

EigenGP: Gaussian Process Models with Adaptive Eigenfunctions

Hao Peng

Department of Computer Science
 Purdue University
 West Lafayette, IN 47907, USA
 pengh@purdue.edu

Yuan Qi

Departments of Computer Science and Statistics
 Purdue University
 West Lafayette, IN 47907, USA
 alanqi@cs.purdue.edu

Abstract

Gaussian processes (GPs) provide a nonparametric representation of functions. However, classical GP inference suffers from high computational cost for big data. In this paper, we propose a new Bayesian approach, EigenGP, that learns both basis dictionary elements—eigenfunctions of a GP prior—and prior precisions in a sparse finite model. It is well known that, among all orthogonal basis functions, eigenfunctions can provide the most compact representation. Unlike other sparse Bayesian finite models where the basis function has a fixed form, our eigenfunctions live in a reproducing kernel Hilbert space as a finite linear combination of kernel functions. We learn the dictionary elements—eigenfunctions—and the prior precisions over these elements as well as all the other hyperparameters from data by maximizing the model marginal likelihood. We explore computational linear algebra to simplify the gradient computation significantly. Our experimental results demonstrate improved predictive performance of EigenGP over alternative sparse GP methods as well as relevance vector machines.

1 Introduction

Gaussian processes (GPs) are powerful nonparametric Bayesian models with numerous applications in machine learning and statistics. GP inference, however, is costly. Training the exact GP regression model with N samples is expensive: it takes an $O(N^2)$ space cost and an $O(N^3)$ time cost. To address this issue, a variety of approximate sparse GP inference approaches have been developed [Williams and Seeger, 2001; Csató and Opper, 2002; Snelson and Ghahramani, 2006; Lázaro-Gredilla *et al.*, 2010; Williams and Barber, 1998; Titsias, 2009; Qi *et al.*, 2010; Higdon, 2002; Cressie and Johannesson, 2008]—for example, using the Nyström method to approximate covariance matrices [Williams and Seeger, 2001], optimizing a variational bound on the marginal likelihood [Titsias, 2009] or grounding the GP on a small set of (blurred) basis points [Snelson and Ghahramani, 2006; Qi *et al.*, 2010]. An elegant unifying view for various sparse

GP regression models is given by Quiñonero-Candela and Rasmussen [2005].

Among all sparse GP regression methods, a state-of-the-art approach is to represent a function as a sparse finite linear combination of pairs of trigonometric basis functions, a sine and a cosine for each spectral point; thus this approach is called sparse spectrum Gaussian process (SSGP) [Lázaro-Gredilla *et al.*, 2010]. SSGP integrates out both weights and phases of the trigonometric functions and learns all hyperparameters of the model (frequencies and amplitudes) by maximizing the marginal likelihood. Using global trigonometric functions as basis functions, SSGP has the capability of approximating any stationary Gaussian process model and been shown to outperform alternative sparse GP methods—including fully independent training conditional (FITC) approximation [Snelson and Ghahramani, 2006]—on benchmark datasets. Another popular sparse Bayesian finite linear model is the relevance vector machine (RVM) [Tipping, 2000; Faul and Tipping, 2001]. It uses kernel expansions over training samples as basis functions and selects the basis functions by automatic relevance determination [MacKay, 1992; Faul and Tipping, 2001].

In this paper, we propose a new sparse Bayesian approach, EigenGP, that learns both functional dictionary elements—eigenfunctions—and prior precisions in a finite linear model representation of a GP. It is well known that, among all orthogonal basis functions, eigenfunctions provide the most compact representation. Unlike SSGP or RVMs where the basis function has a fixed form, our eigenfunctions live in a reproducing kernel Hilbert space (RKHS) as a finite linear combination of kernel functions with their weights learned from data. We further marginalize out weights over eigenfunctions and estimate all hyperparameters—including basis points for eigenfunctions, lengthscales, and precision of the weight prior—by maximizing the model marginal likelihood (also known as evidence). To do so, we explore computational linear algebra and greatly simplify the gradient computation for optimization (thus our optimization method is totally different from RVM optimization methods). As a result of this optimization, our eigenfunctions are data dependent and make EigenGP capable of accurately modeling nonstationary data. Furthermore, by adding an additional kernel term in our model, we can turn the finite model into an infinite model to model the prediction uncertainty better—that

is, it can give nonzero prediction variance when a test sample is far from the training samples.

EigenGP is computationally efficient. It takes $O(NM)$ space and $O(NM^2)$ time for training on with M basis functions, which is same as SSGP and more efficient than RVMs (as RVMs learn weights over N , not M , basis functions.). Similar to FITC and SSGP, EigenGP focuses on predictive accuracy at low computational cost, rather than on faithfully converging towards the full GP as the number of basis functions grows. (For the latter case, please see the approach [Yan and Qi, 2010] that explicitly minimizes the KL divergence between exact and approximate GP posterior processes.)

The rest of the paper is organized as follows. Section 2 describes the background of GPs. Section 3 presents the EigenGP model and an illustrative example. Section 4 outlines the marginal likelihood maximization for learning dictionary elements and the other hyperparameters. In Section 5, we discuss related work. Section 6 shows regression results on multiple benchmark regression datasets, demonstrating improved performance of EigenGP over Nyström [Williams and Seeger, 2001], RVM, FITC, and SSGP.

