

---

# GAUSSIAN PROCESS REGRESSION WITH LOCAL EXPLANATION

---

A PREPRINT

**Yuya Yoshikawa**

Software Technology and Artificial Intelligence Research Laboratory  
Chiba Institute of Technology  
yoshikawa@stair.center

**Tomoharu Iwata**

NTT Communication Science Laboratories  
tomoharu.iwata.gy@hco.ntt.co.jp

November 26, 2020

## ABSTRACT

Gaussian process regression (GPR) is a fundamental model used in machine learning. Owing to its accurate prediction with uncertainty and versatility in handling various data structures via kernels, GPR has been successfully used in various applications. However, in GPR, how the features of an input contribute to its prediction cannot be interpreted. Herein, we propose GPR with local explanation, which reveals the feature contributions to the prediction of each sample, while maintaining the predictive performance of GPR. In the proposed model, both the prediction and explanation for each sample are performed using an easy-to-interpret locally linear model. The weight vector of the locally linear model is assumed to be generated from multivariate Gaussian process priors. The hyperparameters of the proposed models are estimated by maximizing the marginal likelihood. For a new test sample, the proposed model can predict the values of its target variable and weight vector, as well as their uncertainties, in a closed form. Experimental results on various benchmark datasets verify that the proposed model can achieve predictive performance comparable to those of GPR and superior to that of existing interpretable models, and can achieve higher interpretability than them, both quantitatively and qualitatively.

**Keywords** Gaussian processes · interpretable machine learning · locally linear models · explainability · feature relevance

## 1 Introduction

Gaussian processes (GPs) have been well studied for constructing probabilistic models as priors of nonlinear functions in the machine learning (ML) community. They have demonstrated great success in various problem settings, such as regression [Rasmussen, 2003, Wilson et al., 2012], classification [Rasmussen, 2003, Csató et al., 2000], time-series forecasting [Roberts et al., 2013], and black-box optimization [Snoek et al., 2012]. A fundamental model on GPs is Gaussian process regression (GPR) [Rasmussen, 2003]; owing to its high predictive performances and versatility in using various data structures via kernels, it has been used in not only the ML community, but also in various other research areas, such as finance [Gonzalvez et al., 2019], geostatistics [Camps-Valls et al., 2016], material science [Zhang et al., 2020] and medical science [Cheng et al., 2017, Futoma, 2018].

GPR is defined on an infinite-dimensional feature space via kernel functions. Therefore, it requires the values of the kernels defined on pairs of samples, i.e., a covariance matrix of the samples as an input, rather than the samples themselves. Owing to the nonlinearity of the kernel, GPR enables nonlinear predictions. In terms of interpretability, the covariance is useful for understanding the relationship between the samples; however, since the kernels make the features invisible, GPR cannot explain which features contribute to the predictions, like linear regression models. Therefore, it prevents us from judging whether the predictions by GPR are reasonable and performed by fair decision.

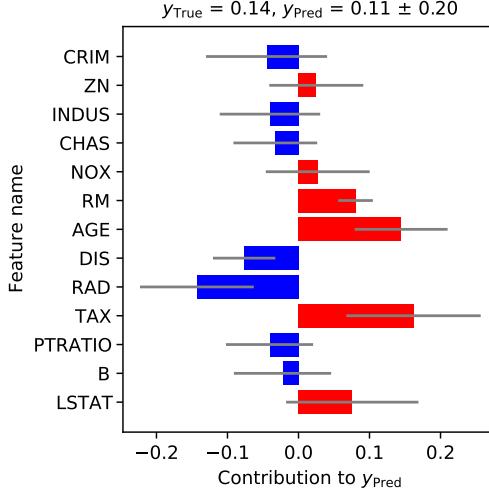


Figure 1: Example of explanation for prediction by GPX for a sample on the Boston housing dataset [Harrison Jr and Rubinfeld, 1978]. Red and blue bars indicate positive and negative predicted feature contributions, respectively, and error bars indicates predicted standard deviation of the feature contributions. We provide further examples and feature description in Appendix A.

For the interpretability of ML, several methodologies that explain the features that contribute to the outputs of prediction models, including GPR, have been proposed; in this case, the prediction models are often regarded as black-boxes [Molnar, 2019, Chen et al., 2018]. Their representative methods are local interpretable model-agnostic explanations (LIME) [Ribeiro et al., 2016] and Shapley additive explanations (SHAP) [Lundberg and Lee, 2017], which approximate the prediction for each test sample by a locally linear explanation model. Since the weights of the learned explanation model represent feature contributions to the prediction, they can assist ML practitioners and scientists to understand the behavior of the prediction models and the functionality of the features in the prediction. However, some limitations exist in these methods. First, because the forms of the prediction and explanation models differ, it is unsure whether the estimated feature contributions reflect those of the prediction model. Second, because the explanation model is learned on each test sample, it may not obtain consistent explanations on similar samples.

To overcome the aforementioned limitations, we propose a novel framework for GP-based regression models, *Gaussian process regression with local explanation*, called *GPX*, which reveals the feature contributions to the prediction for each sample, while maintaining the predictive performance of GPR. In GPX, both the prediction and explanation for each sample are performed using an easy-to-interpret locally linear model. Therefore, no gap exists between the prediction and explanation. The weight vector of the locally linear model is assumed to be generated from multivariate GP priors [Álvarez et al., 2012]. As the multivariate GP priors have a covariance function defined as kernels on the samples, GPX ensures that similar samples have similar weights. The hyperparameters of GPX are estimated by maximizing the marginal likelihood, in which the weight vectors for all the training samples are integrated out. For a test sample, the predictions with their uncertainties of the target variable and weight vector are obtained by computing their predictive distributions. The explanation for the predicted target variable is provided using the estimated weight vector with uncertainty, as shown in Figure 1. Depicting the explanation with uncertainty helps users of GPX judge the reasonability of the predicted weights.

In experiments, we evaluated GPX both qualitatively and quantitatively in terms of predictive performance and interpretability on various benchmark datasets. The experimental results show that 1) GPX can achieve predictive errors comparable to GPR and lower errors compared with existing interpretable methods, 2) it can outperform model-agnostic interpretable methods and locally linear methods in terms of three interpretability measurements, and 3) the feature contributions produced by GPX are appropriate.

## 2 Related Work

Linear regression models are simple types of interpretable models, as their weight for each feature directly represents the contribution to the output of the models if the feature has a binary value, where the contribution helps us to interpret

the effectiveness of the feature on a task. For real-valued features, the contribution for each feature is calculated as the product of the feature value and its corresponding weight. A number of studies introducing various regularizations have been conducted to produce methods such as the ridge regression [Hoerl and Kennard, 1970] and lasso [Tibshirani, 1996] methods. These models have a *global* weight vector that is shared across all samples. In kernel methods, automatic relevance determination (ARD), which considers the global relevance of each feature contained in kernel functions, is widely used [Neal, 2012, Wipf and Nagarajan, 2008]. The above approaches are beneficial in understanding global effectiveness of features. However, since linear regression models often does not fit real complicated problems, resulting in low predictive accuracy, the estimated weights may be also unreliable. In addition, these approaches cannot estimate weights/relevances appropriate for individual samples, which means that they cannot cope with the need of the significant changes of the weights/relevances among the samples. For example, as shown in the result on Digits dataset (Figure 3), this is crucial in image classification tasks that important pixels (i.e., features) and their weights change depending on individual images (i.e., samples).

On the other hand, some *locally* linear models for regression have been proposed, such as the network lasso [Hallac et al., 2015] and localized lasso [Yamada et al., 2017], which have a weight vector for each sample. Therefore, these methods can avoid the drawbacks of globally linear models. To receive the benefit, we focus on generating predictions with explanations using locally linear models. In the network and localized lasso, the weights of locally linear models are estimated via optimization with network-based regularization, where the network must be defined on samples in advance. If the network is not provided, as assumed in standard regression problems, we can construct a  $k$ -nearest-neighbor graph of samples to create a network, where  $k$  is a hyperparameter that must be optimized via cross validation. Meanwhile, GPX can estimate weights and their uncertainties without constructing graphs by assuming that weights are determined by functions generated from GPs.

For GP models, [Paananen et al., 2019] proposed feature selection methods that quantify the feature relevances by measuring the difference between the predictive posterior for each sample and that for its vicinity. The feature relevance can be used for evaluating the importance of the feature in prediction. Meanwhile, the feature contribution produced by GPX indicates whether the feature makes a positive or negative contribution to an individual prediction and how much it contributes.

With regard to research on deep neural networks (DNNs), a number of studies have been conducted on making predictions generated by DNNs interpretable [Chen et al., 2019, Arras et al., 2017, Ying et al., 2019]. Some of these studies have developed methods that make interpretable predictions by generating locally linear models for each sample using DNNs [Melis and Jaakkola, 2018, Schwab et al., 2019, Yoshikawa and Iwata, 2020]. These concepts inspired our study, but we formalize our model without DNNs. To the best of our knowledge, our study is the first to develop a GP-based regression model with local explanations. Compared to the DNN-based locally linear models, GPX has mainly two benefits: 1) GPX can produce the feature contributions with uncertainty, and 2) GPX can be straightforwardly used for tasks in which GPR can be used advantageously, such as Bayesian optimization [Snoek et al., 2012, Golovin et al., 2017].

[Wilson et al., 2012] proposed Gaussian process regression networks, which have similar structure with multi-layer perceptron. Herein, the weights of the networks are generated from GPs. Although their idea is related to GPX, their work did not focus on improving the explainability of the predictions by GPR.

### 3 Proposed Model

In this section, we describe the proposed model, i.e., *Gaussian process regression with local explanation*, called *GPX*.

