1, \ldots, x_{\mathrm{N}}\} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$, and $\{y_1, \ldots, y_{\mathrm{M}}\} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$, where all parameters are unknown.

Declare $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$. Then for,

$$\hat{\mu}_1 := \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} x_n \qquad \hat{\mu}_2 := \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} y_m$$

$\hat{\mu}_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i^2}{\mathrm{N}}), \forall i \in \{1, 2\}$, and $\hat{\mu}_1 - \hat{\mu}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{\mathrm{N}} + \frac{\sigma_2^2}{\mathrm{M}})$.

Hence $t \overset{H_0}{\sim} \mathrm{t}_{\,\nu}$, $\nu := \frac{(s_1^2/\mathrm{N} + s_2^2/\mathrm{M})^2}{\frac{(s_1^2/\mathrm{N})^2}{\mathrm{N}-1} + \frac{(s_2^2/\mathrm{M})^2}{\mathrm{M}-1}}$, where,

$$t := \frac{(\hat{\mu}_1 - \hat{\mu}_2) - 0}{\sqrt{\frac{s_1^2}{\mathrm{N}} + \frac{s_2^2}{\mathrm{M}}}}, \quad s_1^2 := \frac{\sum_{i=1}^{\mathrm{N}} (x_i - \hat{\mu}_1)^2}{\mathrm{N} - 1}, \quad s_2^2 := \frac{\sum_{i=1}^{\mathrm{M}} (y_i - \hat{\mu}_2)^2}{\mathrm{M} - 1}.$$

t-test

Have $\{x_1, \ldots, x_{\mathrm{N}}\} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$, and $\{y_1, \ldots, y_{\mathrm{M}}\} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$, where all parameters are unknown.

Declare $H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$. Then for,

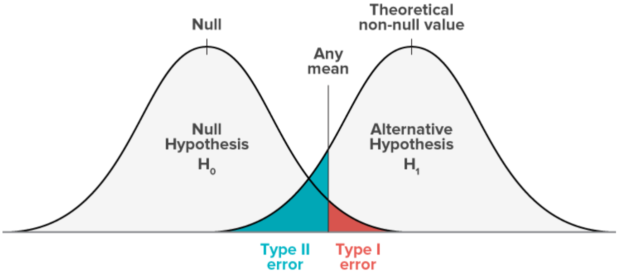
$$\hat{\mu}_1 := \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} x_n \qquad \hat{\mu}_2 := \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} y_m$$

$\hat{\mu}_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i^2}{\mathrm{N}}), \forall i \in \{1, 2\}$, and $\hat{\mu}_1 - \hat{\mu}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{\mathrm{N}} + \frac{\sigma_2^2}{\mathrm{M}})$.

Hence $t \overset{H_0}{\sim} \mathrm{t}_{\,\nu}$, $\nu := \frac{(s_1^2/\mathrm{N} + s_2^2/\mathrm{M})^2}{\frac{(s_1^2/\mathrm{N})^2}{\mathrm{N}-1} + \frac{(s_2^2/\mathrm{M})^2}{\mathrm{M}-1}}$, where,

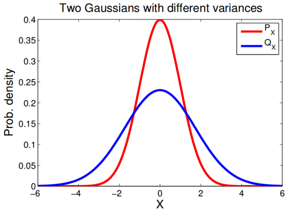
$$t := \frac{(\hat{\mu}_1 - \hat{\mu}_2) - 0}{\sqrt{\frac{s_1^2}{\mathrm{N}} + \frac{s_2^2}{\mathrm{M}}}}, \quad s_1^2 := \frac{\sum_{i=1}^{\mathrm{N}} (x_i - \hat{\mu}_1)^2}{\mathrm{N} - 1}, \quad s_2^2 := \frac{\sum_{i=1}^{\mathrm{M}} (y_i - \hat{\mu}_2)^2}{\mathrm{M} - 1}.$$

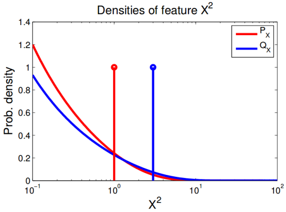
t-test



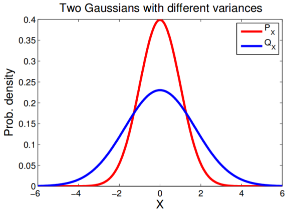
http://grasshopper.com/blog/the-errors-of-ab-testing-your-conclusions-can-make-things-worse/

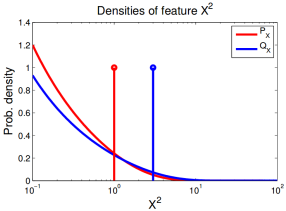
More examples

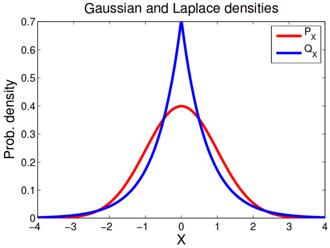




More examples







Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Recap

Kernel mean embedding,

$$\mathbb{E}_{\mathrm{p}(\boldsymbol{x})}[f(\boldsymbol{x})] = \langle f, \mu_{\mathrm{p}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

For a characteristic kernel k and the corresponding RKHS $\mathcal{H}$,

$$\mu_{\mathrm{p}} = \mu_{\mathrm{q}} \text{ iff } \mathrm{p} = \mathrm{q}\,.$$

This means we might be able to distinguish distributions by comparing the corresponding kernel mean embeddings.

Recap

Kernel mean embedding,

$$\mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})}[f(\boldsymbol{x})] = \langle f, \mu_{\mathrm{p}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

For a characteristic kernel k and the corresponding RKHS $\mathcal{H}$,

$$\mu_{\mathrm{p}} = \mu_{\mathrm{q}} \text{ iff } \mathrm{p} = \mathrm{q}\,.$$

This means we might be able to distinguish distributions by comparing the corresponding kernel mean embeddings.

