Bayesian Learning of Kernel Embeddings

Seth Flaxman
Department of Statistics
UNIVERSITY OF OXFORD

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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and corresponding RKHS $\mathcal{H}_k$.
Feature space representation: $\phi(x) = k(\cdot, x)$.

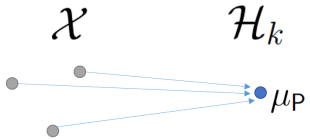
$$h : \mathcal{X} \to \mathbb{R} \text{ where } h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k}, \quad \forall x \in \mathcal{X}, \forall h \in \mathcal{H}_k$$

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

$$\mu_{\mathrm{P}} = \int k(\cdot, x)\, \mathrm{P}(dx).$$

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

$$\int h(x)\mathrm{P}(dx) = \langle h, \mu_{\mathrm{P}}\rangle_{\mathcal{H}_k}$$



Estimating kernel mean embeddings

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

$$\widehat{\mu_{\mathrm{P}}} = \mu_{\widehat{\mathrm{P}}} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i),$$

Spectral kernel mean shrinkage estimator (S-KMSE) of **?**:

$$\check{\mu}_\lambda = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1}\widehat{\mu_{\mathrm{P}}},$$

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 λ is a regularization parameter.

Statistical testing with kernel embeddings

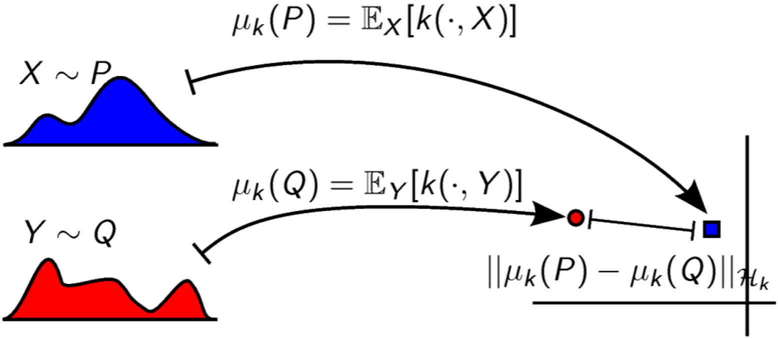


Figure: Given a kernel k and probability measures P and Q, the maximum mean discrepancy (MMD) between P and Q (**?**) is defined as the RKHS distance $\|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k}$ between their embeddings. [Figure credit: Heiko Strathmann.]

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 $\{x_i\}_{i=1}^n \sim \mathrm{P}$ of observations in $\mathbb{R}^D$ for an unknown P, we wish to:

- Infer the kernel embedding $\mu_{\mathrm{P},\theta} = \int k_\theta(\cdot, x)\,\mathrm{P}(dx)$ for a given kernel k_θ, given observations.
- Infer the kernel hyperparameters θ, given observations.

θ determines k_θ 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 $\mu_{\mathrm{P},\theta}$ 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.

Nuclear dominance [Fortet 1974, Lukić & Beder 2001, Pillai et al 2007] makes this precise.

Prior: an approach that does work

We define a GP prior over μ_θ 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 ν is any finite measure on $\mathcal{X}$.

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

r_θ is the convolution of a kernel with itself with respect to ν, so r_θ can be thought of as a smoother version of k_θ.

Likelihood

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

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

Evaluate $\widehat{\mu_\theta}$ 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 θ.

Likelihood continued

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

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

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

$$\sqrt{n}(\widehat{\mu_\theta}(x) - \mu_\theta(x)) \xrightarrow{D} \mathcal{N}(0, \mathrm{Var}_{X\sim \mathrm{P}}[k_\theta(X, x)]).$$

Posterior inference

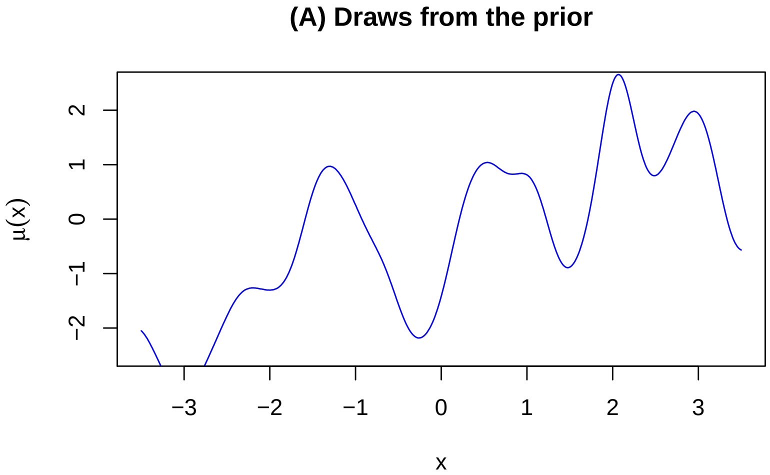
Standard GP results (**?**) yield the posterior distribution:

$$
\begin{aligned}
[\mu_\theta(x_1), \ldots, \mu_\theta(x_n)]^\top \mid [\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \theta \\
\sim \mathcal{N}(R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}[\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \\
R_\theta - R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}R_\theta),
\end{aligned}
$$