2 Background of Gaussian Processes

We denote N independent and identically distributed samples as $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}_N$, where \mathbf{x}_i is a D dimensional input (i.e., explanatory variables) and y_i is a scalar output (i.e., a response), which we assume is the noisy realization of a latent function f at \mathbf{x}_i .

A Gaussian process places a prior distribution over the latent function f . Its projection \mathbf{f}_x at $\{\mathbf{x}_i\}_{i=1}^N$ defines a joint Gaussian distribution $p(\mathbf{f}_x) = \mathcal{N}(\mathbf{f}|\mathbf{m}^0, \mathbf{K})$, where, without any prior preference, the mean \mathbf{m}^0 are set to $\mathbf{0}$ and the covariance function $k(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ encodes the prior notion of smoothness. A popular choice is the anisotropic squared exponential covariance function: $k(\mathbf{x}, \mathbf{x}') = a_0 \exp(-(\mathbf{x} - \mathbf{x}')^T \text{diag}(\boldsymbol{\eta})(\mathbf{x} - \mathbf{x}'))$, where the hyperparameters include the signal variance a_0 and the lengthscales $\boldsymbol{\eta} = \{\eta_d\}_{d=1}^D$, controlling how fast the covariance decays with the distance between inputs. Using this covariance function, we can prune input dimensions by shrinking the corresponding lengthscales based on the data (when $\eta_d = 0$, the d -th dimension becomes totally irrelevant to the covariance function value). This pruning is known as Automatic Relevance Determination (ARD) and therefore this covariance is also called the ARD squared exponential. Note that the covariance function value remains the same when $(\mathbf{x}' - \mathbf{x})$ is the same – regardless where \mathbf{x}' and \mathbf{x} are. This thus leads to a *stationary* GP model. For nonstationary data, however, a stationary GP model is a misfit. Although nonstationary GP models have been developed and applied to real world applications, they are often limited to low-dimensional problems, such as applications in spatial statistics [Paciorek and Schervish, 2004]. Constructing general nonstationary GP models remains a challenging task.

For regression, we use a Gaussian likelihood function $p(y_i|f) = \mathcal{N}(y_i|f(\mathbf{x}_i), \sigma^2)$, where σ^2 is the variance of the observation noise. Given the Gaussian process prior over f and the data likelihood, the exact posterior process

is $p(f|\mathcal{D}, \mathbf{y}) \propto GP(f|0, k) \prod_{i=1}^N p(y_i|f)$. Although the posterior process for GP regression has an analytical form, we need to store and invert a N by N matrix, which has the computational complexity $O(N^3)$, rendering GP unfeasible for big data analytics.

3 Model of EigenGP

To enable fast inference and obtain a nonstationary covariance function, our new model EigenGP projects the GP prior in an eigensubspace. Specifically, we set the latent function

$$f(\mathbf{x}) = \sum_{j=1}^M \alpha_j \phi^j(\mathbf{x}) \quad (1)$$

where $M \ll N$ and $\{\phi^j(\mathbf{x})\}$ are eigenfunctions of the GP prior. We assign a Gaussian prior over $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_M]$, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{w}))$, so that f follows a GP prior with zero mean and the following covariance function

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^M w_j \phi^j(\mathbf{x}) \phi^j(\mathbf{x}'). \quad (2)$$

To compute the eigenfunctions $\{\phi^j(\mathbf{x})\}$, we can use the Galerkin projection to approximate them by Hermite polynomials [Marzouk and Najm, 2009]. For high dimensional problems, however, this approach requires a tensor product of univariate Hermite polynomials that dramatically increases the number of parameters.

To avoid this problem, we apply the Nyström method [Williams and Seeger, 2001] that allows us to obtain an approximation to the eigenfunctions in a high dimensional space efficiently. Specifically, given inducing inputs (i.e. basis points) $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_M]$, we replace

$$\int k(\mathbf{x}, \mathbf{x}') \phi^j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda_j \phi^j(\mathbf{x}') \quad (3)$$

by its Monte Carlo approximation

$$\frac{1}{M} \sum_{i=1}^M k(\mathbf{x}, \mathbf{b}_i) \phi^j(\mathbf{b}_i) \approx \lambda_j \phi^j(\mathbf{x}) \quad (4)$$

Then, by evaluating this equation at \mathbf{B} so that we can estimate the values of $\phi^j(\mathbf{b}_i)$ and λ_j , we obtain the j -th eigenfunction $\phi^j(\mathbf{x})$ as follows

$$\phi^j(\mathbf{x}) = \frac{\sqrt{M}}{\lambda_j^{(M)}} \mathbf{k}(\mathbf{x}) \mathbf{u}_j^{(M)} = \mathbf{k}(\mathbf{x}) \mathbf{u}_j \quad (5)$$

where $\mathbf{k}(\mathbf{x}) \triangleq [k(\mathbf{x}, \mathbf{b}_1), \dots, k(\mathbf{x}, \mathbf{b}_M)]$, $\lambda_j^{(M)}$ and $\mathbf{u}_j^{(M)}$ are the j -th eigenvalue and eigenvector of the covariance function evaluated at \mathbf{B} , and $\mathbf{u}_j = \sqrt{M} \mathbf{u}_j^{(M)} / \lambda_j^{(M)}$. Note that, for simplicity, we have chosen the number of the inducing inputs to be the same as the number of the eigenfunctions; in practice, we can use more inducing inputs while computing only the top M eigenvectors. As shown in (5), our eigenfunction lives in a RKHS as a linear combination of the kernel functions evaluated at \mathbf{B} with weights \mathbf{u}_j .