Recap

Kernel mean embedding,

$$\mathop{\mathbb{E}}_{\mathrm{p}\,(\boldsymbol{x})}[f(\boldsymbol{x})] = \langle f, \mu_{\mathrm{p}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

For a characteristic kernel k and the corresponding RKHS $\mathcal{H}$,

$$\mu_{\mathrm{p}} = \mu_{\mathrm{q}} \text{ iff } \mathrm{p} = \mathrm{q}\,.$$

This means we might be able to distinguish distributions by comparing the corresponding kernel mean embeddings.

Maximum Mean Discrepancy (MMD)

Measure distance between mean embeddings by the worst case difference of expected values [9],

$$\begin{aligned}\mathrm{MMD}(\mathrm{p}, \mathrm{q}, \mathcal{H}) &:= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left(\mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})}[f(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{y})}[f(\boldsymbol{y})] \right) \\ &= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left(\langle \mu_{\mathrm{p}} - \mu_{\mathrm{q}}, f \rangle_{\mathcal{H}} \right)\end{aligned}$$

Notice that $\langle \mu_{\mathrm{p}} - \mu_{\mathrm{q}}, f \rangle_{\mathcal{H}} \leq \left\| \mu_{\mathrm{p}} - \mu_{\mathrm{q}} \right\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$, with equality iff $f \propto \mu_{\mathrm{p}} - \mu_{\mathrm{q}}$. Hence,

$$\mathrm{MMD}(\mathrm{p}, \mathrm{q}, \mathcal{H}) = \left\| \mu_{\mathrm{p}} - \mu_{\mathrm{q}} \right\|_{\mathcal{H}}$$

Maximum Mean Discrepancy (MMD)

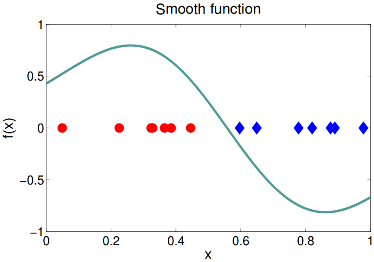
Measure distance between mean embeddings by the worst case difference of expected values [9],

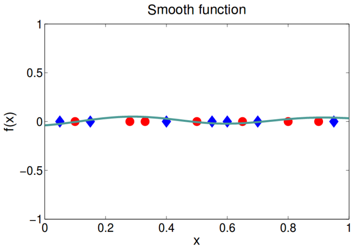
$$\begin{aligned}\mathrm{MMD}(\mathrm{p}, \mathrm{q}, \mathcal{H}) &:= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left(\mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})}[f(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{y})}[f(\boldsymbol{y})] \right) \\ &= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left(\langle \mu_{\mathrm{p}} - \mu_{\mathrm{q}}, f \rangle_{\mathcal{H}} \right)\end{aligned}$$

Notice that $\langle \mu_{\mathrm{p}} - \mu_{\mathrm{q}}, f \rangle_{\mathcal{H}} \leq \left\| \mu_{\mathrm{p}} - \mu_{\mathrm{q}} \right\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$, with equality iff $f \propto \mu_{\mathrm{p}} - \mu_{\mathrm{q}}$. Hence,

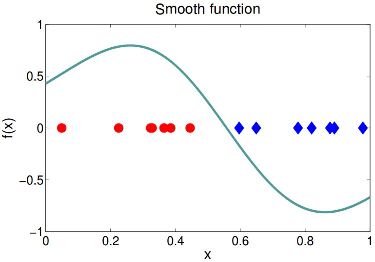
$$\mathrm{MMD}(\mathrm{p}, \mathrm{q}, \mathcal{H}) = \left\| \mu_{\mathrm{p}} - \mu_{\mathrm{q}} \right\|_{\mathcal{H}}$$

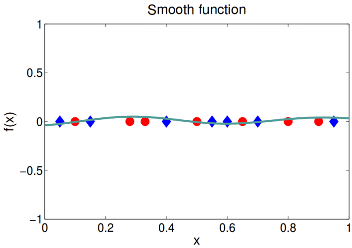
Witness function

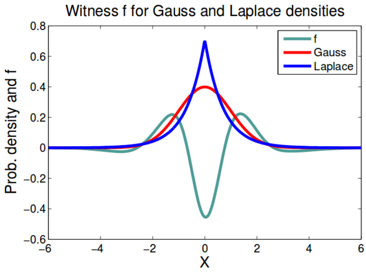




Witness function







Estimation

Recall: empirical kernel mean embedding estimator,

$$\hat{\mu}_{\mathrm{p}} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k\left(x_n, \cdot\right)$$

We can estimate the square of MMD by substituting the empirical estimator. For $\{x_i\}_{i=1}^{\mathrm{N}} \overset{iid}{\sim} \mathrm{p}$ and $\{y_i\}_{i=1}^{\mathrm{N}} \overset{iid}{\sim} \mathrm{q}$,

$$\begin{aligned}\widehat{\mathrm{MMD}}^2 =& \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}} \sum_{j \neq i}^{\mathrm{N}} \left(k\left(x_i, x_j\right) + k\left(y_i, y_j\right)\right) \\ & - \frac{1}{\mathrm{N}^2} \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} \left(k\left(x_i, y_j\right) + k\left(x_j, y_i\right)\right)\end{aligned}$$

Estimation

Recall: empirical kernel mean embedding estimator,

$$\hat{\mu}_{\mathrm{p}} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} k\left(x_n, \cdot\right)$$

We can estimate the square of MMD by substituting the empirical estimator. For $\{x_i\}_{i=1}^{\mathrm{N}} \overset{iid}{\sim} \mathrm{p}$ and $\{y_i\}_{i=1}^{\mathrm{N}} \overset{iid}{\sim} \mathrm{q}$,

$$\begin{aligned}\widehat{\mathrm{MMD}}^2 =& \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}} \sum_{j \neq i}^{\mathrm{N}} \left(k\left(x_i, x_j\right) + k\left(y_i, y_j\right)\right) \\ &- \frac{1}{\mathrm{N}^2} \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} \left(k\left(x_i, y_j\right) + k\left(x_j, y_i\right)\right)\end{aligned}$$

