

Model selection for Gaussian processes utilizing sensitivity of posterior predictive distribution

Topi Paananen

Juho Piironen

Michael Riis Andersen

Aki Vehtari

Aalto University, Department of Computer Science

{topi.paananen, juho.piironen, aki.vehtari}@aalto.fi, michael.riis@gmail.com

Abstract

We propose two novel methods for simplifying Gaussian process (GP) models by examining the predictions of a full model in the vicinity of the training points and thereby ordering the covariates based on their predictive relevance. Our results on synthetic and real world data sets demonstrate improved variable selection compared to automatic relevance determination (ARD) in terms of consistency and predictive performance. We expect our proposed methods to be useful in interpreting and understanding complex Gaussian process models.

1 INTRODUCTION

Gaussian processes (GPs) (Rasmussen and Williams, 2006) offer a flexible nonparametric method for regression in the Bayesian framework. Often the goal of regression problems is not only to learn the relationship of the predictor and target variables, but to also assess the relevance of the inputs. A relevant input variable is one with a high predictive power on the target variable (Vehtari et al., 2012). Selection of relevant variables is important for two reasons: firstly, it makes the model more easily interpretable and understandable. Secondly, it may reduce future measurement costs by reducing the number of explanatory variables needed. Covariate relevance determination methods are beneficial for a wide range of applications in statistical inference, such as disease risk prediction (Peltola et al., 2014) and the analysis of the length of care episode after hip fracture (Riihimäki et al., 2010).

In this paper we propose two novel variable selection methods for Gaussian process models. Both methods utilize the posterior predictive distribution of the full

model at the training points to estimate the relevance of covariates. The first method assesses the sensitivity of predictions to perturbations in the inputs. This is done by computing the Kullback-Leibler divergence from the predictive distribution in a training point to a point moved slightly in one dimension. Averaged over all training points, this is used as the measure of relevance for that covariate. The second method examines the predictive mean given by the GP model and computes its total variance along a covariate. This measure is computed along every covariate at every training point. The idea of going through the training points in order to assess the effect of a predictor to the target variable is related to average predictive comparison (Gelman and Pardoe, 2007).

Our experiments indicate that the proposed methods lead to improved performance compared to automatic relevance determination (ARD) without significantly increasing the computational complexity. They also generate a more consistent ordering for the covariates based on their predictive relevance, and avoid choosing nonlinear but redundant covariates.

We also discuss the evaluation of covariate relevances locally using Kullback-Leibler divergence. By calculating the KL divergence relevance measure in each training point, its distribution along each input can be used to find relevant regions of the input space. The relevance distribution provides a more robust way for assessing the relevance of a covariate than a single value produced by automatic relevance determination.

First, we review Gaussian processes and how the relevance of covariates is inferred automatically using ARD. Secondly, we introduce our proposed feature selection methods. Finally, we examine the predictive performance and consistency of the methods using simulated and real world data sets.

2 BACKGROUND

This section shortly reviews Gaussian processes and presents the principle of automatic relevance determination as well as the problems associated with it.

2.1 Gaussian Process Regression

Gaussian process (GP) models (Rasmussen and Williams, 2006) are a nonparametric class of models that define a distribution over functions. The prior of GPs is thus specified directly on the underlying latent function $f(\mathbf{x})$. The form and smoothness of functions generated by a GP is determined by its covariance function $k(\mathbf{x}, \mathbf{x}')$, which defines the covariance between latent function values at points \mathbf{x} and \mathbf{x}' . A typical choice for the prior is one with zero mean:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f} | 0, \mathbf{K}), \quad (1)$$

where \mathbf{K} is the covariance matrix between the latent function values at the training inputs $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ such that $\mathbf{K}_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

If we assume that the observed target values have independent Gaussian noise on top of the underlying function values, $y^{(i)} \sim \mathcal{N}(f(\mathbf{x}^{(i)}), \sigma^2)$, we can write the joint distribution of the observed outputs \mathbf{y} and the latent function values \mathbf{f}_* at test inputs \mathbf{X}_* as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I} & \mathbf{K}_*^\top \\ \mathbf{K}_* & \mathbf{K}_{**} \end{bmatrix}\right). \quad (2)$$

Here, \mathbf{K}_* is the covariance matrix between the latent values at test and training inputs, and \mathbf{K}_{**} between the test inputs only. The predictive distribution of the latent values at the test points is then given by conditioning on the observed values \mathbf{y} :

$$\begin{aligned} \mathbf{f}_* | \mathbf{y} &\sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \\ \boldsymbol{\mu}_* &= \mathbf{K}_* (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_* (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_*^\top. \end{aligned} \quad (3)$$