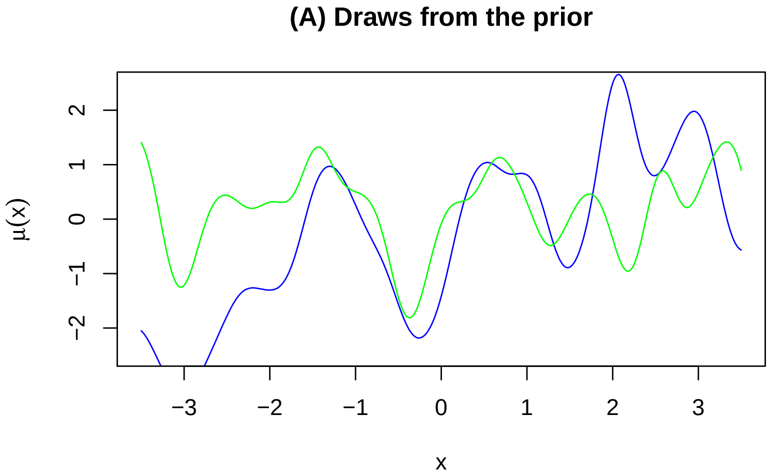
where R_θ is the matrix such that its (i,j)-th element is $r_\theta(x_i, x_j)$.

For squared exponential kernel k_θ, easy to derive r_θ in closed form.

