Bayesian Learning of Kernel Embeddings

Seth Flaxman
Department of Statistics

joint work with Dino Sejdinovic, John P. Cunningham, and Sarah Filippi

UAI 2016
Overview

New probabilistic model for learning **kernel mean embeddings**:

- **Bayesian Kernel Embedding** combines a Gaussian process prior over RKHS with conjugate likelihood
- Yields closed form Bayesian posterior
- Hyperparameter learning through sampling or by maximizing a closed form marginal pseudolikelihood
- Yields a Bayesian viewpoint on estimation of kernel mean embeddings and covariance operators for unsupervised settings such as **Maximum Mean Discrepancy** (MMD) and **Hilbert-Schmidt Independence Criterion** (HSIC)
Kernel embeddings

\( \mathcal{X} = \mathbb{R}^D \) Kernel \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) and corresponding RKHS \( \mathcal{H}_k \).

Feature space representation: \( \phi(x) = k(\cdot, x) \).

\[ h : \mathcal{X} \rightarrow \mathbb{R} \text{ where } h(x) = \langle h, k(\cdot, x) \rangle_{\mathcal{H}_k}, \quad \forall x \in \mathcal{X}, \forall h \in \mathcal{H}_k \]

For probability measure \( P \) on \( \mathcal{X} \), define kernel embedding in \( \mathcal{H}_k \):

\[ \mu_P = \int k(\cdot, x) P(dx). \]

\( \mu_P \in \mathcal{H}_k \) uniquely represents \( P \) for characteristic kernels (captures all moments), and gives expectations of RKHS functions:

\[ \int h(x) P(dx) = \langle h, \mu_P \rangle_{\mathcal{H}_k} \]
Estimating kernel mean embeddings

Given iid samples $x_1, \ldots, x_n$, empirical estimator:

$$\hat{\mu}_P = \hat{\mu}_{\hat{P}} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i),$$

Spectral kernel mean shrinkage estimator (S-KMSE) of $\mu$:

$$\hat{\mu}_\lambda = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1} \hat{\mu}_P,$$

where $\hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i) \otimes k(\cdot, x_i)$ is the empirical covariance operator on $\mathcal{H}_k$, and $\lambda$ is a regularization parameter.
Statistical testing with kernel embeddings

Figure: Given a kernel $k$ and probability measures $P$ and $Q$, the maximum mean discrepancy (MMD) between $P$ and $Q$ ($\mu$) is defined as the RKHS distance $\|\mu_P - \mu_Q\|_{\mathcal{H}_k}$ between their embeddings. [Figure credit: Heiko Strathmann.]
Uses of kernel embeddings

For an overview, see Muandet et al. survey [2016]

- Statistical testing: two sample testing, (conditional) independence testing
- Learning with kernels: kernel Bayes’ rule, kernel EP, kernel ABC, etc.
- Kernel PCA and kernel CCA
- Distribution regression
- Many causal inference approaches, e.g. Zhang et al. [UAI 2012], Lopez-Paz et al. [ICML 2015], Flaxman et al. [ACM TIST 2015]

Note: randomized explicit feature expansions (e.g. random Fourier features) mean these methods are scalable and do not require the kernel trick.
How to set hyperparameters?

\[ k(x, x') = e^{-\frac{||x-x'||^2}{2\ell^2}} \]

- Supervised settings
- Classical approaches
- Gaussian processes
- Unsupervised settings: “median heuristic”:

  \[
  \text{lengthscale } \ell = \text{median}(||x_i - x_j||_2)
  \]
Problem statement

Given a parametric family of kernels \( \{k_\theta(\cdot, \cdot)\}_{\theta \in \Theta} \), a dataset \( \{x_i\}_{i=1}^n \sim P \) of observations in \( \mathbb{R}^D \) for an unknown \( P \), we wish to:

▶ Infer the kernel embedding \( \mu_{P,\theta} = \int k_\theta(\cdot, x) P(dx) \) for a given kernel \( k_\theta \), given observations.

▶ Infer the kernel hyperparameters \( \theta \), given observations.

\( \theta \) determines \( k_\theta \) which determines \( \mathcal{H}_k \) so at a high level, we are trying to learn a good feature representation.

For Bayesian posterior learning, need both a prior over \( \mu_{P,\theta} \) and a likelihood.
Prior: an approach that does not work!

Let $h \sim \mathcal{GP}(0, k_\theta(\cdot, \cdot))$.

Then $P(h \in \mathcal{H}_k) = 0$ [Parzen 1963, Wahba 1990, Lukić & Beder 2001].

Why? Because $\|h\|_{\mathcal{H}_k}$ is not finite. Proof in Appendix.

Intuition: $f \in \mathcal{H}_k$ is smoother then $h$.

