Bayesian Learning of Kernel Embeddings

Seth Flaxman
flaxman@stats.ox.ac.uk
Department of Statistics
University of Oxford

Dino Sejdinovic
dino.sejdinovic@stats.ox.ac.uk
Department of Statistics
University of Oxford

John P. Cunningham
jpc2181@columbia.edu
Department of Statistics
Columbia University

Sarah Filippi
filippi@stats.ox.ac.uk
Department of Statistics
University of Oxford

Abstract

Kernel methods are one of the mainstays of machine learning, but the problem of kernel learning remains challenging, with only a few heuristics and very little theory. This is of particular importance in methods based on estimation of kernel mean embeddings of probability measures. For characteristic kernels, which include most commonly used ones, the kernel mean embedding uniquely determines its probability measure, so it can be used to design a powerful statistical testing framework, which includes nonparametric two-sample and independence tests. In practice, however, the performance of these tests can be very sensitive to the choice of kernel and its lengthscale parameters. To address this central issue, we propose a new probabilistic model for kernel mean embeddings, the Bayesian Kernel Embedding model, combining a Gaussian process prior over the Reproducing Kernel Hilbert Space containing the mean embedding with a conjugate likelihood function, thus yielding a closed form posterior over the mean embedding. The posterior mean of our model is closely related to recently proposed shrinkage estimators for kernel mean embeddings, while the posterior uncertainty is a new, interesting feature with various possible applications. Critically for the purposes of kernel learning, our model gives a simple, closed form marginal pseudolikelihood of the observed data given the kernel hyperparameters. This marginal pseudolikelihood can either be optimized to inform the hyperparameter choice or fully Bayesian inference can be used.

1 INTRODUCTION

A large class of popular and successful machine learning methods rely on kernels (positive semidefinite functions), including support vector machines, kernel ridge regression, kernel PCA (Schölkopf and Smola, 2002), Gaussian processes (Rasmussen and Williams, 2006), and kernel-based hypothesis testing (Gretton et al., 2005, 2008, 2012a). A key component for many of these methods is that of estimating kernel mean embeddings and covariance operators of probability measures based on data. The use of simple empirical estimators has been challenged recently (Muandet et al., 2016) and alternative, better-behaved frequentist shrinkage strategies have been proposed. In this article, we develop a Bayesian framework for estimation of kernel mean embeddings, recovering desirable shrinkage properties as well as allowing quantification of full posterior uncertainty. Moreover, the developed framework has an additional extremely useful feature. Namely, a persistent problem in kernel methods is that of kernel choice and hyperparameter selection, for which no general-purpose strategy exists. When a large dataset is available in a supervised setting, the standard approach is to use cross-validation. However, in unsupervised learning and kernel-based hypothesis testing, cross-validation is not straightforward to apply and yet the choice of kernel is critically important. Our framework gives a tractable closed-form marginal pseudolikelihood of the data allowing direct hyperparameter optimization as well as fully Bayesian posterior inference through integrating over the kernel hyperparameters. We emphasise that this approach is fully unsupervised: it is based solely on the modelling of kernel mean embeddings – going beyond marginal likelihood based approaches in, e.g., Gaussian process regression – and is thus broadly applicable in situations, such as kernel-based hypothesis testing, where the hyperparameter choice has thus far been mainly driven by heuristics.

In Section 2 we provide the necessary background on Reproducing Kernel Hilbert Spaces (RKHS) as well as describe some related works. In Section 3 we develop our Bayesian Kernel Embedding model, showing a rigorous Gaussian process prior formulation for an RKHS. In Section 4 we show how to perform kernel learning and posterior inference with our model. In Section 5 we empirically evaluate our model, arguing that our Bayesian Kernel Learning (BKL) objective should be considered as a

"drop-in" replacement for heuristic methods of choosing kernel hyperparameters currently in use, especially in unsupervised settings such as kernel-based testing. We close in Section 6 with a discussion of various applications of our approach and future work.

2 BACKGROUND AND RELATED WORK

2.1 KERNEL EMBEDDINGS OF PROBABILITY MEASURES

For any positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$. RKHS is an (often infinite-dimensional) space of functions $h : \mathcal{X} \to \mathbb{R}$ where evaluation can be written as an inner product, and in particular $h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k}$ for all $h \in \mathcal{H}_k, x \in \mathcal{X}$. Given a probability measure P on $\mathcal{X}$, its kernel embedding into $\mathcal{H}_k$ is defined as:

$$\mu_{\mathsf{P}} = \int k(\cdot, x)\, \mathsf{P}(dx). \tag{1}$$

Embedding μ_{P} is an element of $\mathcal{H}_k$ and serves as a representation of P akin to a characteristic function. It represents expectations of RKHS functions in the form of an inner product $\int h(x)\mathsf{P}(dx) = \langle h, \mu_{\mathsf{P}}\rangle_{\mathcal{H}_k}$. For a broad family of kernels termed *characteristic* (Sriperumbudur et al., 2011), every probability measure has a unique embedding – thus, such embeddings completely determine their probability measures and capture all of the moment information. This yields a framework for constructing nonparametric hypothesis tests for the two-sample problem and for independence, which are consistent against all alternatives (Gretton et al., 2008, 2012a) – we review this framework in the next section.

2.2 KERNEL MEAN EMBEDDING AND HYPOTHESIS TESTING

Given a kernel k and probability measures P and Q, the maximum mean discrepancy (MMD) between P and Q (Gretton et al., 2012a) is defined as the squared RKHS distance $\|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{H}_k}$ between their embeddings. A related quantity is the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2005, 2008), a nonparametric dependence measure between random variables X and Y on domains $\mathcal{X}$ and $\mathcal{Y}$ respectively, defined as the squared RKHS distance $\|\mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X\mathsf{P}_Y}\|^2_{\mathcal{H}_\kappa}$ between the embeddings of the joint distribution P_{XY} and of the product of the marginals $\mathsf{P}_X\mathsf{P}_Y$ with respect to a kernel $\kappa : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ on the product space. Typically, κ factorises, i.e. $\kappa((x, y), (x', y')) = k(x, x')l(y, y')$. The empirical versions of MMD and HSIC are used as test statistics for the two-sample ($\mathbf{H}_0 : \mathsf{P} = \mathsf{Q}$ vs. $\mathbf{H}_1 : \mathsf{P} \neq \mathsf{Q}$) and independence ($\mathbf{H}_0 : X \perp\!\!\!\perp Y$ vs. $\mathbf{H}_1 : X \not\perp\!\!\!\perp Y$) tests, respectively. With the help of the approximations to the asymptotic distribution under the null hypothesis, corresponding p-values can be computed (Gretton et al., 2012a). In addition, the so-called "witness function" which is proportional to $\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}$ can be used to assess where the difference between the distributions arises.

2.3 KERNEL MEAN EMBEDDING ESTIMATORS

For a set of i.i.d. samples $x_1, \ldots, x_n$, the kernel mean embedding is typically estimated by its empirical version

$$\widehat{\mu_{\mathsf{P}}} = \mu_{\widehat{\mathsf{P}}} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i), \tag{2}$$

from which various associated quantities, including the estimators of the squared RKHS distances between embeddings needed for kernel-based hypothesis tests, follow. As an empirical mean in an infinite-dimensional space, (2) is affected by Stein's phenomenon, as overviewed by Muandet et al. (2013) who also propose alternative shrinkage estimators similar to the well known James-Stein estimator. Improvements of test power using such shrinkage estimators are reported by Ramdas and Wehbe (2015). Connections between the James-Stein estimator and empirical Bayes procedures are classical (Efron and Morris, 1973), and thus a natural question to consider is whether a Bayesian formulation of the problem of kernel embedding estimation would yield similar shrinkage properties. In this paper, we will give a Bayesian perspective of the problem of kernel embedding estimation. In particular, we will construct a flexible model for underlying probability measures based on Gaussian measures in RKHSs which allows derivation of a full posterior distribution of μ_{P}, recovering similar shrinkage properties to Muandet et al. (2013), as discussed in Section 4.2. The model will give us a further advantage, however – as the marginal likelihood of the data given the kernel parameter can be derived leading to an informed choice of kernel parameters.

