Bayesian Interpretations of RKHS Embedding Methods

David Duvenaud
Department of Engineering
University of Cambridge
dkd23@cam.ac.uk

Abstract

In recent years, RKHS embeddings have been used to propose new two-sample tests, independence tests, approximate integration, inference and message-passing algorithms. Starting from a correspondence between the Maximum Mean Discrepancy and the posterior variance of the integral of a Gaussian process, we derive corresponding Bayesian interpretations of these methods. Using these interpretations, we then shed light on the original frequentist procedures.

1 Introduction

2 Reproducing Kernel Hilbert Space Embeddings

Well-known reproducing property [Saitoh, 1988]: If $\mathcal{F}$ is an RKHS on $\mathbb{R}^D$, with $k(x, x')$ the associated kernel, and $\psi(x)$ be the feature map of k, where

$$\psi(x) = k(\cdot, x) \tag{1}$$

then

$$f(x) = \langle f, \psi(x) \rangle \quad \forall f \in \mathcal{F}, \forall x \in \mathbb{R}^D \tag{2}$$

3 Maximum Mean Discrepancy

For selecting pseudosamples, herding relies on an objective based on the maximum mean discrepancy (Sriperumbudur et al., 2010). MMD measures the divergence between two distributions, p and q with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$\mathrm{MMD}_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right| \tag{3}$$

Intuitively, if two distributions are close in the MMD sense, then no matter which function f we choose from $\mathcal{F}$, the difference in its integral over p or q should be small. A particularly interesting case is when the function class $\mathcal{F}$ is functions of unit norm from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. In this case, the MMD between two distributions can be conveniently expressed using

expectations of the associated kernel $k(x, x')$ only (Sriperumbudur et al., 2010):

$$MMD^2_{\mathcal{H}}(p, q) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}} \left| \int f(x)p(x)dx - \int f(x)q(x)dx \right|^2 \tag{4}$$

$$= \|\mu_p - \mu_q\|^2_{\mathcal{H}} \tag{5}$$

$$= \iint k(x, y)p(x)p(y)dxdy - 2 \iint k(x, y)p(x)q(y)dxdy + \iint k(x, y)q(x)q(y)dxdy, \tag{6}$$

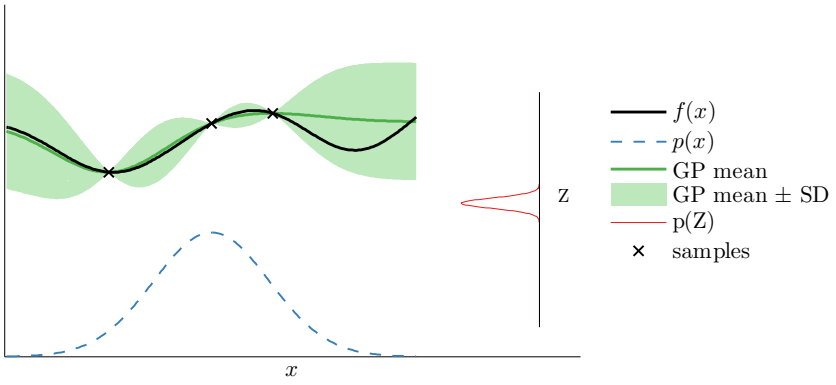
where in the above formula

$$\mu_p = \int \phi(x)p(x)dx \in \mathcal{H} \tag{7}$$

denotes the *mean element* associated with the distribution p. For characteristic kernels, such as the Gaussian kernel, the mapping between a distribution and its mean element is bijective. As a consequence, $\mathrm{MMD}_{\mathcal{H}}(p, q) = 0$ if and only if $p = q$, making it a powerful measure of divergence.

4 Integrals of functions drawn from Gaussian process

[In this section, show that you can integrate under a GP prior in closed form]



5 The Link between MMD and GP Integrals

[This section will show that the posterior variance of the difference between the integral of a GP with respect to two different distributions is equal to MMD. This generalizes a result from (Huszár & Duvenaud, 2012)]

Proposition 1. *Given that f is drawn from a standard GP prior, the expected squared difference in the integral between $f(x)$ against $p(x)$ minus the integral of $f(x)$ against $q(x)$ is equal to the squared Maximum Mean Discrepancy between p and q.*

Figure 1: A graphical intepretation of the kernel two-sample test: The test statistic measures expected difference in integrals between the two empirical distributions.

