Subspace-Induced Gaussian Processes

Zilong Tan
Department of Computer Science
Duke University
Email: ztan@cs.duke.edu

Sayan Mukherjee
Departments of Statistical Science
Computer Science, Mathematics,
Biostatistics & Bioinformatics
Duke University
Email: sayan@stat.duke.edu

February 22, 2018

Abstract

We present a new Gaussian process (GP) regression model where the covariance kernel is indexed or parameterized by a sufficient dimension reduction subspace of a reproducing kernel Hilbert space. The covariance kernel will be low-rank while capturing the statistical dependency of the response to the covariates, this affords significant improvement in computational efficiency as well as potential reduction in the variance of predictions. We develop a fast Expectation-Maximization algorithm for estimating the parameters of the subspace-induced Gaussian process (SIGP). Extensive results on real data show that SIGP can outperform the standard full GP even with a low rank-m, $m \leq 3$, inducing subspace.

1 Introduction

Gaussian processes (GPs) [Doob, 1944] have been used extensively for non-parametric regression and density estimation in the statistics and machine learning literature [Stein, 1999, Rasmussen, 2006]. The key utility of models based on GPs is that nonlinear interpolation and spatial structure can be naturally modeled. The covariance kernel of the GP is the crucial component that specifies the structure of a GP by characterizing the class of random nonlinearities specified by the Gaussian process. Another perspective on GP models is that the covariance kernel corresponds to a feature space and this feature space encodes the class of nonlinear functions the GP can realize. Typically, the covariance kernel is a positive definite function defined on the input space where for two points $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$ $\operatorname{Cov}\left[f\left(\boldsymbol{x}\right), f\left(\boldsymbol{z}\right)\right] = \kappa\left(\boldsymbol{x}, \boldsymbol{z}\right)$ and κ is positive definite.

In this paper, we present a new GP regression model, referred to as the Subspace-Induced Gaussian Process (SIGP), where the covariance kernel is not specified as a function of two input points, but is specified by function values at the two input points, $\kappa\left(\boldsymbol{x}, \boldsymbol{z}\right) = \operatorname{Cov}[\psi\left(\boldsymbol{x}\right), \psi\left(\boldsymbol{z}\right)]$. The stochastic model for the random function ψ will be specified by a distribution over functions in a reproducing kernel Hilbert space (RKHS). SIGP can be viewed as the dual of the standard GP, and enjoys computational advantages due to the low-rank nature of the covariance kernel. In addition, the SIGP has performance advantages over standard GP regression models as there is a natural way of learning the hyperparameters of the kernel function. We will demonstrate using several real datasets that SIGP induced by a low-rank RKHS performs competitively in prediction compared to state-of-the-art GP inference methods as well as support vector machines (SVMs).

The proposed SIGP model is motivated from an approximation theory perspective of RKHS [Paulsen and Raghupathi, 1999]. The target function f will be approximated by a random function drawn from a distribution induced by the RKHS of the observed data, see [Pillai et al., 2007] for details on distributions over kernel models and RKHS. If we represent the random functions as a linear combination of the feature maps, $f(\boldsymbol{x}) = \sum_{i=1}^{n} c_i \phi(\boldsymbol{x}_i)$, then specifying a multivariate normal distribution over the coefficients $\boldsymbol{c}$ induces a distribution over the RKHS. The mean vector of the multivariate normal determines the coefficients of $\mathbb{E}(f)$ which has a natural interpretation due to the reproducing property of the RKHS. Thus, SIGP is fundamentally different from the standard GP that associates Gaussian random variables to the data in the input space $\mathcal{X}$ (see e.g., [Williams, 1997, Rasmussen, 2006]). While the covariance of GP is determined by the input data, the covariance of SIGP depends on the function values at the input data. As an advantage, SIGP inherently has a low-rank covariance structure due to the nearly exponential eigenvalue decay of the kernel operator in the RKHS [Belkin, 2018].

This inherent low-rank property of the SIGP covariance admits both efficient parameter estimation (§ 4) for the GP as well as effective parameterization of the covariance kernel. We will specify the covariance structure of the SIGP on a sufficient dimension reduction (SDR) subspace [Li, 1991, Adragni and Cook, 2009, Wu et al., 2013] of the full RKHS induced by data. The SDR for regression is originally defined in the input space, and aims to find an m-dimensional subspace $\mathcal{B}$ such that the projection of covariate vector $\boldsymbol{x} \in \mathbb{R}^p$, $p \geq m$, onto $\mathcal{B}$ captures the statistical dependency of the response y on $\boldsymbol{x}$, $Y \perp\!\!\!\perp \boldsymbol{X} \mid \mathcal{B}^\top \boldsymbol{X}$. We consider the SDR framework for functions in an RKHS to obtain a rank-m subspace of the full RKHS which preserves the information in the RKHS relevant to predicting y. The SIGP will specify a distribution over functions in the subspace.

The hyperparameters of the SIGP covariance consists of the SDR subspace represented in terms of the induced feature maps as well as the covariance of the multivariate normal. For simplicity, we assume that the kernel matrix of the RKHS is given; however, inferring the parameters of the kernel is also feasible as in the standard GP setting. To estimate the hyperparameters of the covariance, we first derive a computationally-efficient likelihood specification over SDR subspaces based on a result from [Cook and Forzani, 2009]. Then, an Expectation-Maximization (EM) algorithm is developed to estimate the covariance structure. These algorithms leverage the low-rank property of the SIGP covariance induced by a rank-m RKHS, and has computational complexity $O\left(n^2 m\right)$. Our approach differs from the sparse inference methods which use a subset of the training points [Smola and Bartlett, 2001, Seeger et al., 2003, Snelson and Ghahramani, 2006a, Hensman et al., 2013, 2015] to enable faster computation. The construction of the SIGP also can use sparse inference methods to improve computational complexity to $O\left(s^2 m\right)$ with s pseudo-inputs. We leave this as future work.

Contribution The main contribution of this work is a new GP regression model, SIGP, where the target function f is randomly drawn from a distribution over functions in an RKHS that is data dependent, and f is a linear combination of feature maps on the data with the coefficients of the linear combination specified by a multivariate normal. We show the computational advantage of SIGP that arises from leveraging the low-rank nature of the covariance. Fast parameter estimation algorithms for SIGP using the likelihood over SDR subspaces are also developed. Finally, we report competitive performance of SIGP on several real datasets.