Inserting (5) into (1) we obtain

$$f(\mathbf{x}) = \sum_{j=1}^M \alpha_j \sum_{i=1}^M u_{ij} k(\mathbf{x}, \mathbf{b}_i) \quad (6)$$

This equation reveals a two-layer structure of EigenGP. The first layer linearly combines multiple kernel functions to generate each eigenfunction ϕ^j . The second layer takes these eigenfunctions as the basis functions to generate the function value f . Note that f is a Bayesian linear combination of $\{\phi^j\}$ where the weights α are integrated out to avoid overfitting.

All the model hyperparameters are learned from data. Specifically, for the first layer, to learn the eigenfunctions $\{\phi_i\}$, we estimate the inducing inputs \mathbf{B} and the kernel hyperparameters (such as lengthscales $\boldsymbol{\eta}$ for the ARD kernel) by maximizing the model marginal likelihood. For the second layer, we marginalize out α to avoid overfitting and maximize the model marginal likelihood to learn the hyperparameter \mathbf{w} of the prior.

With the estimated hyperparameters, the prior over f is nonstationary because its covariance function in (2) varies at different regions of \mathbf{x} . This comes at no surprise since the eigenfunctions are tied with $p(\mathbf{x})$ in (3). This nonstationarity reflects the fact that our model is adaptive to the distribution of the explanatory variables \mathbf{x} .

Note that to recover the full uncertainty captured by the kernel function k , we can add the following term into the kernel function of EigenGP:

$$\delta(\mathbf{x} - \mathbf{x}') (k(\mathbf{x}, \mathbf{x}') - \tilde{k}(\mathbf{x}, \mathbf{x}')) \quad (7)$$

where $\delta(a) = 1$ if and only if $a = 0$. Compared to the original EigenGP model, which has a finite degree of freedom, this modified model has the infinite number of basis functions (assuming k has an infinite number of basis functions as the ARD kernel). Thus, this model can accurately model the uncertainty of a test point even when it is far from the training set. We derive the optimization updates of all the hyperparameters for both the original and modified EigenGP models. But according to our experiments, the modified model does not improve the prediction accuracy over the original EigenGP (it even reduces the accuracy sometimes.). Therefore, we will focus on the original EigenGP model in our presentation for its simplicity. Before we present details about hyperparameter optimization, let us first look at an illustrative example on the effect of hyperparameter optimization, in particular, the optimization of the inducing inputs \mathbf{B} .

3.1 Illustrative Example

For this example, we consider a toy dataset used by the FITC algorithm [Snelson and Ghahramani, 2006]. It contains 200 one-dimensional training samples. We use 5 basis points ($M = 5$) and choose the ARD kernel. We then compare the basis functions $\{\phi^j\}$ and the corresponding predictive distributions in two cases. For the first case, we use the kernel width η learned from the full GP model as the kernel width for EigenGP, and apply K-means to set the basis points \mathbf{B} as cluster centers. The idea of using K-means to set the basis points has been suggested by Zhang *et al.* [2008] to minimize an error bound for the Nyström approximation. For

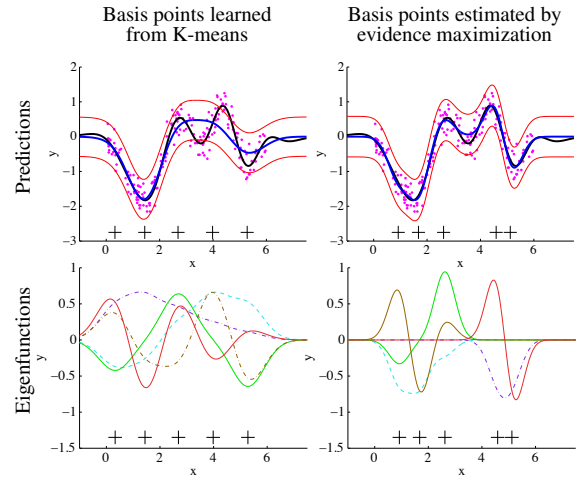


Figure 1: Illustration of the effect of optimizing basis points (i.e., inducing inputs). In the first row, the pink dots represent data points, the blue and red solid curves correspond to the predictive mean of EigenGP and \pm two standard deviations around the mean, the black curve corresponds to the predictive mean of the full GP, and the black crosses denote the basis points. In the second row, the curves in various colors represent the five eigenfunctions of EigenGP.

the second case, we optimize η , \mathbf{B} and \mathbf{w} by maximizing the marginal likelihood of EigenGP.