Proof. The proof involves invoking the representer theorem, using bilinearity of scalar products and the fact that if f is a standard Gaussian process then $\forall g \in \mathcal{H} : \langle f, g \rangle \sim \mathcal{N}(0, \|g\|_{\mathcal{H}}^2)$:

$$\mathbb{V}_{f\sim GP}\left[\int f(x)p(x)dx - \int f(x)q(x)dx\right] \tag{8}$$

$$= \mathbb{E}_{f\sim GP}\left(\int f(x)p(x)dx - \int f(x)q(x)dx\right)^2 \tag{9}$$

$$= \mathbb{E}_{f\sim GP}\left(\int \langle f, \phi(x)\rangle\, p(x)dx - \int \langle f, \phi(x)\rangle\, q(x)dx\right)^2 \tag{10}$$

$$= \mathbb{E}_{f\sim GP}\left\langle f, \int \phi(x)p(x)dx - \int \phi(x)q(x)dx\right\rangle^2 \tag{11}$$

$$= \mathbb{E}_{f\sim GP}\left\langle f, \mu_p - \mu_q\right\rangle^2 \tag{12}$$

$$= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \tag{13}$$

$$= \mathrm{MMD}^2(p, q) \tag{14}$$

□

5.1 Other links

Energy distances and parzen window estimators. [Smola's talk]

6 Kernel Herding and Bayesian Quadrature

In this section we show that Bayesian Quadrature Kernel herding (Chen et al., 2010)

Proposition 2. *The expected variance in the Bayesian quadrature estimate ϵ_{BQ}^2 is the maximum mean discrepancy between the target distribution $p(x)$ and $q_{\mathrm{BQ}}(x) = \sum_{n=1}^{N} w_{\mathrm{BQ}}^{(n)} \delta_{x_n}(x)$*

Proof. Because the true integral is $\int f(x)p(x)dx$ and the BQ estimator is given by $\sum_i f(x_i) w_{BQ}^{(i)}$, Proposition 1 implies that thier expected difference is given by $\mathrm{MMD}^2(p, q_{BQ})$. □

6.1 Empirical estimator

[Re-derive the empirical estimator]

7 Kernel Two-sample Tests

The kernel two-sample test was introducted by Gretton et al. (2005, 2008). It was also proven [cite] that if $X \sim p, Y \sim q$, then $\mathrm{MMD}(X, Y)$ is an unbiased estimator of $\mathrm{MMD}(p, q)$. Furthermore, the empirical MMD converges to the true MMD uniformly for any p, q.

related to Kernel ICA Bach & Jordan (2003), which normalizes to lessen the effect of the marginals.

7.1 Bayesian Two-sample tests

A Bayesian two-sample test was proposed by Borgwardt & Ghahramani (2009). The nonparametric version of their test places Dirichlet process mixture of Gaussian priors on $p(x)$ and $q(x)$, and computes the Bayes factor comparing the hypotheses $\mathcal{H}_0 : p(x) = q(x)$ and $\mathcal{H}_1 : p(x) \neq q(x)$.

This approach is appealing because it directly compares the likelihood of the two cases that two-sample tests are designed to compare. However, the DP-MoG likelihood ratio cannot be computed in closed form.

In comparison, the test statistic given by $\mathrm{MMD}(p, q)$ is an indirect measure by which to compare the two hypotheses of interest, but it does have a closed form solution.

8 Hilbert-Schmidt Independence Criterion and GP Integrals

New test statistic based on infinite-dimensional Frobenius norm of cross-covariance matrix of features of x and y: [Gretton et. al, 2005]

$$\mathrm{HSIC}(p(x,y), k_x, k_y) = ||C_{xy}||^2_{HS} \tag{15}$$

$$= \mathbb{E}_{x,x',y,y'}\left[k_x(x,x')k_y(y,y)\right] + \mathbb{E}_{x,x'}\left[k_x(x,x')\right]\mathbb{E}_{y,y'}\left[k_y(y,y')\right] \tag{16}$$

$$- 2\mathbb{E}_{x,y}\left[\mathbb{E}_{x'}\left[k_x(x,x')\right]\mathbb{E}_{y'}\left[k_y(y,y')\right]\right]$$