2.4 SELECTION OF KERNEL PARAMETERS

In supervised kernel methods like support vector machines, leave-one-out or k-fold crossvalidation is an effective and widely used method for kernel selection, and the myriad papers on multiple kernel learning (e.g. Bach et al. (2004); Sonnenburg et al. (2006); Gönen and Alpaydın (2011)) assume that some loss function is available and thus focus on effective ways of learning combinations of kernels. In the related but distinct world of smoothing kernels and kernel density estimation, there are a variety of long-standing approaches to bandwidth selection, again based on a loss function (in this case, mean integrated squared error is a

popular choice (Bowman, 1985), and there is even a formula giving the optimal smoothing parameter asymptotically, see Rosenblatt (1956); Parzen (1962)) but we are not aware of work linking this literature to methods based on positive definite/RKHS kernels we study here. Separately, Gaussian process learning can be undertaken by maximizing the marginal likelihood, which has a convenient closed form. This is noteworthy for its success and general applicability even for learning complicated combinations of kernels (Duvenaud et al., 2013) or rich kernel families (Wilson and Adams, 2013). Our approach has the same basic design as that of Gaussian process learning, yet it is applicable to learning kernel embeddings, which falls outside the realm of supervised learning.

As noted in Gretton et al. (2012b), the choice of the kernel k is critically important for the power of the tests presented in Section 2.2. However, no general, theoretically-grounded approaches for kernel selection in this context exist. The difficulty is that, unlike in supervised kernel methods, a simple cross-validation approach for the kernel parameter selection is not possible. What would be an ideal objective function – asymptotic test power – cannot be computed due to a complicated asymptotic null distribution. Moreover, even if we were able to estimate the power by performing tests on "training data" for each of the individual candidate kernels, in order to account for multiple comparisons, this training data would have to be disjoint from the one on which the hypothesis test is performed, which is clearly wasteful of power and appropriate only in the type of large-scale settings discussed in Gretton et al. (2012b). For these reasons, most users of kernel hypothesis tests in practice resort to using a parameterized kernel family such as squared exponential, and setting the length-scale parameter based on the "median heuristic."

The exact origins of the median heuristic are unclear (interestingly, it does not appear in the book that is most commonly cited as its source, Schölkopf and Smola (2002)) but it may have been derived from Takeuchi et al. (2006) and has precursors in classical work on bandwidth selection for kernel density estimation (Bowman, 1985). Note that there are two versions of the median heuristic in the literature: in both versions, given a set of observations $x_1, \ldots, x_n$ we calculate $\ell = \text{median}(\|x_i - x_j\|_2)$ and then one version (e.g. Mooij et al. (2015)) uses the Gaussian RBF / squared exponential kernel parameterized as $k(x, x') = \exp(-\frac{\|x-x'\|^2}{\ell^2})$ and the second version (e.g. Muandet et al. (2014)) uses the parameterization $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2\ell^2})$. Some recent work has highlighted the situations in which the median heuristic can lead to poor performance (Gretton et al., 2012b). Cases in which the median heuristic performs quite well and also cases in which it performs quite poorly are discussed in (Reddi et al., 2015; Ramdas et al., 2015). We note that the median heuristic has also been used as a default value for supervised learning tasks (e.g. for the SVM implementation in R package `kernlab`) or when cross-validation is simply too expensive.

Outside of kernel methods, the same basic conundrum arises in spectral clustering in the choice of the parameters for the similarity graph (Von Luxburg, 2007, Section 8.1) and it is implicitly an issue in any unsupervised statistical method based on distances or dissimilarities, like the distance covariance (which is in fact equivalent to HSIC with a certain family of kernel functions (Sejdinovic et al., 2013)), or even the choice of the number of neighbors k in k-nearest neighbors algorithms.

3 OUR MODEL: BAYESIAN KERNEL EMBEDDING

Below, we will work with a parametric family of kernels $\{k_\theta(\cdot,\cdot)\}_{\theta\in\Theta}$. Given a dataset $\{x_i\}_{i=1}^n \sim \mathsf{P}$ of observations in $\mathbb{R}^D$ for an unknown probability distribution P, we wish to infer the kernel embedding $\mu_{\mathsf{P},\theta} = \int k_\theta(\cdot, x)\,\mathsf{P}(dx)$ for a given kernel k_θ in the parametric family. Moreover, we wish to construct a model that will allow inference of the kernel hyperparameter θ as well. Note that the two goals are related, since θ determines the space in which the embedding $\mu_{\mathsf{P},\theta}$ lies. When it is obvious from context, we suppress the dependence of the embeddings on the underlying measure P, writing μ_θ to emphasize the dependence on θ. Similarly, we will use $\widehat{\mu_\theta}$ to denote the simple empirical estimator from Eq. (2), which depends on a fixed sample $\{x_i\}_{i=1}^n$.

Our Bayesian Kernel Embedding (BKE) approach consists in specifying a prior on the kernel mean embedding μ_θ and a likelihood function linking it to the observations through the empirical estimator $\widehat{\mu_\theta}$. This will then allow us to infer the posterior distribution of the kernel mean embedding. The hyperparameter θ can itself have a prior, with the goal of learning a posterior distribution over the hyperparameter space.

3.1 PRIOR

A given hyperparameter θ (which can itself have a prior distribution), parameterizes a kernel k_θ and a corresponding RKHS $\mathcal{H}_{k_\theta}$. While it is tempting to define a $\mathcal{GP}(0, k_\theta(\cdot,\cdot))$ prior on μ_θ, this is problematic since draws from such prior would almost surely fall outside $\mathcal{H}_k$ (Wahba, 1990). Therefore, we define a GP prior over μ_θ as follows:

$$\mu_\theta \mid \theta \sim \mathcal{GP}(0, r_\theta(\cdot,\cdot))\ , \tag{3}$$

$$r_\theta(x, y) := \int k_\theta(x, u) k_\theta(u, y)\nu(du)\ . \tag{4}$$

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 (Lukić and Beder, 2001; Pillai et al., 2007) of k_θ over r_θ for any stationary kernel k_θ and more broadly whenever $\int k_\theta(x,x)\nu(dx) < \infty$. For completeness, we provide details of this construction in the Appendix in Section A.2. Since Eq. (4) is the convolution of a kernel with itself with respect to ν, for typical kernels k_θ, the resulting kernel r_θ can be thought of as a smoother version of k_θ. A particularly convenient choice for $\mathcal{X} = \mathbb{R}^D$ is to take ν to be proportional to a Gaussian measure in which case r_θ can be computed analytically for a squared exponential kernel k_θ. The derivation is given in the Appendix in Section A.3, where we further show that if we set ν to be proportional to an isotropic Gaussian measure with a large variance parameter, r_θ becomes very similar to a squared exponential kernel with lengthscale $\theta\sqrt{2}$.

3.2 LIKELIHOOD

We need a likelihood linking the kernel mean embedding μ_θ to the observations $\{x_i\}_{i=1}^n$. We define the likelihood via the empirical mean embedding estimator of Eq. (2), $\widehat{\mu_\theta}$ which depends on $\{x_i\}_{i=1}^n$ and θ. Consider evaluating $\widehat{\mu_\theta}$ at some $x \in \mathbb{R}^D$ (which need not be one of our observations). The result is a real number giving an empirical estimate of $\mu_\theta(x)$ based on $\{x_i\}_{i=1}^n$ and θ. We link the empirical estimate, $\widehat{\mu_\theta}(x)$, to the corresponding modeled estimate, $\mu_\theta(x)$ using a Gaussian distribution with variance τ^2/n:

$$p(\widehat{\mu_\theta}(x)|\mu_\theta(x)) = \mathcal{N}(\widehat{\mu_\theta}(x); \mu_\theta(x), \tau^2/n), \quad x \in \mathcal{X}. \tag{5}$$

Our motivation for choosing this likelihood comes from the Central Limit Theorem. For a fixed location x, $\widehat{\mu_\theta}(x) = \frac{1}{n}\sum_{i=1}^n k_\theta(x_i, x)$ is an average of i.i.d. random variables so it satisfies:

$$\sqrt{n}(\widehat{\mu_\theta}(x) - \mu_\theta(x)) \xrightarrow{D} \mathcal{N}(0, \mathrm{Var}_{X\sim\mathsf{P}}[k_\theta(X,x)]). \tag{6}$$

We note that considering a heteroscedastic variance dependent on x in (5) would be a straightforward extension to our model, but we do not pursue this idea further here, i.e. while τ^2 can depend both on θ and x, we treat it as a single hyperparameter in the model.

3.3 JUSTIFICATION FOR THE MODEL

There are various ways to understand the construction of our hierarchical model. $\{x_i\}_{i=1}^n$ are drawn iid from P, which we do not have access to. We could estimate P directly (e.g. with a Gaussian mixture model) obtaining $\hat{\mathsf{P}}$, and then estimate $\mu_{\theta,\hat{\mathsf{P}}}$. But since density estimation is challenging in high dimensions, we posit a generative model for μ_θ directly.