2.2 Automatic Relevance Determination

A widely used covariance function for Gaussian process regression is the squared exponential (SE) with separate length-scale parameters l_i for each of the p input dimensions

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{i=1}^p \frac{(x_i - x'_i)^2}{l_i^2}\right). \quad (4)$$

While the common hyperparameter σ_f determines the overall variability, the separate length-scale parameters l_i allow the functions to vary at different

scales along the covariates. The use of separate length-scales for each covariate is also known as automatic relevance determination (ARD), where the predictive relevance of a covariate is inferred from the inverse of its length-scale parameter. Use of length-scales as a measure of relevance has two problems: 1) length-scale parameters are not well identified (Zhang, 2004) which increases variance of the relevance measure, and 2) ARD severely overestimates the relevance of nonlinear covariates over linear ones of equal relevance in the squared error sense (Piironen and Vehtari, 2016).

3 PREDICTING FEATURE RELEVANCES USING LOCAL VARIATIONS

This section describes the two proposed methods for improved ordering of covariates based on local variations in the posterior predictive distribution. We first present the principle of both methods and then discuss their computational complexity.

3.1 KL Divergence as a Relevance Measure

The idea of our first method is to utilize Kullback-Leibler divergence of the predictive distribution at the training points to predict the most relevant covariates of a full model, thus being able to simplify it without losing too much predictive capacity. An estimate for the relevance of a covariate x_j is achieved by computing the KL divergence from the posterior predictive distribution in a training point to a point moved by some small amount Δ with respect to the covariate x_j only. Averaging over all training points produces an average estimate for the predictive relevance of that covariate.

In GP models with a Gaussian likelihood assumption, the posterior predictive distribution is also a Gaussian. The KL divergence from one univariate normal distribution $\mathcal{N}_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ to another $\mathcal{N}_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ is (Kullback, 1959)

$$\text{KL}(\mathcal{N}_1 || \mathcal{N}_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}. \quad (5)$$

As this is a quadratic function of the difference in their means, it is natural to use square root of the KL divergence as a measure of distance between the distributions. More specifically, by relating the measure to the total-variation distance of Pinsker's inequality, it is reasonable to use the measure (Simpson et al., 2017)

$$d(\mathcal{N}_1 || \mathcal{N}_2) = \sqrt{2 \text{KL}(\mathcal{N}_1 || \mathcal{N}_2)}. \quad (6)$$

Because we are considering the divergence from a training point to a nearby point, we have specified the direction unambiguously and do not have to worry about the asymmetry of KL divergence. In order to reduce the dependency on the perturbation distance Δ , we have defined the final measure of relevance as

$$r(\mathcal{N}_1 || \mathcal{N}_2) = \frac{\sqrt{2 \text{KL}(\mathcal{N}_1 || \mathcal{N}_2)}}{\Delta}. \quad (7)$$

Let us further rationalize our choice of relevance measure given in equation (7). Consider a dataset (\mathbf{x}, \mathbf{y}) of scalar inputs giving rise to a posterior GP that at some point x_0 has a slope k_0 and a variance that is constant along the input. Let us denote the mean and variance of the predictive distribution at this local point x_0 as μ_0 and σ_0^2 , thus $p(\tilde{y} | \mathbf{x}, \mathbf{y}, x_0) = \mathcal{N}(\tilde{y} | \mu_0, \sigma_0^2)$. Using equation (5), we get the KL divergence from the distribution at x_0 to the distribution at a nearby point $x_0 + \Delta$ as

$$\begin{aligned} \text{KL}(\mathcal{N}_0 || \mathcal{N}_\Delta) &= \frac{\sigma_0^2 + (\mu_0 - (\mu_0 + k_0 \Delta))^2}{2\sigma_0^2} - \frac{1}{2} \\ &= \frac{k_0^2 \Delta^2}{2\sigma_0^2}. \end{aligned} \quad (8)$$

The relevance measure of equation (7) then reduces to k_0/σ_0 . Intuitively, this is a reasonable measure of local predictive relevance, as it is simply the derivative of the predictive mean weighted by the inverse of the standard deviation.

3.1.1 Choice of Perturbation Distance

The choice of the perturbation distance Δ has to be determined based on the distribution of inputs. According to our investigations, the proposed method is not sensitive to the size of the perturbation. The results of this paper are calculated with Δ approximately 0.0001 times the standard deviation of the inputs, and we did not observe any differences when varying the distance for two orders of magnitude above and below this value. However, very small values should be avoided because of potential numerical errors.

3.2 Variance of the Predictive Mean

In this section we present another procedure for ordering covariates based on their predictive relevance. The idea of this method is to use the variance of the posterior predictive mean of a full GP along each covariate as an estimate for the relevance of the covariate.