Notation We use bold lowercase letters to represent vectors, bold capital letters for matrices, calligraphic letters to refer to sets. The i-th row and j-th column of a matrix $\boldsymbol{M}$ are denoted by $\boldsymbol{M}_{i:}$ and $\boldsymbol{M}_j$, respectively. In addition, we write $\boldsymbol{M}_\perp$ for the orthogonal complement to $\boldsymbol{M}$, and $\det(\cdot)$ for the matrix determinant. The (column) vector of the diagonal entries of $\boldsymbol{M}$ is written as $\operatorname{diag}(\boldsymbol{M})$. $\boldsymbol{I}_n$ and $\mathbf{1}_n$ represent respectively the n-by-n identity matrix and a (column) n-dimensional vector of all ones. Moreover, we follow the conventional notation $\widehat{\boldsymbol{a}}$ to denote an estimate of $\boldsymbol{a}$.

Organization The remainder of the paper is organized as follows. Section 2 reviews the function-space view of GP, and introduces the SIGP. Section 3 describes how to base the SIGP on a subspace of the full RKHS, and derives the likelihood function for estimating the SDR subspace. Parameter estimation algorithms are developed in § 4. Section 5 provides the predictive distribution of SIGP as well as the computational techniques for efficient implementation. Experiment results are reported in § 6, and § 7 concludes this paper.

2 Subspace-Induced Gaussian Processes

We first review GPs from a function-space perspective and then provide a dual view which motivates the Subspace-Induced Gaussian process (SIGP).

In a reproducing kernel Hilbert space (RKHS), each data point $\boldsymbol{x}$ can be associated with a feature map $\phi(\boldsymbol{x})$. Here, $\phi(\boldsymbol{x}_j)$ can be either finite- or infinite-dimensional, and denote by $\phi_i(\boldsymbol{x}_j)$ the i-th coordinate of the feature map associated with $\boldsymbol{x}_j$. At any point $\boldsymbol{x}_i$, the standard GP formulation assumes that the observed deviation $\epsilon_i \coloneqq f(\boldsymbol{x}_i) - \mathbb{E}(f(\boldsymbol{x}_i))$ is a linear combination $\epsilon_i = \sum_j \alpha_j \phi_j(\boldsymbol{x}_i)$, with $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$ (see e.g., [Williams, 1997]). Thus, the standard GP has

$$\mathbb{E}_{\boldsymbol{\alpha}}(\epsilon_i) = 0 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\alpha}}(\epsilon_i \epsilon_j) = \phi(\boldsymbol{x}_i)^\top \boldsymbol{\Sigma}_\alpha \phi(\boldsymbol{x}_j).$$

The SIGP is motivated by a dual view from function approximation in RKHS. The target function f is approximated with a linear combination $\sum_i b_i \phi(\boldsymbol{x}_i)$and the evaluation of f in the RKHS at $\boldsymbol{x}$ is given by the inner product $\langle f, \phi(\boldsymbol{x}) \rangle$. Unlike the GP formulation, which imposes a Gaussian random coefficient to each entry of the feature map, one can consider that f is randomly drawn from a distribution over the functions in the RKHS. The coefficients of f in terms of the feature maps have a joint multivariate normal distribution. Specifically, we have $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{\Sigma}_\beta)$, and the observed deviation is given by $\epsilon_i = f(\boldsymbol{x}_i) - \mathbb{E}(f(\boldsymbol{x}_i)) = \sum_{j=1}^n \overline{\beta}_j \langle \phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}_i) \rangle$, where $\overline{\boldsymbol{\beta}}$ is the centered $\boldsymbol{\beta}$, and $\mathbb{E}(f(\boldsymbol{x}_i)) = \sum_{j=1}^n b_j \langle \phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}_i) \rangle$ assuming a zero intercept without loss of generality. The distribution of ϵ_i is normal with $\mathbb{E}_{\boldsymbol{\beta}}(\epsilon_i) = 0$ and $\mathbb{E}_{\boldsymbol{\beta}}(\epsilon_i \epsilon_j) = (\boldsymbol{K}\boldsymbol{\Sigma}_\beta \boldsymbol{K})_{ij}$, where $K_{ij} = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$ is the kernel matrix. It should be noted, the covariance between two points under the above assumption also depends on the evaluation of the feature maps at the points.

One advantage of this formulation is that $\boldsymbol{\epsilon}$ in SIGPs can be well approximated in terms of a low-rank subspace of the RKHS due to the nearly exponential decay of eigenvalues for certain kernels, including the popular radial kernels [Belkin, 2018]. Therefore, the covariance of SIGP is inherently low-rank, and admits fast parameter inference. In particular, we consider writing $\boldsymbol{\epsilon}$ in terms of an m-rank subspace of the RKHS spanned by the basis

$$\left[\sum_i W_{i1} \phi(\boldsymbol{x}_i) \quad \sum_i W_{i2} \phi(\boldsymbol{x}_i) \quad \cdots \quad \sum_i W_{im} \phi(\boldsymbol{x}_i) \right].$$

Here, $\boldsymbol{W}$ is the coefficient matrix of the basis functions. Note that the subspace is still an RKHS with the same reproducing kernel $k(\boldsymbol{x}, \boldsymbol{z}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z}) \rangle$ for all input data points $\boldsymbol{x}$ and $\boldsymbol{z}$. Ideally, $\boldsymbol{W}$ should be chosen such that the projection of $\phi(\boldsymbol{x}_i)$ onto the subspace captures the statistical dependency of $\boldsymbol{\epsilon}$ on $\phi(\boldsymbol{x}_i)$ for all i. The desired subspace is a sufficient dimension reduction (SDR) subspace [Li, 1991, Adragni and Cook, 2009]. A full discussion of $\boldsymbol{W}$ is given in the next subsection. The resulting distribution of the SIGP is as follows:

$$\boldsymbol{\epsilon} = \boldsymbol{K}\boldsymbol{W}\overline{\boldsymbol{\beta}} \quad \text{and} \quad \operatorname{var}(\boldsymbol{\epsilon}) = \boldsymbol{K}\boldsymbol{W}\boldsymbol{\Sigma}_\beta \boldsymbol{W}^\top \boldsymbol{K}. \tag{1}$$

Written in standard GP notation,

$$f(\cdot) \sim \mathcal{GP}\left(\boldsymbol{K}(\cdot, \boldsymbol{X})\boldsymbol{b}, \boldsymbol{K}(\cdot, \boldsymbol{X})\boldsymbol{W}\boldsymbol{\Sigma}_\beta \boldsymbol{W}^\top \boldsymbol{K}(\boldsymbol{X}, \cdot)\right).$$

Here, we define $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{Z})_{ij} \coloneqq \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{z}_j) \rangle$. Like the GP, the SIGP can be used as a prior for Bayesian inference, and does not depend on the training data [Rasmussen, 2004]. However, the mean function and covariance function can be chosen in light of the training data. The covariance of SIGP has two free parameters, $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_\beta$. The selection of $\boldsymbol{W}$ is based on a likelihood function over SDR subspaces, discussed in the next section. The parameter $\boldsymbol{\Sigma}_\beta$ is chosen via an EM algorithm given in § 4.