Beginning at the top of the hierarchy, we have a fixed or random hyperparameter θ, which immediately defines k_θ and the corresponding RKHS $\mathcal{H}_{k_\theta}$. Then, we introduce a GP prior over μ_θ to ensure that $\mu_\theta \in \mathcal{H}_{k_\theta}$. A few realizations of μ_θ drawn from our prior are shown in Figure 1 (A), for an illustrative one-dimensional example where the prior is a Gaussian process with squared exponential kernel with lengthscale $\theta = 0.25$. Small values of θ yield rough functions and large values of θ yield smooth functions. Next, we need to define the likelihood, which links these

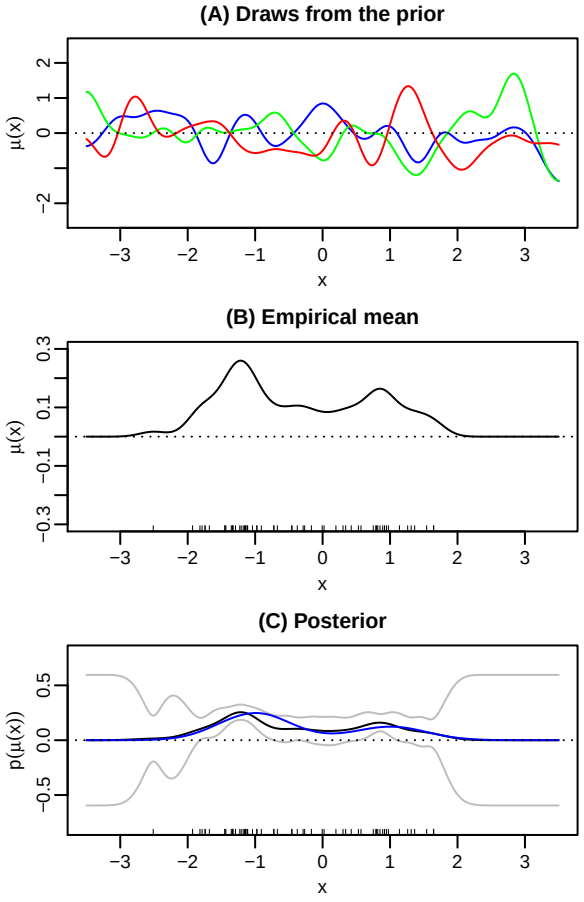

Figure 1: An illustration of the Bayesian Kernel Embedding model, where k_θ is a squared exponential kernel with lengthscale 0.1. Three draws of μ_θ from the prior are shown in (A). The empirical mean estimator $\widehat{\mu_\theta}$, which is the link function for the likelihood, is shown in (B) with the observations shown as a rug plot. In (C), the posterior mean embedding (black line) with uncertainty intervals (gray lines) is shown, as is the true mean embedding (blue line) based on the true data generating process (a mixture of Gaussians) and the same k_θ.

draws from the prior to the observations $\{x_i\}_{i=1}^n$. Since μ_θ is an infinite dimensional element in a Hilbert space and $\{x_i\}_{i=1}^n \in \mathcal{X}$ we need to transform the observations so that we can put a probability distribution over them. We use the empirical estimate of the mean embedding $\widehat{\mu_\theta}$ as our link function. Given a few observations, $\widehat{\mu_\theta}$ is shown in

Figure 1 (B). Our likelihood links $\widehat{\mu_\theta}$ to μ_θ at the observation locations $\{x_i\}_{i=1}^n$ by assuming a squared loss function, i.e. Gaussian errors. As mentioned above, the motivation is the Central Limit Theorem, but also the convenient conjugate form that a Gaussian process with Gaussian likelihood yields. A plot of the posterior over the mean embedding is shown in Figure 1 (C). A few points are worth noting: since the empirical estimator is already quite smooth (notice its similarity to a kernel density estimate), the posterior mean embedding is only slightly smoother than the empirical mean embedding. Notice that unlike kernel density estimation, there is no requirement that the kernel mean embedding be non-negative, thus explaining the posterior uncertainty intervals which are below zero.

Our original motivation for considering a Bayesian model for kernel mean embeddings was to see whether there was a coherent Bayesian formulation that corresponded to the shrinkage estimators in Muandet et al. (2013), while also enabling us to learn the hyperparameters. The first difficulty we faced was how to define a valid prior over the RKHS and a reasonable likelihood function. Our choices are by no means definitive, and we hope to see further development in this area in the future. The second difficulty was that of developing a method for inferring hyperparameters, to which we turn in the next section.

4 BAYESIAN KERNEL LEARNING

In this section we show how to perform learning and inference in the Bayesian Kernel Embedding model introduced in the previous section. Our model inherits various attractive properties from the Gaussian process framework (Rasmussen and Williams, 2006). First, we derive the posterior and posterior predictive distributions for the kernel mean embedding in closed form due to the conjugacy of our model, and show the relationship with previously proposed shrinkage estimators. We then derive the tractable marginal likelihood of the observations given the hyperparameters allowing for efficient MAP estimation or posterior inference for hyperparameters.

4.1 POSTERIOR AND POSTERIOR PREDICTIVE DISTRIBUTIONS

Similarly to GP models, the posterior mean of μ_θ is available in closed form due to the conjugacy of Gaussians. Perhaps given our data we wish to infer μ_θ at a new location $x^* \in \mathbb{R}^D$. Given a value of the hyperparameter θ we can calculate the posterior distribution of μ_θ as well as the posterior predictive distribution $p(\mu_\theta(x^*)|\widehat{\mu_\theta}, \theta)$.

Standard GP results (Rasmussen and Williams, 2006) yield the posterior distribution as:

$$\begin{aligned}[\mu_\theta(x_1), \ldots, \mu_\theta(x_n)]^\top \mid [\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \theta \\ \sim \mathcal{N}(R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}[\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \\ R_\theta - R_\theta(R_\theta + (\tau^2/n)I_n)^{-1}R_\theta),\end{aligned} \tag{7}$$

where R_θ is the $n \times n$ matrix such that its (i, j)-th element is $r_\theta(x_i, x_j)$. The posterior predictive distribution at a new location x^* is:

$$\begin{aligned}\mu_\theta(x^*)^\top \mid [\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \theta \\ \sim \mathcal{N}(R_\theta^{*\top}(R_\theta + (\tau^2/n)I_n)^{-1}[\widehat{\mu_\theta}(x_1), \ldots, \widehat{\mu_\theta}(x_n)]^\top, \\ r_\theta^{**} - R_\theta^{*\top}(R_\theta + (\tau^2/n)I_n)^{-1}R_\theta^*)\end{aligned} \tag{8}$$

where $R_\theta^* = [r_\theta(x^*, x_1), \ldots r_\theta(x^*, x_n)]^\top$ and $r_\theta^{**} = r_\theta(x^*, x^*)$.

As in standard GP inference, the time complexity is $\mathcal{O}(n^3)$ due to the matrix inverses and the storage is $\mathcal{O}(n^2)$ to store the $n \times n$ matrix R_θ.

4.2 RELATION TO THE SHRINKAGE ESTIMATOR

The spectral kernel mean shrinkage estimator (S-KMSE) of Muandet et al. (2013) for a fixed kernel k is defined as:

$$\check{\mu}_\lambda = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\mu}, \tag{9}$$

where $\hat{\mu} = \sum_{i=1}^n k(\cdot, x_i)$ is the empirical embedding, $\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) \otimes k(\cdot, x_i)$ is the empirical covariance operator on $\mathcal{H}_k$, and λ is a regularization parameter. (Muandet et al., 2013, Proposition 12) shows that $\check{\mu}_\lambda$ can be expressed as a weighted kernel mean $\check{\mu}_\lambda = \sum_{i=1}^n \beta_i k(\cdot, x_i)$, where

$$\begin{aligned}\beta &= \frac{1}{n}(K + n\lambda I)^{-1}K\mathbf{1} \\ &= (K + n\lambda I)^{-1}[\widehat{\mu}(x_1), \ldots, \widehat{\mu}(x_n)]^\top.\end{aligned}$$

Now, evaluating S-KMSE at any point x^* gives

$$\begin{aligned}\check{\mu}_\lambda(x^*) &= \sum_{i=1}^n \beta_i k(x^*, x_i) \\ &= K_*^\top(K + n\lambda I)^{-1}[\widehat{\mu}(x_1), \ldots, \widehat{\mu}(x_n)]^\top,\end{aligned}$$

where $K_* = [k(x^*, x_1), \ldots, k(x^*, x_n)]^\top$. Thus, the posterior mean in Eq. (7) recovers the S-KMSE estimator (Muandet et al., 2013), where the regularization parameter is related to the variance in the likelihood model (5), with a difference that in our case the kernel k_θ used to compute the empirical embedding is not the same as the kernel r_θ used to compute the kernel matrices. We note that our method has various advantages over the frequentist estimator $\check{\mu}_\lambda$:

we have a closed-form uncertainty estimate, while we are not aware of a principled way of calculating the standard error of the frequentist estimators of embeddings. Our model also leads to a method for learning the hyperparameters, which we discuss next.