The results are shown in Figure 1. The first row demonstrates that, by optimizing the hyperparameters, EigenGP achieves the predictions very close to what the full GP achieves—but using only 5 basis functions. In contrast, when the basis points are set to the cluster centers by K-means, EigenGP leads to the prediction significantly different from that of the full GP and fails to capture the data trend, in particular, for $x \in (3, 5)$. The second row of Figure 1 shows that K-means sets the basis points almost evenly spaced on y , and accordingly the five eigenfunctions are smooth *global* basis functions whose shapes are not directly linked to the function they fit. Evidence maximization, by contrast, sets the basis points unevenly spaced to generate the basis functions whose shapes are more *localized* and adaptive to the function they fit; for example, the eigenfunction represented by the red curve well matches the data on the right.

4 Learning Hyperparameters

In this section we describe how we optimize all the hyperparameters, denoted by $\boldsymbol{\theta}$, which include \mathbf{B} in the covariance function (2), and all the kernel hyperparameters (e.g., a_0 and $\boldsymbol{\eta}$). To optimize $\boldsymbol{\theta}$, we maximize the marginal likelihood (i.e. evidence) based on a conjugate Newton method¹. We explore two strategies for evidence maximization. The first one is sequential optimization, which first fixes \mathbf{w} while updating all the other hyperparameters, and then optimizes \mathbf{w} while fixing the other hyperparameters. The second strategy is to optimize

¹We use the code from <http://www.gaussianprocess.org/gpml/code/matlab/doc>

all the hyperparameters jointly. Here we skip the details for the more complicated joint optimization and describe the key gradient formula for sequential optimization. The key computation in our optimization is for the log marginal likelihood and its gradient:

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2} \ln |\mathbf{C}_N| - \frac{1}{2} \mathbf{y}^T \mathbf{C}_N^{-1} \mathbf{y} - \frac{N}{2} \ln(2\pi), \quad (8)$$

$$d \ln p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2} [\text{tr}(\mathbf{C}_N^{-1} d\mathbf{C}_N) - \text{tr}(\mathbf{C}_N^{-1} \mathbf{y} \mathbf{y}^T \mathbf{C}_N^{-1} d\mathbf{C}_N)] \quad (9)$$

where $\mathbf{C}_N = \tilde{\mathbf{K}} + \sigma^2 \mathbf{I}$, $\tilde{\mathbf{K}} = \Phi \text{diag}(\mathbf{w}) \Phi^T$, and $\Phi = \{\phi^m(\mathbf{x}_n)\}$ is an N by M matrix. Because the rank of $\tilde{\mathbf{K}}$ is $M \ll N$, we can compute $\ln |\mathbf{C}_N|$ and \mathbf{C}_N^{-1} efficiently with the cost of $O(M^2 N)$ via the matrix inversion and determinant lemmas. Even with the use of the matrix inverse lemma for the low-rank computation, a naive calculation would be very costly. We apply identities from computational linear algebra [Minka, 2001; de Leeuw, 2007] to simplify the needed computation dramatically.

To compute the derivative with respect to \mathbf{B} , we first notice that, when \mathbf{w} is fixed, we have

$$\mathbf{C}_N = \tilde{\mathbf{K}} + \sigma^2 \mathbf{I} = \mathbf{K}_{XB} \mathbf{K}_{BB}^{-1} \mathbf{K}_{BX} + \sigma^2 \mathbf{I} \quad (10)$$

where \mathbf{K}_{XB} is the cross-covariance matrix between the training data \mathbf{X} and the inducing inputs \mathbf{B} , and \mathbf{K}_{BB} is the covariance matrix on \mathbf{B} .

For the ARD squared exponential kernel, utilizing the following identities, $\text{tr}(\mathbf{P}^T \mathbf{Q}) = \text{vec}(\mathbf{P})^T \text{vec}(\mathbf{Q})$ and $\text{vec}(\mathbf{P} \circ \mathbf{Q}) = \text{diag}(\text{vec}(\mathbf{P})) \text{vec}(\mathbf{Q})$, where $\text{vec}(\cdot)$ vectorizes a matrix into a column vector, and \circ represents the Hadamard product, we can derive the derivative of the first trace term in (9)

$$\begin{aligned} \frac{\text{tr}(\mathbf{C}_N^{-1} d\mathbf{C}_N)}{d\mathbf{B}} &= 4\mathbf{R} \mathbf{X}^T \text{diag}(\boldsymbol{\eta}) - 4(\mathbf{R} \mathbf{1} \mathbf{1}^T) \circ (\mathbf{B}^T \text{diag}(\boldsymbol{\eta})) \\ &\quad - 4\mathbf{S} \mathbf{B}^T \text{diag}(\boldsymbol{\eta}) + 4(\mathbf{S} \mathbf{1} \mathbf{1}^T) \circ (\mathbf{B}^T \text{diag}(\boldsymbol{\eta})) \end{aligned} \quad (11)$$

where $\mathbf{1}$ is a column vector of all ones, and

$$\mathbf{R} = (\mathbf{K}_{BB}^{-1} \mathbf{K}_{BX} \mathbf{C}_N^{-1}) \circ \mathbf{K}_{BX} \quad (12)$$

$$\mathbf{S} = (\mathbf{K}_{BB}^{-1} \mathbf{K}_{BX} \mathbf{C}_N^{-1} \mathbf{K}_{XB} \mathbf{K}_{BB}^{-1}) \circ \mathbf{K}_{BB} \quad (13)$$