In order to efficiently estimate the variance, we will model the distribution of the input variables. By assuming that the inputs have a joint Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the conditional distribution of one covariate, given the value of the others, is also Gaussian. If the joint distribution of the input variables, with covariate x_j separated, is denoted as

$$\begin{bmatrix} \mathbf{x}_{-j} \\ x_j \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_{-j} \\ \mu_j \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{-j,-j} & \boldsymbol{\sigma}_{j,-j} \\ \boldsymbol{\sigma}_{-j,j} & \sigma_{j,j} \end{bmatrix} \right), \quad (9)$$

the conditional distribution of the covariate x_j is given by

$$\begin{aligned} x_j | \mathbf{x}_{-j} &\sim \mathcal{N}(\mu_a, \sigma_a^2), \\ \mu_a &= \mu_j + \boldsymbol{\sigma}_{j,-j} \boldsymbol{\Sigma}_{-j,-j}^{-1} (\mathbf{x}_{-j} - \boldsymbol{\mu}_{-j}), \\ \sigma_a^2 &= \sigma_{j,j} - \boldsymbol{\sigma}_{j,-j} \boldsymbol{\Sigma}_{-j,-j}^{-1} \boldsymbol{\sigma}_{-j,j}. \end{aligned} \quad (10)$$

The subscript j refers to selecting the row or column j from $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$, whereas the subscript $-j$ refers to excluding them. Using equation (10), we can estimate the relevance of covariate the x_j at an arbitrary point by computing the variance of the predictive mean along the covariate with Gauss-Hermite quadrature. A good estimate for the overall relevance of each covariate is achieved by repeating this procedure at each training point.

In order to compute equation (10), the full sample mean and sample covariance matrix must be estimated from the training inputs. For the sample covariance matrix, the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}$ of the true covariance matrix is used:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^{(i)} - \bar{\mathbf{x}})(\mathbf{x}^{(i)} - \bar{\mathbf{x}})^\top, \quad (11)$$

where $\bar{\mathbf{x}}$ is the sample mean vector. Here, we will only consider data sets where the number of data points is greater than the number of input dimensions. In the absence of linearly dependent components in the inputs, the resulting sample covariance matrix will be positive definite and its inverse can be computed using the Cholesky decomposition. In order to increase numerical stability in data sets with similar input points, a small diagonal term was added to ill-conditioned sample covariance matrices.

As reviewed in Section 2.1, the predictive distribution of latent function values in GP models is Gaussian with mean $\boldsymbol{\mu}_*$ given in equation (3). Variance of the predictive mean along a covariate x_j is then given by integrating over the conditional Gaussian

of equation (10)

$$\begin{aligned} \text{Var}[\mu_{*,j}(x_j)] &= \int \mu_{*,j}^2(x_j) \mathcal{N}(x_j | \mu_a, \sigma_a^2) dx_j \\ &\quad - \left(\int \mu_{*,j}(x_j) \mathcal{N}(x_j | \mu_a, \sigma_a^2) dx_j \right)^2. \end{aligned} \quad (12)$$

With a change of variables $k = (x_j - \mu_a)/(\sqrt{2}\sigma_a)$, the variance takes the form

$$\begin{aligned} \text{Var}[\mu_{*,j}(x_j)] &= \int \mu_{*,j}^2(\sqrt{2}\sigma_a k + \mu_a) \frac{e^{-k^2}}{\sqrt{\pi}} dk \\ &\quad - \left(\int \mu_{*,j}(\sqrt{2}\sigma_a k + \mu_a) \frac{e^{-k^2}}{\sqrt{\pi}} dk \right)^2 \\ &\approx \pi^{-1/2} \sum_{i=1}^n w_i \mu_{*,j}^2(\sqrt{2}\sigma_a k_i + \mu_a) \\ &\quad - \pi^{-1} \left(\sum_{i=1}^n w_i \mu_{*,j}(\sqrt{2}\sigma_a k_i + \mu_a) \right)^2. \end{aligned} \quad (13)$$

where w_i and k_i are the weights and evaluation points of the Gauss-Hermite quadrature.

3.3 Complexity

The exact inference with Gaussian processes has complexity $\mathcal{O}(n^3)$ for a data set with n observations. This hinders their applicability especially in large data sets. Once a full GP model is fitted, ordering covariates using ARD comes about automatically, requiring no additional computations. By a projection approach (Piironen and Vehtari, 2016), the covariates can be ordered more effectively, but the drawback is an increase in complexity to $\mathcal{O}(p^2 n^3)$, where p is the dimension of the inputs.