4.3 INFERENCE OF THE KERNEL PARAMETERS

In this section we focus on hyperparameter learning in our model. For the purposes of hyperparameter learning, we want to integrate out the kernel mean embedding μ_θ and consider the probability of our observations $\{x_i\}_{i=1}^n$ given the hyperparameters θ. In order to link our generative model directly to the observations, we use a pseudolikelihood approach as discussed in detail below.

We use the term pseudolikelihood because the model in this section will not correspond to the likelihood of the infinite dimensional empirical embedding; rather it will rely on the evaluations of the empirical embedding at a finite set of points. Let us fix a set of points $z_1, \ldots, z_m$ in $\mathcal{X} \subset \mathbb{R}^D$, with $m \geq D$. These points are not treated as random, and the inference method we develop does not require any specific choice of $\{z_j\}_{j=1}^m$. However, to ensure that there is a reasonable variability in the values of $k(x_i, z_j)$, these points should be placed in the high density regions of P. The simplest approach is to use a small held out portion of the data (with $m \ll n$ but $m \geq D$). Now, when we evaluate $\widehat{\mu_\theta}$ at these points, our modelling assumption from (5) on vector $\widehat{\mu_\theta}(\mathbf{z}) = [\widehat{\mu_\theta}(z_1), \ldots, \widehat{\mu_\theta}(z_m)]$ can be written as

$$\widehat{\mu_\theta}(\mathbf{z})|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \frac{\tau^2}{n} I_m\right). \tag{10}$$

However, as $\widehat{\mu_\theta}(z_j) = \frac{1}{n}\sum_{i=1}^n k_\theta(X_i, z_j)$ and all the terms $k_\theta(X_i, z_j)$ are independent given μ_θ, by Cramér's decomposition theorem, this modelling assumption is for the mapping $\phi_\mathbf{z} : \mathbb{R}^D \mapsto \mathbb{R}^m$, given by

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

equivalent to:

$$\phi_\mathbf{z}(X_i)|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right). \tag{11}$$

Applying the change of variable $x \mapsto \phi_\mathbf{z}(x)$ and using the generalization of the change-of-variables formula to non-square Jacobian matrices as described in (Ben-Israel, 1999), we obtain a distribution for x conditionally on μ_θ and θ:

$$p(x|\mu_\theta, \theta) = p\left(\phi_\mathbf{z}(x)|\mu_\theta(\mathbf{z})\right) \mathrm{vol}\left[J_\theta(x)\right], \tag{12}$$

where $J_\theta(x) = \left[\frac{\partial k_\theta(x, z_i)}{\partial x^{(j)}}\right]_{ij}$ is an $m \times D$ matrix, and

$$\begin{aligned}\mathrm{vol}\left[J_\theta(x)\right] &= \left(\det\left[J_\theta(x)^\top J_\theta(x)\right]\right)^{1/2} \\ &= \left(\det\left[\sum_{l=1}^m \frac{\partial k_\theta(x, z_l)}{\partial x^{(i)}} \frac{\partial k_\theta(x, z_l)}{\partial x^{(j)}}\right]_{ij}\right)^{1/2} \\ &=: \gamma_\theta(x)\,. \end{aligned} \tag{13}$$

The notation $\gamma_\theta(x)$ highlights the dependence on both θ and x. An explicit calculation of $\gamma_\theta(x)$ for squared exponential kernels is described in Section 4.4.

By the conditional independence of $\{\phi_\mathbf{z}(X_i)\}_{i=1}^n$ given μ_θ, we obtain the pseudolikelihood of all n observations:

$$\begin{aligned} p(x_1, \ldots, x_n|\mu_\theta, \theta) &= \prod_{i=1}^n \mathcal{N}\left(\phi_\mathbf{z}(x_i); \mu_\theta(\mathbf{z}), \tau^2 I_m\right) \gamma_\theta(x_i) \\ &= \mathcal{N}\left(\phi_\mathbf{z}(\mathbf{x}); \mathbf{m}_\theta(\mathbf{z}), \tau^2 I_{mn}\right) \prod_{i=1}^n \gamma_\theta(x_i), \end{aligned} \tag{14}$$

where

$$\phi_\mathbf{z}(\mathbf{x}) = \left[\phi_\mathbf{z}(x_1)^\top \cdots \phi_\mathbf{z}(x_n)^\top\right]^\top = \mathrm{vec}\left\{K_{\theta,\mathbf{zx}}\right\} \in \mathbb{R}^{mn}$$

and in the mean vector $\mathbf{m}_\theta(\mathbf{z}) = \left[\mu_\theta(\mathbf{z})^\top \cdots \mu_\theta(\mathbf{z})^\top\right]^\top$, $\mu_\theta(\mathbf{z})$ repeats n times. Under the prior (3), this mean vector has mean $\mathbf{0}$ and covariance $\mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}}$ where $R_{\theta,\mathbf{zz}}$ is the $m \times m$ matrix such that its (i, j)-th element is $r_\theta(z_i, z_j)$. Combining this prior and the pseudolikelihood in (14), we have the marginal pseudolikelihood:

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \int p(x_1, \ldots, x_n|\mu_\theta, \theta) p(\mu_\theta|\theta) d\mu_\theta \\ &= \int \mathcal{N}\left(\phi_\mathbf{z}(\mathbf{x}); \mathbf{m}_\theta(\mathbf{z}), \tau^2 I_{mn}\right) \left[\prod_{i=1}^n \gamma_\theta(x_i)\right] p(\mu_\theta|\theta) d\mu_\theta \\ &= \mathcal{N}\left(\phi_\mathbf{z}(\mathbf{x}); \mathbf{0}, \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right) \prod_{i=1}^n \gamma_\theta(x_i). \end{aligned} \tag{15}$$

While the marginal pseudolikelihood in Eq. (15) involves a computation of the likelihood for an mn-dimensional normal distribution, the Kronecker structure of the covariance matrix allows efficient computation as described in Appendix A.4. The complexity for calculating this likelihood is $\mathcal{O}(m^3 + mn)$ (dominated by the inversion of $R_{\theta,\mathbf{zz}} + (\tau^2/n) I_m$). The Jacobian term depends on the parametric form of k_θ, but a typical cost as shown in Section 4.4 for the squared exponential kernel is $\mathcal{O}(nD^3 + nmD^2)$. In this case, the computation of matrices $R_{\theta,\mathbf{zz}}$ and $\phi_\mathbf{z}(\mathbf{x}) = \mathrm{vec}\left\{K_{\theta,\mathbf{zx}}\right\}$ is $\mathcal{O}(m^2 D)$ and $\mathcal{O}(mnD)$ respectively.

Just as in GP modeling, the marginal pseudolikelihood can be maximized directly for maximum likelihood II (also

known as empirical Bayes) estimation, in which we look for a single best $\hat{\theta}$, or it can be used to construct an efficient MCMC sampler from the posterior of θ.

4.4 EXPLICIT CALCULATIONS FOR SQUARED EXPONENTIAL (RBF) KERNEL

Consider the isotropic squared exponential kernel with lengthscale matrix $\theta^2 I_D$ defined by

$$k_\theta(x, y) = \exp(-.5(x - y)^\top \theta^{-2} I_D (x - y)). \quad (16)$$

In this case, we can analytically calculate $r_\theta(x, y)$, exact form is given in the Appendix in Section A.3.

The partial derivatives of $k_\theta(x, y)$ with respect to $x^{(i)}$ for $i = 1, \dots D$ can be easily derived as

$$\frac{\partial k_\theta(x, y)}{\partial x^{(i)}} = k_\theta(x, y) \frac{x^{(i)} - y^{(i)}}{\theta^2}$$

and therefore the Jacobian from Eq. (13) is equal to

$$\gamma_\theta(x) = \left(\det \left[\sum_{l=1}^{m} k_\theta(x, z_l)^2 \, \frac{(x^{(i)} - z_l^{(j)})^2}{\theta^4} \right]_{ij} \right)^{1/2}. \quad (17)$$

The computation of the matrix is $\mathcal{O}(mD^2)$ and the determinant is $\mathcal{O}(D^3)$. Since we must calculate $\gamma_\theta(x_i)$ for each x_i, the overall time complexity is $\mathcal{O}(nD^3 + nmD^2)$.

5 EXPERIMENTS

We demonstrate our approach on two synthetic datasets and one example on real data, focusing on two-sample testing with MMD and independence testing with HSIC. First, we use our Bayesian Kernel Embedding model and learn the kernel hyperparameters with maximum likelihood II, optimizing the marginal likelihood. Second, we take a fully Bayesian approach to inference and learning with our model. Finally, we apply the PC algorithm for causal structure discovery to a real dataset. The PC algorithm relies on a series of independence tests; we use HSIC with the lengthscales set with Bayesian Kernel Learning.