We consider a scalar-valued regression problem. Suppose that training data  $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{z}_i)\}_{i=1}^n$  containing  $n$  samples is provided.  $\mathbf{x}_i \in \mathcal{X}$  is an original input representing the  $i$ th sample, where  $\mathcal{X}$  is an original input space. Although a typical representation for  $\mathbf{x}_i$  is a vector, it can be any data representation on which kernel functions are defined, such as graphs [Vishwanathan et al., 2010] and sets [Muandet et al., 2012, Yoshikawa et al., 2014].  $y_i \in \mathbb{R}$  is a target variable for the sample.  $\mathbf{z}_i \in \mathbb{R}^d$  is a  $d$ -dimensional vector of simplified representation for  $\mathbf{x}_i$ . Because GPX explains the prediction via a simplified representation, the meaning of each dimension of  $\mathbf{z}_i$  should be easily understood by humans, e.g., tabular data and bag-of-words representation for text.  $\mathbf{z}_i$  is an optional input; therefore, if  $\mathbf{x}_i$  can be used as a simplified representation, one can define  $\mathbf{z}_i = \mathbf{x}_i$ . Let us denote  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ ,  $\mathbf{y} = (y_i)_{i=1}^n \in \mathbb{R}^n$  and  $\mathbf{Z} = (\mathbf{z}_i)_{i=1}^n \in \mathbb{R}^{n \times d}$ .

In GPX, both the prediction of target variables  $\mathbf{y}$  and their explanations are performed via easy-to-interpret locally linear models, i.e., target variable  $y_i$  for the  $i$ th sample is assumed to be obtained using locally linear function  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ , defined as follows:

$$f_i(\mathbf{z}_i) = \mathbf{w}_i^\top \mathbf{z}_i + \epsilon_y, \quad (1)$$

where  $\mathbf{w}_i \in \mathbb{R}^d$  is a  $d$ -dimensional weight vector for the  $i$ th sample, and  $\epsilon_y \sim \mathcal{N}(0, \sigma_y^2)$  is a Gaussian noise with variance  $\sigma_y^2 > 0$ . Here, the explanation for the  $i$ th sample is obtained using either weight vector  $\mathbf{w}_i$  or feature contributions  $\phi_i = (w_{il}z_{il})_{l=1}^d$ .

Estimating  $\mathbf{W} = (\mathbf{w}_i)_{i=1}^n \in \mathbb{R}^{n \times d}$  without any constraints is an ill-posed problem because the number of free parameters in  $\mathbf{W}$ ,  $nd$ , is larger than that of target variable  $n$ . To avoid this problem in GPX, we assume that functions determining  $\mathbf{W}$  are generated from a multivariate GP. More specifically, weight vector  $\mathbf{w}_i$  for the  $i$ th sample is obtained as follows:

$$\mathbf{w}_i = \mathbf{g}(\mathbf{x}_i) + \boldsymbol{\epsilon}_w, \quad (2)$$

where  $\boldsymbol{\epsilon}_w \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I}_d)$  is a  $d$ -dimensional Gaussian noise with variance  $\sigma_w^2 > 0$ , and  $\mathbf{I}_d$  is an identity matrix of order  $d$ . Here, vector-valued function  $\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^d$  is a function that determines the weight vector for each sample, and each element of  $\mathbf{g}$  is generated from a univariate GP independently, as follows:

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_d(\mathbf{x}))^\top, \quad (3)$$

where

$$g_l(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k_\theta(\mathbf{x}, \mathbf{x}')), \quad (4)$$

where  $m(\mathbf{x})$  is the mean function, and  $k_\theta(\mathbf{x}, \mathbf{x}')$  is the covariance function with set of parameters  $\theta$ . Herein, we use zero mean function for  $m(\mathbf{x})$ . Covariance function  $k_\theta(\mathbf{x}, \mathbf{x}')$  is a kernel function defined on two inputs  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ . For example, one can use a scaled RBF kernel with parameters  $\theta = \{\theta_1, \theta_2\}$  as the kernel function when  $\mathbf{x}, \mathbf{x}'$  are vectors, defined as follows:

$$k_\theta(\mathbf{x}, \mathbf{x}') = \theta_1 \exp\left(-\frac{1}{\theta_2} \|\mathbf{x} - \mathbf{x}'\|_2^2\right) \quad (\theta_1, \theta_2 > 0). \quad (5)$$

By using  $\mathbf{g}$  generated as such, GPX ensures that two similar samples, i.e., those having a large kernel value, have similar weight vectors.

We let  $\mathbf{G} = (\mathbf{g}(\mathbf{x}_i)^\top)_{i=1}^n \in \mathbb{R}^{n \times d}$ . Based on the generative process above, the joint distribution of GPX is written as follows:

$$p(\mathbf{y}, \mathbf{W}, \mathbf{G} | \mathbf{X}, \mathbf{Z}) = p(\mathbf{G} | \mathbf{X}) \prod_{i=1}^n p(y_i | \mathbf{w}_i, \mathbf{z}_i) p(\mathbf{w}_i | \mathbf{G}_{i,\cdot}), \quad (6)$$

where

$$p(y_i | \mathbf{w}_i, \mathbf{z}_i) = \mathcal{N}(y_i | \mathbf{w}_i^\top \mathbf{z}_i, \sigma_y^2), \quad (7)$$

$$p(\mathbf{w}_i | \mathbf{G}_{i,\cdot}) = \mathcal{N}(\mathbf{w}_i | \mathbf{g}(\mathbf{x}_i), \sigma_w^2 \mathbf{I}_d), \quad (8)$$

$$p(\mathbf{G} | \mathbf{X}) = \prod_{l=1}^d \mathcal{N}(\mathbf{G}_{\cdot,l} | \mathbf{0}, \mathbf{K}), \quad (9)$$

where  $\mathbf{G}_{i,\cdot}$  and  $\mathbf{G}_{\cdot,l}$  denote the  $i$ th row and  $l$ th column vectors of  $\mathbf{G}$ , respectively, and  $\mathbf{K} = (k_\theta(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$  is a Gram matrix of order  $n$ , which is identical to the requirement in GPR.

## 4 Training and Prediction

In this section, we describe the derivation of the marginal likelihood of GPX, the hyperparameter estimation for GPX, and the derivation of the predictive distributions of target variables and weight vectors for test samples.

### Marginal likelihood

To ease the derivation of the marginal likelihood, we first modified the formulation of the joint distribution (6), while maintaining its mathematical meanings, as follows:

$$\begin{aligned} p(\mathbf{y}, \mathbf{W}, \mathbf{G} | \mathbf{X}, \mathbf{Z}) \\ = \mathcal{N}(\text{vec}(\mathbf{G}) | \mathbf{0}, \bar{\mathbf{K}}) \mathcal{N}(\text{vec}(\mathbf{W}) | \text{vec}(\mathbf{G}), \sigma_w^2 \mathbf{I}_{nd}) \times \mathcal{N}(\mathbf{y} | \bar{\mathbf{Z}} \text{vec}(\mathbf{W}), \sigma_y^2 \mathbf{I}_{nd}), \end{aligned} \quad (10)$$

where  $\bar{\mathbf{K}}$  is a block diagonal matrix of order  $nd$  whose block is  $\mathbf{K}$ , and  $\text{vec}(\cdot)$  is a function that flattens the input matrix in a column-major order. Here,

$$\bar{\mathbf{Z}} = (\text{diag}(\mathbf{Z}_{\cdot,1}), \text{diag}(\mathbf{Z}_{\cdot,2}), \dots, \text{diag}(\mathbf{Z}_{\cdot,d})) \in \mathbb{R}^{n \times nd}, \quad (11)$$

where  $\text{diag}(\cdot)$  is a diagonal matrix whose diagonal elements possess the values of the input vector. In (6),  $d$  functions that output  $n$ -dimensional column vectors in  $\mathbf{W}$  are generated from GPs; however, in (10), it is rewritten such that a single function that outputs an  $nd$ -dimensional flatten vector  $\text{vec}(\mathbf{W})$  is generated from a single GP. Consequently, the likelihood of target variables  $\mathbf{y}$  can be rewritten as a single multivariate normal distribution.

Subsequently, we derived the marginal likelihood by integrating out  $\mathbf{G}$  and  $\mathbf{W}$  in (10). Owing to the property of normal distributions, it can be obtained analytically, as follows:

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}, \mathbf{Z}) &= \iint p(\mathbf{y}, \mathbf{W}, \mathbf{G} | \mathbf{X}, \mathbf{Z}) d\mathbf{W} d\mathbf{G} \\ &= \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{C}), \end{aligned} \quad (12)$$

where

$$\begin{aligned} \mathbf{C} &= \sigma_y^2 \mathbf{I}_n + \bar{\mathbf{Z}} (\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd}) \bar{\mathbf{Z}}^\top \\ &= \sigma_y^2 \mathbf{I}_n + (\mathbf{K} + \sigma_w^2 \mathbf{I}_n) \odot \mathbf{Z} \mathbf{Z}^\top. \end{aligned} \quad (13)$$

### Hyperparameter estimation

If  $k_\theta(\mathbf{x}, \mathbf{x}')$  is differentiable with respect to  $\theta$ , all the hyperparameters, i.e.,  $\theta$ ,  $\sigma_w$ , and  $\sigma_y$ , can be estimated by maximizing the logarithm of the marginal likelihood with respect to them for the training data using gradient-based optimization methods, e.g., L-BFGS [Liu and Nocedal, 1989].