Proposition 3. *The HSIC is equivalent to the variance in the squared difference of integrals of functions drawn from a GP prior against the joint distribution, and the product of the marginal distributions.*

Proof. We use the identities that for a zero-mean GP,

$$\mathbb{E}_{f\sim\mathrm{GP}}\left[f(x,y)\right] = k(x,y) \tag{17}$$

$$\mathbb{E}_{f\sim\mathrm{GP}}\left[f(x,x',y,y')\right] = k(x,x',y,y') \tag{18}$$

$$\mathbb{V}_{f\sim\mathrm{GP}}\left[\iint f(x,y)p_{xy}(x,y)dxdy - \iint f(x',y')p_x(x')p_y(y')dxdy\right] \tag{19}$$

$$= \mathbb{E}_{f\sim\mathrm{GP}}\left[\left(\iint f(x,y)p_{xy}(x,y)dxdy - \iint f(x',y')p_x(x')p_y(y')dxdy\right)^2\right] \tag{20}$$

$$= \mathbb{E}_{f\sim\mathrm{GP}}\left[\iint f(x,y)p_{xy}(x,y)dxdy \iint f(x',y')p_{xy}(x',y')dx'dy'\right] \tag{21}$$

$$- 2\mathbb{E}_{f\sim\mathrm{GP}}\left[\iint f(x,y)p_{xy}(x,y)dxdy \iint f(x',y')p_x(x')p_y(y')dx'dy'\right]$$

$$+ \mathbb{E}_{f\sim\mathrm{GP}}\left[\iint f(x,y)p_x(x')p_y(y')dxdy \iint f(x',y')p_x(x')p_y(y')dx'dy'\right]$$

$$= \iiiint k(x,x',y,y')p_{xy}(x,y)p_{xy}(x',y')dxdydx'dy' \tag{22}$$

$$+ \iiiint k(x,x',y,y')p_x(x)p_x(x')dxdx'p_y(y)p_y(y')dydy'$$

$$- 2\iiiint k(x,x',y,y')p_x(x')dx'p_y(y')dyp_x(x)p_y(y)dxdy \tag{23}$$

Figure 2: A graphical intepretation of the HSIC: The expected difference in integrals under the joint empirical distribution, and the product of marginal empirical distributions.

Then, using the assumption that $k(x, x', y, y') = k_x(x, x')k_y(y, y')$, we can continue:

$$= \iiiint k_x(x, x')k_y(y, y')p_{xy}(x, y)p_{xy}(x', y')dxdydx'dy' \tag{24}$$
$$+ \iint k_x(x, x')p_x(x)p_x(x')dxdx' \iint k_y(y, y')p_y(y)p_y(y')dydy'$$
$$- 2 \iint \left[\int k_x(x, x')p_x(x')dx'\right] \left[\int k_y(y, y')p_y(y')dy\right] p_x(x)p_y(y)dxdy$$
$$= \mathbb{E}_{x,x',y,y'} \left[k_x(x, x')k_y(y, y)\right] + \mathbb{E}_{x,x'} \left[k_x(x, x')\right] \mathbb{E}_{y,y'} \left[k_y(y, y')\right] \tag{25}$$
$$- 2\mathbb{E}_{x,y} \left[\mathbb{E}_{x'} \left[k_x(x, x')\right] \mathbb{E}_{y'} \left[k_y(y, y')\right]\right]$$
$$= \text{HSIC}(p(x, y), k_x, k_y) \tag{26}$$

□

8.1 Generalization of HSIC

If we can come up with a more sensible kernel $k(x, x', y, y')$ than the factored kernel $k(x, x', y, y') = k_x(x, x')k_y(y, y')$, we obtain a more general solution:

$$GHSIC = \iiiint k(x, x', y, y')p_{xy}(x, y)p_{xy}(x', y')dxdydx'dy' \tag{27}$$
$$+ \iiiint k(x, x', y, y')p_x(x)p_x(x')dxdx'k_y(y, y')p_y(y)p_y(y')dydy'$$
$$- 2 \iiiint k(x, x', y, y')p_x(x')dx'p_y(y')dyp_x(x)p_y(y)dxdy$$
$$= \mathbb{E}_{x,x',y,y'} \left[k(x, x', y, y')p_{xy}(x, y)p_{xy}(x', y')\right] \tag{28}$$
$$+ \mathbb{E}_{x,x',y,y'} \left[k(x, x', y, y')p_x(x)p_x(x')p_y(y)p_y(y')\right]$$
$$- 2\mathbb{E}_{x,x',y,y'} \left[k\right](x, x', y, y')p_x(x')p_y(y')p_x(x)p_y(y)$$