Choosing lengthscales with the median heuristic is often a very bad idea. In the case of two sample testing, Gretton et al. (2012b) showed that MMD with the median heuristic failed to reject the null hypothesis when comparing samples from a grid of isotropic Gaussians to samples from a grid of non-isotropic Gaussians. We repeated this experiment by considering a distribution P of a mixture of bivariate Gaussians centered on a grid with diagonal covariance and unit variance and a distribution Q of a mixture of bivariate Gaussians centered at the same locations but with rotated covariance matrices with a ratio ϵ of largest to smallest covariance eigenvalues.

As illustrated in Figures 2(A) and (B), for small values of ϵ both distributions are very similar whereas the distinction between P and Q becomes more apparent as ϵ increases. For different values of ϵ, we sample 100 observations from each mixture component, yielding 900 observations from P and 900 observations from Q and then perform a two-sample test ($\mathbf{H}_0 : \mathsf{P} = \mathsf{Q}$ vs. $\mathbf{H}_1 : \mathsf{P} \neq \mathsf{Q}$) using the MMD empirical estimate with an isotropic squared exponential kernel with one hyperparameter, the lengthscale. The type II error (i.e. probability that the test fails to reject the null hypothesis that $\mathsf{P} = \mathsf{Q}$ at $\alpha = 0.05$) is shown in Figure 2(C) for differently skewed covariances (ϵ from 0.5 to 15) when the median heuristic is chosen to select the kernel lengthscale or when using the Bayesian Kernel Learning. In this example, the median heuristic picks a kernel with a large lengthscale, since the median distance between points is large. With this large lengthscale MMD always fails to reject at $\alpha = 0.05$ even for simple cases where ϵ is large. When we use Bayesian Kernel Learning and optimize the marginal likelihood of Eq. (15) for $\tau^2 = 1$ (our results were not sensitive to the choice of this parameter, but in the fully Bayesian case below we show that we can learn it) we found the maximum marginal likelihood at a lengthscale of 0.85. With this choice of lengthscale, MMD correctly rejects the null hypothesis at $\alpha = 0.05$ even for very hard situations when $\epsilon = 2$. We observe that when ϵ is smaller than 2, the type II error of MMD is very high for both choices of lengthscale, because the two distributions P and Q are so similar that the test always retains the null hypothesis. In Figure 2(D) we illustrate the BKL marginal likelihood across a range of lengthscales. Interestingly, there are multiple local optima and the median heuristic lies between the two main modes. The plot indicates that multiple scales may be of interest for this dataset, which makes sense given that the true data generating process is a mixture model. This insight can be incorporated into the Bayesian Kernel Embedding framework by expanding our model, as discussed below. In Figure 2(E) we used the BKE posterior to estimate the witness function $\mu_{P,\theta} - \mu_{Q,\theta}$. This function is large in magnitude in the locations where the two distributions differ. For ease of visualization we do not try to include posterior uncertainty intervals, but these are readily available from our model, and we show them for a 1-dimensional case below.

Our model does not just provide a better way of choosing lengthscales. We can also use it in a fully Bayesian context, where we place priors over the hyperparameters θ and τ^2, and then integrate them out to learn a posterior distribution over the mean embedding. Switching to one dimension, we consider a distribution $\mathsf{P} = \mathcal{N}(0, 1)$ and a distribution $\mathsf{Q} = \text{Laplace}(0, \sqrt{.5})$. The densities are shown in Figure 3(A). Notice that the first two moments of these distributions are equal. To create a synthetic dataset we sampled n observations from each distribution, and then combined them together into a sample of size $2n$, follow-

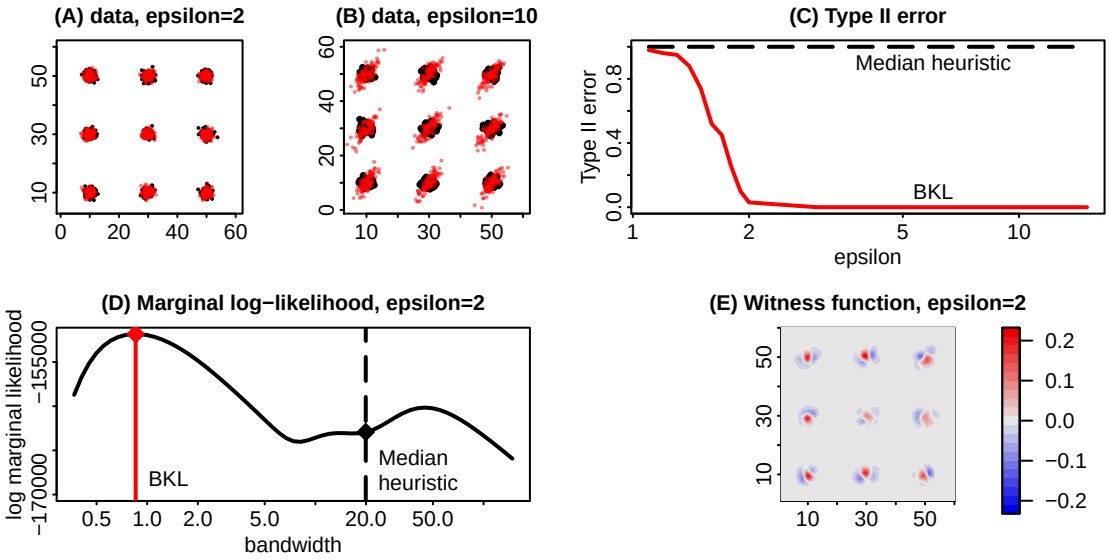

Figure 2: Two sample testing on a challenging simulated data set: comparing samples from a grid of isotropic Gaussians (black dots) to samples from a grid of non-isotropic Gaussians (red dots) with a ratio ϵ of largest to smallest covariance eigenvalues. Panels (A) and (B) illustrate such samples for two values of ϵ. (C) Type II error as a function of ϵ for significant level $\alpha = 0.05$ following the median heuristic or the BKL approach to choose the lengthscale. (D) BKL marginal log-likelihood across a range of lengthscales. It is maximised for a lengthscale of 0.85 whereas the median heuristic suggests a value of 20. (E) Witness function for the difficult case where $\epsilon = 2$ using the BKL lengthscale.

ing the strategy in the previous experiment to learn a single lengthscale and kernel mean embedding for the combined dataset. We ran a Hamiltonian Monte Carlo sampler (HMC) with NUTS (Stan source code is in the Appendix in Section B) for the Bayesian Kernel Embedding model with a squared exponential kernel, placing a Gamma$(1, 1)$ prior on the lengthscale θ of the kernel and a Gamma$(1, 1)$ prior on τ^2. We ran 4 chains for 400 iterations, discarding 200 iterations as warmup, with the chains starting at different random initial values. Standard convergence and mixing diagnostics were good ($\hat{R} \approx 1$), so we considered the result to be 800 draws from the posterior distribution. Recall that for fixed hyperparameters θ and τ^2 we can obtain a posterior distribution over $\mu_{P,\theta}$ and $\mu_{Q,\theta}$. For each of our 800 draws, we drew a sample from these two distributions and then calculated the witness function as the difference, thus obtaining a random function drawn from the posterior distribution over $\mu_{P,\theta} - \mu_{Q,\theta}$ (where in practice we evaluate this function at a fine grid for plotting purposes). We thus obtained the full posterior distribution over the witness function, integrating over the kernel hyperparameter. We followed this procedure twice to create a dataset with $n = 50$ and a dataset with $n = 400$. In Figure 3(B) we see that the witness function for the small dataset is not able to distinguish between the distributions as it rarely excludes 0. (Note that our model has the function 0 as its prior, which corresponds to the null hypothesis that the two distributions are equal. This could easily be changed to incorporate any relevant prior information.). As shown in Figure 3(C), with more data the witness function is able to distinguish between the two distributions, mostly excluding 0.

Finally, we consider the ozone dataset analyzed in Breiman and Friedman (1985), consisting of daily measurements of ozone concentration and eight related meteorological variables. Following the approach in Flaxman et al. (2015), we first pre-whiten the data to control for underlying temporal autocorrelation, then we use a combination of Gaussian process regression followed by HSIC to test for conditional independence. Each time we run HSIC, we set the kernel hyperparameters using Bayesian Kernel Learning. The graphical model that we learn is shown in Figure 4. The directed edge from the temperature variable to ozone is encouraging, as higher temperatures favor ozone formation through a variety of chemical processes which are not represented by variables in this dataset (Bloomer et al., 2009; Sillman, 1999). Note that this edge was not present in the graphical model in Flaxman et al. (2015) in which the median heuristic was used.