### Predictive distributions

For a new test sample  $(\mathbf{x}_*, \mathbf{z}_*)$ , our goal is to infer the predictive distributions of target variable  $y_*$  and weight vector  $\mathbf{w}_*$ . First, the predictive distribution of  $y_*$  is obtained similarly as in the standard GPR, as follows:

$$p(y_* | \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \mathcal{N}(y_* | \mathbf{c}_*^\top \mathbf{C}^{-1} \mathbf{y}, c_{**} - \mathbf{c}_*^\top \mathbf{C}^{-1} \mathbf{c}_*), \quad (14)$$

where  $\mathbf{c}_* = (k_\theta(\mathbf{x}_*, \mathbf{x}_i) \mathbf{z}_*^\top \mathbf{z}_i)_{i=1}^n \in \mathbb{R}^n$  and  $c_{**} = \sigma_y^2 + (k_\theta(\mathbf{x}_*, \mathbf{x}_*) + \sigma_w^2) \mathbf{z}_*^\top \mathbf{z}_* \in \mathbb{R}$ .

Second, the predictive distribution of  $\mathbf{w}_*$  is obtained by solving the following integral:

$$p(\mathbf{w}_* | \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \int p(\mathbf{w}_* | \mathbf{W}, \mathbf{X}, \mathbf{x}_*) p(\mathbf{W} | \mathcal{D}) d\mathbf{W}, \quad (15)$$

$$p(\mathbf{w}_* | \mathbf{W}, \mathbf{X}, \mathbf{x}_*) = \mathcal{N}(\mathbf{w}_* | \text{Avec}(\mathbf{W}), \bar{\mathbf{c}}_{**} - \mathbf{A}\bar{\mathbf{k}}_*), \quad (16)$$

$$p(\mathbf{W} | \mathcal{D}) = \mathcal{N}(\text{vec}(\mathbf{W}) | \sigma_y^{-2} \mathbf{S} \bar{\mathbf{Z}}^\top \mathbf{y}, \mathbf{S}), \quad (17)$$

where we define  $\mathbf{A} = \bar{\mathbf{k}}_*^\top (\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd})^{-1}$ ,  $\mathbf{S} = \mathbf{L} - \mathbf{L} \bar{\mathbf{Z}}^\top \mathbf{C}^{-1} \bar{\mathbf{Z}} \mathbf{L}$ ,  $\mathbf{L} = \bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd}$ , and  $\bar{\mathbf{c}}_{**} = (k_\theta(\mathbf{x}_*, \mathbf{x}_*) + \sigma_w^2) \mathbf{I}_d$ .  $\bar{\mathbf{k}}_*$  is an  $nd$ -by- $d$  block matrix, where each block is an  $n$ -by-1 matrix, and  $(l, l)$ -block of the block matrix is  $(k_\theta(\mathbf{x}_*, \mathbf{x}_i))_{i=1}^n$  for  $l = 1, 2, \dots, d$ , and the other blocks are zero matrices. Solving the integral analytically according to the property of the normal distributions, we obtain

$$p(\mathbf{w}_* | \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \mathcal{N}(\mathbf{w}_* | \sigma_y^{-2} \mathbf{A} \bar{\mathbf{Z}}^\top \mathbf{y}, \bar{\mathbf{c}}_{**} - \mathbf{A} \bar{\mathbf{k}}_* + \mathbf{A} \mathbf{S} \mathbf{A}^\top). \quad (18)$$

We provide the detailed derivation of predictive distributions (14) and (18) in Appendix B.

The marginal likelihood (13) and the predictive distribution for  $y_*$  (14) are similar to those of GPR, except that GPX can obtain the predictive distribution for  $\mathbf{w}_*$  (18). Since GPX can be used with the same input as GPR if  $\mathbf{Z} = \mathbf{X}$ , it can be employed in existing ML models, instead of GPR.

### Computational efficiency

As with ordinary GPR, the computational cost of GPX is dominated by the inverse computation. The computation of  $\mathbf{A}$  requires inverting a square matrix of order  $nd$ ,  $\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd}$ . However, because the matrix is block diagonal and every diagonal block comprises  $\mathbf{K} + \sigma_w^2 \mathbf{I}_n$ , a square matrix of order  $n$ ,  $\mathbf{A}$  can be obtained by inverting  $\mathbf{K} + \sigma_w^2 \mathbf{I}_n$  only once. The remaining inverse matrix  $\mathbf{C}^{-1}$  is of order  $n$ . Therefore, all the inverse matrices appearing in GPX can be obtained using a naive implementation with a computational complexity of  $\mathcal{O}(n^3)$ , which is the same as that in GPR. To significantly reduce the computational cost, efficient computation methods for GPR, such as the inducing variable method [Titsias, 2009] and KISS-GP [Wilson and Nickisch, 2015], can be used for GPX. In addition, because  $\bar{\mathbf{k}}_*$ ,  $\bar{\mathbf{K}}$  and  $\bar{\mathbf{Z}}$  are sparse matrices, one can obtain the predictive distributions efficiently using libraries for sparse matrix computation.

## 5 Experiments

In this section, we demonstrate the effectiveness of the proposed model, GPX, quantitatively and qualitatively, by comparing various interpretable models. Through a quantitative evaluation, we evaluated the models based on the following perspectives:

- **Accuracy:** How accurate is the prediction of the interpretable model?
- **Faithfulness:** Are feature contributions indicative of “true” importance?
- **Sufficiency:** Do  $k$ -most important features reflect the prediction?
- **Stability:** How consistent are the explanations for similar or neighboring examples?

In addition, we qualitatively evaluated whether the feature contributions produced by the models were appropriate by visualizing them. Subsequently, we experimentally compared the computational efficiency of the models.

All the experiments were done with a computer with Intel Xeon Gold 6132 2.6GHz CPU with 16 cores, and 120GB of main memory.

### 5.1 Preparation

#### Datasets

We used eight datasets in the UCI machine learning repository [Dua and Graff, 2017], referred to as Digits, Abalone, Diabetes, Boston, Fish, Wine, Paper and Drug in our experiments. We provide the details of the datasets in Appendix C. Digits dataset is originally a classification dataset for recognizing handwritten digits from 0 to 9. To use it as a regression problem, we transformed the labels into target variables  $y$  of scalar values, i.e., the target variables for the labels from 0 to 4 were  $-1$ , and those for the remaining labels were  $1$ . With Paper and Drug datasets whose samples were represented as sentences, the original input  $\mathbf{X}$  and the simplified input  $\mathbf{Z}$  differed, i.e., we used the 512-dimensional sentence vectors obtained using Sentence Transformers [Reimers and Gurevych, 2020] as  $\mathbf{X}$ , while we used bag-of-words binary vectors for the sentences as  $\mathbf{Z}$ . Each of the remaining datasets had the same  $\mathbf{X}$  and  $\mathbf{Z}$ . In all the datasets, the values of  $\mathbf{X}$  and  $y$  were standardized before training and prediction. For a quantitative evaluation of each dataset, we evaluated the average scores over five experiments performed on different training/test splittings, where the training set was 80% of the entire dataset, whereas the remaining was the test set.

#### GPX setup

In GPX, we consistently used a scaled RBF kernel defined as (5). The hyperparameters of GPX were estimated based on the method described in Section 4, where they were initialized with  $\theta_1 = 1.0$ ,  $\sigma_y = 0.1$  and  $\sigma_w = 0.1$ . In addition, we initialized bandwidth parameter  $\theta_2$  using median heuristics [Garreau et al., 2017].

#### Comparing methods

We compared GPX with several methods with globally or locally linear weights that can be used as interpretable feature contributions or relevances for predictions. Lasso [Tibshirani, 1996] and Ridge [Hoerl and Kennard, 1970] are standard linear regression models with  $\ell_1$  and  $\ell_2$  regularizers, respectively, where their weights are globally shared across all samples. ARD is a GPR model with ARD kernel that identifies the relevance of each feature. The network lasso (“Network” for short) is a locally linear model that regularizes the weights of nodes such that neighboring nodes in a network have similar weights [Hallac et al., 2015]. In our case, each node represents a sample, and the network is a  $k$ -nearest neighbor graph on the samples based on the cosine similarity on  $\mathbf{X}$ . The localized lasso (“Localized” for short) is an extension of the network lasso; it can estimate the sparse and exclusive weights of each sample by further incorporating an  $\ell_{1,2}$  regularizer into the network lasso [Yamada et al., 2017].

To compare model-agnostic interpretable methods with GPX in terms of explainability for prediction, we used LIME [Ribeiro et al., 2016] and Kernel SHAP [Lundberg and Lee, 2017], which produce a locally linear model for each test sample to explain the prediction by a black-box prediction model. For a fair comparison, we used GPR with RBF kernel as the prediction model. In addition, we used a Kullback-Leibler (KL) divergence-based feature selection method for GPR with ARD kernel (“KL” for short) [Paananen et al., 2019]. The hyperparameters of GPR were estimated by maximizing marginal likelihood, similarly as for GPX. Meanwhile, those of the remaining comparing methods were optimized by grid search. We provide the detailed description of the comparing methods in Appendix D.

Table 1: Average and standard deviation of mean squared errors (MSEs) for the predictions of target variables on each dataset (lower scores are better). Values in bold typeface are not statistically different (at 5% level) from the best performing method in each row according to a paired t-test. As GPR is not an interpretable model, we only included its performances in this table to show that those of GPX and GPR are similar.