Estimation

Proof sketch:

$$\mathrm{MMD}^2(\mathrm{p}, \mathrm{q}, \mathcal{H}) = \left\|\mu_{\mathrm{p}} - \mu_{\mathrm{q}}\right\|_{\mathcal{H}}^2 = \left\|\mu_{\mathrm{p}}\right\|_{\mathcal{H}}^2 + \left\|\mu_{\mathrm{q}}\right\|_{\mathcal{H}}^2 - 2\langle \mu_{\mathrm{p}}, \mu_{\mathrm{q}} \rangle_{\mathcal{H}}$$

where,

$$\begin{aligned}\left\|\mu_{\mathrm{p}}\right\|_{\mathcal{H}}^2 = \langle \mu_{\mathrm{p}}, \mu_{\mathrm{p}} \rangle_{\mathcal{H}} = \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \mu_{\mathrm{p}}(x) = \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \langle \mu_{\mathrm{p}}, k(x, \cdot) \rangle_{\mathcal{H}} \\ = \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \mathop{\mathbb{E}}_{\tilde{x}\sim \mathrm{p}} k(x, \tilde{x}) \approx \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}} \sum_{j\neq i}^{\mathrm{N}} k(x_i, x_j)\end{aligned}$$

Similarly for the other terms.

Estimation

Proof sketch:

$$\mathrm{MMD}^2(\mathrm{p}, \mathrm{q}, \mathcal{H}) = \left\|\mu_\mathrm{p} - \mu_\mathrm{q}\right\|_\mathcal{H}^2 = \left\|\mu_\mathrm{p}\right\|_\mathcal{H}^2 + \left\|\mu_\mathrm{q}\right\|_\mathcal{H}^2 - 2\langle\mu_\mathrm{p}, \mu_\mathrm{q}\rangle_\mathcal{H}$$

where,

$$\begin{aligned}\left\|\mu_\mathrm{p}\right\|_\mathcal{H}^2 &= \langle\mu_\mathrm{p}, \mu_\mathrm{p}\rangle_\mathcal{H} = \underset{x\sim\mathrm{p}}{\mathbb{E}}\, \mu_\mathrm{p}(x) = \underset{x\sim\mathrm{p}}{\mathbb{E}} \langle\mu_\mathrm{p}, k(x,\cdot)\rangle_\mathcal{H} \\ &= \underset{x\sim\mathrm{p}}{\mathbb{E}}\ \underset{\tilde{x}\sim\mathrm{p}}{\mathbb{E}}\, k(x,\tilde{x}) \approx \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}} \sum_{j\neq i}^{\mathrm{N}} k(x_i, x_j)\end{aligned}$$

Similarly for the other terms.

Estimation

Proof sketch:

$$\mathsf{MMD}^2(\mathrm{p}, \mathrm{q}, \mathcal{H}) = \left\|\mu_{\mathrm{p}} - \mu_{\mathrm{q}}\right\|_{\mathcal{H}}^2 = \left\|\mu_{\mathrm{p}}\right\|_{\mathcal{H}}^2 + \left\|\mu_{\mathrm{q}}\right\|_{\mathcal{H}}^2 - 2\langle \mu_{\mathrm{p}}, \mu_{\mathrm{q}} \rangle_{\mathcal{H}}$$

where,

$$\begin{aligned}\left\|\mu_{\mathrm{p}}\right\|_{\mathcal{H}}^2 = \langle \mu_{\mathrm{p}}, \mu_{\mathrm{p}} \rangle_{\mathcal{H}} &= \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \mu_{\mathrm{p}}(x) = \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \langle \mu_{\mathrm{p}}, k(x, \cdot) \rangle_{\mathcal{H}} \\ &= \mathop{\mathbb{E}}_{x\sim \mathrm{p}} \mathop{\mathbb{E}}_{\tilde{x}\sim \mathrm{p}} k(x, \tilde{x}) \approx \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}} \sum_{j\neq i}^{\mathrm{N}} k(x_i, x_j)\end{aligned}$$

Similarly for the other terms.

Considerations

- The distribution of $\widehat{\mathrm{MMD}}^2$ under H_0 assymptotically approaches an infinite sum of shifted chi-squared random variables multiplied by eigenvalues of the RKHS.
  - Approximations to the sampling distribution [1, 10, 11, 12].
- Calculation of the naive $\widehat{\mathrm{MMD}}^2$ estimator is $\mathcal{O}(\mathrm{N}^2)$.
  - Linear time approximations [2, 26].
- Performance is dependent on choice of the kernel. There is no universally best performing kernel.
  - Previously heuristical, recently replaced by hyperparameter optimisation based on test power, or Bayesian evidence [4, 24].
- Empirical kernel mean estimator might be suboptimal.
  - Better estimators exist [4, 16, 17].

Considerations

- The distribution of $\widehat{\mathrm{MMD}}^2$ under H_0 assymptotically approaches an infinite sum of shifted chi-squared random variables multiplied by eigenvalues of the RKHS.
  - Approximations to the sampling distribution [1, 10, 11, 12].
- Calculation of the naive $\widehat{\mathrm{MMD}}^2$ estimator is $\mathcal{O}(\mathrm{N}^2)$.
  - Linear time approximations [2, 26].
- Performance is dependent on choice of the kernel. There is no universally best performing kernel.
  - Previously heuristical, recently replaced by hyperparameter optimisation based on test power, or Bayesian evidence [4, 24].
- Empirical kernel mean estimator might be suboptimal.
  - Better estimators exist [4, 16, 17].

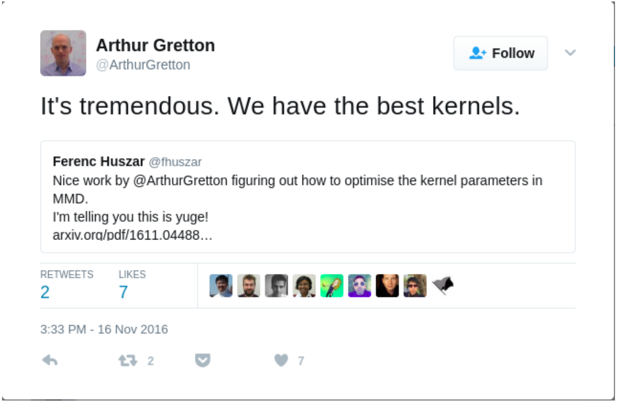
Considerations

- The distribution of $\widehat{\mathrm{MMD}}^2$ under H_0 assymptotically approaches an infinite sum of shifted chi-squared random variables multiplied by eigenvalues of the RKHS.
  - Approximations to the sampling distribution [1, 10, 11, 12].
- Calculation of the naive $\widehat{\mathrm{MMD}}^2$ estimator is $\mathcal{O}(\mathrm{N}^2)$.
  - Linear time approximations [2, 26].
- Performance is dependent on choice of the kernel. There is no universally best performing kernel.
  - Previously heuristical, recently replaced by hyperparameter optimisation based on test power, or Bayesian evidence [4, 24].
- Empirical kernel mean estimator might be suboptimal.
  - Better estimators exist [4, 16, 17].