The complexity of Gaussian process inference arises from the unavoidable matrix inversion. However, the same inverse can be used for making an arbitrary number of predictions at new test points, achieved by solving triangular systems, which are only $\mathcal{O}(n^2)$ in complexity. The Kullback-Leibler divergence method proposed for ordering covariates requires computing $2p + 1$ predictions at every training point, giving it a total complexity of $\mathcal{O}(p \cdot n \cdot n^2) = \mathcal{O}(pn^3)$. One prediction is made at the training point, which is compared to two predictions for every input dimension that are a distance Δ above and below the training point.

The variance method, on the other hand, requires computing as many predictions as the chosen number of quadrature points for every dimension and training point. This number can be chosen to be a small constant, thus keeping the total complexity at $\mathcal{O}(pn^3)$.

In addition to this, the method requires computing the inverse of the sample covariance submatrix of the inputs, $\Sigma_{-j,-j}$, for each of the p covariates. Taking advantage of the positive definiteness of the full covariance matrix, the Cholesky decomposition of it, $\mathcal{O}(p^3)$ in complexity, needs to be computed only once per training set. Then the Cholesky decomposition for each submatrix $\Sigma_{-j,-j}$ is obtained with a rank one update from the full covariance matrix, resulting in p rank one updates of complexity $\mathcal{O}(p^2)$. Thus, the full complexity of the variance method is $\mathcal{O}(pn^3 + p^3)$.

4 EXPERIMENTS

4.1 Toy Model

We mimic the argument against ARD made by Piironen and Vehtari (2016) by considering a similar toy model with eight covariates with varying degrees of nonlinearity. Besides uniformly distributed inputs, we also consider normally distributed inputs and define the toy model as follows:

$$\begin{aligned} y &= f_1(x_1) + \dots + f_8(x_8) + \varepsilon, \\ f_j(x_j) &= A_j \sin(\phi_j x_j), \quad j = 1, \dots, 8, \\ \varepsilon &\sim \mathcal{N}(0, 0.3^2), \end{aligned} \quad (14)$$

where $x_j \sim \mathcal{U}(-1, 1)$ or $x_j \sim \mathcal{N}(0, 0.4^2)$, respectively. The sine coefficients ϕ_j are equally spaced between $\pi/10$ and π , and the scaling factors A_j are such that the variance of each $f_j(x_j)$ is one, depending on the distribution of the corresponding input variable x_j . The functions f_j are presented in Figure 1 for uniformly distributed inputs (black) and normally distributed inputs (red).

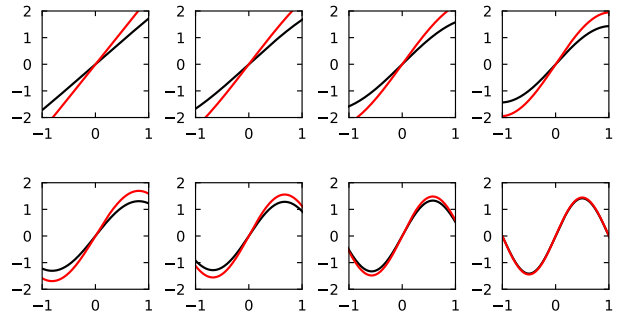


Figure 1: Latent functions $f_j(x_j)$, $j = 1, \dots, 8$ of the two toy models. Black represents the model with uniform inputs and red represents normally distributed inputs, each function scaled to unit variance according to its corresponding input distribution.

For both the toy models, we constructed a Gaussian process model with a covariance function being the squared exponential (4) with an added constant term, sampled 300 training points and computed the output values y according to equation (14). Using the full model with hyperparameters optimized to the maximum of marginal likelihood, we calculated the relevance of each covariate either directly using ARD, or by averaging the KL and VAR relevance estimates from each training point. The Gauss-Hermite integrations were computed using 11 quadrature points. The averaged results of 200 repetitions are presented in Figure 2 for the two models with inputs distributed uniformly (top) and normally (bottom). Input 1 is the most linear one and input 8 is the most nonlinear.

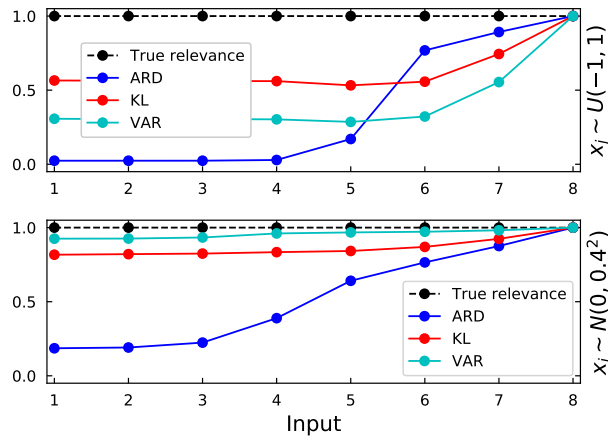


Figure 2: Relevance estimates for eight covariates in the two toy models (14) with uniformly distributed inputs (top) and normally distributed inputs (bottom). The estimates are computed with ARD (blue), KL divergence (red), and variance of the predictive mean (cyan). The results are averaged over 200 data realizations and scaled so that the most relevant covariate has a relevance of one. Error bars representing 95% confidence intervals are indistinguishable.