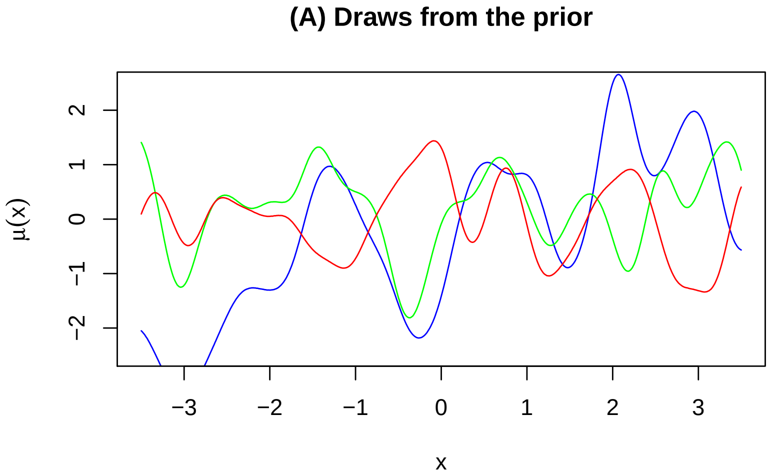
Illustration



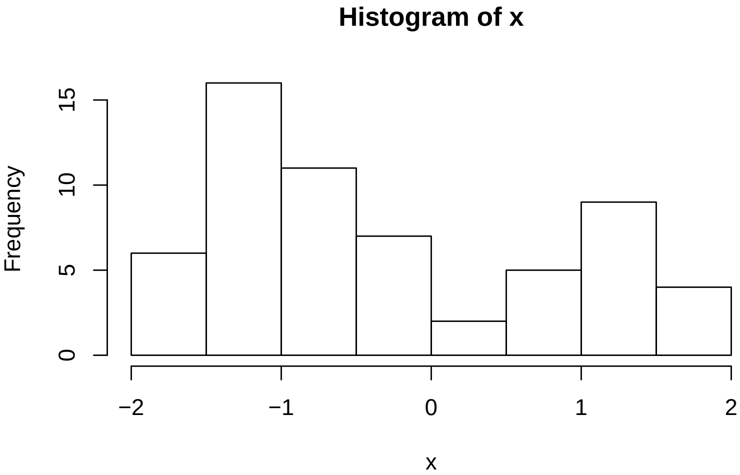
Illustration



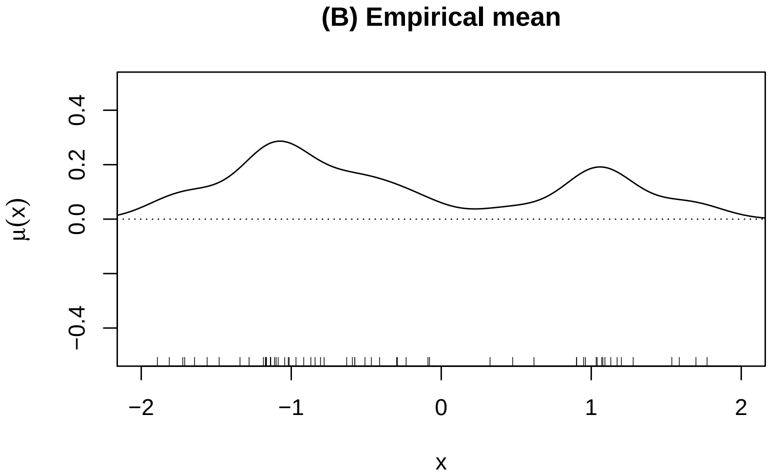
Illustration



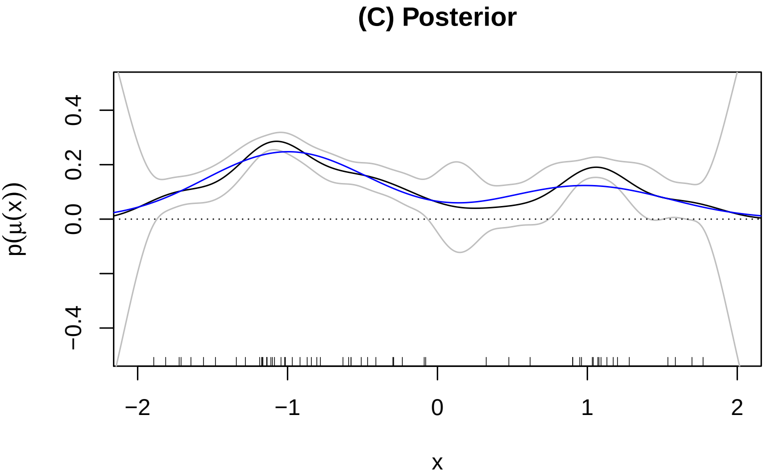
Illustration



Illustration



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 $\mathcal{X} \subset \mathbb{R}^D$, with $m \geq D$.
- Consider change of variables from mapping $\phi_{\mathbf{z}} : \mathbb{R}^D \mapsto \mathbb{R}^m$, given by

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

- By Cramér's decomposition theorem our model is equivalent to:

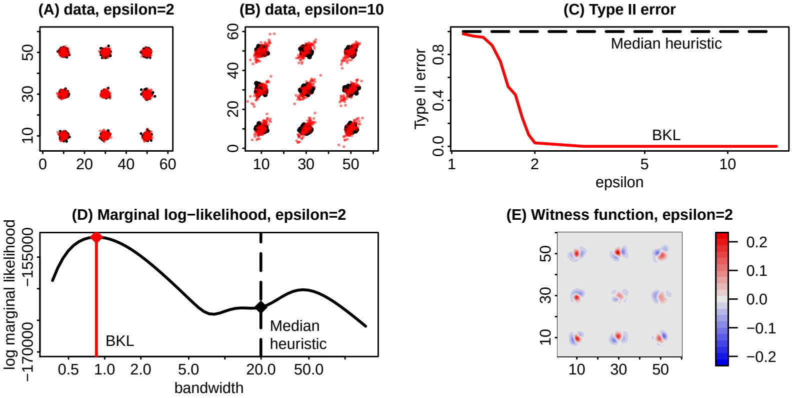
$$\phi_{\mathbf{z}}(X_i)|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right). \tag{1}$$

- Applying the change of variable $x \mapsto \phi_{\mathbf{z}}(x)$ we obtain:

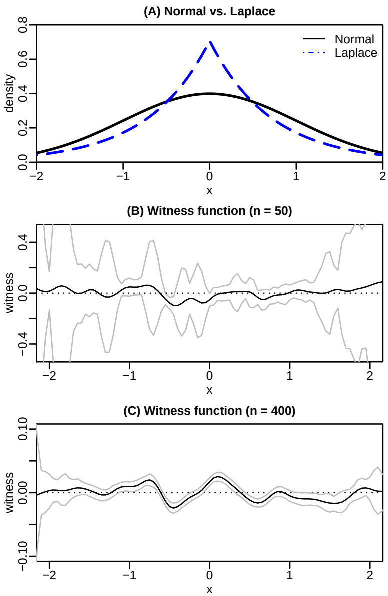
$$p(x|\mu_\theta, \theta) = p\left(\phi_{\mathbf{z}}(x)|\mu_\theta(\mathbf{z})\right) \operatorname{vol}\left[J_\theta(x)\right], \tag{2}$$

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



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
- New paper on arXiv, "Probabilistic Integration and Intractable Distributions" [Oates et al.] using Bayesian Kernel Embedding.
- Come see poster for more details

Thanks!

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