	GPX (ours)	Localized	Network	ARD	Lasso	Ridge	GPR
Digits	<b>0.078 ± 0.010</b>	0.135 ± 0.042	0.163 ± 0.033	0.258 ± 0.371	0.399 ± 0.028	0.398 ± 0.024	0.074 ± 0.008
Abalone	<b>0.428 ± 0.036</b>	0.519 ± 0.023	0.534 ± 0.016	<b>0.428 ± 0.026</b>	0.477 ± 0.052	0.477 ± 0.053	0.427 ± 0.034
Diabetes	<b>0.493 ± 0.041</b>	0.610 ± 0.062	0.667 ± 0.091	<b>0.492 ± 0.042</b>	0.504 ± 0.039	0.503 ± 0.040	0.490 ± 0.048
Boston	<b>0.116 ± 0.053</b>	0.208 ± 0.070	0.233 ± 0.062	<b>0.115 ± 0.047</b>	0.293 ± 0.078	0.284 ± 0.081	0.116 ± 0.052
Fish	<b>0.370 ± 0.061</b>	0.479 ± 0.051	0.523 ± 0.054	<b>0.376 ± 0.065</b>	0.437 ± 0.055	0.437 ± 0.055	0.375 ± 0.066
Wine	<b>0.579 ± 0.048</b>	0.715 ± 0.024	0.908 ± 0.051	<b>0.628 ± 0.053</b>	0.723 ± 0.039	0.712 ± 0.044	0.605 ± 0.046
Paper	<b>0.806 ± 0.054</b>	0.981 ± 0.058	0.919 ± 0.088	<b>0.815 ± 0.092</b>	0.821 ± 0.047	0.936 ± 0.057	0.762 ± 0.087
Drug	<b>0.835 ± 0.027</b>	1.011 ± 0.027	1.072 ± 0.019	<b>0.845 ± 0.033</b>	0.875 ± 0.037	0.911 ± 0.036	0.844 ± 0.033

Table 2: Average and standard deviation of faithfulness scores on each dataset (higher scores are better). This table can be interpreted similarly as Table 1. For Wine dataset, we could not measure the scores of the methods except for GPX owing to computational time limitations. For Paper and Drug datasets, we did not evaluate the scores as changes in  $Z$  cannot reflect  $X$ .

	GPX (ours)	GPR+LIME	GPR+SHAP	KL	Localized	Network	ARD
Digits	<b>0.888 ± 0.003</b>	0.384 ± 0.038	0.651 ± 0.013	0.757 ± 0.022	0.300 ± 0.029	0.352 ± 0.021	-0.059 ± 0.012
Abalone	<b>0.898 ± 0.017</b>	0.775 ± 0.024	<b>0.914 ± 0.007</b>	0.551 ± 0.024	0.432 ± 0.042	0.497 ± 0.033	0.288 ± 0.107
Diabetes	<b>0.966 ± 0.008</b>	0.844 ± 0.027	0.928 ± 0.010	0.558 ± 0.050	0.340 ± 0.074	0.365 ± 0.086	0.593 ± 0.027
Boston	<b>0.898 ± 0.026</b>	0.693 ± 0.035	<b>0.869 ± 0.009</b>	0.518 ± 0.080	0.562 ± 0.075	0.525 ± 0.039	0.195 ± 0.042
Fish	<b>0.902 ± 0.016</b>	0.672 ± 0.043	0.826 ± 0.033	0.428 ± 0.012	0.480 ± 0.046	0.394 ± 0.090	0.190 ± 0.067
Wine	<b>0.749 ± 0.027</b>	0.575 ± 0.009	<b>0.734 ± 0.008</b>	0.234 ± 0.028	0.270 ± 0.013	0.321 ± 0.017	-0.006 ± 0.019

## 5.2 Results

### Accuracy

First, we demonstrate the predictive performances of GPX and the comparing methods in Table 1. GPX achieved the lowest predictive errors on all the datasets, compared to the other globally or locally linear models. In addition, their predictive errors were comparable to that of GPR on all the datasets. This result indicates that GPR can be replaced by GPX to achieve similar predictive performances.

### Faithfulness

Assessing the correctness of the estimated contribution of each feature to a prediction requires a reference “true” contribution for comparison. As this is rarely available, a typical approach for measuring the faithfulness of the contributions produced by interpretable models is to rely on the proxy notion of the contributions: observing the effect of removing features on the model’s prediction. Following previous studies [Melis and Jaakkola, 2018, Bhatt et al., 2020], we computed the faithfulness score by removing features one-by-one, measuring the differences between the original predictions and the predictions from the inputs without the removed features, and calculating the correlation between the differences and the contributions of the removed features.

Table 2 shows the faithfulness scores of GPX and the comparing methods. Here, we denote the results of LIME and Kernel SHAP using GPR as the black-box prediction model by GPR+LIME and GPR+SHAP, respectively. We found that GPX achieved the best faithfulness scores on all the datasets. As GPX predicts and explains using a single locally linear model for each test sample, when removing a feature from the input, the contribution of the feature is subtracted from the prediction directly. Meanwhile, because GPR+LIME, GPR+SHAP and KL have different prediction and explanation models, a gap may exist between the estimated contribution in the explanation model and the latent contribution in the prediction. Because the predictions by GPX and GPR were performed using similar calculations, their faithfulness differences were likely due to the gap. With ARD, it cannot estimate feature relevances appropriate for each sample, as the feature relevances are shared over samples; therefore, it produced relatively low faithfulness scores.

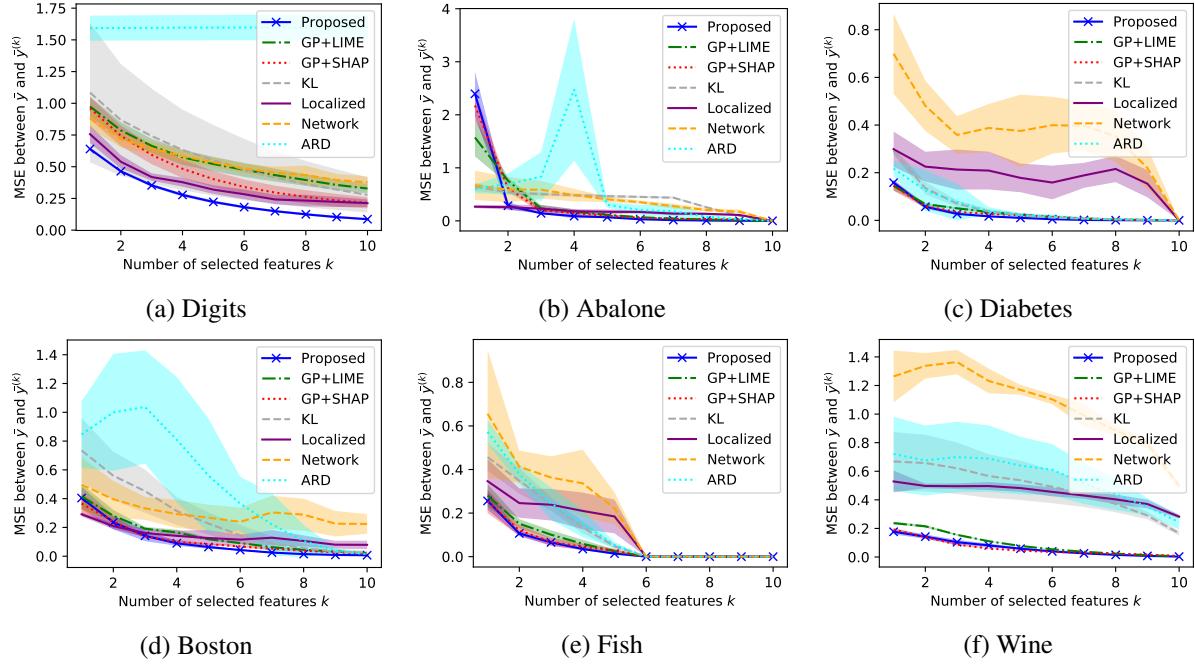


Figure 2: Average sufficiency scores on Digits, Abalone, Diabetes, Boston, Fish and Wine datasets (lower scores are better). The filled area on each line indicates its standard deviation. For Paper and Drug datasets, we did not evaluate the scores as changes in  $Z$  cannot reflect  $X$ .

Table 3: Average and standard deviation of stability scores on each dataset (lower scores are better). This table can be interpreted similarly as Table 1. We did not measure the scores for models that have global feature relevances/weights, such as ARD, as their scores are obviously zero according to (19).

	GPX (ours)	GPR+LIME	GPR+SHAP	KL	Localized	Network
Digits	<b>1.153 ± 0.015</b>	2.410 ± 0.079	2.274 ± 0.110	2.058 ± 1.043	1.996 ± 0.098	1.989 ± 0.222
Abalone	<b>3.094 ± 0.249</b>	11.809 ± 0.453	11.625 ± 0.362	4.485 ± 0.911	15.439 ± 1.193	15.772 ± 0.948
Diabetes	<b>1.164 ± 0.065</b>	1.870 ± 0.119	1.400 ± 0.147	1.401 ± 0.113	2.833 ± 0.085	2.693 ± 0.123
Boston	<b>1.452 ± 0.081</b>	4.180 ± 0.426	2.176 ± 0.304	2.956 ± 0.221	3.794 ± 1.215	3.467 ± 0.799
Fish	<b>1.497 ± 0.075</b>	> 10 <sup>5</sup>	1.634 ± 0.057	3.289 ± 0.536	5.746 ± 0.563	5.625 ± 0.591
Wine	<b>1.531 ± 0.113</b>	> 10 <sup>6</sup>	> 10 <sup>5</sup>	5.255 ± 0.906	3.989 ± 0.306	4.033 ± 0.282
Paper	<b>0.004 ± 0.002</b>	5.535 ± 0.161	4.548 ± 0.098	<b>0.026 ± 0.052</b>	6.482 ± 0.161	6.986 ± 0.442
Drug	<b>0.067 ± 0.014</b>	16.976 ± 0.399	12.037 ± 0.513	11.547 ± 0.265	16.908 ± 0.532	17.341 ± 0.589

## Sufficiency

In general, the inputs contain many irrelevant features that do not contribute to the predictions, and discovering important features in all the features is difficult for users of the models. Therefore, a desirable property of the interpretable models is that it can assign high contributions only for important features that affect the predictions well. To quantify how each method satisfies the property, we define the sufficiency score at  $k$ , where  $k$  is the number of important features. In particular, the sufficiency score at  $k$  was computed by identifying  $k$  important features in the descending order of the absolute values of their estimated contributions, predicting from the inputs having only  $k$  important features, and comparing them against the original predictions. Because the number of important features varied according to the sample and dataset, we evaluated them at  $k = 1, 2, \dots, 10$ .