Figure 2 demonstrates that in the toy model with uniform inputs, all three methods prefer nonlinear inputs over linear ones to some extent. However, the preference in our methods is not as severe as with ARD, which assigns relevance values close to zero for half of the covariates. The bottom figure, representing the toy model with Gaussian distributed inputs, shows that our methods generate almost equal relevance values. Overall, our methods are notably better than ARD in identifying the true equal relevance of the covariates with varying degrees of nonlinearity.

4.2 Real World Data

In predictive feature selection, finding out the relevance sequence for the covariates is more important than the differences in the relevance values considered in Figure 2. To this end, we tested the performance of the three methods on four data sets obtained from the UCI machine learning repository¹. The data sets are summarized in Table 1. For each method, a Gaussian process model with a sum of constant and squared exponential (4) kernels as a covariance function was used. The model was first fitted with all p covariates included, and then a submodel was fitted with only part of the most relevant covariates included, according to the relevance ordering given by the particular method. We performed 50 repetitions, each time splitting the data into random training and test sets with the number of training points shown in Table 1. Both the full model and submodels were trained on the training set, and the predictive performance of the methods was evaluated by computing the mean log predictive densities (MLPDs) using the independent test set.

Table 1: Summary of real world dataset parameters: number of covariates p , data points n_{tot} , and training points used n .

Dataset	p	n_{tot}	n
Concrete	7	103	80
Boston housing	13	506	300
Automobile	38	193	150
Crime	102	1992	400

The mean log predictive densities of the test sets are presented in Figure 3 as a function of the number of covariates included in the submodel. A plot for each data set contains results when the covariates are sorted using ARD (blue), KL divergence (red), and variance of the predictive mean (cyan). The results are obtained by maximizing the hyperparameter posterior distribution, with half- t distribution as the prior for the noise and signal magnitudes, and inverse-gamma distribution for the length-scales. The inverse-gamma was chosen because it has a sharp left tail that penalizes very small length-scales, but its long right tail allows the length-scales to become large (Stan Development Team, 2017). The plots for the Automobile and Crime sets are shown only up to a point where the predictive performance saturates. The full model MLPD, presented as a horizontal

¹<https://archive.ics.uci.edu/ml/index.html>

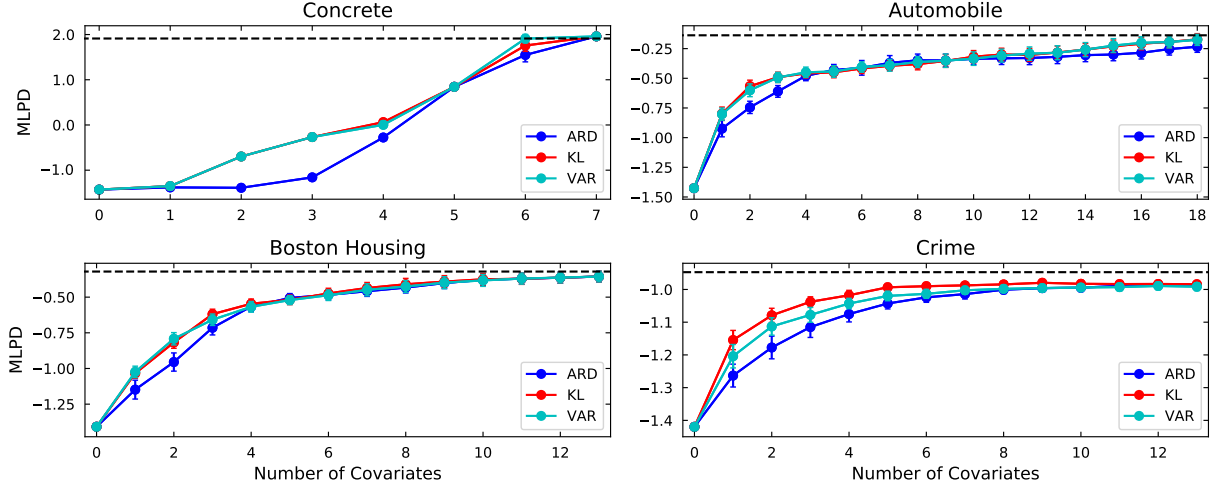


Figure 3: Mean log predictive densities (MLPDs) of the test sets with 95% confidence intervals for submodels as a function of covariates included in the submodel. Blue depicts covariates sorted using ARD, red and cyan depict our KL divergence and variance methods, respectively. The dashed horizontal line depicts the MLPD of the full model with hyperparameters sampled using HMC.

line, was computed by sampling 100 values for every training set from the hyperparameter posterior using Hamiltonian Monte Carlo (HMC) (Duane et al., 1987).