3 Selecting the Subspace

The key observation is that the distribution of the SIGP satisfies the dimension reduction subspace assumption, i.e., the conditional distribution $\boldsymbol{x} \mid y$ is normal (Theorem 1 of [Cook and Forzani, 2009]). It follows that the log-likelihood for an SDR subspace spanned by the columns of $\boldsymbol{B}$ in $\mathbb{R}^p$ can be expressed as [Cook and Forzani, 2009]:

$$l(\boldsymbol{B}) \propto -\frac{1}{2}\sum_y n_y \log\det\left(\boldsymbol{B}^\top \operatorname{var}(\boldsymbol{x} \mid y)\boldsymbol{B}\right) + \frac{n}{2}\log\det\left(\boldsymbol{B}^\top \operatorname{var}(\boldsymbol{x})\boldsymbol{B}\right), \tag{2}$$

where $\operatorname{var}(\boldsymbol{x})$ and $\operatorname{var}(\boldsymbol{x} \mid y)$ are the sample variances, and n_y denotes the number of samples with response value y [1]. Note that this log-likelihood is defined for subspaces in $\mathbb{R}^p$, not the RKHS. Additionally, the log-likelihood (2) can be computationally expensive due to the $O(n)$ non-convex $\log\det(\cdot)$ terms. To overcome both limitations, we propose an equivalent, but more practical alternative.

The subspace likelihood We propose the following equivalent (up to a constant) log-likelihood over SDR subspaces spanned by the columns of $\boldsymbol{B}$, abbreviated as $g(\boldsymbol{B})$:

$$g(\boldsymbol{B}) \coloneqq -\frac{n}{2}\log\frac{\det\left[\boldsymbol{B}^\top \mathbb{E}(\operatorname{var}(\boldsymbol{x} \mid y))\boldsymbol{B}\right]}{\det\left(\boldsymbol{B}^\top \operatorname{var}(\boldsymbol{x})\boldsymbol{B}\right)}. \tag{3}$$

The above log-likelihood $g(\boldsymbol{B})$ is rather cheap to compute. Both variance components need to be evaluated only once from data. In particular, a simple method as used in sliced inverse regression (SIR) [Li, 1991] to estimate $\mathbb{E}(\operatorname{var}(\boldsymbol{x} \mid y))$ is taking the weighted average of variances within $\boldsymbol{X}$ slices

[1] This can be easily generalized to a continuous response by slicing the range of the response, see e.g., [Li, 1991].

based on the range of y. Compared to (2), the proposed (3) has only two $\log\det(\cdot)$ terms, and has a closed-form maximizer, see Theorem 1 and Lemma 1. Thus, (3) is arguably more favorable in practice. Below, we first provide some intuitions behind (3), and then prove that it is equivalent to (2).

First, $g(\boldsymbol{B}) = g(\boldsymbol{B}')$ whenever $\boldsymbol{B}$ and $\boldsymbol{B}'$ are bases of the same subspace, i.e., $\boldsymbol{B} = \boldsymbol{B}'\boldsymbol{C}$ for some invertible $\boldsymbol{C} \in \mathbb{R}^{m\times m}$. Thus, the particular basis of the inducing subspace in SIGP is not important as long as it represents the same subspace. Second, when $\boldsymbol{B}$ is a vector, corresponding to the class of single index models, $g(\boldsymbol{B}) \propto -\log(1 - s(\boldsymbol{B}))$, where $s(\cdot)$ is the maximizing objective of sliced inverse regression (SIR) [Li, 1991], which coincides with the objective of Fisher Linear Discriminant Analysis (LDA) for two groups. This shows that maximizing the log-likelihood (3) yields consistent SDR subspace as given by SIR. Theorem 1 formalizes the consistency for a general $\boldsymbol{B}$.

Theorem 1. *The column space of a maximizer $\boldsymbol{B}_\star$ to $g(\boldsymbol{B})$ coincides with the SDR subspace given by SIR.*

Proof. The SIR solves for the e.d.r. directions $\boldsymbol{b}_i$ via the following generalized eigenvalue decomposition:

$$\operatorname{var}(\mathbb{E}(\boldsymbol{x} \mid y))\,\boldsymbol{b}_i = \lambda_i \operatorname{var}(\boldsymbol{x})\,\boldsymbol{b}_i.$$

From Eve's law, this is equivalent to

$$\begin{aligned} & [\operatorname{var}(\boldsymbol{x}) - \mathbb{E}(\operatorname{var}(\boldsymbol{x} \mid y))]\,\boldsymbol{b}_i = \lambda_i \operatorname{var}(\boldsymbol{x})\,\boldsymbol{b}_i \\ \iff \quad & (1-\lambda_i)\operatorname{var}(\boldsymbol{x})\,\boldsymbol{b}_i = \mathbb{E}[\operatorname{var}(\boldsymbol{y} \mid x)]\,\boldsymbol{b}_i \\ \iff \quad & [\mathbb{E}(\operatorname{var}(\boldsymbol{y} \mid x))]^{-1}\operatorname{var}(\boldsymbol{x})\,\boldsymbol{b}_i = \frac{1}{1-\lambda_i}\boldsymbol{b}_i. \end{aligned}$$

Note that $\lambda_i < 1$. Thus, the e.d.r. directions which are the $\boldsymbol{b}_i$ corresponding to the largest λ_i are also the leading eigenvectors of $[\mathbb{E}(\operatorname{var}(\boldsymbol{y} \mid x))]^{-1}\operatorname{var}(\boldsymbol{x})$. The fact that these leading eigenvectors form a basis of $\boldsymbol{B}_\star$ follows from Lemma 1. □

Lemma 1. *For positive definite matrices $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{n\times n}$, the column space of an optimal full-rank $\boldsymbol{B}_* \in \mathbb{R}^{n\times m}$, $m \le n$, for*

$$\min_{\boldsymbol{B}} \frac{\det\left(\boldsymbol{B}^\top\boldsymbol{M}\boldsymbol{B}\right)}{\det\left(\boldsymbol{B}^\top\boldsymbol{N}\boldsymbol{B}\right)} \tag{4}$$

coincides with the span of the d leading eigenvectors of $\boldsymbol{M}^{-1}\boldsymbol{N}$.