Figure 2 shows the sufficiency scores of GPX and the comparing methods. Independent of the  $k$  values, GPX outperformed the others on Digits dataset, whereas GPX, GPR+LIME, and GPR+SHAP produced the best sufficiency scores on Diabetes, Fish and Wine datasets. These results indicate that GPX was appropriately assigned high contributions for the important features. On Abalone and Boston datasets, GPX was slightly inferior to the localized lasso at  $k = 1, 2$ , although GPX outperformed it at  $k \geq 3$ . This is because the localized lasso has a regularizer that induces sparse weights. This result suggests that GPX can be further improved by employing the mechanism for generating sparse weights.

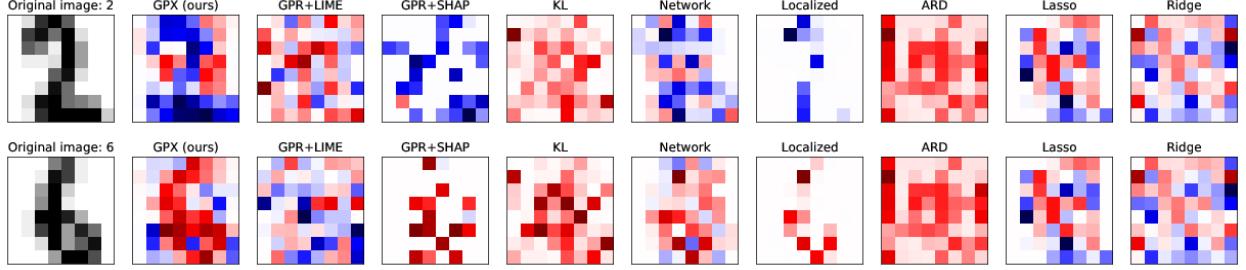


Figure 3: Examples of estimated weights of each model on Digits dataset. The upper row shows the weights for the sample with digit two ( $y = -1$ ), whereas the bottom one displays those for the sample with digit six ( $y = 1$ ). Red and blue denote positive and negative weights, respectively, and their color strengths represent their magnitudes.

## Stability

To generate meaningful explanations, interpretable methods must be robust against local perturbations from the input, as explanations that are sensitive to slight changes in the input may be regarded as inconsistent by users. In particular, flexible models such as locally linear models might be sensitive to such changes for achieving better predictions. As with the work by [Melis and Jaakkola, 2018], we used the following quantity for measuring the stability of the estimated weights for test sample  $(\mathbf{x}_*, \mathbf{z}_*)$ , as follows:

$$L(\mathbf{x}_*, \mathbf{z}_*) = \max_{\mathbf{x}'_*, \mathbf{z}'_* \in \mathcal{B}_\epsilon(\mathbf{x}_*)} \frac{\|\mathbf{w}'_* - \mathbf{w}_*\|_2}{\|\mathbf{z}'_* - \mathbf{z}_*\|_2}, \quad \mathcal{B}_\epsilon(\mathbf{x}_*) = \{(\mathbf{x}'_*, \mathbf{z}'_*) \in \mathcal{D}_{\text{te}} \mid \frac{1}{m} \|\mathbf{x}'_* - \mathbf{x}_*\|_2 < \epsilon\}, \quad (19)$$

where,  $\mathcal{D}_{\text{te}} = \{(\mathbf{x}_*, \mathbf{z}_*)\}$  is a set of test samples;  $\mathbf{w}_*$  and  $\mathbf{w}'_*$  are the standardized estimated weights associated with test samples  $(\mathbf{x}_*, \mathbf{z}_*)$  and  $(\mathbf{x}'_*, \mathbf{z}'_*)$ , respectively;  $\epsilon > 0$  is a parameter that determines neighboring samples;  $m$  is the dimensionality of  $\mathbf{x}_*$ . We set  $\epsilon = 0.05$  in our experiments. Intuitively, the stability score will be high when the estimated weights for the sample and its neighboring samples are similar. Subsequently, we computed the stability score on a dataset by averaging the quantity (19) on all the test samples in the dataset.

Table 3 shows the stability scores on each dataset. GPX achieved the best stability scores on all the datasets. With GPR+LIME and GPR+SHAP, their stability scores were lower than that of GPX, although the prediction powers of GPX and GPR were comparable. This would be because LIME and Kernel SHAP estimated the weights independently over the test samples. The stability score of KL was as good as that of GPX only on Paper dataset; however, on the other datasets, KL was inferior to than GPX.

## Qualitative comparison

Finally, we qualitatively compared the estimated weights using GPX and the comparing methods on Digits dataset, in which the appropriate contributions for predictions were apparent. For this comparison, we rescaled the inputs  $\mathbf{X}$  and  $\mathbf{Z}$  to be within  $[0, 1]$ .

Figure 3 shows the estimated weights on two samples. We provide the results for all the digits in Appendix E. On this dataset, the appropriate weights can be obtained by assigning weights having the same sign with the target variable to black pixels. We found that the methods except for GPX and the localized lasso could not estimate reasonable weights. Meanwhile, the weights estimated by GPX and the localized lasso were appropriate, although they exhibited different characteristics, i.e., dense weights from GPX, whereas sparse ones from the localized lasso. The task determines the better explanation; however, as showing important regions rather than pixels is meaningful for images, the estimated weights using GPX would be easier to interpret on Digits dataset. Furthermore, the degree of sparsity in the localized lasso can be changed as a hyperparameter; if the value of the hyperparameter is zero, the localized lasso is identical to the network lasso. However, because the estimated weights using the network lasso were inappropriate, those using GPX cannot be mimicked by the localized lasso.

## Computational efficiency

Table 4 shows the computational times of each of the methods on Digits dataset. First, the training time of GPX was much the same as those of GPR and ARD, and significantly faster than those of the localized and network lasso. Since the localized and network lasso requires the hyperparameter search, the their actual training times were about 48 and 12 times longer than the times shown in the table, respectively. Second, the prediction time of GPX was significantly

Table 4: Training, prediction and total times (seconds) of each method on the Digits dataset. The training times of GPR+LIME/SHAP and KL are those of GPR and ARD, respectively. The prediction times of GPR+LIME/SHAP and KL are those of producing explanations for all the test samples.

	GPX (ours)	GPR+LIME	GPR+SHAP	KL	Localized	Network	ARD
Training	5.74	5.53	5.76	6.57	105.84	106.16	6.18
Prediction	24.57	155.54	2643.06	156.19	1.59	1.61	0.27
Total	30.31	161.07	2648.82	162.76	107.43	107.77	6.45

faster than those of GPR+LIME/SHAP and KL. This is because GPX does not require learning model parameters at prediction phase. On the other hand, since GPR+LIME/SHAP learns an explanation model for each sample at that time, and KL requires producing a lot of predictions with slight changes in the value of each dimension of the input, their prediction times lead to increase.

## 6 Conclusion

We proposed a GP-based regression model with sample-wise explanations. The proposed model assumes that each sample has a locally linear model, which is used for both prediction and explanation, and the weight vector of the locally linear model are generated from multivariate GP priors. The hyperparameters of the proposed models were estimated by maximizing the marginal likelihood, in which all the weight vectors were integrated out. Subsequently, for a test sample, the proposed model predicted its target variable and weight vector with uncertainty. In the experiments, we confirmed that the proposed model outperformed the existing globally and locally linear models and achieved comparable performances with the standard GPR in terms of predictive performance, and the proposed model was superior to the existing methods, including model-agnostic interpretable methods, in terms of three interpretability measurements. Then, we showed that the feature weights estimated by the proposed model were appropriate as the explanation.

In future studies, we will confirm the effectiveness of the proposed model by applying its concept into various problems in which GPs have been successfully used, such as classification, black-box optimization, and time-series forecasting. In addition, we will extend the proposed model for further improvements in interpretability, e.g., by employing the mechanism of inducing sparsity for the weight vectors.

## Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 18K18112.

## A Feature Description and Additional Examples for the Boston Housing Dataset

The Boston housing dataset, referred to as “Boston” in our experiments, contains information collected by the U.S. Census Service regarding housing in the area of Boston, Massachusetts [Harrison Jr and Rubinfeld, 1978] and is used for predicting house prices based on the information. Table 5 lists the names of the features and their descriptions for the Boston housing dataset.

Figure 4 presents four examples of feature contributions estimated by GPX. We found that each of these examples has different feature contributions, although some of the features, such as “AGE” and “DIS,” had consistent positive or negative contributions, respectively.

## B Detailed Derivation of Predictive Distributions

In this appendix, we describe the derivation of predictive distributions in detail. For a new test sample  $(\mathbf{x}_*, \mathbf{z}_*)$ , our goal is to infer the predictive distributions of the target variable  $y_*$  and weight vector  $\mathbf{w}_*$ .

Table 5: Feature names and their descriptions for the Boston housing dataset.