Note that we can compute $\mathbf{K}_{BX} \mathbf{C}_N^{-1}$ efficiently via low-rank updates. Also, $\mathbf{R} \mathbf{1} \mathbf{1}^T$ ($\mathbf{S} \mathbf{1} \mathbf{1}^T$) can be implemented efficiently by first summing over the columns of \mathbf{S} (\mathbf{R}) and then copying it multiple times—without any multiplication operation. To obtain $\frac{\text{tr}(\mathbf{C}_N^{-1} \mathbf{y} \mathbf{y}^T \mathbf{C}_N^{-1} d\mathbf{C}_N)}{d\mathbf{B}}$, we simply replace \mathbf{C}_N^{-1} in (12) and (13) by $\mathbf{C}_N^{-1} \mathbf{y} \mathbf{y}^T \mathbf{C}_N^{-1}$.

Using similar derivations, we can obtain the derivatives with respect to the lengthscale $\boldsymbol{\eta}$, a_0 and σ^2 respectively.

To compute the derivative with respect to \mathbf{w} , we can use the formula $\text{tr}(\mathbf{P} \text{diag}(\mathbf{w}) \mathbf{Q}) = \mathbf{1}^T (\mathbf{Q}^T \circ \mathbf{P}) \mathbf{w}$ to obtain the two trace terms in (9) as follows:

$$\frac{\text{tr}(\mathbf{C}_N^{-1} d\mathbf{C}_N)}{d\mathbf{w}} = \mathbf{1}^T (\Phi \circ (\mathbf{C}_N^{-1} \Phi)) \quad (14)$$

$$\frac{\text{tr}(\mathbf{C}_N^{-1} \mathbf{y} \mathbf{y}^T \mathbf{C}_N^{-1} d\mathbf{C}_N)}{d\mathbf{w}} = \mathbf{1}^T (\Phi \circ (\mathbf{C}_N^{-1} \mathbf{y} \mathbf{y}^T \mathbf{C}_N^{-1} \Phi)) \quad (15)$$

For either sequential or joint optimization, the overall computational complexity is $O(\max(M^2, D)N)$ where D is the data dimension.

5 Related Work

Our work is closely related to the seminal work by [Williams and Seeger, 2001], but they differ in multiple aspects. First, we define a valid probabilistic model based on an eigen-decomposition of the GP prior. By contrast, the previous approach [Williams and Seeger, 2001] aims at a low-rank approximation to the finite covariance/kernel matrix used in GP training—from a numerical approximation perspective—and its predictive distribution is not well-formed in a probabilistic framework (*e.g.*, it may give a negative variance of the predictive distribution.). Second, while the Nyström method simply uses the first few eigenvectors, we maximize the model marginal likelihood to adjust their weights in the covariance function. Third, exploring the clustering property of the eigenfunctions of the Gaussian kernel, our approach can conduct semi-supervised learning, while the previous one cannot. The semi-supervised learning capability of EigenGP is investigated in another paper of us. Fourth, the Nyström method lacks a principled way to learn model hyperparameters including the kernel width and the basis points while EigenGP does not.

Our work is also related to methods that use kernel principle component analysis (PCA) to speed up kernel machines [Hoegaerts *et al.*, 2005]. However, for these methods it can be difficult—if not impossible—to learn important hyperparameters including kernel width for each dimension and inducing inputs (not a subset of the training samples). By contrast, EigenGP learns all these hyperparameters from data based on gradients of the model marginal likelihood.

6 Experimental Results

In this section, we compare EigenGP² and alternative methods on synthetic and real benchmark datasets. The alternative methods include the sparse GP methods—FITC, SSGP, and the Nyström method—as well as RVMs. We implemented the Nyström method ourselves and downloaded the software implementations for the other methods from their authors' websites. For RVMs, we used the fast fixed point algorithm [Faul and Tipping, 2001]. We used the ARD kernel for all the methods except RVMs (since they do not estimate the lengthscales in this kernel) and optimized all the hyperparameters via evidence maximization. For RVMs, we chose the squared exponential kernel with the same lengthscale for all the dimensions and applied a 10-fold cross-validation on the training data to select the lengthscale. On large real data, we used the values of $\boldsymbol{\eta}$, a_0 , and σ^2 learned from the full GP on a subset that was 1/10 of the training data to initialize all the methods except RVMs. For the rest configurations, we used the default setting of the downloaded software packages. For our own model, we denote the versions with sequential and joint optimization as EigenGP and EigenGP*, respectively. To evaluate the test performance of each method, we measure the Normalized Mean Square Error (NMSE) and the Mean

²The implementation is available at: <https://github.com/haopeng/EigenGP>

Negative Log Probability (MNLP), defined as:

$$\text{NMSE} = \frac{\sum_i (y_i - \mu_i)^2 / \sum_i (\mu_i - \bar{y})^2}{2N} \quad (16)$$

$$\text{MNLP} = \frac{1}{2N} \sum_i \left[\left(\frac{y_i - \mu_i}{\sigma_i} \right)^2 + \ln \sigma_i^2 + \ln 2\pi \right] \quad (17)$$

where y_i , μ_i and σ_i^2 are the response value, the predictive mean and variance for the i -th test point respectively, and \bar{y} is the average response value of the training data.