6 DISCUSSION

We developed a framework for Bayesian learning of kernel embeddings of probability measures. It is primarily designed for unsupervised settings, and in particular for kernel-based hypothesis testing. In these settings, one re-

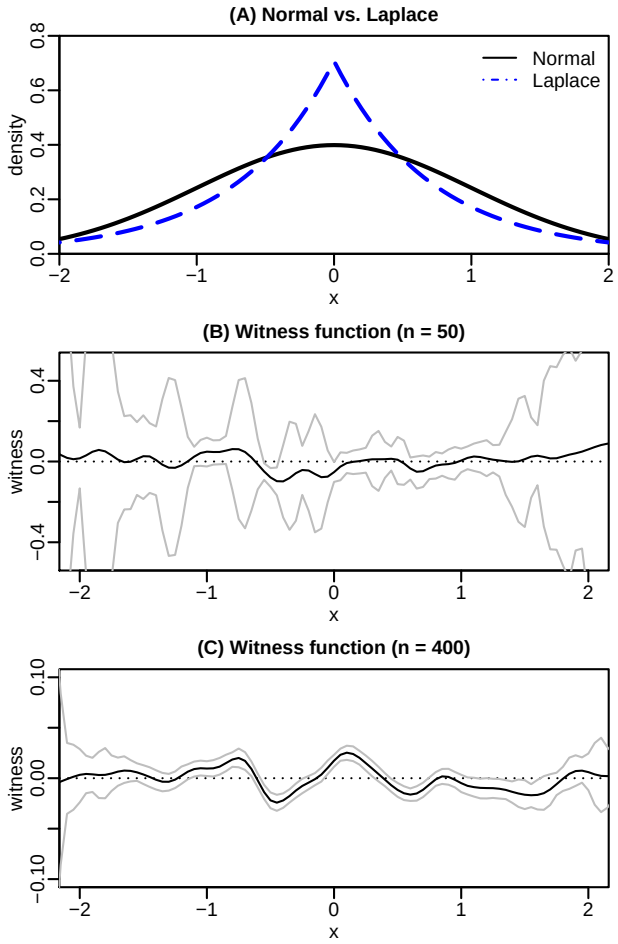

Figure 3: The true data generating process is shown in (A) where two samples of size n are drawn from distributions with equal means and variances. We then fit our Bayesian Kernel Embedding model, with priors over the hyperparameters θ and τ^2 to obtain a posterior over the witness function for two-sampling testing. The witness function indicates the model's posterior estimates of where the two distributions differ (when the witness function is zero, it indicates no difference between the distributions). Posterior means and 80% uncertainty intervals are shown. In (B) the small sample size means that the model does not effectively distinguish between samples from a normal and a Laplace distribution, while in (C) larger samples enable the model to find a clear difference, with much of the uncertainty envelope excluding 0.

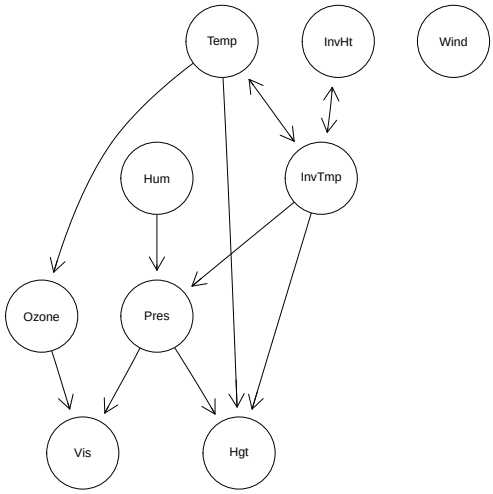

Figure 4: Graphical model representing an equivalence class of DAGs for the Ozone dataset from Breiman and Friedman (1985), learned using the PC algorithm following the approach in Flaxman et al. (2015) with HSIC to test for independence. We used BKL to set hyperparameters of HSIC. Singly directed edges represent causal links, while bidirected edges represent edges that the algorithm failed to orient. The causal edge from temperature to ozone accords with scientific understanding, and was not present in the graphical model learned in Flaxman et al. (2015) which employed the median heuristic.

lies critically on a good choice of kernel and our framework yields a new method, termed Bayesian Kernel Learning, to inform this choice. We only explored learning the length-scale of the squared exponential kernel, but our method extends to the case of richer kernels with more hyperparameters. We conceive of Bayesian Kernel Learning as a drop-in replacement for selecting the kernel hyperparameters in settings where cross-validation is unavailable. A sampling-based Bayesian approach is also demonstrated, enabling integration over kernel hyperparameters, and e.g., obtaining the full posterior distribution over the witness function in two-sample testing.

While our method is designed for unsupervised settings, there are various reasons it might be helpful in supervised settings or in applied Bayesian modelling more generally. With the rise of large-scale kernel methods, it has become possible to apply, e.g. SVMs or GPs to very large datasets. But even with efficient methods, it can be very costly to run cross-validation over a large space of hyperparameters. In practice, when, e.g. large scale approximations based on random Fourier features (Rahimi and Recht, 2007) are used, we have not seen much attention paid to kernel learning – the features are often just one part of a complicated pipeline, so again the median heuristic is often employed. For these reasons, we think that the developed method for Bayesian Kernel Learning would be a judicious alternative. Moreover, it would be straightforward to develop scalable approximate versions of Bayesian Kernel Learning itself.

7 Acknowledgments

SRF was supported by the ERC (FP7/617071) and EPSRC (EP/K009362/1). Thanks to Wittawat Jitkrittum, Krikamol Muandet, Sayan Mukherjee, Jonas Peters, Aaditya Ramdas, Alex Smola, and Yee Whye Teh for helpful discussions.

References

Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

Adi Ben-Israel. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

Bryan J. Bloomer, Jeffrey W. Stehr, Charles A. Piety, Ross J. Salawitch, and Russell R. Dickerson. Observed relationships of ozone air pollution with temperature and emissions. *Geophysical Research Letters*, 36(9), 2009. ISSN 1944-8007. L09803.

Adrian W Bowman. A comparative study of some kernel-based nonparametric density estimators. *Journal of Statistical Computation and Simulation*, 21(3-4):313–327, 1985.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1166–1174, 2013.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Seth R Flaxman, Daniel B Neill, and Alexander J Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2015.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12: 2211–2268, 2011.

A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schoelkopf, and A. Smola. A kernel statistical test of independence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2008. MIT Press.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012a.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012b.

Milan N. Lukić and Jay H. Beder. Stochastic Processes with Sample Paths in Reproducing Kernel Hilbert Spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, pages 1–96, 2015.

K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel Mean Shrinkage Estimators. *Journal of Machine Learning Research (forthcoming)*, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Kernel mean estimation and Stein's effect. *arXiv preprint arXiv:1306.0842*, 2013.

Krikamol Muandet, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean estimation via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Natesh S Pillai, Qiang Wu, Feng Liang, Sayan Mukherjee, and Robert L Wolpert. Characterizing the function space for bayesian kernel models. *Journal of Machine Learning Research*, 8(8), 2007.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184, 2007.

Aaditya Ramdas and Leila Wehbe. Nonparametric independence testing for small sample sizes. *24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, 2006.

Sashank J Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *AISTATS*, 2015.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 (3):832–837, 1956.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization and beyond*. the MIT Press, 2002.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Sanford Sillman. The relation between ozone, no x and hydrocarbons in urban and polluted rural environments. *Atmospheric Environment*, 33(12):1821–1845, 1999.

R. Silverman. Locally stationary random processes. *IRE Transactions on Information Theory*, 3(3):182–187, September 1957.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264, 2006.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics*

and computing, 17(4):395–416, 2007.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

Andrew G Wilson and Ryan P Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1067–1075, 2013.

A Some derivations for Bayesian Kernel Embedding

A.1 Notation

Consider a dataset $x_1, \ldots, x_n \in \mathbb{R}^D$ and suppose that there exists some unknown probability distribution P for which the x_i are i.i.d.:

$$x_i \sim \mathsf{P} \ . \tag{18}$$

Denote by μ_θ the RKHS mean embedding element for a given kernel $k_\theta(\cdot,\cdot)$ with hyperparameter $\theta \in \mathbb{R}^Q$ and by $\widehat{\mu_\theta}(\cdot)$ the empirical mean embedding

$$\widehat{\mu_\theta}(\cdot) := \frac{1}{n}\sum_{i=1}^{n} k_\theta(x_i, \cdot) \ . \tag{19}$$

We posit as our model that μ_θ has a GP prior with covariance r_θ, where

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

where ν is a finite measure on $\mathbb{R}^D$ thus ensuring that $\mu_\theta \in \mathcal{H}_{k_\theta}$ when drawn from the prior

$$\mu_\theta | \theta \sim \mathcal{GP}(0, r_\theta(\cdot,\cdot)) \ . \tag{20}$$

In addition, we model the link between the population mean embedding and the empirical mean embedding functions at a given location x as follows

$$p(\widehat{\mu_\theta}(x) | \mu_\theta(x)) = \mathcal{N}(\widehat{\mu_\theta}(x); \mu_\theta(x), \tau^2/n) \tag{21}$$

where τ^2 is another hyperparameter.