Feature name	Description
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per \$10,000
PTRATIO	Pupil-teacher ratio by town
B	$1000(B_k - 0.63)^2$ where $B_k$ is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in \$1000's

### Predictive distribution of $y_*$

The predictive distribution of  $y_*$  is obtained similarly to the standard GPR [Rasmussen, 2003]. In Section 4, we demonstrated that the marginal distribution of training target variables  $\mathbf{y}$  for GPX is defined as

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}) &= \int \int p(\mathbf{y}, \mathbf{W}, \mathbf{G} \mid \mathbf{X}, \mathbf{Z}) d\mathbf{W} d\mathbf{G} \\ &= \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{C}), \end{aligned} \quad (20)$$

where  $\mathbf{C} = \sigma_y^2 \mathbf{I}_n + (\mathbf{K} + \sigma_w^2 \mathbf{I}_n) \odot \mathbf{Z} \mathbf{Z}^\top$ . According to (20), the joint marginal distribution of  $\mathbf{y}$  and  $y_*$  is defined as

$$p(\mathbf{y}, y_* \mid \mathbf{X}, \mathbf{Z}, \mathbf{x}_*, \mathbf{z}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \mid \mathbf{0}, \begin{bmatrix} \mathbf{C} & \mathbf{c}_* \\ \mathbf{c}_*^\top & c_{**} \end{bmatrix}\right), \quad (21)$$

where  $\mathbf{c}_* = (k_\theta(\mathbf{x}_*, \mathbf{x}_i) \mathbf{z}_*^\top \mathbf{z}_i)_{i=1}^n \in \mathbb{R}^n$ , and  $c_{**} = \sigma_y^2 + (k_\theta(\mathbf{x}_*, \mathbf{x}_*) + \sigma_w^2) \mathbf{z}_*^\top \mathbf{z}_* \in \mathbb{R}$ . The predictive distribution of  $y_*$  is the conditional distribution of  $y_*$  given  $\mathbf{y}$  with training and testing inputs. Therefore, it can be obtained by applying the formula of conditional distributions for normal distributions [Petersen and Pedersen, 2012, Eq. (354)] to (21) as follows:

$$p(y_* \mid \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \mathcal{N}(y_* \mid \mathbf{c}_*^\top \mathbf{C}^{-1} \mathbf{y}, c_{**} - \mathbf{c}_*^\top \mathbf{C}^{-1} \mathbf{c}_*). \quad (22)$$

### Predictive distribution of $w_*$

The predictive distribution of  $\mathbf{w}_*$  can be obtained by solving the following equation:

$$p(\mathbf{w}_* \mid \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \int p(\mathbf{w}_* \mid \mathbf{W}, \mathbf{X}, \mathbf{x}_*) p(\mathbf{W} \mid \mathcal{D}) d\mathbf{W}, \quad (23)$$

where the first integrand  $p(\mathbf{w}_* \mid \mathbf{W}, \mathbf{X}, \mathbf{x}_*)$  is the conditional distribution of  $\mathbf{w}_*$  and the second integrand  $p(\mathbf{W} \mid \mathcal{D})$  is the posterior distribution of  $\mathbf{W}$ . The conditional distribution of  $\mathbf{w}_*$  is derived similarly to the conditional distribution

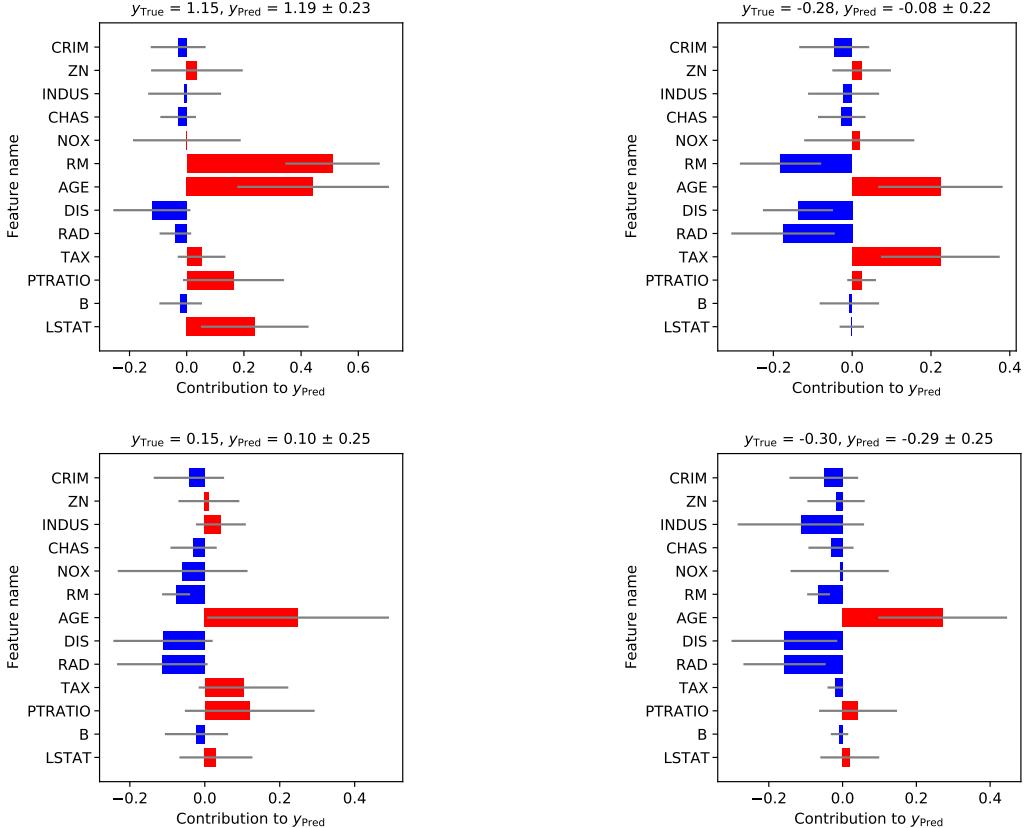


Figure 4: Examples of feature contributions estimated by GPX for the Boston housing dataset. Here, the error bars denote the standard deviations or feature contributions.

of  $y_*$  (22). The distribution of  $\mathbf{W}$  in which the functions  $\mathbf{G}$  are integrated out is given by

$$\begin{aligned} p(\mathbf{W} | \mathbf{X}) &= \int p(\mathbf{W} | \mathbf{G})p(\mathbf{G} | \mathbf{X})d\mathbf{G} \\ &= \prod_{l=1}^d \mathcal{N}(\mathbf{W}_{\cdot,l} | \mathbf{0}, \mathbf{K} + \sigma_w^2 \mathbf{I}_n), \end{aligned} \quad (24)$$

where  $\mathbf{W}_{\cdot,l}$  is the  $l$ th column vector of  $\mathbf{W}$ . According to this, the joint distribution of  $\mathbf{W}$  and  $\mathbf{w}_*$  is defined as

$$p(\mathbf{W}, \mathbf{w}_* | \mathbf{X}, \mathbf{x}_*) = \prod_{l=1}^d \mathcal{N}\left(\begin{bmatrix} \mathbf{W}_{\cdot,l} \\ w_{*,l} \end{bmatrix} \mid \mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_w^2 \mathbf{I}_n & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} \end{bmatrix}\right), \quad (25)$$

where we let  $\mathbf{k}_* = (k_\theta(\mathbf{x}_*, \mathbf{x}_i))_{i=1}^n$  and  $k_{**} = k_\theta(\mathbf{x}_*, \mathbf{x}_*) + \sigma_w^2$ . Subsequently, we can obtain the conditional distribution of  $\mathbf{w}_*$  by applying the formula of conditional distributions for normal distributions [Petersen and Pedersen, 2012, Eq. (354)] to (25) as follows:

$$\begin{aligned} p(\mathbf{w}_* | \mathbf{W}, \mathbf{X}, \mathbf{x}_*) &= \prod_{l=1}^d \mathcal{N}(w_{*,l} \mid \mathbf{k}_*^\top (\mathbf{K} + \sigma_w^2 \mathbf{I}_n)^{-1} \mathbf{W}_{\cdot,l}, k_{**} - \mathbf{k}_*^\top (\mathbf{K} + \sigma_w^2 \mathbf{I}_n)^{-1} \mathbf{k}_*). \end{aligned} \quad (26)$$

Here, we can rewrite (26) as a single  $d$ -dimensional multivariate normal distribution as follows:

$$\begin{aligned} p(\mathbf{w}_* | \mathbf{W}, \mathbf{X}, \mathbf{x}_*) &= \mathcal{N}\left(\mathbf{w}_* \mid \bar{\mathbf{k}}_*^\top (\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd})^{-1} \text{vec}(\mathbf{W}), \bar{\mathbf{c}}_{**} - \bar{\mathbf{k}}_*^\top (\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd})^{-1} \bar{\mathbf{k}}_*\right), \end{aligned} \quad (27)$$

where  $\bar{\mathbf{K}}$  is a block diagonal matrix of order  $nd$  whose block is  $\mathbf{K}$ ,  $\text{vec}(\cdot)$  is a function that flattens the input matrix in column-major order, and  $\bar{\mathbf{c}}_{**} = (k_\theta(\mathbf{x}_*, \mathbf{x}_*) + \sigma_w^2) \mathbf{I}_d$ .  $\bar{\mathbf{k}}_*$  is an  $nd$ -by- $d$  block matrix, where each block is an  $n$ -by-1 matrix the and  $(l, l)$ -block of the block matrix is  $(k_\theta(\mathbf{x}_*, \mathbf{x}_i))_{i=1}^n$  for  $l = 1, 2, \dots, d$ , while the other blocks are zero matrices. By letting  $\mathbf{A} = \bar{\mathbf{k}}_*^\top (\bar{\mathbf{K}} + \sigma_w^2 \mathbf{I}_{nd})^{-1}$ , we obtain

$$p(\mathbf{w}_* | \mathbf{W}, \mathbf{X}, \mathbf{x}_*) = \mathcal{N}(\mathbf{w}_* | \text{Avec}(\mathbf{W}), \bar{\mathbf{c}}_{**} - \mathbf{A}\bar{\mathbf{k}}_*). \quad (28)$$