Proof. Since $\boldsymbol{M}$ is positive definite and hence invertible, we can let $\boldsymbol{B} = \boldsymbol{M}^{-1/2}\boldsymbol{T}$ and let $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{R}$ be the QR decomposition. $\boldsymbol{B}$ is full-rank, so is $\boldsymbol{R}$. The objective of (4) is can be written as

$$\frac{1}{\det\left(\boldsymbol{Q}^\top\boldsymbol{M}^{-1/2}\boldsymbol{N}\boldsymbol{M}^{-1/2}\boldsymbol{Q}\right)}$$

The minimum is attained by letting the columns of $\boldsymbol{Q}$ be the leading eigenvectors of $\boldsymbol{M}^{-1/2}\boldsymbol{N}\boldsymbol{M}^{-1/2}$. Now, the columns of $\boldsymbol{M}^{-1/2}\boldsymbol{Q}$ are the leading eigenvectors of $\boldsymbol{M}^{-1}\boldsymbol{N}$. Thus, $\boldsymbol{B} = \boldsymbol{M}^{-1/2}\boldsymbol{T} = \left(\boldsymbol{M}^{-1/2}\boldsymbol{Q}\right)\boldsymbol{R}$ has the same column space as the leading eigenspace of $\boldsymbol{M}^{-1}\boldsymbol{N}$. □

We will provide the log-likelihood g_H over SDR subspaces of an RKHS in (6). Since g_H is essentially the same as g except working with a basis of the RKHS, Theorem 1 applies to g_H.

Theorem 2 states that the proposed log-likelihood (3) is equivalent to (2).

Theorem 2. *Under the normality assumption of $\boldsymbol{x} \mid y$, the proposed SDR log-likelihood (3) is equivalent to (2) up to a constant.*

Proof. By the definition of an SDR subspace (Proposition 1(i) of [Cook and Forzani, 2009]), we have

$$\boldsymbol{B}_\perp^\top \mathrm{var}\left(\boldsymbol{x} \mid y\right)^{-1} = \boldsymbol{B}_\perp^\top \mathbb{E}^{-1}\left(\mathrm{var}\left(\boldsymbol{x} \mid y\right)\right). \tag{5}$$

Since $g\left(\boldsymbol{B}\right)$ is invariant to linear transformations of $\boldsymbol{B}$, we can assume that $\boldsymbol{B}$ is semi-orthogonal.

We will use the following result from [Rao, 1973]: Let $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be of rank n and let $\boldsymbol{B} \in \mathbb{R}^{p\times(p-n)}$ be of rank $p-n$ such that $\boldsymbol{A}^\top\boldsymbol{B} = \boldsymbol{0}$. Then

$$\boldsymbol{\Sigma} = \boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top + \boldsymbol{\Sigma}\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}.$$

Note that if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are semi-orthogonal, then due to the above result:

$$\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} - \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{B} = \left(\boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)^{-1}.$$

Additionally, $\det\left(\boldsymbol{\Sigma}\right)$ is equivalently written as

$$\begin{aligned}
&\det\left(\begin{bmatrix}\boldsymbol{B} & \boldsymbol{B}_\perp\end{bmatrix}^\top \boldsymbol{\Sigma} \begin{bmatrix}\boldsymbol{B} & \boldsymbol{B}_\perp\end{bmatrix}\right) \\
=& \det\left(\begin{bmatrix}\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} & \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp \\ \boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B} & \boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp\end{bmatrix}\right) \\
=& \det\left(\boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp\right)\det\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} - \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp\left(\boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp\right)^{-1}\boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B}\right).
\end{aligned}$$

We obtain $\det\left(\boldsymbol{B}_\perp^\top\boldsymbol{\Sigma}\boldsymbol{B}_\perp\right) = \det\left(\boldsymbol{\Sigma}\right)\det\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)$. Combining this result and (5) completes the proof. □

Extension to RKHS The definition of (3) can be extended to RKHS. The columns of $\boldsymbol{B}$ in (3) now represent functionals in the RKHS, forming the subspace $\mathcal{B}$ of the RKHS. Let $\boldsymbol{K}$ be the kernel matrix for the RKHS, and suppose that $\boldsymbol{B}_i = \sum_j \boldsymbol{W}_{ji}\phi\left(\boldsymbol{x}_j\right)$. Define $\phi \coloneqq \left(\phi\left(\boldsymbol{x}_1\right), \phi\left(\boldsymbol{x}_2\right), \cdots, \phi\left(\boldsymbol{x}_n\right)\right)$, we have

$$\boldsymbol{B} = \phi\boldsymbol{W} \quad \text{and} \quad \boldsymbol{K} = \phi^\top\phi.$$

Suppose that $\boldsymbol{X}$ and $\boldsymbol{y}$ are sorted by the value of y, the following holds

$$\mathrm{var}\left(\phi\left(\boldsymbol{x}\right)\right) = \frac{\phi\phi^\top}{n} - \phi\frac{\boldsymbol{1}_n\boldsymbol{1}_n^\top}{n^2}\phi^\top, \qquad \mathbb{E}\left(\mathrm{var}\left(\boldsymbol{x} \mid y\right)\right) = \phi\left[\frac{1}{n}\boldsymbol{I}_n - \frac{1}{n}\mathrm{diag}\left(\frac{\boldsymbol{1}_{n_h}\boldsymbol{1}_{n_h}^\top}{n_h}\right)\right]\phi^\top,$$

where n_h denotes the number of data points for class h in classification problems, or the size of the h-th slice based on the range of y in regression problems (see [Li, 1991]). Combined with (3), we arrive at the log-likelihood over SDR subspaces of the RKHS:

$$g_H(\boldsymbol{B}) = -\frac{n}{2}\log\frac{\det\left[\boldsymbol{W}^\top(\boldsymbol{A}+\eta\boldsymbol{I})\boldsymbol{W}\right]}{\det\left(\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}\right)}, \quad \text{with} \tag{6}$$

$$\boldsymbol{A} = \boldsymbol{K}\left[\boldsymbol{I}_n - \operatorname{diag}\left(\frac{\mathbf{1}_{n_h}\mathbf{1}_{n_h}}{n_h}\right)\right]\boldsymbol{K} \quad \text{and} \quad \boldsymbol{C} = \boldsymbol{K}\left(\boldsymbol{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)\boldsymbol{K}. \tag{7}$$

A small $\eta > 0$ is added in the numerator to enhance the numerical stability. One should distinguish (6) from (kernel) SIR in that: (i) g_H is the (unnormalized) log-likelihood over SDR subspaces $\boldsymbol{B}$, and (ii) both the numerator and denominator are determinants. In the above definition, $\boldsymbol{B}$ is parameterized by the coefficient matrix $\boldsymbol{W}$.

4 Learning an SIGP

Training a GP involves setting the mean and covariance given data. This is generally done by setting the hyperparameters to maximize the marginal likelihood [Rasmussen, 2004]. In this section, we develop an Expectation-Maximization (EM) algorithm for this task. As we discussed in § 3, the SIGP satisfies the SDR subspace assumption (see also the linear design condition in SIR [Li, 1991]). This implies that we can infer $\boldsymbol{\Sigma}_\beta$ and σ^2 from the data projection onto the SDR subspace, which is given by $\boldsymbol{K}\boldsymbol{W}$. The desired $\boldsymbol{W}$ is obtained by maximizing the RKHS SDR log-likelihood g_H in (6), and consists of the m leading eigenvectors of $\boldsymbol{A}^{-1}\boldsymbol{C}$ from Lemma 1.