6.1 Approximation Quality on Synthetic Data

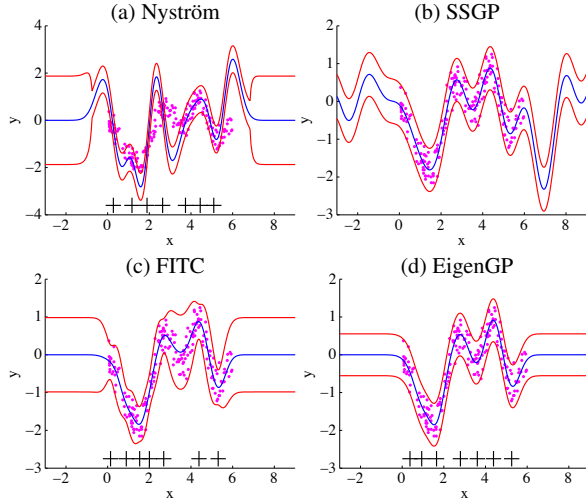


Figure 2: Predictions of four sparse GP methods. Pink dots represent training data; blue curves are predictive means; red curves are two standard deviations above and below the mean curves, and the black crosses indicate the inducing inputs.

As in Section 3.1, we use the synthetic data from the FITC paper for the comparative study. To let all the methods have the same computational complexity, we set the number of inducing inputs $M = 7$. The results are summarized in Figure 2. For the Nyström method, we used the kernel width learned from the full GP and applied K-means to choose the basis locations [Zhang *et al.*, 2008]. Figure 2(a) shows that it does not fit well. Figure 2(b) demonstrates that the prediction of SSGP oscillates outside the range of the training samples, probably due to the fact that the sinusoidal components are *global* and span the whole data range (increasing the number of basis functions would improve SSGP’s predictive performance, but increase the computational cost.). As shown by Figure 2(c), FITC fails to capture the turns of the data accurately for x near 4 while EigenGP can.

Using the full GP predictive mean as the label for $x \in [-1, 7]$ (we do not have the true y values in the test data), we compute the NMSE and MNLP of all the methods. The average results from 10 runs are reported in Table 1 (dataset 1). We have the results from two versions of the Nyström method. For the first version, the kernel width is learned from the full GP and the basis locations are chosen by K-means as before; for the second version, denoted as Nyström*, its hyperparameters are learned by evidence maximization. Note

Table 1: NMSE and MNLP on synthetic data

| NMSE | | |
|----------|-------------------|-----------------|
| Method | dataset 1 | dataset 2 |
| Nyström | 39 ± 18 | 1526 ± 769 |
| Nyström* | 2.41 ± 0.53 | 2721 ± 370 |
| FITC | 0.02 ± 0.005 | 0.50 ± 0.04 |
| SSGP | 0.54 ± 0.01 | 0.22 ± 0.05 |
| EigenGP | 0.006 ± 0.001 | 0.06 ± 0.02 |
| EigenGP* | 0.009 ± 0.002 | 0.06 ± 0.02 |

| MNLP | | |
|----------|------------------|-----------------|
| Method | dataset 1 | dataset 2 |
| Nyström | 645 ± 56 | 2561 ± 1617 |
| Nyström* | 7.39 ± 1.66 | 40 ± 5 |
| FITC | -0.07 ± 0.01 | 0.88 ± 0.05 |
| SSGP | 1.22 ± 0.03 | 0.87 ± 0.09 |
| EigenGP | -0.33 ± 0.00 | 0.40 ± 0.07 |
| EigenGP* | -0.31 ± 0.01 | 0.44 ± 0.07 |

that the evidence maximization algorithm for the Nyström approximation is novel too—developed by us for the comparative analysis. Table 1 shows that both EigenGP and EigenGP* approximate the mean of the full GP model more accurately than the other methods, in particular, several orders of magnitude better than the Nyström method.

Furthermore, we add the difference term (7) into the kernel function and denote this version of our algorithm as EigenGP⁺. It gives better predictive variance when far from the training data but its predictive mean is slightly worse than the version without this term (7); the NMSE and MNLP of EigenGP⁺ are 0.014 ± 0.001 and -0.081 ± 0.004 . Thus, on the other datasets, we only use the versions without this term (EigenGP and EigenGP*) for their simplicity and effectiveness. We also examine the performance of all these methods with a higher computational complexity. Specifically, we set $M = 10$. Again, both versions of the Nyström method give poor predictive distributions. And SSGP still leads to extra wavy patterns outside the training data. FITC, EigenGP and EigenGP⁺ give good predictions. Again, EigenGP⁺ gives better predictive variance when far from the training data, but with a similar predictive mean as EigenGP.

Finally, we compare the RVM with EigenGP* on this dataset. While the RVM gives NMSE = 0.048 in 2.0 seconds, EigenGP* achieves NMSE = 0.039 ± 0.017 in 0.33 ± 0.04 second with $M = 30$ (EigenGP performs similarly), both faster and more accurate.