A.2 Priors over RKHS

The results in this section have appeared in the literature before, but as they are not well known or collected in one place, we have included them for completeness. A similar discussion appears in Pillai et al. (2007), but without the construction of explicit GP priors over the RKHSs which we provide below.

It is well known that the sample paths of a GP with kernel k are almost surely outside RKHS $\mathcal{H}_k$, the result known as Kallianpur's 0-1 law (**?**Wahba, 1990). It is easiest to demonstrate this by considering a Mercer's expansion Rasmussen and Williams (2006, Section 4.3) of kernel k given by

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \tag{22}$$

for the eigenvalue-eigenfunction pairs $\{(\lambda_i, e_i)\}_{i=1}^n$. Then, a representation of $f \sim \mathcal{GP}(0, k)$ is given by $f = \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i e_i$, where $\{Z_i\}_{i=1}^{\infty}$ are independent and identically distributed standard normal random variables. However,

$$\|f\|_{\mathcal{H}_k}^2 = \sum_{i=1}^{\infty} \frac{\lambda_i Z_i^2}{\lambda_i} = \sum_{i=1}^{\infty} Z_i^2 = \infty, \quad a.s. \tag{23}$$

so $f \notin \mathcal{H}_k$ almost surely. This issue is often sidelined in the literature, cf. e.g. (Rasmussen and Williams, 2006, Section 6.1) – in GP regression, it is not necessary to ensure that the prior on the regression function is supported on $\mathcal{H}_k$ (the posterior mean will still lie in $\mathcal{H}_k$, however). However, since the object of our interest, kernel embedding, is by construction an element of $\mathcal{H}_k$ - we opt for an approach where the prior is indeed specified over the correct space. Fortunately, it is straightforward to construct a kernel r such that the realizations from a GP with kernel r are almost surely inside RKHS $\mathcal{H}_k$. For this, we will need notions of dominance and nuclear dominance for kernel functions.

Definition 1. *Kernel k is said to dominate kernel r (written $k \succ r$) if $\mathcal{H}_r \subseteq \mathcal{H}_k$.*

Lukić and Beder (2001, Theorem 1.1) characterise dominance $k \succ r$ via the existence of a certain positive, continuous and self-adjoint operator $L : \mathcal{H}_k \to \mathcal{H}_k$ for which

$$r(x, x') = \langle L[k(\cdot, x)], k(\cdot, x') \rangle_{\mathcal{H}_k}, \qquad \forall x, x' \in \mathcal{X}. \tag{24}$$

When L is also a trace class operator, dominance is termed *nuclear*, and denoted $k \succ\succ r$. The following theorem from Lukić and Beder (2001, Theorem 7.2) then fully characterises kernels that lead to valid GP priors over RKHS $\mathcal{H}_k$.

Theorem 1. *Let $\mathcal{H}_k$ be separable and let $m \in \mathcal{H}_k$. Then $\mathcal{GP}(0, r(\cdot,\cdot))$ has trajectories in $\mathcal{H}_k$ with probability 1 if and only if $k \succ\succ r$.*

Thus, we just need to specify a trace-class, positive, continuous and self-adjoint operator $L : \mathcal{H}_k \to \mathcal{H}_k$ and compute $\langle L[k(\cdot,x)], k(\cdot,x')\rangle_{\mathcal{H}_k}$. A convenient choice for a given bounded continuous kernel k can be defined as follows. Take the convolution operator $S_k : L^2(\mathcal{X};\nu) \to \mathcal{H}_k$ with respect to a finite measure ν, defined as

$$[S_k f](x) = \int f(u)k(x,u)\nu(du). \tag{25}$$

It is well known that the adjoint of S_k is the inclusion of $\mathcal{H}_k$ into L^2 (Steinwart and Christmann, 2008, Section 4.3). Thus, we let $L = S_k S_k^*$, which is the (uncentred) covariance operator $L = \int k(\cdot,u) \otimes k(\cdot,u)\nu(du)$ of ν. As a covariance operator, L is then positive, continuous and self-adjoint. It is also trace-class in most cases of interest – and in particular whenever $\int k(u,u)\nu(du) < \infty$ (Steinwart and Christmann, 2008, Theorem 4.27), and thus for every stationary kernel provided that ν is a finite measure. This leads to

$$\begin{aligned} r(x,x') &= \langle S_k S_k^*[k(\cdot,x)], k(\cdot,x')\rangle_{\mathcal{H}_k} \\ &= \langle S_k^*[k(\cdot,x)], S_k^* k(\cdot,x')\rangle_{L^2(\mathcal{X};\nu)} \\ &= \int k(x,u)k(u,x')\nu(du), \end{aligned}$$

so r can be simply computed as a convolution of k with itself, and we can use $\mathcal{GP}(0, r(\cdot,\cdot))$ as a prior over $\mathcal{H}_k$.

A.3 Covariance function r_θ

In this subsection, we derive the covariance function r_θ for squared exponential kernels. Consider a squared exponential kernel on $\mathcal{X} = \mathbb{R}^D$ with full covariance matrix Σ_θ defined by

$$k_\theta(x,y) = \exp\left(-\frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y)\right), \quad x,y \in \mathbb{R}^D. \tag{26}$$

While we have required in A.2 that ν is a finite measure for the covariance operator to be trace class when working with stationary kernels, let us for simplicity first consider the instructive case when ν is the Lebesgue measure. Then, we have

$$\begin{aligned} r_\theta(x,y) &= \int k_\theta(x,u)k_\theta(u,y)du \\ &= \int \exp\left(-\frac{1}{2}\left((x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u)\right)\right)du \end{aligned}$$

Note that

$$(x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u) = 2\left(u - \frac{x+y}{2}\right)^T\Sigma_\theta^{-1}\left(u - \frac{x+y}{2}\right) + \frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y)\ .$$

Then

$$\begin{aligned} r_\theta(x,y) &= \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right)\int \exp\left(-\frac{1}{2}\left(u - \frac{x+y}{2}\right)^T\left(\frac{1}{2}\Sigma_\theta\right)^{-1}\left(u - \frac{x+y}{2}\right)\right)du \\ &= \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right) \times (2\pi)^{D/2}|\Sigma_\theta/2|^{1/2} \\ &= \pi^{D/2}\ |\Sigma_\theta|^{1/2}\ \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right)\ . \end{aligned}$$

Thus r_θ is proportional to another squared exponential kernel with covariance $2\Sigma_\theta$. For the special case where the covariance matrix Σ_θ is diagonal – let $\Sigma_\theta = \theta I_D$ and $\theta = (\theta^{(1)}, \ldots, \theta^{(D)})^T$– we have

$$r_\theta(x,y) = \pi^{D/2}\ \left(\prod_{d=1}^D \theta^{(d)}\right)^{1/2}\ \exp\left(-\frac{1}{2}(x-y)^T(2\theta I_D)^{-1}(x-y)\right). \tag{27}$$

Now, take $\nu(du) = \exp\left(-\frac{\|u\|_2^2}{2\eta^2}\right) du$, i.e., ν is a finite measure and is proportional to a Gaussian measure on $\mathbb{R}^d$. In that case, we have

$$\begin{aligned} r_\theta(x,y) &= \int k_\theta(x,u)k_\theta(u,y)\nu(du) \\ &= \int \exp\left(-\frac{1}{2}\underbrace{\left((x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u) + \eta^{-2}u^\top u\right)}_{\mathsf{A}}\right) du. \end{aligned}$$

From standard Gaussian integration rules, it follows that

$$\mathsf{A} = \frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y) + (u-\mathsf{m})^\top \mathsf{S}^{-1}(u-\mathsf{m}) + \left(\frac{x+y}{2}\right)^\top \left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)$$

where $\mathsf{m} = \mathsf{S}^{-1}\Sigma_\theta^{-1}(x+y)$ and $\mathsf{S} = (2\Sigma_\theta^{-1} + \eta^{-2}I_D)^{-1}$. Therefore

$$\begin{aligned} r_\theta(x,y) &= (2\pi)^{D/2}|\mathsf{S}|^{1/2}\exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y) - \frac{1}{2}\left(\frac{x+y}{2}\right)^\top\left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)\right) \\ &= (2\pi)^{D/2}\left|2\Sigma_\theta^{-1} + \eta^{-2}I_D\right|^{-1/2}\exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right) \\ &\qquad \times \exp\left(-\frac{1}{2}\left(\frac{x+y}{2}\right)^\top\left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)\right). \end{aligned}$$

Thus, we see that r_θ has a nonstationary component that penalises the norm of $\left(\frac{x+y}{2}\right)$. This is reminiscent of the well known locally stationary covariance functions (Silverman, 1957). However, for large values of η, the nonstationary component becomes negligible and r_θ reverts to being proportional to a standard squared exponential kernel with covariance $2\Sigma_\theta$, just like in the case of Lebesgue measure. We note that any choice of $\eta > 0$ gives a valid prior over $\mathcal{H}_k$. Treating η as another hyperparameter to be learned would be an interesting direction for future research.