Let $\mu(\boldsymbol{X})$ denote the mean function, SIGP models the observed response as $\boldsymbol{y} = \mu(\boldsymbol{X}) + \boldsymbol{K}\boldsymbol{W}\overline{\boldsymbol{\beta}} + \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}\right)$ is the noise variable with variance $\sigma^2\boldsymbol{I}$. Recall that $\overline{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ is an unobserved latent variable, it is natural to consider an EM algorithm for estimating $\boldsymbol{\Sigma}_\beta$ and σ^2. Denote by $\mathcal{P} := \left\{\sigma^2, \boldsymbol{\Sigma}_\beta\right\}$ the set of parameters, we can state the log-likelihood, dropping constants that are irrelevant, as

$$\log p\left(\boldsymbol{y}, \overline{\boldsymbol{\beta}} \,\middle|\, \boldsymbol{X}; \widehat{\mathcal{P}}\right) \propto -\frac{1}{2}\log\det\widehat{\boldsymbol{\Sigma}}_\beta - \frac{n}{2}\log\widehat{\sigma}^2 - \frac{1}{2}\overline{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}_\beta^{-1}\overline{\boldsymbol{\beta}} - \frac{\gamma^\top\gamma}{2\widehat{\sigma}^2}. \tag{8}$$

Denote by $\boldsymbol{G} := \boldsymbol{K}\boldsymbol{W}\boldsymbol{\Sigma}_\beta\boldsymbol{W}^\top\boldsymbol{K} + \sigma^2\boldsymbol{I}$ the marginal variance, the posterior $p\left(\overline{\boldsymbol{\beta}} \,\middle|\, \boldsymbol{y}; \widehat{\mathcal{P}}\right)$ is given by the following multivariate normal density

$$\mathcal{N}\left(\widehat{\boldsymbol{\Sigma}}_\beta\boldsymbol{W}^\top\boldsymbol{K}\widehat{\boldsymbol{G}}^{-1}(\boldsymbol{y} - \mu(\boldsymbol{X})), \widehat{\boldsymbol{\Sigma}}_\beta - \widehat{\boldsymbol{\Sigma}}_\beta\boldsymbol{W}^\top\boldsymbol{K}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\Sigma}}_\beta\right). \tag{9}$$

The conditional mean can be equivalently expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\sigma}^2\widehat{\boldsymbol{\Sigma}}_\beta^{-1} + \boldsymbol{W}^\top\boldsymbol{K}^2\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{K}(\boldsymbol{y} - \mu(\boldsymbol{X})). \tag{10}$$

Algorithm 1 EM algorithm for learning an SIGP

```
**Input:** data $\{\boldsymbol{X}, \boldsymbol{y}\}$, RKHS rank $m$, kernel matrix $\boldsymbol{K}$
**Output:** estimated covariance parameters $\widehat{\mathcal{P}}$, mean function parameter $\widehat{\boldsymbol{\alpha}}$
Initialize $\boldsymbol{W}$, $\widehat{\sigma}^2$
**repeat**
    Compute $\widehat{\boldsymbol{G}}^{-1}$ using (13)
    Estimate $\widehat{\boldsymbol{\alpha}}$ via (14) based on $\widehat{\boldsymbol{G}}^{-1}$ and data
    Compute $\boldsymbol{r}$ with the fitted mean function
    Obtain the estimators $\widehat{\beta}$, $\widehat{\boldsymbol{\Sigma}}_\beta$, and $\widehat{\sigma}^2$ via (10), (11), and (12), respectively
    Compute the log-likelihood (8)
**until** increase in log-likelihood is less than a given threshold.
```

E-step Let $\boldsymbol{r} \coloneqq \boldsymbol{y} - \mu(\boldsymbol{X})$. From (8) and (9), the following expectation is computed in the E-step:

$$\begin{aligned}
&\mathbb{E}_{\overline{\boldsymbol{\beta}}\sim p(\overline{\boldsymbol{\beta}}|\boldsymbol{y};\widehat{\mathcal{P}})}\left[\log p\left(\boldsymbol{y}, \overline{\boldsymbol{\beta}} \,\middle|\, \boldsymbol{X}; \mathcal{P}\right)\right] \\
&\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log\det\boldsymbol{\Sigma}_\beta - \frac{1}{2}\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}_\beta^{-1}\widehat{\boldsymbol{\beta}} - \frac{1}{2}\operatorname{tr}\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}_\beta^{-1} - \frac{1}{2\sigma^2}\mathbb{E}_{\overline{\boldsymbol{\beta}}\sim p(\overline{\boldsymbol{\beta}}|\boldsymbol{y};\widehat{\mathcal{P}})}\left(\boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right) \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log\det\boldsymbol{\Sigma}_\beta - \frac{1}{2}\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}_\beta^{-1}\widehat{\boldsymbol{\beta}} - \frac{1}{2}\operatorname{tr}\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}_\beta^{-1} \\
&\quad - \frac{1}{2\sigma^2}\left(\left\|\boldsymbol{r} - \boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\beta}}\right\|^2 + \operatorname{tr}\boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{S}}\boldsymbol{W}^\top\boldsymbol{K}\right),
\end{aligned}$$

with

$$\widehat{\boldsymbol{S}} \coloneqq \left(\widehat{\boldsymbol{\Sigma}}_\beta^{-1} + \widehat{\sigma}^{-2}\boldsymbol{W}^\top\boldsymbol{K}^2\boldsymbol{W}\right)^{-1} = \widehat{\boldsymbol{\Sigma}}_\beta - \widehat{\boldsymbol{\Sigma}}_\beta\boldsymbol{W}^\top\boldsymbol{K}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\Sigma}}_\beta.$$

M-step The M-step maximizes the expectation from the E-step. Taking the partial derivative of the expectation with respect to $\boldsymbol{\Sigma}_\beta^{-1}$ and setting it to zero yields

$$\widehat{\boldsymbol{\Sigma}}_\beta \leftarrow \widehat{\boldsymbol{\Sigma}}_\beta + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top - \widehat{\boldsymbol{\Sigma}}_\beta\boldsymbol{W}^\top\boldsymbol{K}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\Sigma}}_\beta. \tag{11}$$

Similarly, setting the partial derivative with respect to σ^2 to zero gives

$$\begin{aligned}
\widehat{\sigma}^2 &\leftarrow \frac{1}{n}\left[\left\|\boldsymbol{r} - \boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\beta}}\right\|^2 + \operatorname{tr}\left(\boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{S}}\boldsymbol{W}^\top\boldsymbol{K}\right)\right] \\
&= \widehat{\sigma}^2 + \frac{1}{n}\left[\left\|\boldsymbol{r} - \boldsymbol{K}\boldsymbol{W}\widehat{\boldsymbol{\beta}}\right\|^2 - \widehat{\sigma}^4\operatorname{tr}\left(\widehat{\boldsymbol{G}}^{-1}\right)\right].
\end{aligned} \tag{12}$$

In the GP, $\boldsymbol{r}$ also involves the parameters of the mean function. Estimation of these parameters can be performed independently in the above M-step. We assume that $\boldsymbol{K}$ is given. If not, the above M-step needs addition updates for the parameters of $\boldsymbol{K}$. Alternatively, $\boldsymbol{K}$ can be chosen using cross-validation. Algorithm 1 gives the pseudo-code for parameter estimation.