To derive the posterior distribution of  $\mathbf{W}$ ,  $p(\mathbf{W} | \mathcal{D})$ , we first consider the joint distribution of  $\mathbf{W}$  and  $\mathcal{D}$ . This distribution is straightforwardly obtained as

$$p(\mathbf{W}, \mathcal{D}) = \prod_{l=1}^d \mathcal{N}(\mathbf{W}_{\cdot, l} | \mathbf{0}, \mathbf{K}) \prod_{i=1}^n \mathcal{N}(y_i | \mathbf{w}_i^\top \mathbf{z}_i, \sigma_y^2), \quad (29)$$

which can be rewritten as

$$p(\mathbf{W}, \mathcal{D}) = \mathcal{N}(\text{vec}(\mathbf{W}) | \mathbf{0}, \bar{\mathbf{K}}) \mathcal{N}(\mathbf{y} | \bar{\mathbf{Z}} \text{vec}(\mathbf{W}), \sigma_y^2 \mathbf{I}_n), \quad (30)$$

where  $\bar{\mathbf{Z}} = (\text{diag}(\mathbf{Z}_{\cdot, 1}), \text{diag}(\mathbf{Z}_{\cdot, 2}), \dots, \text{diag}(\mathbf{Z}_{\cdot, d})) \in \mathbb{R}^{n \times nd}$ . By applying the formula of conditional distributions of normal distributions [Bishop, 2006, Eqs. (2.113)–(2.117)] to (30), we can obtain

$$p(\mathbf{W} | \mathcal{D}) = \mathcal{N}(\text{vec}(\mathbf{W}) | \sigma_y^{-2} \Sigma \bar{\mathbf{Z}}^\top \mathbf{y}, \Sigma), \quad (31)$$

where

$$\Sigma = \left( \bar{\mathbf{K}}^{-1} - \bar{\mathbf{Z}}^\top (\sigma_y^{-2} \mathbf{I}_n) \bar{\mathbf{Z}} \right)^{-1}. \quad (32)$$

Here, the computation of  $\Sigma$  requires inverting a square matrix of order  $nd$  with a computational complexity of  $\mathcal{O}(n^3 d^3)$ . By using the Woodbury identity [Petersen and Pedersen, 2012, Eq. (156)] to compute this inversion efficiently, we can transform  $\Sigma$  into  $\mathbf{S} = \bar{\mathbf{K}} - \bar{\mathbf{K}} \bar{\mathbf{Z}}^\top \mathbf{D}^{-1} \bar{\mathbf{Z}} \bar{\mathbf{K}}$ , which requires inverting a matrix of order  $n$ ,  $\mathbf{D} = \sigma_y^2 \mathbf{I}_n + \mathbf{K} \odot \mathbf{Z} \mathbf{Z}^\top$ . Consequently, we obtain

$$p(\mathbf{W} | \mathcal{D}) = \mathcal{N}(\text{vec}(\mathbf{W}) | \sigma_y^{-2} \mathbf{S} \bar{\mathbf{Z}}^\top \mathbf{y}, \mathbf{S}). \quad (33)$$

From (28) and (33), one can see that (23) can be represented by the following equation:

$$p(\mathbf{w}_* | \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \int \mathcal{N}(\mathbf{w}_* | \text{Avec}(\mathbf{W}), \bar{\mathbf{c}}_{**} - \mathbf{A}\bar{\mathbf{k}}_*) \mathcal{N}(\text{vec}(\mathbf{W}) | \sigma_y^{-2} \mathbf{S} \bar{\mathbf{Z}}^\top \mathbf{y}, \mathbf{S}) d\mathbf{W}. \quad (34)$$

This integral can be obtained in a closed form, as shown in [Bishop, 2006, Eqs. (2.113)–(2.117)]. Therefore, we can obtain the predictive distribution of  $\mathbf{w}_*$  as follows:

$$p(\mathbf{w}_* | \mathbf{x}_*, \mathbf{z}_*, \mathcal{D}) = \mathcal{N}(\mathbf{w}_* | \sigma_y^{-2} \mathbf{A} \mathbf{S} \bar{\mathbf{Z}}^\top \mathbf{y}, \bar{\mathbf{c}}_{**} - \mathbf{A}\bar{\mathbf{k}}_* + \mathbf{A} \mathbf{S} \mathbf{A}^\top). \quad (35)$$

## C Specification of Datasets

We considered eight datasets from the UCI machine learning repository [Dua and Graff, 2017], which were referred to as Digits<sup>1</sup>, Abalone<sup>2</sup>, Diabetes<sup>3</sup>, Boston [Harrison Jr and Rubinfeld, 1978], Fish<sup>4</sup>, Wine<sup>5</sup>, Paper<sup>6</sup>, and Drug<sup>7</sup> in our experiments.

The first six datasets consisted of tabular data. We treated the original inputs  $\mathbf{X}$  and simplified inputs  $\mathbf{Z}$  identically in our experiments. The Digits dataset was originally developed as a classification dataset for recognizing handwritten digits from zero to nine. As described in Section 5.1, we used this dataset for a regression problem by transforming the digit labels into binary values of 1 or -1. Here, we used only the testing set from the original Digits dataset because that is how scikit-learn [Pedregosa et al., 2011] distributes this dataset. The Abalone dataset is a dataset for predicting the

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits>

<sup>2</sup><https://archive.ics.uci.edu/ml/datasets/Abalone>

<sup>3</sup><https://archive.ics.uci.edu/ml/datasets/diabetes>

<sup>4</sup><https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity>

<sup>5</sup><https://archive.ics.uci.edu/ml/datasets/wine+quality>

<sup>6</sup><https://archive.ics.uci.edu/ml/datasets/Paper+Reviews>

<sup>7</sup>[https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+\(Druglib.com\)](https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+(Druglib.com))

Table 6: Specification of datasets.

	Digits	Abalone	Diabetes	Boston	Fish	Wine	Paper	Drug
$n$	1,797	4,177	442	506	908	6,497	399	3,989
$d$	64	10	10	13	6	11	2,990	2,429

age of abalone based on physical measurements. The Diabetes dataset is a dataset for predicting the onset of diabetes based on diagnostic measures. The Boston dataset is a dataset for predicting house prices, as described in Appendix A. The Fish dataset is a dataset for predicting acute aquatic toxicity toward the fish pimephales promelas for a set of chemicals. The Wine dataset is a dataset for predicting the quality of white and red wines based on physicochemical tests. The remaining two datasets are text datasets. The Paper dataset is a dataset for predicting evaluation scores for papers based on review texts written mainly in Spanish. The Drug dataset is a drug review dataset for predicting 10-star ratings for drugs based on patient review texts. For each dataset,  $\mathbf{X}$  and  $\mathbf{Z}$  are different. Specifically, we used the 512-dimensional sentence vectors obtained using sentence transformers [Reimers and Gurevych, 2020] as  $\mathbf{X}$  and used bag-of-words binary vectors of the sentences as  $\mathbf{Z}$ , where the cutoff frequencies for words were set to two and five for the Paper and Drug datasets, respectively. Table 6 lists the number of samples  $n$  and number of features  $d$  in each dataset.

## D Detailed Description of Comparing Methods

In this appendix, we describe the implementation and hyperparameter search methods used for comparing methods.

We implemented GPR using PyTorch v1.5.0<sup>8</sup>. All hyperparameters for GPR were estimated by maximizing marginal likelihood [Rasmussen, 2003], where we initialized the hyperparameters to the same values as those for GPX. For Lasso and Ridge, we used the implementations provided by scikit-learn [Pedregosa et al., 2011]. The hyperparameters that regularize the strengths of the  $\ell_1$  and  $\ell_2$  regularizers in Lasso and Ridge, respectively, were optimized through a grid search using functions provided by scikit-learn (i.e., `sklearn.linear_model.LassoCV` and `sklearn.linear_model.RidgeCV`) with the default options. The search range for the hyperparameters for Lasso was limited to within 100 grid points such that the ratio of its minimum value to its maximum value was capped at 0.001, while that for Ridge was limited to within a range of {0.1, 1, 10}. For ARD, we implemented it as with GPR. Herein, the kernel function used in ARD is defined as

$$k_{\theta}(\mathbf{x}, \mathbf{x}') = \alpha \exp \left( -\frac{1}{2} \sum_{l=1}^d (x_l - x'_l)^2 / \theta_l \right) \quad (\alpha, \theta_l > 0), \quad (36)$$

where  $\alpha$  is a scale parameter;  $\theta_l$  is the relevance of the  $l$ th feature;  $\theta = \{\alpha, \theta_1, \theta_2, \dots, \theta_d\}$  is a set of parameters to be estimated as with GPR. For the localized lasso, we used the original implementation written in Python<sup>9</sup>. The hyperparameters and their search ranges for the localized lasso are the strength of network regularization  $\lambda_1 \in \{1, 3, 5, 7\}$ , strength of the  $\ell_{1,2}$  regularizer  $\lambda_2 \in \{0.01, 0.1, 1, 10\}$ , and  $k \in \{5, 10, 15\}$  for the  $k$ -nearest-neighbor graph. The hyperparameters were optimized through a grid search. The network lasso is a special case of the localized lasso. If  $\lambda_2$  for the localized lasso is zero, then the localized lasso is identical to the network lasso. Therefore, we used the implementation of the localized lasso and set  $\lambda_2 = 0$  for the network lasso. The hyperparameter search for the network lasso was the same as that for the localized lasso, except for the setting of  $\lambda_2$ . For LIME and Kernel SHAP, we used the original implementations<sup>10</sup>. For KL, we implemented it by mimicking its original implementation<sup>11</sup>. Herein, we set amount of perturbation  $\Delta = 0.001$  throughout our experiments.