Considerations

- The distribution of $\widehat{\mathrm{MMD}}^2$ under H_0 assymptotically approaches an infinite sum of shifted chi-squared random variables multiplied by eigenvalues of the RKHS.
  - Approximations to the sampling distribution [1, 10, 11, 12].
- Calculation of the naive $\widehat{\mathrm{MMD}}^2$ estimator is $\mathcal{O}(\mathrm{N}^2)$.
  - Linear time approximations [2, 26].
- Performance is dependent on choice of the kernel. There is no universally best performing kernel.
  - Previously heuristical, recently replaced by hyperparameter optimisation based on test power, or Bayesian evidence [4, 24].
- Empirical kernel mean estimator might be suboptimal.
  - Better estimators exist [4, 16, 17].

Optimised MMD and model criticism



Arthur Gretton's twitter.

Optimised MMD and model criticism

MMD tends to lose test power with increasing dimensionality. [18]

Pick the kernel such that the test power is maximised, [24]

$$\mathrm{p}_{H_1}\left(\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2}{\sqrt{V_m}} > \frac{\hat{c}_\alpha/m - \mathrm{MMD}^2}{\sqrt{V_m}}\right)$$

$$\xrightarrow{m\to\infty} 1 - \Phi\left(\frac{c_\alpha}{m\sqrt{V_m}} - \frac{\mathrm{MMD}^2}{\sqrt{V_m}}\right)$$

where $\hat{c}_\alpha$ is an estimator of the theoretical rejection threshold c_α.

Optimised MMD and model criticism

MMD tends to lose test power with increasing dimensionality. [18]

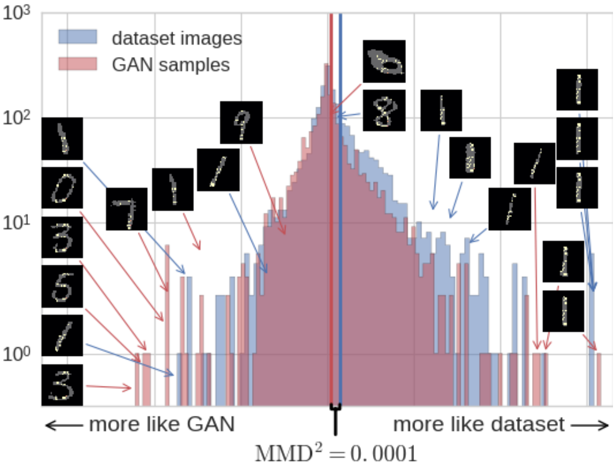
Pick the kernel such that the test power is maximised, [24]

$$\mathrm{p}_{H_1}\left(\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2}{\sqrt{V_m}} > \frac{\hat{c}_\alpha/m - \mathrm{MMD}^2}{\sqrt{V_m}}\right)$$

$$\xrightarrow{m\to\infty} 1 - \Phi\left(\frac{c_\alpha}{m\sqrt{V_m}} - \frac{\mathrm{MMD}^2}{\sqrt{V_m}}\right)$$

where $\hat{c}_\alpha$ is an estimator of the theoretical rejection threshold c_α.

Optimised MMD and model criticism



Witness function evaluated on a test set, ARD weights highlighted [24].

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Integral Probability metrics

Have two probability measure p, and q. Integral Probability Metric (IPM) defines a discrepancy measure,

$$d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) := \sup_{f \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \big(f(\boldsymbol{x})\big) - \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{y})} \big(f(\boldsymbol{y})\big) \right|$$

where the space of functions $\mathcal{H}$ must be rich enough such that $d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) = 0$ iff $\mathrm{p} = \mathrm{q}$.

$f^* := \mathrm{argmax}_{f \in \mathcal{H}} \left| \mathbb{E}_{\mathrm{p}}(f(\boldsymbol{x})) - \mathbb{E}_{\mathrm{q}}(f(\boldsymbol{y})) \right|$ is the *witness function*.

MMD is an IPM where $\mathcal{H}$ is the unit ball in characteristic RKHS.

Integral Probability metrics

Have two probability measure p, and q. Integral Probability Metric (IPM) defines a discrepancy measure,

$$d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) := \sup_{f \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left(f(\boldsymbol{x})\right) - \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{y})} \left(f(\boldsymbol{y})\right) \right|$$

where the space of functions $\mathcal{H}$ must be rich enough such that $d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) = 0$ iff $\mathrm{p} = \mathrm{q}$.

$f^* := \mathrm{argmax}_{f \in \mathcal{H}} \left| \mathbb{E}_{\mathrm{p}}(f(\boldsymbol{x})) - \mathbb{E}_{\mathrm{q}}(f(\boldsymbol{y})) \right|$ is the *witness function*.

MMD is an IPM where $\mathcal{H}$ is the unit ball in characteristic RKHS.

Integral Probability metrics

Have two probability measure p, and q. Integral Probability Metric (IPM) defines a discrepancy measure,

$$d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) := \sup_{f \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left(f(\boldsymbol{x})\right) - \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{y})} \left(f(\boldsymbol{y})\right) \right|$$

where the space of functions $\mathcal{H}$ must be rich enough such that $d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) = 0$ iff $\mathrm{p} = \mathrm{q}$.

$f^* := \mathrm{argmax}_{f \in \mathcal{H}} \left| \mathbb{E}_{\mathrm{p}}(f(\boldsymbol{x})) - \mathbb{E}_{\mathrm{q}}(f(\boldsymbol{y})) \right|$ is the *witness function*.