Computational complexity Both the above E-step and M-step have the computational complexity $O\left(n^2 m\right)$. In practice, m can be very small, e.g., $m \leq 3$, as we will show for several real datasets. Note that $\boldsymbol{G}^{-1}$ is not computed directly, but uses the more efficient Woodbury identity

$$\boldsymbol{G}^{-1} = \sigma^{-2}\left[\boldsymbol{I} - \boldsymbol{P}\left(\sigma^2 \boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{P}^\top \boldsymbol{P}\right)^{-1} \boldsymbol{P}^\top\right] \tag{13}$$

with $\boldsymbol{P} := \boldsymbol{K}\boldsymbol{W}$. This can be computed within $O\left(n^2 m\right)$ time rather than $O\left(n^3\right)$ time for the direct inverse.

Determining the subspace rank m For the SIGP, by construction m depends on how well the rank-m subspace of the RKHS can approximate the true target function f. As the eigenvalue of the operator corresponding to the radial kernel decays almost exponentially [Belkin, 2018], a small m can be sufficient in practice. Methods for determining the dimensionality of SDR subspaces may be applied [Li, 1991, Schott, 1994, Cook and Forzani, 2009].

5 SIGP for Regression

We now consider regression using the SIGP. The prediction of the response variable in a test set $\boldsymbol{y}_t$ as a function of the feature matrix $\boldsymbol{X}_t$ is analogous to the standard GP, except the following covariance matrix is used

$$\operatorname{cov}\left(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}\right) = \boldsymbol{K}\left(\boldsymbol{x}_t, \boldsymbol{X}\right) \boldsymbol{W} \boldsymbol{\Sigma}_\beta \boldsymbol{W}^\top \boldsymbol{K}.$$

Denote by $\boldsymbol{M}_{ZY} := \boldsymbol{K}\left(\boldsymbol{Z}, \boldsymbol{X}\right) \boldsymbol{W} \boldsymbol{\Sigma}_\beta \boldsymbol{W}^\top \boldsymbol{K}\left(\boldsymbol{X}, \boldsymbol{Y}\right)$, the $\boldsymbol{y}_t$ prediction for test data $\boldsymbol{X}_t$ is given by

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{y}_t\right) &= \mu\left(\boldsymbol{X}_t\right) + \boldsymbol{M}_{X_t X} \boldsymbol{G}^{-1}\left(\boldsymbol{y} - m\left(\boldsymbol{X}\right)\right) \\ \operatorname{var}\left(\boldsymbol{y}_t\right) &= \boldsymbol{M}_{X_t X_t} + \sigma^2 \boldsymbol{I} - \boldsymbol{M}_{X_t X} \boldsymbol{G}^{-1} \boldsymbol{M}_{X X_t},\end{aligned}$$

where $\mu\left(\boldsymbol{X}_t\right)$ is the same mean function as defined in standard GPs, and $\boldsymbol{G}^{-1}$ is the inverse of the marginal variance (13).

Often times, only the diagonal entries of $\operatorname{var}\left(\boldsymbol{y}_t\right)$ are needed to determine the confidence interval. These diagonal entries can be computed efficiently. Denote by $\boldsymbol{P}_Z = \boldsymbol{K}\left(\boldsymbol{Z}, \boldsymbol{X}\right) \boldsymbol{W} \boldsymbol{\Sigma}_\beta^{1/2}$, then we have $\boldsymbol{M}_{ZY} = \boldsymbol{P}_Z \boldsymbol{P}_Y^\top$. It is easy to see $\operatorname{diag}\left(\boldsymbol{M}_{X_t X_t}\right)_i = \left\|\left(\boldsymbol{P}_{X_t}\right)_{i:}\right\|^2$ and

$$\operatorname{diag}\left(\boldsymbol{M}_{X_t X} \boldsymbol{G}^{-1} \boldsymbol{M}_{X X_t}\right)_i = \operatorname{diag}\left(\boldsymbol{P}_{X_t}\left(\sigma^2\left(\boldsymbol{P}_X^\top \boldsymbol{P}_X\right)^{-1} + \boldsymbol{I}\right)^{-1} \boldsymbol{P}_{X_t}^\top\right)_i.$$

Thus, the i-th desired diagonal entry is written

$$\operatorname{diag}\left(\operatorname{var}\left(\boldsymbol{y}_t\right)\right)_i = \left\|\left[\boldsymbol{P}_{X_t}\left(\sigma^{-2} \boldsymbol{P}_X^\top \boldsymbol{P}_X + \boldsymbol{I}\right)^{-1/2}\right]_{i:}\right\|^2 + \sigma^2.$$

A mean function for SIGP To use SIGP for regression, a mean function such as a linear model or the kernel Ridge regression model is required. The parameter of the mean function can be inferred together with the SIGP covariance parameters as described in Algorithm 1. In the following, we consider a particular mean function $\mu(\boldsymbol{Z}) = \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - c\mathbf{1}$, where $\boldsymbol{Z}$ can be either $\boldsymbol{X}$ or $\boldsymbol{K}$. This mean function has two parameters $\boldsymbol{\alpha}$ and c, which are estimated through

$$\min_{\boldsymbol{\alpha},c} \quad (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - c\mathbf{1})^\top \boldsymbol{G}^{-1} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - c\mathbf{1}) + \lambda \boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha}.$$

Here, λ is the regularization parameter, and $\boldsymbol{H}$ is chosen based on the inner product. In particular, one can let $\boldsymbol{H}$ to be the identity matrix or $\boldsymbol{K}$ depending on the choice of $\boldsymbol{Z}$. The optimal $\boldsymbol{\alpha}_\star$ and $c_\star$ satisfy

$$\left(\boldsymbol{Z}^\top \boldsymbol{G}^{-1}\boldsymbol{Z} + \lambda\boldsymbol{H}\right)\boldsymbol{\alpha}_\star = \boldsymbol{Z}^\top \boldsymbol{G}^{-1}(\boldsymbol{y} - c_\star\mathbf{1})$$

$$c_\star = \frac{\mathbf{1}^\top \boldsymbol{G}^{-1}\boldsymbol{y} - \mathbf{1}^\top \boldsymbol{G}^{-1}\boldsymbol{Z}\boldsymbol{\alpha}_\star}{\mathbf{1}^\top \boldsymbol{G}^{-1}\mathbf{1}}.$$