6.2 Prediction Quality on Nonstationary Data

We then compare all the sparse GP methods on an one-dimensional nonstationary synthetic dataset with 200 training and 500 test samples. The underlying function is $f(x) = x \sin(x^3)$ where $x \in (0, 3)$ and the standard deviation of the white noise is 0.5. This function is nonstationary in the sense that its frequency and amplitude increase when x increases from 0. We randomly generated the data 10 times and set the number of basis points (functions) to be 14 for all the com-

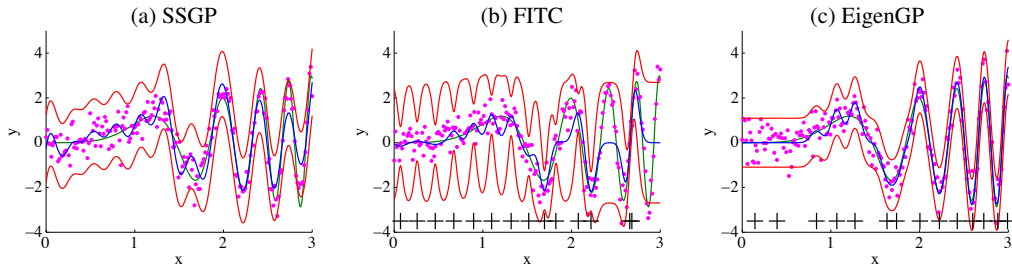


Figure 3: Predictions on nonstationary data. The pink dots correspond to noisy data around the true function $f(x) = x \sin(x^3)$, represented by the green curves. The blue and red solid curves correspond to the predictive means and \pm two standard deviations around the means. The black crosses near the bottom represent the estimated basis points for FITC and EigenGP.

petitive methods. Using the true function value as the label, we compute the means and the standard errors of NMSE and MNLP as in Table 1 (dataset 2). For the Nyström method, the marginal likelihood optimization leads to much smaller error than the K-means based approach. However, both of them fare poorly when compared with the alternative methods. Table 1 also shows that EigenGP and EigenGP* achieve a striking $\sim 25,000$ fold error reduction compared with Nyström*, and a ~ 10 -fold error reduction compared with the second best method, SSGP. RVMs gave NMSE 0.0111 ± 0.0004 with 1.4 ± 0.05 seconds, averaged over 10 runs, while the results of EigenGP* with $M = 50$ are NMSE 0.0110 ± 0.0006 with 0.89 ± 0.1042 seconds (EigenGP gives similar results).

We further illustrate the predictive mean and standard deviation on a typical run in Figure 3. As shown in Figure 3(a), the predictive mean of SSGP contains reasonable high frequency components for $x \in (2, 3)$ but, as a stationary GP model, these high frequency components give extra wavy patterns in the left region of x . In addition, the predictive mean on the right is smaller than the true one, probably affected by the small dynamic range of the data on the left. Figure 3(b) shows that the predictive mean of FITC at $x \in (2, 3)$ has lower frequency and smaller amplitude than the true function—perhaps influenced by the low-frequency part on the left $x \in (0, 2)$. Actually because of the low-frequency part, FITC learns a large kernel width η ; the average kernel width learned over the 10 runs is 207.75. This large value affects the quality of learned basis points (e.g., lacking of basis points for the high frequency region on the right). By contrast, using the same initial kernel width as FITC, EigenGP learns a suitable kernel width—on average, $\eta = 0.07$ —and provides good predictions as shown in Figure 3(c).

6.3 Accuracy vs. Time on Real Data

To evaluate the trade-off between prediction accuracy and computational cost, we use three large real datasets. The first dataset is California Housing [Pace and Barry, 1997]. We randomly split the 8 dimensional data into 10,000 training and 10,640 test points. The second dataset is Physicochemical Properties of Protein Tertiary Structures (PPPTS) which can be obtained from Lichman [2013]. We randomly split the 9 dimensional data into 20,000 training and 25,730 test points. The third dataset is Pole Telecomm that was used in Lázaro-Gredilla *et al.* [2010]. It contains 10,000 training

and 5000 test samples, each of which has 26 features. We set $M = 25, 50, 100, 200, 400$, and the maximum number of iterations in optimization to be 100 for all methods.

The NMSE, MNLP and the training time of these methods are shown in Figure 4. In addition, we ran the Nyström method based on the marginal likelihood maximization, which is better than using K-means to set the basis points. Again, the Nyström method performed orders of magnitude worse than the other methods: with $M = 25, 50, 100, 200, 400$, on California Housing, the Nyström method uses 146, 183, 230, 359 and 751 seconds for training, respectively, and gives the NMSE 917, 120, 103, 317, and 95; on PPPTS, the training times are 258, 304, 415, 853 and 2001 seconds and the NMSEs are 1.7×10^5 , 6.4×10^4 , 1.3×10^4 , 8.1×10^3 , and 8.1×10^3 ; and on Pole Telecomm, the training times are 179, 205, 267, 478, and 959 seconds and the NMSEs are 2.3×10^4 , 5.8×10^3 , 3.8×10^3 , 4.5×10^2 and 84. The MNLPs are consistently large, and are omitted here for simplicity.

For RVMs, we include cross-validation in its training time because choosing an appropriate kernel width is crucial for RVM. Since RVM learns the number of basis functions automatically from the data, in Figure 4 it shows a single result for each dataset. EigenGP achieves the lowest prediction error using shorter time. Compared with EigenGP based on the sequential optimization, EigenGP* achieves similar errors, but takes longer because the joint optimization is more expensive.