The results show that in all four data sets, both of our proposed methods generate a slightly better ordering for the covariates than ARD does, resulting in improved predictive performance. The improvement is most distinct in the first three or four covariates in all the data sets. This is because ARD picks the most nonlinear covariates first by definition, but our methods are able to identify covariates that are more relevant for prediction, albeit more linear. After the initial improvement, the ordering in the latter covariates is never worse than for ARD. Despite making the assumption of normally distributed inputs, the VAR method performs well even in the automobile data set which has multiple binary variables.

4.2.1 Ordering Consistency

The length-scale parameters of the squared exponential covariance function (4) are often weakly identified. Especially when a length-scale becomes large compared to the scale of the data, the generated function is essentially linear with respect to the corresponding covariate. Thus increasing the length-scale further does not significantly alter the likelihood.

The weak identifiability of the length-scale increases the variation in the relevance estimate given by ARD. To assess the variation, we examined how consistently

each method determines the relevance sequence of covariates between different random training sets in real data sets. For every covariate, we computed the relevance ordinal numbers given by each of the 50 training sets from the Concrete and Boston Housing data sets. For the Concrete data set, we plotted the relevance ordinal numbers of all seven covariates, and for the Housing data, we chose the 6 covariates that were the most relevant in the sense that they were picked early on average by the VAR method. The plots of the ordinal numbers given by ARD, KL, and VAR methods are presented in Figures 4 and 5. The markers in each ordinal number slot are jittered horizontally to visualize the number of points in each slot.

Figures 4 and 5 show that both of our proposed methods produce the relevance sequence more consistently compared to ARD. The plot of the Concrete data effectively displays the fact that the improved predictive performance is partly a result of improved consistency. For example, the better performance in the submodel with 6 covariates in Figure 3 is strictly the result of choosing covariate 5 more often than covariate 6. The Housing data plot in Figure 5 shows that while both our methods pick covariate 5 as the most relevant in a majority of training samples, ARD is less consistent, choosing covariates 12, 7, and 4 almost equiprobably. After the first choice, there is variance in every method, but it is largest with automatic relevance determination.

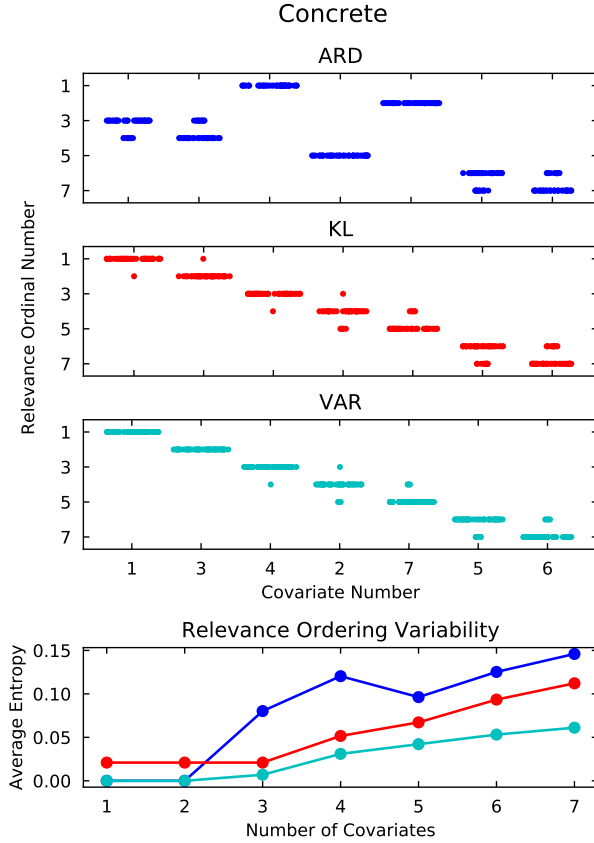


Figure 4: A scatter plot representing the frequency of relevance ordinal numbers between different training sets given to the 7 covariates in the Concrete data set. The average cumulative entropy in the bottom plot depicts the variability in the covariate ordering choices.