Let $\boldsymbol{L} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^\top \boldsymbol{G}^{-1}\left(\mathbf{1}^\top \boldsymbol{G}^{-1}\mathbf{1}\right)^{-1}$, we obtain

$$\boldsymbol{\alpha}_\star = \left(\boldsymbol{Z}^\top \boldsymbol{G}^{-1}\boldsymbol{L}\boldsymbol{Z} + \lambda\boldsymbol{H}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{G}^{-1}\boldsymbol{L}\boldsymbol{y}. \tag{14}$$

The estimators $\boldsymbol{\alpha}_\star$ and $c_\star$ can be computed while training the SIGP in Algorithm 1.

In the case that $\boldsymbol{Z} = \boldsymbol{X}$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is high-dimensional, i.e., $p \gg n$, the computation (14) can be expensive. Alternatively, we consider the following dual estimator in place of (14):

$$\boldsymbol{\alpha}_\star = \boldsymbol{H}^{-1}\boldsymbol{Z}^\top \left(\boldsymbol{Z}\boldsymbol{H}^{-1}\boldsymbol{Z}^\top + \lambda\left(\boldsymbol{G}^{-1}\boldsymbol{L} + \tau\boldsymbol{I}\right)^{-1}\right)^{-1}\boldsymbol{y},$$

for a small given number $\tau > 0$. Computing the dual estimator takes $O\left(n^3\right)$ as opposed to $O\left(p^3\right)$ for (14).

6 Evaluation on real and simulated data

In this section, we evaluate the performance of the SIGP on both synthetic toy datasets as well as seven real datasets of varying size. Both binary classification and regression tasks are considered. We first illustrate on the toy datasets the successful fitting of the SIGP to 1D data, as well as the difference in the contours generated by GP and SIGP classification. On real benchmark datasets, we compare the performance of SIGP and other popular nonlinear methods, including state-of-the-art GP inference methods like FITC [Snelson and Ghahramani, 2006a] and the support vector machine (SVM).

We used the GP implementation in GPML toolbox [Rasmussen and Nickisch, 2010] which is generally considered to be amongst the best implementation of these algorithms. The SVM results are based on the `fitcsvm` from Matlab. All methods use the radial basis kernel, where the parameters for SVM and SIGP were obtained by cross-validation on the training data. We also standardized the input data. The testbed for these experiments was a Linux workstation with Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz and 256GB RAM.

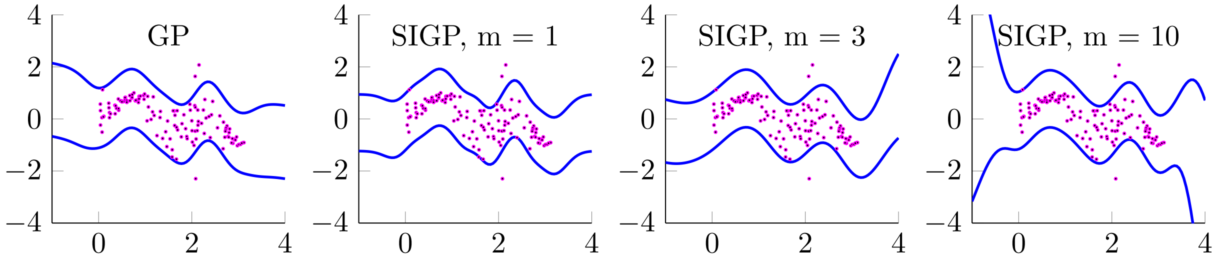

Figure 1: The predictive $2 - \sigma$ confidence intervals for the held-out validation data of `Synthetic`. The figure shows successful learning of the SIGP on the 1D data. In particular, the value of m in the SIGP can affect the uncertainty in prediction.

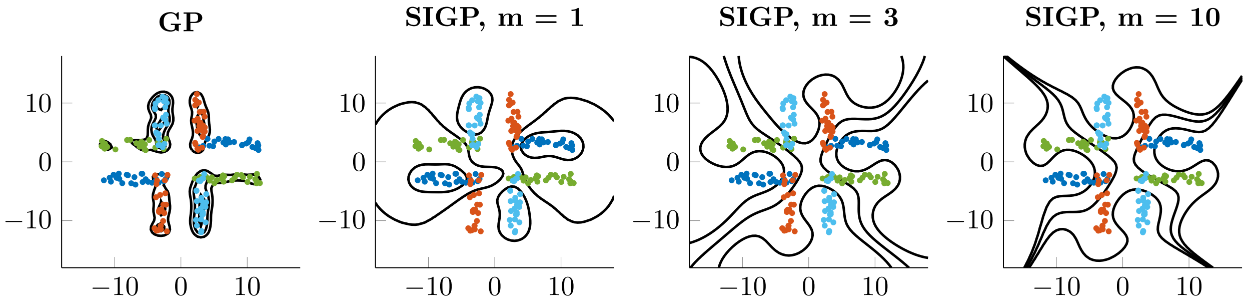

Figure 2: Classification of four classes of 2D points marked by different colors. The curves are the contours corresponding to the decision boundaries. The figure shows that both GP and SIGP correctly classify the points. GP achieves the best fitness, but can potentially overfit.

The main finding of these experiments is that SIGP consistently outperforms the standard GP in classification, even with a very low rank $m \leq 3$, m is the rank of the projected RKHS. Increasing m may not necessarily increase the predictive performance of SIGP. We emphasize that the covariance of the SIGP at a given point depends on the value of the induced functions at that point. Nevertheless, SIGP achieves competitive predictive performance in regression, measured by the negative log predictive density.

6.1 Simulation on toy datasets

Two toy datasets were used for the simulation. The first is a small 1D `Synthetic` dataset from WCCI-2006 Predictive Uncertainty in Environmental Modeling Competition. The dataset consists of 256 training patterns and 128 validation patterns. We use this dataset to illustrate successful learning of SIGP by comparing the predictive distribution of GP and SIGP on the validation set. The second toy dataset consists of 2D points from fours classes. We compare the contours generated by GP and SIGP corresponding to the decision boundaries.

The predictive distributions for `Synthetic` on the validation set is shown in Figure 1. Both the GP and the SIGP successfully capture the trends as well as uncertainty of the held-out validation data points. Increasing the subspace rank m takes into account the value of an extended number of functions in the RKHS while modeling the covariance. This is reflected in the prediction uncertainty.