8.2 Empirical estimator

The empricial estimate of HSIC is given by:

$$\left|\left|\hat{\Sigma}_{YX}^{(N)}\right|\right|_{HS}^2 = \frac{1}{N^2} \tag{29}$$

9 Mean Embeddings and Conditional Mean Embeddings

Many results in the MMD literature discuss the *mean embedding* of a distribution in a RKHS. What is the corresponding Bayesian intepretation? Based on a result due to Muandet & Ghahramani (2012), we can say that if $f \sim \text{GP}$,

$$\mu_{p(x)} = \int \phi(x)p(x)dx \tag{30}$$
$$= \int k(x, \cdot)p(x)dx \tag{31}$$
$$= \mathbb{E}_{f\sim\text{GP}} \left[\int f(x)f(\cdot)p(x)dx\right] \tag{32}$$
$$= \mathbb{E}_{f\sim\text{GP}} \left[f(\cdot) \int f(x)p(x)dx\right] \tag{33}$$

Thus we can interpret the mean embedding of $p(x)$ as the expected covariance of each point of the function with its integral with respect to $p(x)$.

10 Kernel Conditional Independence Tests

Kernel Conditional Independence Tests Fukumizu et al. (2008)

Cross-covariance operator is defined as:Proposition 1

$$\langle f, \Sigma_{YX} g \rangle = \text{cov}[f(X), g(Y)] = \mathbb{E}f(X)g(Y) - \mathbb{E}f(X)\mathbb{E}g(Y) \tag{34}$$

Normalized conditional cross-covariance operator:

$$V_{YX|Z} = V_{YZ}V_{ZX} \tag{35}$$

conditional covariance operator:

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \tag{36}$$

Attempted interpretation of the unnormalized version of their test:

$$I^{CONDU} = \mathbb{V}_{f\sim\text{GP}}\left[\iint f(x,y,z,z')p_{xy}(x,y|z)dxdy - \iint f(x',y')p_x(x'|z)p_y(y'|z)dxdy\right] \tag{37}$$

$$= \mathbb{E}_{z\sim p(z)}\left[\text{MMD}(p_{xy}(x,y|z)dxdy, p_x(x'|z)p_y(y'|z)\right] \tag{38}$$

$$\tag{39}$$

We should probably think of conditional independence given z as a function of z, but we can summarize this function by its expectation.

The naive test is identically zero under the empirical distribution:

$$BCIT_1 = \mathbb{E}_{z\sim p(z)}\left[\mathbb{V}_{f\sim\text{GP}}\left[\iint f(x,y)p_{xy}(x,y|z)dxdy - \iint f(x',y')p_x(x'|z)p_y(y'|z)dxdy\right]\right] \tag{40}$$

$$= \mathbb{E}_{z\sim p(z)}\left[\text{MMD}(p_{xy}(x,y|z)dxdy, p_x(x'|z)p_y(y'|z)\right] \tag{41}$$

$$= \frac{1}{N}\sum_i \iiiint k(x,x',y,y',z,z)p(x,y|z)p(x',y'|z)dxdydx'dy' \tag{42}$$

$$+ \iiiint k(x,x',y,y',z,z)p(x|z)p(x'|z)dxdx'p(y|z)p(y'|z)dydy'$$

$$- 2\iiiint k(x,x',y,y',z,z)p(x'|z)dx'p(y'|z)dyp(x|z)p(y|z)dxdy$$

$$= 0 \tag{43}$$

using the empirical distribution for $p(x,y|z)$, which will be multiplied by zero whenever $p(z) = 0$.

10.1 Bayesian Interpretation of HS norm

The HS-Norm of operator A is given by $||A||_{HS}^2 = \sum_i, j \langle \psi_j, A\phi_i \rangle_{\mathcal{H}_2}^2$.