For both data sets, the more consistent covariate ordering of our methods correlates well with the improved predictive performance in Figure 3. This highlights the problems stated earlier for automatic relevance determination, namely the higher variance and that it overly prefers inputs with a nonlinear response, sacrificing predictive relevance as a consequence.

Below the scatter plots of Figures 4 and 5, the average variability of the covariate decisions between 50 training samples as a function of the number of covariates are shown. The plots were obtained by computing the cumulative entropy of each consecutive choice of covariates to add for the submodel, which is divided by the total number of covariates chosen to depict the average. The entropy plots demonstrate the fact that ARD has more variability

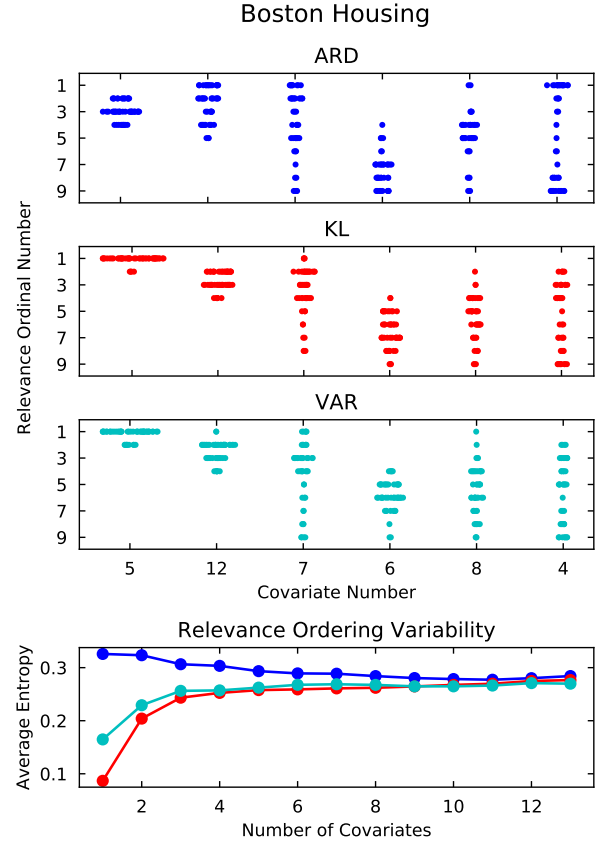


Figure 5: A scatter plot representing the frequency of relevance ordinal numbers between different training sets given to 6 relevant covariates in the Boston Housing data set. The average cumulative entropy in the bottom plot depicts the variability in the covariate ordering choices.

on average in generating the relevance sequence. For the Housing data set, the variability is greatest in the first choices, and for the Concrete data, in the latter choices. For the Automobile and Crime data sets, the differences in consistency were smaller, but the average variability of all submodel choices was the largest for ARD in all four data sets.

4.3 Estimation of Locally Relevant Covariates

In some cases, a covariate might have strong predictive relevance in some region, while being quite irrelevant on average. In some applications, the identification of such locally relevant covariates is important. Consider a hypothetical regression problem, where the covariates represent measurements to be made on a patient, and the dependent variable rep-

resents the progression of a disease. The information that some measurement has little relevance on average, but for some patients it is a clear indication of how far the disease has progressed, may provide essential information for medical professionals.

Because our methods compute relevance estimates at each training point, we get a distribution of relevance values in the regions of the input space where training points are located. The relevance values can then be addressed individually to assess the relevance of covariates in a subspace of the inputs. Using the toy model with normally distributed inputs (14), presented in Figure 1, we computed the distribution of the KL relevance values of the eighth covariate from a sample of 300 training points. The distribution is plotted in Figure 6 together with a sketch of the original latent function $f_8(x_8)$. While there is some noise, the distribution neatly captures both the flat and steep regions of the original sinusoidal function, exhibiting low and high relevance values, respectively.

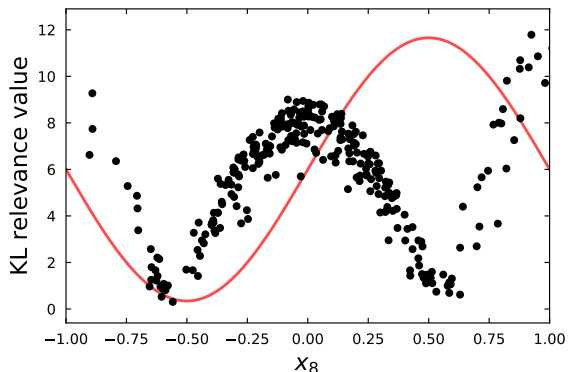


Figure 6: The distribution of the KL relevance values for the eighth covariate computed from a sample of 300 points from the toy model. The red curve is a sketch of the original latent function $f_8(x_8)$ and is not scaled properly.