MMD is an IPM where $\mathcal{H}$ is the unit ball in characteristic RKHS.

Stein Discrepancy

Construct $\mathcal{H}$ such that $\mathbb{E}_{\mathrm{q}}(f(\boldsymbol{y})) = 0, \forall f \in \mathcal{H}$.

$$d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) = S_{\mathrm{q}}(\mathrm{p}, \mathcal{T}, \mathcal{H}) := \sup_{f \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[(\mathcal{T} f)(\boldsymbol{x}) \right] \right|$$

where $\mathcal{T}$ is a real–valued operator and $\mathbb{E}_{\mathrm{q}}[(\mathcal{T} f)(\boldsymbol{y})] = 0, \forall f \in \mathcal{H}$.

A standard choice of $\mathcal{T}$ when $\boldsymbol{x} = x \in \mathbb{R}$ is the *Stein's operator*,

$$(\mathcal{T} f)(x) = \mathcal{A}_{\mathrm{q}} f(x) := s_{\mathrm{q}}(x) f(x) + \nabla_x f(x)$$

Stein Discrepancy

Construct $\mathcal{H}$ such that $\mathbb{E}_{\mathrm{q}}(f(\boldsymbol{y})) = 0, \forall f \in \mathcal{H}$.

$$d_{\mathcal{H}}(\mathrm{p}, \mathrm{q}) = S_{\mathrm{q}}(\mathrm{p}, \mathcal{T}, \mathcal{H}) := \sup_{f \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[(\mathcal{T} f)(\boldsymbol{x}) \right] \right|$$

where $\mathcal{T}$ is a real–valued operator and $\mathbb{E}_{\mathrm{q}}[(\mathcal{T} f)(\boldsymbol{y})] = 0, \forall f \in \mathcal{H}$.

A standard choice of $\mathcal{T}$ when $\boldsymbol{x} = x \in \mathbb{R}$ is the *Stein's operator*,

$$(\mathcal{T} f)(x) = \mathcal{A}_{\mathrm{q}} f(x) := s_{\mathrm{q}}(x) f(x) + \nabla_x f(x)$$

Kernelised Stein Discrepancy (KSD)

Assume that $\mathrm{p}(\boldsymbol{x})$, and $\mathrm{q}(\boldsymbol{y})$ are differentiable continuous density functions; q might be unnormalised [3, 15].

Use the Stein's operator in the unit ball of the product RKHS $\mathcal{F}^{\mathrm{D}}$,

$$
\begin{aligned}
S_{\mathrm{q}}(\mathrm{p}, \mathcal{A}_{\mathrm{q}}, \mathcal{F}^{\mathrm{D}}) &:= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{q}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[\mathrm{Tr}\left(f(\boldsymbol{x}) s_{\mathrm{q}}(\boldsymbol{x})^{\mathrm{T}} + \nabla^2 f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \sum_{d=1}^{\mathrm{D}} \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[f_d(\boldsymbol{x}) s_{\mathrm{q},d}(\boldsymbol{x}) + \frac{\partial f_d(\boldsymbol{x})}{\partial \boldsymbol{x}_d} \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \langle f, \boldsymbol{\beta}_{\mathrm{q}} \rangle_{\mathcal{F}^{\mathrm{D}}} \right| = \left\| \boldsymbol{\beta}_{\mathrm{q}} \right\|_{\mathcal{F}^{\mathrm{D}}} \ (\text{s.t. } \boldsymbol{\beta}_{\mathrm{q}} \in \mathcal{F}^{\mathrm{D}})
\end{aligned}
$$

$\langle f, g \rangle_{\mathcal{F}^{\mathrm{D}}} := \sum_{d=1}^{\mathrm{D}} \langle f_d, g_d \rangle_{\mathcal{F}}$, $\boldsymbol{\beta}_{\mathrm{q}} := \mathbb{E}_{\mathrm{p}}[k(\boldsymbol{x}, \cdot) u_{\mathrm{q}}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)]$.

Kernelised Stein Discrepancy (KSD)

Assume that $\mathrm{p}(\boldsymbol{x})$, and $\mathrm{q}(\boldsymbol{y})$ are differentiable continuous density functions; q might be unnormalised [3, 15].

Use the Stein's operator in the unit ball of the product RKHS $\mathcal{F}^{\mathrm{D}}$,

$$
\begin{aligned}
S_{\mathrm{q}}(\mathrm{p}, \mathcal{A}_{\mathrm{q}}, \mathcal{F}^{\mathrm{D}}) &:= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{q}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[\mathrm{Tr}\left(f(\boldsymbol{x}) s_{\mathrm{q}}(\boldsymbol{x})^{\mathrm{T}} + \nabla^2 f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \sum_{d=1}^{\mathrm{D}} \mathop{\mathbb{E}}_{\mathrm{p}(\boldsymbol{x})} \left[f_d(\boldsymbol{x}) s_{\mathrm{q},d}(\boldsymbol{x}) + \frac{\partial f_d(\boldsymbol{x})}{\partial \boldsymbol{x}_d} \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \langle f, \boldsymbol{\beta}_{\mathrm{q}} \rangle_{\mathcal{F}^{\mathrm{D}}} \right| = \left\| \boldsymbol{\beta}_{\mathrm{q}} \right\|_{\mathcal{F}^{\mathrm{D}}} \text{ (s.t. } \boldsymbol{\beta}_{\mathrm{q}} \in \mathcal{F}^{\mathrm{D}})
\end{aligned}
$$

$\langle f, g \rangle_{\mathcal{F}^{\mathrm{D}}} := \sum_{d=1}^{\mathrm{D}} \langle f_d, g_d \rangle_{\mathcal{F}}$, $\boldsymbol{\beta}_{\mathrm{q}} := \mathbb{E}_{\mathrm{p}}\left[k(\boldsymbol{x}, \cdot) u_{\mathrm{q}}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot) \right]$.