A.4 Fast computation of the marginal pseudolikelihood

The marginal pseudolikelihood in Eq. (15) requires computation of the likelihood for an mn-dimensional normal distribution

$$\mathcal{N}\left(\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}; \mathbf{0},\ \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right).$$

However, the Kronecker product structure in the covariance matrix $C = \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}$ allows efficient computation. We denote with $R_{\theta,\mathbf{zz}} = Q\Lambda Q^\top$ the eigendecomposition of the matrix $R_{\theta,\mathbf{zz}}$ with $\Lambda = \text{diag}\left[\lambda_1, \ldots, \lambda_m\right]$. Note that $\mathbf{1}_n\mathbf{1}_n^\top$ is a rank-one matrix with the eigenvalue equal to n. Therefore C has top m eigenvalues equal to $n\lambda_i + \tau^2$, $i = 1, \ldots, m$, and the remaining $n(m-1)$ all equal to τ^2. Thus, the log-determinant is simply

$$\log\det C = \sum_{i=1}^m \log(n\lambda_i + \tau^2) + m(n-1)\log\tau^2 = \log\det\left[R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right] + m\log n + m(n-1)\log\tau^2. \quad (28)$$

Further, we need to compute $\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}^\top C^{-1}\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}$. By completing $b_1 = n^{-1/2}\mathbf{1}_n$ to an orthonormal basis $\{b_1, \ldots, b_n\}$ of $\mathbb{R}^n$ and forming the corresponding matrix $B = [b_1 \cdots b_n]$, and denoting by n an $n \times n$ matrix with $\mathsf{n}_{11} = n$ and $\mathsf{n}_{ij} = 0$ elsewhere, we have that

$$C^{-1} = (B \otimes Q)(\mathsf{n} \otimes \Lambda + \tau^2 I_{nm})^{-1}(B \otimes Q)^\top. \quad (29)$$

We now simply need to apply Kronecker identity $(B^\top \otimes Q^\top)\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\} = \text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}} B\right\}$, to obtain

$$\begin{aligned} &\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}^\top C^{-1}\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\} = \text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}}B\right\}^\top (\mathsf{n} \otimes \Lambda + \tau^2 I_{nm})^{-1}\text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}}B\right\} \\ &= \sum_{j=1}^m \frac{n^{-1}\left[Q^\top K_{\theta,\mathbf{zx}}\mathbf{1}_n\right]_j^2}{n\lambda_j + \tau^2} + \frac{1}{\tau^2}\sum_{i=2}^n\sum_{j=1}^m \left[Q^\top K_{\theta,\mathbf{zx}} b_i\right]_j^2. \end{aligned} \quad (30)$$

For the first term, we have

$$\sum_{j=1}^{m} \frac{n^{-1}\left[Q^\top K_{\theta,\mathbf{zx}}\mathbf{1}_n\right]_j^2}{n\lambda_j+\tau^2} = \sum_{j=1}^{m} \frac{\left[Q^\top \hat{\mu}(\mathbf{z})\right]_j^2}{\lambda_j+\tau^2/n} = \sum_{j=1}^{m} \frac{\mathrm{Tr}\left[\hat{\mu}(\mathbf{z})\hat{\mu}(\mathbf{z})^\top q_j q_j^\top\right]}{\lambda_j+\tau^2/n}$$
$$= \hat{\mu}(\mathbf{z})^\top \left(R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right)^{-1} \hat{\mu}(\mathbf{z}). \tag{31}$$

And for the second term:

$$\begin{aligned}
\frac{1}{\tau^2}\sum_{i=2}^{n}\sum_{j=1}^{m}\left[Q^\top K_{\theta,\mathbf{zx}} b_i\right]_j^2 &= \frac{1}{\tau^2}\sum_{j=1}^{m}\sum_{i=2}^{n}\left[q_j^\top K_{\theta,\mathbf{zx}} b_i\right]^2 \\
&= \frac{1}{\tau^2}\sum_{j=1}^{m}\left\{\left\|K_{\theta,\mathbf{xz}} q_j\right\|^2 - n\left(q_j^\top \hat{\mu}(\mathbf{z})\right)^2\right\} \\
&= \frac{1}{\tau^2}\left\|K_{\theta,\mathbf{xz}}\right\|_F^2 - \frac{n}{\tau^2}\left\|\hat{\mu}\left(\mathbf{z}\right)\right\|^2 .
\end{aligned} \tag{32}$$

Altogether, the log-likehood is given by

$$\begin{aligned}
\log\left\{\mathcal{N}\left(\mathrm{vec}\left\{K_{\theta,\mathbf{zx}}\right\}; \mathbf{0},\ \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right)\right\} &= -\frac{1}{2}\Bigg\{\log\det\left[R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right] \\
&\quad + \hat{\mu}(\mathbf{z})^\top\left(R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right)^{-1}\hat{\mu}(\mathbf{z}) \\
&\quad + \frac{1}{\tau^2}\left\|K_{\theta,\mathbf{xz}}\right\|_F^2 - \frac{n}{\tau^2}\left\|\hat{\mu}\left(\mathbf{z}\right)\right\|^2 \\
&\quad + m\log n + m(n-1)\log\tau^2 + mn\log(2\pi)\Bigg\}.
\end{aligned} \tag{33}$$

B Source for Stan model

```
functions {
  // phi should be m x n
  real kron_multi_normal(matrix K,matrix R,matrix Q1,vector e1,int m,int n,real sigma2) {
    vector[m*n] e;
    matrix[m,m] Q2;
    vector[m] e2;
    vector[m] ones;
    vector[m*n] mv2;
    real mvp;
    real logdet;
    Q2 <- eigenvectors_sym(R);
    e2 <- eigenvalues_sym(R);
    for(j in 1:m) {
      ones[j] <- 1;
      for(i in 1:n)
        e[(j-1)*n + i] <- 1/(e1[i] * e2[j] + sigma2);
    }
    mv2 <- to_vector((transpose(Q2) * transpose(K)) * Q1);
    mvp <- sum(mv2 .* e .* mv2);
    logdet <- sum(log(e2 .* (ones * n) + ones * sigma2)) + m * (n-1) * log(sigma2);

    return( - .5 * logdet - .5 * mvp);
  }
}

data {
  int<lower=1> n;
  int<lower=1> m;
  vector[n] x;
```

```
  vector[m] u;
}

transformed data {
  matrix[n,m] xu_dist2;
  matrix[m,m] u_dist2;
  matrix[n,n] ones;
  vector[n] zeros;
  matrix[n,n] Q1;
  vector[n] e1;

  for (i in 1:n) {
    zeros[i] <- 0;
    e1[i] <- 0;
    for (j in 1:n)
      ones[i,j] <- 1;
    for(j in 1:m)
      xu_dist2[i, j] <- square(x[i] - u[j]);
  }
  for(i in 1:m) {
    for(j in 1:m)
      u_dist2[i,j] <- square(u[i] - u[j]);
  }
  e1[1] <- n;
  Q1 <- eigenvectors_sym(ones);
}

parameters {
  real<lower=0> lengthscale;
  real<lower=0> sigma2;
}
transformed parameters {
  matrix[m,m] R;
  matrix[n,m] J;
  matrix[n,m] K;

// R <- lengthscale * sqrt(pi()) *
  R <- exp(- u_dist2/(4*lengthscale^2));
  K <- exp(- xu_dist2/(2*lengthscale^2));
  J <- K .* K .* xu_dist2 / lengthscale^4;
}

model {
  for(i in 1:n) // Jacobian
    increment_log_prob(log(.5 * sum(J[i])));

  increment_log_prob(kron_multi_normal(K, R, Q1, e1, m, n, sigma2));
  lengthscale ~ gamma(1,1);
  sigma2 ~ gamma(1,1);
}
```