11 Related Work

The MMD has been related to energy distances Sejdinovic et al. (2012a,b), proving that energy distances [define] are equivalent to MMD. Thus, we can claim as a corollary of Proposition 1 that energy distances are also equivalent to the variance of difference of integrals.

12 Open Questions

Can the Bayesian interpretation be extended to covariate shift?

What is the intepretation of normalization in the kernel conditional dependence work?

Instead of low-rank approximation, can we use other sparse GP methods?

13 Applications

13.1 Semi-supervised classification or ICA

We can learn the kernel such that the empirical distribution of labeled examples from the same class are forced to be together, and those from different classes are forced to be apart. Can we then interpret f, the function drawn from the GP prior? Would f then be the "class membership" function?

13.2 The witness function

Does the witness function have a Bayesian interpretation? It equals $w(x') = \int (p(x) - q(x))k(x, x')dx$, which is the difference between p and q convolved with the kernel function.

$$\mathrm{MMD}^2(p, q) \tag{44}$$

$$= \mathbb{V}_{f\sim GP}\left[\int f(x)p(x)dx - \int f(x)q(x)dx\right] \tag{45}$$

$$= \mathbb{V}_{f\sim GP}\left[\int f(x)\left[p(x) - q(x)\right]dx\right] \tag{46}$$

$$= \mathbb{E}_{f\sim GP}\left[\left(\int f(x)\left[p(x) - q(x)\right]dx\right)^2\right] \tag{47}$$

$$= \mathbb{E}_{f\sim GP}\left[\left(\int f(x)\left[p(x) - q(x)\right]dx\right)\left(\int f(x')\left[p(x') - q(x')\right]dx'\right)\right] \tag{48}$$

$$= \mathbb{E}_{f\sim GP}\left[\int\int f(x)f(x')\left[p(x) - q(x)\right]\left[p(x') - q(x')\right]dxdx'\right] \tag{49}$$

$$= \int\int k(x, x')\left[p(x) - q(x)\right]\left[p(x') - q(x')\right]dxdx' \tag{50}$$

$$= \int w(x)\left[p(x') - q(x')\right]dx' \tag{51}$$

So, the expectation of the witness function is the MMD.

14 Summary

Table 1: Almost-equivalent Methods

Frequentist Method		Bayesian Method	
$\mathcal{O}(N^2)$	Kernel herding	$\mathcal{O}(N^3)$	Bayesian Quadrature
$\mathcal{O}(N^2)$	Hilbert-Schmidt Independence Criterion	$\mathcal{O}(N^3)$	Variance of difference of integrals
$\mathcal{O}(N^2)$	Kernel Two-sample test	$\mathcal{O}(N^3)$	Variance of difference of integrals

Table 2: Related Methods

Frequentist Method	Bayesian Method
Kernel Regression (Nadayara-Watson)	GP Regression
Functional ANOVA, HKL	Additive Gaussian Processes
Kernel Bayes' Rule	Bayesian Quadrature for Ratios
Conditional Mean Embeddings	Gaussian Process Dynamic Programming
Kernel Message Passing	???

Acknowledgments

We would like to thank Miguel Hernandez-Labato, Roger Grosse, Philipp Hennig, and Michael Osborne for helpful suggestions and advice.

References

Bach, F.R. and Jordan, M.I. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.

Borgwardt, K.M. and Ghahramani, Z. Bayesian two-sample tests. *arXiv preprint arXiv:0906.4032*, 2009.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. UAI, 2010.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. 2008.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. Kernel methods for measuring independence. *The Journal of Machine Learning Research*, 6:2075–2129, 2005.

Gretton, A., Borgwardt, K., Rasch, M.J., Scholkopf, B., and Smola, A.J. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*, 2008.

Huszár, Ferenc and Duvenaud, David. Optimally-weighted herding is Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, 2012.

Muandet, Krikamol and Ghahramani, Zoubin. personal communication, 2012.

Sejdinovic, D., Gretton, A., Sriperumbudur, B., and Fukumizu, K. Hypothesis testing using pairwise distances and associated kernels. *arXiv preprint arXiv:1205.0411*, 2012a.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *arXiv preprint arXiv:1207.6076*, 2012b.

Sriperumbudur, Bharath K., Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert R.G. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 99:1517–1561, August 2010. ISSN 1532-4435.