KSD vs. alternative GoF tests

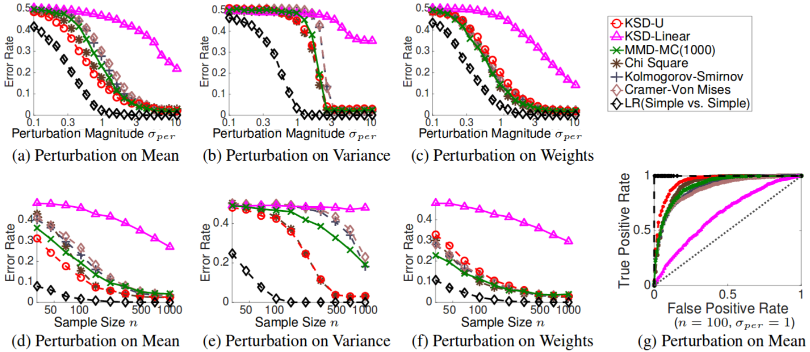


Figure 1. Results on 1D Gaussian mixture. (a)-(c) The error rates of different methods vs. the perturbation magnitude σ_{per} when perturbing the mean, variance and mixture weights, respectively; we use a fixed sample size of $n = 100$. (d)-(f) the error rates vs. the sample size n, with fixed perturbation magnitude $\sigma_{per} = 1$. We find that the type I errors of all the methods are well controlled under 0.05, and hence the reported error rates are essentially type II errors. (g) The ROC curve with mean perturbation, $n = 100$, $\sigma_{per} = 1$.

Variational Inference using KSD

Natural idea: minimise KSD between the true and an approximate posterior distribution $\rightarrow$ a particular case of *Operator Variational Inference*. [19]

Quick detour: For the true posterior $\mathrm{p}(\boldsymbol{x}) = \widetilde{\mathrm{p}}(\boldsymbol{x})/Z_{\mathrm{p}}$ and an approximation $\mathrm{q}(\boldsymbol{x}) = \widetilde{\mathrm{q}}(\boldsymbol{x})/Z_{\mathrm{q}}$,

$$
\begin{aligned}
S(\mathrm{p}, \mathrm{q}) &= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{p}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{p}} f(\boldsymbol{x}) - \mathcal{A}_{\mathrm{q}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[f(\boldsymbol{x})^{\mathrm{T}} (s_{\mathrm{p}}(\boldsymbol{x}) - s_{\mathrm{q}}(\boldsymbol{x})) \right] \right| \\
&= \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \mathop{\mathbb{E}}_{\mathrm{q}(\tilde{\boldsymbol{x}})} k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) (s_{\mathrm{p}}(\boldsymbol{x}) - s_{\mathrm{q}}(\boldsymbol{x}))^{\mathrm{T}} (s_{\mathrm{p}}(\tilde{\boldsymbol{x}}) - s_{\mathrm{q}}(\tilde{\boldsymbol{x}}))
\end{aligned}
$$

$\rightarrow \mathrm{q}(\boldsymbol{x}) =\propto \mathrm{q}(\boldsymbol{x}, \epsilon)$, e.g. $\mathrm{q}(\boldsymbol{x}, \epsilon) = \mathcal{N}(\boldsymbol{x} \mid \mathcal{T}_{\boldsymbol{\theta}}(\epsilon), \sigma^2 \boldsymbol{I})\, \mathcal{N}(\epsilon \mid \boldsymbol{0}, \boldsymbol{I})$

Variational Inference using KSD

Natural idea: minimise KSD between the true and an approximate posterior distribution $\rightarrow$ a particular case of *Operator Variational Inference*. [19]

Quick detour: For the true posterior $\mathrm{p}(\boldsymbol{x}) = \widetilde{\mathrm{p}}(\boldsymbol{x})/Z_{\mathrm{p}}$ and an approximation $\mathrm{q}(\boldsymbol{x}) = \widetilde{\mathrm{q}}(\boldsymbol{x})/Z_{\mathrm{q}}$,

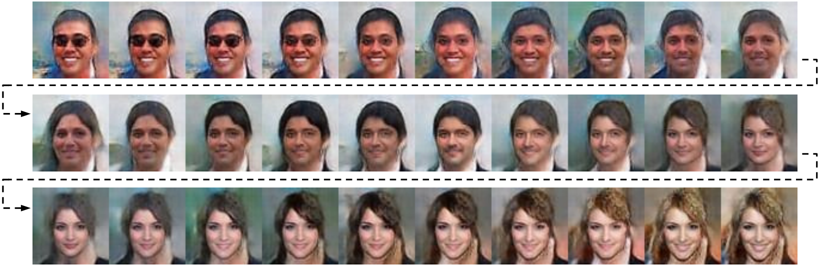
$$
\begin{aligned}
S(\mathrm{p}, \mathrm{q}) &= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{p}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[\mathrm{Tr}\left(\mathcal{A}_{\mathrm{p}} f(\boldsymbol{x}) - \mathcal{A}_{\mathrm{q}} f(\boldsymbol{x}) \right) \right] \right| \\
&= \sup_{f \in \mathcal{F}^{\mathrm{D}}, \|f\|_{\mathcal{F}^{\mathrm{D}}} \leq 1} \left| \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \left[f(\boldsymbol{x})^{\mathrm{T}} (s_{\mathrm{p}}(\boldsymbol{x}) - s_{\mathrm{q}}(\boldsymbol{x})) \right] \right| \\
&= \mathop{\mathbb{E}}_{\mathrm{q}(\boldsymbol{x})} \mathop{\mathbb{E}}_{\mathrm{q}(\tilde{\boldsymbol{x}})} k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) (s_{\mathrm{p}}(\boldsymbol{x}) - s_{\mathrm{q}}(\boldsymbol{x}))^{\mathrm{T}} (s_{\mathrm{p}}(\tilde{\boldsymbol{x}}) - s_{\mathrm{q}}(\tilde{\boldsymbol{x}}))
\end{aligned}
$$

$\rightarrow \mathrm{q}(\boldsymbol{x}) = \propto \mathrm{q}(\boldsymbol{x}, \boldsymbol{\epsilon})$, e.g. $\mathrm{q}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{x} \mid \mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}), \sigma^2 \boldsymbol{I})\, \mathcal{N}(\boldsymbol{\epsilon} \mid \mathbf{0}, \boldsymbol{I})$

Learning to Sample using KSD

Two papers [15, 25] on optimising a set of particles, resp. a set of samples from a model (amortisation), repeatedly using KL-optimal perturbations $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \varepsilon_t f(\boldsymbol{x})$.