Prior: an approach that does work

We define a GP prior over $\mu_\theta$ as follows:

$$
\mu_\theta \mid \theta \sim \mathcal{GP}(0, r_\theta(\cdot, \cdot)) ,
$$

$$
r_\theta(x, y) := \int k_\theta(x, u)k_\theta(u, y)\nu(du) .
$$

where $\nu$ is any finite measure on $\mathcal{X}$.

This choice of $r_\theta$ ensures that $\mu_\theta \in \mathcal{H}_{k_\theta}$ with probability 1 by the nuclear dominance of $k_\theta$ over $r_\theta$.

$r_\theta$ is the convolution of a kernel with itself with respect to $\nu$, so $r_\theta$ can be thought of as a smoother version of $k_\theta$. 
Likelihood

Likelihood links $\mu_\theta$ to the observations $\{x_i\}_{i=1}^n$.

Use the empirical mean embedding estimator: $\hat{\mu}_\theta = \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i)$ which depends on $\{x_i\}_{i=1}^n$ and $\theta$.

Evaluate $\hat{\mu}_\theta$ at some $x \in \mathbb{R}^D$.

**Result:** real number giving an empirical estimate of $\mu_\theta(x)$ based on $\{x_i\}_{i=1}^n$ and $\theta$. 
Likelihood continued

Our likelihood links the empirical estimate, $\hat{\mu}_\theta(x)$, to the corresponding modeled estimate, $\mu_\theta(x)$ using a Gaussian distribution with variance $\tau^2/n$:

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

CLT motivation: for fixed $x$, $\hat{\mu}_\theta(x) = \frac{1}{n} \sum_{i=1}^{n} k_\theta(x_i, x)$ is an average of iid random variables so it satisfies:

$$\sqrt{n}(\hat{\mu}_\theta(x) - \mu_\theta(x)) \xrightarrow{D} \mathcal{N}(0, \text{Var}_{X \sim \mathcal{P}}[k_\theta(X, x)]).$$
Posterior inference

Standard GP results (?) yield the posterior distribution:

\[
\begin{align*}
[\mu_\theta(x_1), \ldots, \mu_\theta(x_n)]^\top & | [\widehat{\mu}_\theta(x_1), \ldots, \widehat{\mu}_\theta(x_n)]^\top, \theta \\
& \sim \mathcal{N}(R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}[\widehat{\mu}_\theta(x_1), \ldots, \widehat{\mu}_\theta(x_n)]^\top, \\
& \hspace{1cm} R_\theta - R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}R_\theta),
\end{align*}
\]

where \(R_\theta\) is the matrix such that its \((i,j)\)-th element is \(r_\theta(x_i, x_j)\).

For squared exponential kernel \(k_\theta\), easy to derive \(r_\theta\) in closed form.
Illustration

(A) Draws from the prior

\( \mu(x) \)

\( x \)
(A) Draws from the prior

\[ \mu(x) \]

\[ x \]

\[ -3, -2, -1, 0, 1, 2, 3 \]

\[ -2, -1, 0, 1, 2 \]
(A) Draws from the prior

\[ \mu(x) \]

\[ x \]
Illustration

Histogram of $x$

![Histogram of $x$](image-url)
(B) Empirical mean

Illustration
Bayesian Kernel Learning

- We infer hyperparameters using marginal pseudolikelihood
- We evaluate empirical embedding at a set of points \( z_1, \ldots, z_m \) in \( X \subset \mathbb{R}^D \), with \( m \geq D \).
- Consider change of variables from mapping \( \phi_z : \mathbb{R}^D \mapsto \mathbb{R}^m \), given by
  \[
  \phi_z(x) := [k_\theta(x, z_1), \ldots, k_\theta(x, z_m)] \in \mathbb{R}^m,
  \]
- By Cramér’s decomposition theorem our model is equivalent to:
  \[
  \phi_z(X_i) | \mu_\theta \sim \mathcal{N} (\mu_\theta(z), \tau^2 I_m) .
  \]
- Applying the change of variable \( x \mapsto \phi_z(x) \) we obtain:
  \[
  p(x | \mu_\theta, \theta) = p \left( \phi_z(x) | \mu_\theta(z) \right) \text{vol} [J_\theta(x)] ,
  \]
  where \( J_\theta(x) = \left[ \frac{\partial k_\theta(x,z_i)}{\partial x(j)} \right]_{ij} \) is an \( m \times D \) matrix.
Experiments
Experiments

(A) Normal vs. Laplace

(B) Witness function (n = 50)

(C) Witness function (n = 400)
Conclusion

- Lots of open questions:
  - Refining the model: more realistic likelihood
  - How well does it work in high-dimensions?
  - Scalable learning approaches
  - Can you choose between different kernel classes?
  - Does it help with KPCA, clustering, other unsupervised settings?
  - Fully Bayesian measures of (in)dependence, distance between distributions


- Come see poster for more details

Thanks!
Contact: flaxman@stats.ox.ac.uk
www.sethrf.com