The distribution of covariate relevance values presented in Figure 6 neatly captures one of the novel aspects of our proposed methods. While automatic relevance determination outputs only a single number representing the nonlinearity of a covariate, our methods give a distribution of relevance values. A single nonlinearity value can represent a myriad of different functions, and while some of them can be relevant for the regression problem, others can be very irrelevant. Our methods capture the relevance in the whole space where the training inputs are located. The mean of the produced relevance distri-

bution can be used to assess the average relevance of a covariate, but the distribution can also be used to examine relevances locally. As shown in section 4.2, the averaging already provides an improvement to the assessment of predictive relevance, and the ability to consider the local relevances is a useful extra feature, improving the applicability of the methods.

5 DISCUSSION

We have proposed two new methods for ordering covariates in Gaussian process models based on their predictive relevance. The methods utilize the predictions given by a full model in the vicinity of the training points. Our results on simulated and real world data sets show that the methods produce a slightly improved feature relevance ordering compared to the commonly used automatic relevance determination via length-scale parameters. The difference is greatest in the first few covariates, where ARD overly prefers the inputs with the most nonlinear response, but our methods are able to avoid this. Additionally, our methods were shown to generate the relevance ordering for covariates more consistently. If the task is to generate interpretable submodels, it is important to be confident about which covariates are the most relevant.

The methods proposed here require computing relevance values for each covariate in each point of the training data, thus increasing the computational complexity compared to automatic relevance determination. The variance method, in the form presented here, is also restricted to data sets where the dimensionality is less than the number of data points. Additionally, the assumption of normally distributed inputs may cause errors in data sets that are far from Gaussian. Despite their limitations, we believe the methods proposed here provide additional insight into feature selection and relevance determination for Gaussian process models, leading to more accurate and interpretable regression models.

References

- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222.
- Gelman, A. and Pardoe, I. (2007). Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37(1):23–51.
- Kullback, S. (1959). Statistics and information theory. *J. Wiley and Sons, New York*.

-
- Peltola, T., Havulinna, A. S., Salomaa, V., and Vehtari, A. (2014). Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop-Volume 1218*, pages 79–88. CEUR-WS. org.
- Piironen, J. and Vehtari, A. (2016). Projection predictive model selection for Gaussian processes. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.
- Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.
- Riihimäki, J., Sund, R., and Vehtari, A. (2010). Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach. *Health care management science*, 13(2):170–181.
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.
- Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual*. <http://mc-stan.org>. Version 2.16.0.
- Vehtari, A., Ojanen, J., et al. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

SUPPLEMENTARY MATERIAL

Toy Model with Irrelevant Covariates

In the toy model presented in Section 4.1, all covariates are equally relevant, thus it does not show how the methods treat irrelevant covariates. We also tested an extension of the toy model with 50 covariates, 40 of which had no impact on the target variable, and 10 equally relevant with each other. The 10 relevant covariates range from linear to nonlinear similarly as in the model in the paper. The relevance values for the 50 covariates are presented in Figure 7.

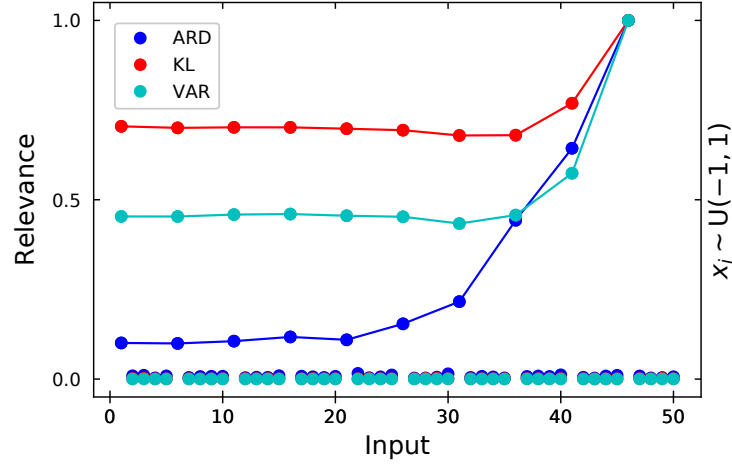


Figure 7: Relevance estimates for 50 covariates in the toy model with 10 equally relevant covariates and 40 irrelevant covariates. The estimates are computed with ARD (blue), KL divergence (red), and variance of the predictive mean (cyan). The 10 relevant covariates are joined with a line, and range from linear (input 1) to nonlinear (input 46). The results are averaged over 50 data realizations and scaled so that the most relevant covariate has a relevance of one.