7 Conclusions

In this paper we have presented a simple yet effective sparse Gaussian process method, EigenGP, and applied it to regression. Despite its similarity to the Nyström method, EigenGP can improve its prediction quality by several orders of magnitude. EigenGP can be easily extended to conduct online learning by either using stochastic gradient descent to update the weights of the eigenfunctions or applying the online VB idea for GPs [Hensman *et al.*, 2013].

Acknowledgments

This work was supported by NSF ECCS-0941043, NSF CAREER award IIS-1054903, and the Center for Science of Information, an NSF Science and Technology Center, under grant agreement CCF-0939370.

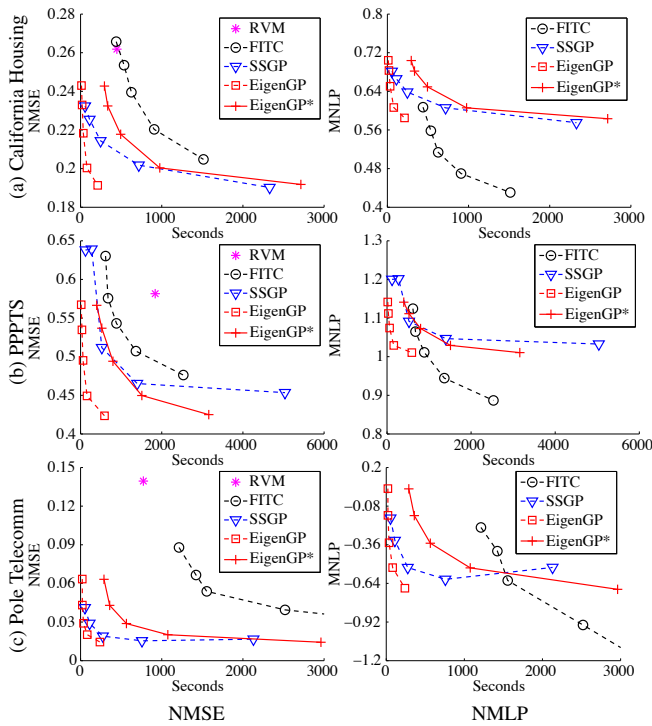


Figure 4: NMSE and MNLP vs. training time. Each method (except RVMs) has five results associated with $M = 25, 50, 100, 200, 400$, respectively. In (c), the fifth result of FITC is out of the range; the actual training time is 3485 seconds, the NMSE 0.033, and the MNLP -1.27 . The values of MNLP for RVMs are 1.30, 1.25 and 0.95, respectively.

References

- [Cressie and Johannesson, 2008] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, February 2008.
- [Csató and Opper, 2002] Lehel Csató and Manfred Opper. Sparse online Gaussian processes. *Neural Computation*, 14:641–668, March 2002.
- [de Leeuw, 2007] Jan de Leeuw. Derivatives of generalized eigen systems with applications. In *Department of Statistics Papers*. Department of Statistics, UCLA, UCLA, 2007.
- [Faul and Tipping, 2001] Anita C. Faul and Michael E. Tipping. Analysis of sparse Bayesian learning. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.
- [Hensman et al., 2013] James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 282–290, 2013.
- [Higdon, 2002] Dave Higdon. Space and space-time modeling using process convolutions. In Clive W. Anderson, Vic Barnett, Philip C. Chatwin, and Abdel H. El-Shaarawi, editors, *Quantitative methods for current environmental issues*, pages 37–56. Springer Verlag, 2002.
- [Hoegaerts et al., 2005] Luc Hoegaerts, Johan A. K. Suykens, Joos Vandewalle, and Bart De Moor. Subset based least squares subspace regression in RKHS. *Neurocomputing*, 63:293–323, January 2005.
- [Lázaro-Gredilla et al., 2010] Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl E. Rasmussen, and Aníbal R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.
- [Lichman, 2013] Moshe Lichman. UCI machine learning repository, 2013.
- [MacKay, 1992] David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.
- [Marzouk and Najm, 2009] Youssef M. Marzouk and Habib N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862 – 1902, 2009.
- [Minka, 2001] Thomas P. Minka. Old and new matrix algebra useful for statistics. Technical report, MIT Media Lab, 2001.
- [Pace and Barry, 1997] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3):291–297, 1997.
- [Paciorek and Schervish, 2004] Christopher J. Paciorek and Mark J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.
- [Qi et al., 2010] Yuan Qi, Ahmed H. Abdel-Gawad, and Thomas P. Minka. Sparse-posterior Gaussian processes for general likelihoods. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.
- [Quiñonero-Candela and Rasmussen, 2005] Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1935–1959, 12 2005.
- [Snelson and Ghahramani, 2006] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*. MIT press, 2006.
- [Tipping, 2000] Michael E. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.
- [Titsias, 2009] Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *The 12th International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.
- [Williams and Barber, 1998] Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.
- [Williams and Seeger, 2001] Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, volume 13. MIT Press, 2001.
- [Yan and Qi, 2010] Feng Yan and Yuan Qi. Sparse Gaussian process regression via ℓ_1 penalization. In *Proceedings of 27th International Conference on Machine Learning*, pages 1183–1190, 2010.
- [Zhang et al., 2008] Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1232–1239. ACM, 2008.