## E Additional Results for the Digits Dataset

Figure 5 presents additional examples of estimated weights for the Digits dataset. We found that the weights estimated by GPX were appropriately assigned such that the regions of black pixels have weights with the same signs as those of the target variables.

In terms of the stability of explanations, estimated weights for the same digit should be similar. Figure 6 presents three examples of estimated weights for the digit two. We found that GPX estimated similar weights for all three examples.

<sup>8</sup><https://pytorch.org/>

<sup>9</sup><https://riken-yamada.github.io/localizedlasso.html>

<sup>10</sup>LIME: <https://github.com/marcotcr/lime>, Kernel SHAP: <https://github.com/slundberg/shap>

<sup>11</sup><https://github.com/topipa/gp-varsel-kl-var>

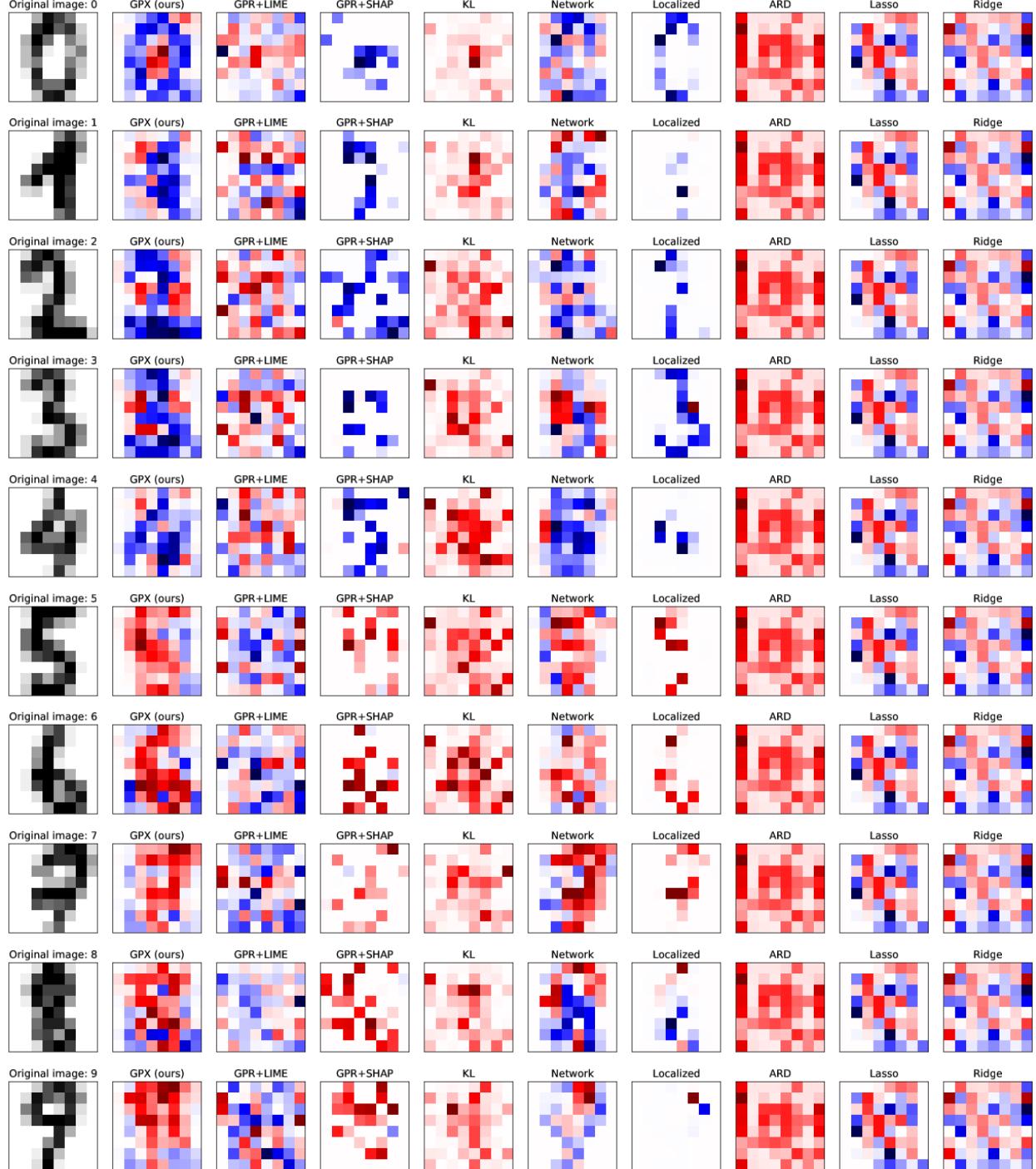


Figure 5: Examples of estimated weights for digits ranging from zero to nine for the Digits dataset. The five upper rows present the weights of samples with digits of zero to four ( $y = -1$ ), whereas the five bottom rows present those for samples with digits from five to nine ( $y = 1$ ). Red and blue denote positive and negative weights, respectively, and their color strengths represent their magnitudes.

## References

- Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

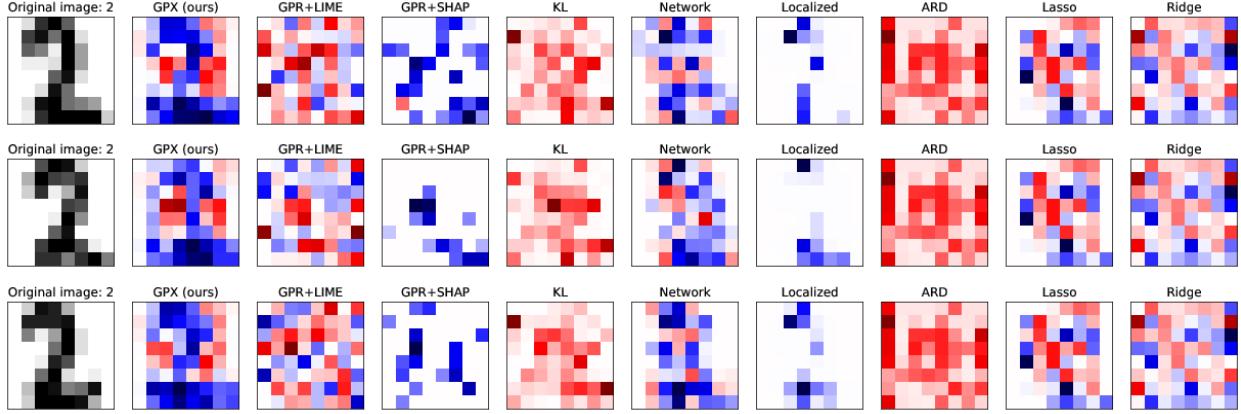


Figure 6: Different examples of estimated weights for digit two for the Digits dataset.

- Andrew Gordon Wilson, David A Knowles, and Zoubin Ghahramani. Gaussian process regression networks. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1139–1146, 2012.
- Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, and Ole Winther. Efficient approaches to gaussian process classification. In *Advances in neural information processing systems*, pages 251–257, 2000.
- Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.
- Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.
- Joan Gonzalvez, Edmond Lezmi, Thierry Roncalli, and Jiali Xu. Financial applications of gaussian processes and bayesian optimization. *arXiv preprint arXiv:1903.04841*, 2019.
- Gustau Camps-Valls, Jochem Verrelst, Jordi Munoz-Mari, Valero Laparra, Fernando Mateo-Jimenez, and Jose Gomez-Dans. A survey on gaussian processes for earth-observation data analysis: A comprehensive investigation. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):58–78, 2016.
- Yichi Zhang, Daniel W Apley, and Wei Chen. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific Reports*, 10(1):1–13, 2020.
- Li-Fang Cheng, Gregory Darnell, Bianca Dumitrescu, Corey Chivers, Michael E Draugelis, Kai Li, and Barbara E Engelhardt. Sparse multi-output gaussian processes for medical time series prediction. *arXiv preprint arXiv:1703.09112*, 2017.
- Joseph Futoma. *Gaussian process-based models for clinical time series in healthcare*. PhD thesis, Duke University, 2018.
- Christoph Molnar. *Interpretable Machine Learning*. 2019. <https://christophm.github.io/interpretable-ml-book/>.
- Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892, 2018.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM, 2016. ISBN 978-1-4503-4232-2. doi:10.1145/2939672.2939778.
- Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.
- Mauricio A Álvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.
- David P Wipf and Srikantan S Nagarajan. A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632, 2008.
- David Hallac, Jure Leskovec, and Stephen Boyd. Network Lasso: Clustering and Optimization in Large Graphs. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, pages 387–396. ACM, 2015. ISBN 978-1-4503-3664-2. doi:10.1145/2783258.2783313.
- Makoto Yamada, Takeuchi Koh, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski. Localized Lasso for High-Dimensional Regression. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 325–333. PMLR, April 2017.
- Topi Paananen, Juho Pironen, Michael Riis Andersen, and Aki Vehtari. Variable selection for gaussian processes via sensitivity analysis of the posterior predictive distribution. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1743–1752, 2019.
- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.
- Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8), 2017.
- Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, pages 9240–9251, 2019.
- David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.
- Patrick Schwab, Djordje Miladinovic, and Walter Karlen. Granger-causal attentive mixtures of experts: Learning important features with neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4846–4853, 2019.
- Yuya Yoshikawa and Tomoharu Iwata. Neural generators of sparse local linear models for achieving both accuracy and interpretability. *arXiv preprint arXiv:2003.06441*, 2020.
- Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017.
- S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.
- Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.
- Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada. Latent support measure machines for bag-of-words data classification. In *Advances in neural information processing systems*, pages 1961–1969, 2014.
- Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.
- Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.
- Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*, 2020. <https://github.com/UKPLab/sentence-transformers>.
- Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

- Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.
- K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012. URL <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>. Version: November 15, 2012.
- Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.