Read Yingzhen's blog post [13] and a recent paper [14]!



Random walk through the latent space of a GAN trained with KSD adversary. [25]

Outline

RKHS

Kernel Mean Embeddings

Characteristic kernels

Two Sample Testing

- MMD
- Kernelised Stein Discrepancy

Kernel Bayes' Rule

Tensor product

If $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces then so is $\mathcal{H} \otimes \mathcal{K}$.

If $\{a_i\}_{i=1}^{\infty}$ spans $\mathcal{H}$ and $\{b_j\}_{j=1}^{\infty}$ spans $\mathcal{K}$ then $\{a_i \otimes b_j\}_{i,j=0}^{\infty}$ spans $\mathcal{H} \otimes \mathcal{K}$.

For any $f, f' \in \mathcal{H}$ and $g, g' \in \mathcal{K}$ we have

$$\langle f \otimes g, f' \otimes g' \rangle = \langle f, f' \rangle \langle g, g' \rangle$$

$\mathcal{H} \otimes \mathcal{K}$ is isomporphic to a space of bounded linear operators $\mathcal{K} \to \mathcal{H}$

$$(f \otimes g)g' = f \langle g, g' \rangle$$

Covariance operators

Covariance operators $C_{XY} : \mathcal{K} \to \mathcal{H}$ are defined as

$$C_{XY} = \mathbb{E}[k_X(X, \cdot) \otimes k_Y(Y, \cdot)]$$
$$C_{XX} = \mathbb{E}[k_X(X, \cdot) \otimes k_X(X, \cdot)]$$

For any $f \in \mathcal{H}$ and $g \in \mathcal{K}$

$$\langle f, C_{XY} g \rangle = \langle C_{YX} f, g \rangle = \mathbb{E}[f(X)g(Y)]$$

Empirical estimators

$$(X_i, Y_i) \sim_{i.i.d.} P(X, Y)$$
$$\hat{C}_{XY} = \frac{1}{N} \sum_{i=1}^{N} k_X(X_i, \cdot) \otimes k_Y(Y_i, \cdot)$$
$$\hat{C}_{XX} = \frac{1}{N} \sum_{i=1}^{N} k_X(X_i, \cdot) \otimes k_X(X_i, \cdot)$$

Conditional embeddings

Let

$$\mu_{X|Y=y} = \mathbb{E}[k_X(X,\cdot)|Y=y]$$

We seek an operator $\mathcal{U}_{X|Y}$ such that for all y

$$\mu_{X|Y=y} = \mathcal{U}_{X|Y} k_Y(y,\cdot)$$

which we call the conditional mean embedding.

Conditional embeddings

Lemma ([21])

For any $f \in \mathcal{H}$ *let* $h(y) = \mathbb{E}[f(X)|Y = y]$ *and assume that* $h \in \mathcal{K}$. *Then*

$$C_{YY}h = C_{YX}f$$

Proof.

Take any $g \in \mathcal{K}$.

$$\langle g, C_{YY}h\rangle = \mathop{\mathbb{E}}_{Y\sim P(Y)}[g(Y)h(Y)] =$$
$$\mathop{\mathbb{E}}_{Y\sim P(Y)}[g(Y)\mathop{\mathbb{E}}_{X\sim P(X|Y)}[f(X)|Y]] = \mathop{\mathbb{E}}_{Y\sim P(Y)}[\mathop{\mathbb{E}}_{X\sim P(X|Y)}[g(Y)f(X)|Y]]$$
$$\mathop{\mathbb{E}}_{X,Y\sim P(X,Y)}[g(Y)f(X)] = \langle g, C_{YX}f\rangle$$

$\square$

Conditional embeddings

Whenever C_{YY} is invertible we have

$$\mathcal{U}_{X|Y} = C_{XY} C_{YY}^{-1}$$

in practice we use a regularized version

$$\mathcal{U}_{X|Y}^{\epsilon} = C_{XY}(C_{YY} + \epsilon I)^{-1}$$

which can be estimated by

$$\hat{\mathcal{U}}_{X|Y}^{\epsilon} = \hat{C}_{XY}(\hat{C}_{YY} + \epsilon I)^{-1}$$

Kernel Bayes' rule

We could compute posterior embedding by

$$\mu_{X|Y=y} = \mathcal{U}_{X|Y} k_Y(y, \cdot)$$

but we may want a different prior.

Setting:

- let $P(X, Y)$ be the joint distribution of the model
- let $Q(X, Y)$ be a different distribution such that $P(Y|X = x) = Q(Y|X = x)$ for all x
- we have a sample $(X_i, Y_i) \sim Q$ and $\tilde{X}_j \sim P(X)$
- we want to estimate the posterior embedding $\mathbb{E}_{X \sim P(X|Y=y)}[k_X(X, \cdot)]$ for some particular y

Kernel sum rule

Sum rule

$$P(X) = \sum_{Y} P(X, Y)$$

Kernel sum rule

$$\mu_X = \mathcal{U}_{X|Y} \mu_Y$$

Corollary

$$C_{XX} = \mathcal{U}_{XX|Y} \mu_Y$$

Kernel product rule

Product rule

$$P(X, Y) = P(X|Y)P(Y)$$

Kernel product rule

$$C_{XY} = \mathcal{U}_{XY} C_{YY}$$

Kernel Bayes' rule

Observe that

$$
\begin{aligned}
C_{XY} &= \mathcal{U}_{Y|X} C_{XX} \\
C_{YY} &= \mathcal{U}_{YY|X} \mu_X
\end{aligned}
$$

and so

$$\mathcal{U}_{X|Y} = C_{XY} C_{YY}^{-1} = (\mathcal{U}_{Y|X} C_{XX})(\mathcal{U}_{YY|X} \mu_X)^{-1}$$

lets us express $\mathcal{U}_{X|Y}$ in terms of $\mathcal{U}_{Y|X}$.