Next, we compare the contours given by the GP and SIGP induced by rank-m subspaces,

Table 1: Classification data sizes.

Data set	n	p	t	Data set	n	p	t
Arcene	200	10000	100	German	1000	24	300
Gisette	7000	5000	1000	Heart	270	13	100
Madelon	2600	500	600	Cancer	699	10	200

Table 2: Comparison of F1 scores for binary classification on benchmark datasets.

Data set	Laplace	KL	EP	FITC[1]	SVM	SIGP		
						$m = 1$	$m = 3$	$m = 10$
Arcene	0.8235	0.8269	0.8235	0.8235	0.8352	**0.8571**	**0.8571**	**0.8571**
Gisette	0.9570	0.9541	0.9571	0.9571	0.9670	**0.9710**	**0.9710**	0.9701
Madelon	0.5695	0.5695	0.5695	0.5695	0.5990	0.6517	**0.6667**	0.6367
German	0.6211	0.6211	0.6211	0.6125	0.6182	**0.6341**	0.6258	0.6258
Heart	0.8409	0.8409	0.8409	0.8506	**0.8605**	**0.8605**	**0.8605**	**0.8605**
Cancer	0.9425	0.9213	0.9438	0.9778	0.9888	**1.0**	**1.0**	**1.0**

[1] Using $\lfloor n/2 \rfloor$ pseudo-inputs.

$m = 1, 3, 10$, of the RKHS. These points are positioned in corner shapes as shown in Figure 2. Each class has 50 points, and is marked by a different color. Note that each corner consists of points from two distinct classes. Figure 2 exhibits unique contour patterns for both methods. Two salient observations are: 1) both methods, including the SIGP induced by a rank-1 subspace of the full RKHS, correctly classify the points; and 2) SIGP tends to be more robust against overfitting compared to GP.

6.2 Results on real datasets

We compare the performance of SIGP and state-of-the-art methods for binary classification as well as regression on several real benchmark datasets.

For the binary classification, we considered six benchmark datasets of varying size from the UCI repository. Table 2 reports the F1 scores for the methods in comparison. The total number of observations (n), number of attributes (p), and number of test cases (t) for each dataset are given in Table 1.

As shown in Table 2, SIGP achieves competitive classification performance compared to state-of-the-art GP inference methods as well as the SVM. A significant improvement was observed on the high-dimensional `Arcene` dataset which has many more predictor variables (10000) than the number of observations (200). These results suggest that the SIGP induced by a low-rank subspace of the RKHS suffices to deliver desirable performance.

Regression For regression, we consider the `Temp` dataset from WCCI-2006 Predictive Uncertainty in Environmental Modeling Competition. The `Temp` datasets consists of 7117 training data points

Table 3: Prediction performance on the held-out validation of `Temp`. FITC-d denotes the FITC method using d pseudo-inputs.

Method	Validation	
	NLPD	MSE
FITC-10	0.1291	0.2752
FITC-100	0.1287	0.2751
FITC-1000	0.1290	0.2752
SIGP, $m = 1$	0.0484	0.2522
SIGP, $m = 3$	0.0660	0.2581
SIGP, $m = 10$	0.0583	0.2552

and 3558 validation patterns, and each observation has 106 features. We report the negative log predictive density (NLPD) as well as mean squared error (MSE) on the held-out validation data. For this task, we compare the SIGP induced by rank-m, $m = 1, 3, 10$, subspaces of the RKHS and GP with the FITC inference which produces better NLPD than the other inference methods [Snelson and Ghahramani, 2006b].

Table 3 states the predictive performance measured by NLPD (smaller is better) as well as MSE on the held-out validation of `Temp`. From Table 3, the number of pseudo-inputs does not seem to have a significant impact on the performance. The best NLPD and MSE both are achieved for SIGP induced by a rank-1 subspace of the RKHS.

7 Conclusions

The proposed SIGP model can be viewed as the dual of the standard GP, where the target function f is randomly drawn from a distribution over the RKHS induced by the input data. In particular, f is a linear combination of the feature maps with the coefficients specified by a multivariate normal. The inherent low-rank nature of the SIGP covariance admits both efficient computation as well as effective representation in terms of an SDR subspace of the RKHS. We developed efficient algorithms for selecting the SIGP mean function and covariance function in the light of the training data. Extensive results on real datasets show that SIGP performs competitively in classification as well as regression.

Code and Data

The datasets as well as Matlab implementation of SIGP are available on the Git repository: `https://github.com/ZilongTan/sigp`.

Acknowledgements

Z.T. is supported by grant NSF CNS-1423128. S.M. would like to acknowledge the support of grants NSF IIS-1546331, NSF DMS-1418261, NSF IIS-1320357, NSF DMS-1045153, and NSF DMS-1613261.

References

K. P. Adragni and R. D. Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

M. Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. *ArXiv e-prints*, 2018. URL https://arxiv.org/abs/1801.03437.

R. D. Cook and L. Forzani. Likelihood-Based Sufficient Dimension Reduction. *Journal of the American Statistical Association*, 104(485):197–208, 2009.

J. L. Doob. The Elementary Gaussian Processes. *Ann. Math. Statist.*, 15(3):229–282, 1944.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *Conference on Uncertainty in Artificial Intellegence (UAI)*, pages 282–290, 2013.

J. Hensman, A. G. de G. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *Artificial Intelligence and Statistics (AISTATS)*, 2015.

K.-C. Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

V. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 1999.

N. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert. Characterizing the Function Space for Bayesian Kernel Models. *J. Mach. Learn. Res.*, 8:1769–1797, 2007.

C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, 1973.

C. Rasmussen. *Gaussian Processes in Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 63–71. Springer, 2004.

C. E. Rasmussen. Gaussian Processes for Machine Learning. MIT Press, 2006.

C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (GPML) toolbox. *J. Mach. Learn. Res.*, 11:3011–3015, 2010.

J. R. Schott. Determining the Dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, 89(425):141–148, 1994.

M. Seeger, C. Williams, and N. Lawrence. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. *Artificial Intelligence and Statistics (AISTATS)*, 2003.

A. J. Smola and P. L. Bartlett. Sparse Greedy Gaussian Process Regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 619–625. 2001.

E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1257–1264. 2006a.

E. Snelson and Z. Ghahramani. Variable Noise and Dimensionality Reduction for Sparse Gaussian Processes. In *Conference on Uncertainty in Artificial Intellegence (UAI)*, 2006b.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, 1999.

C. K. I. Williams. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In *Learning and Inference in Graphical Models*, pages 599–621. Kluwer, 1997.

Q. Wu, F. Liang, and S. Mukherjee. Kernel Sliced Inverse Regression: Regularization and Consistency. *Abstract and Applied Analysis*, 2013, 2013.