Reminder

Setting:

- let $P(X, Y)$ be the joint distribution of the model
- let $Q(X, Y)$ be a different distribution such that $P(Y|X = x) = Q(Y|X = x)$ for all x
- we have a sample $(X_i, Y_i) \sim Q$ and $\tilde{X}_j \sim P(X)$
- we want to estimate the posterior embedding $\mathbb{E}_{X \sim P(X|Y=y)}[k_X(X, \cdot)]$ for some particular y

Kernel Bayes' rule

We have

$$\mathcal{U}^P_{Y|X} = \mathcal{U}^Q_{Y|X}$$
$$\mathcal{U}^P_{X|Y} \neq \mathcal{U}^Q_{X|Y}$$

but

$$\begin{aligned}\mathcal{U}^P_{X|Y} &= (\mathcal{U}^P_{Y|X} C^P_{XX})(\mathcal{U}^P_{YY|X} \mu^P_X)^{-1} \\ &= (\mathcal{U}^Q_{Y|X} C^P_{XX})(\mathcal{U}^Q_{YY|X} \mu^P_X)^{-1}\end{aligned}$$

so

$$\mu^P_{X|Y=y} = (\mathcal{U}^Q_{Y|X} C^P_{XX})(\mathcal{U}^Q_{YY|X} \mu^P_X)^{-1} k_Y(y, \cdot)$$

Kernel Bayes' rule

In practice we use the following estimator

$$\hat{\mu}_{X|Y=y} = (\hat{\mathcal{U}}^{Q}_{Y|X}\hat{C}^{P}_{XX})((\hat{\mathcal{U}}^{Q}_{YY|X}\hat{\mu}^{P}_{X})^2 + \epsilon I)^{-1}(\hat{\mathcal{U}}^{Q}_{YY|X}\hat{\mu}^{P}_{X})k_Y(y,\cdot)$$

which can be written as

$$\hat{\mu}_{X|Y=y} = A(B^2 + \delta I)^{-1}Bk_Y(y,\cdot)$$
$$A = \hat{\mathcal{U}}^{Q}_{Y|X}\hat{C}^{P}_{XX} = \hat{C}^{Q}_{YX}(\hat{C}^{Q}_{XX} + \epsilon I)^{-1}\hat{C}^{P}_{XX}$$
$$B = \hat{\mathcal{U}}^{Q}_{YY|X}\hat{\mu}^{P}_{X} = \hat{C}^{Q}_{YYX}(\hat{C}^{Q}_{XX} + \epsilon I)^{-1}\hat{\mu}^{P}_{X}$$

For $\epsilon = N^{-\frac{1}{3}}$ and $\delta = N^{-\frac{4}{27}}$ it can be shown [7] that

$$\left\|\mu_{X|Y=y} - \hat{\mu}_{X|Y=y}\right\| = O_p(N^{-\frac{4}{27}})$$

Kernel Bayes' rule

When is Kernel Bayes' Rule useful?

- when densities aren't tractable (ABC)
- when you don't know how to write a model but you know how to pick a kernel
- perhaps it can perform better than alternatives even if the above aren't satisfied

Other topics

Other topics combining KMEs with Bayesian inference

- adaptive Metropolis-Hastings using KME [20]
- Hamiltonian Monte Carlo without gradients [23]
- Bayesian estimation of KMEs [4]

References I

M. A. Arcones and E. Gine.
On the Bootstrap of U and V Statistics.
Ann. Statist., 20(2):655–674, 06 1992.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton.
Fast Two-Sample Testing with Analytic Representations of Probability Measures.
ArXiv e-prints, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton.
A Kernel Test of Goodness of Fit.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi.
Bayesian learning of kernel embeddings.
In *UAI*, 2016.

References II

K. Fukumizu, F. R. Bach, and M. I. Jordan.
Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces.
JMLR, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf.
Kernel Measures of Conditional Dependence.
In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 489–496, 2007.

K. Fukumizu, L. Song, and A. Gretton.
Kernel Bayes' rule: Bayesian inference with positive definite kernels.
Journal of Machine Learning Research, 14:3753–3783, 2013.

References III

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola.
A Kernel Method for the Two-sample-problem.
In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 513–520, 2006.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola.
A Kernel Two-sample Test.
JMLR, 13:723–773, Mar. 2012.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur.
A Fast, Consistent Kernel Two-sample Test.
In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 673–681, 2009.

References IV

W. Hoeffding.

Probability Inequalities for Sums of Bounded Random Variables.

Journal of the American Statistical Association, 58(301):13–30, 1963.

N. L. Johnson, S. Kotz, and N. Balakrishnan.

Continuous Univariate Distributions.

Wiley, 1994.

Y. Li.

Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm, August 2016.

Y. Li, R. E. Turner, and Q. Liu.

Approximate Inference with Amortised MCMC.

arXiv preprint arXiv:1702.08343, 2017.

References V

Q. Liu, J. D. Lee, and M. I. Jordan.

A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation.

arXiv preprint arXiv:1602.03253, 2016.

K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf.

Kernel mean shrinkage estimators.

Journal of Machine Learning Research, 17(48):1–41, 2016.

K. Muandet, B. Sriperumbudur, and B. Schölkopf.

Kernel mean estimation via spectral filtering.

In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

References VI

A. Ramdas, S. J. Reddi, B. Poczos, A. Singh, and L. Wasserman.
On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives.
arXiv preprint arXiv:1411.6314, 2014.

R. Ranganath, D. Tran, J. Altosaar, and D. Blei.
Operator Variational Inference.
In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton.
Kernel adaptive Metropolis-Hastings.
In *ICML*, 2014.

References VII

L. Song, J. Huang, A. Smola, and K. Fukumizu.
Hilbert space embeddings of conditional distributions with applications to dynamical systems.
In *ICML*, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet.
Hilbert Space Embeddings and Metrics on Probability Measures.
JMLR, 11:1517–1561, 2010.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton.
Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families.
In *NIPS*, 2015.

References VIII

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton.

Generative models and model criticism via optimized maximum mean discrepancy.

arXiv preprint arXiv:1611.04488, 2016.

D. Wang and Q. Liu.

Learning to Draw Samples: With Application to Amortized MLE for Generative Adversarial Learning.

arXiv preprint arXiv:1611.01722, 2016.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic.

Large-Scale Kernel Methods for Independence Testing.

ArXiv e-prints, 2016.