SUPPLEMENTARY MATERIAL

KL Method Relevance Measure Equations

Gaussian Observation Model

For a Gaussian observation model, the predictive distribution of a Gaussian process model at a single test point is a univariate normal distribution. Let us denote the mean and variance of the predictive distribution at test point $\mathbf{x}^{(i)}$ as $\mu_i = \mathrm{E}[y_*|\mathbf{x}^{(i)}, \mathbf{y}]$ and $\sigma_i^2 = \mathrm{Var}[y_*|\mathbf{x}^{(i)}, \mathbf{y}]$, respectively. Analogously, denote the mean and variance of the predictive distribution at the perturbed point as $\mu_{i,\Delta_j} = \mathrm{E}[y_*|\mathbf{x}^{(i)} + \Delta_j, \mathbf{y}]$ and $\sigma_{i,\Delta_j}^2 = \mathrm{Var}[y_*|\mathbf{x}^{(i)} + \Delta_j, \mathbf{y}]$. The KL divergence between these distributions is

$$\log \frac{\sigma_{i,\Delta_j}}{\sigma_i} + \frac{\sigma_i^2 + (\mu_i - \mu_{i,\Delta_j})^2}{2\sigma_{i,\Delta_j}^2} - \frac{1}{2}.$$

The measure of predictive relevance in equation (2) is then

$$r(i, j, \Delta) = \frac{\sqrt{2}}{\Delta} \sqrt{\log \frac{\sigma_{i,\Delta_j}}{\sigma_i} + \frac{\sigma_i^2 + (\mu_i - \mu_{i,\Delta_j})^2}{2\sigma_{i,\Delta_j}^2} - \frac{1}{2}}.$$

Binary Classification

Consider a binary classification problem modelled with a Gaussian process. The predictive distribution at test point $\mathbf{x}^{(i)}$ is a Bernoulli distribution with success probability denoted as $\pi_* = p(y_* = 1|\mathbf{x}^{(i)}, \mathbf{y})$. The KL divergence between this distribution and the predictive distribution at a perturbed point, with success probability $\pi_{*,\Delta_j} = p(y_* = 1|\mathbf{x}^{(i)} + \Delta_j, \mathbf{y})$, is then

$$\pi_* \log \frac{\pi_*}{\pi_{*,\Delta_j}} + (1 - \pi_*) \log \frac{1 - \pi_*}{1 - \pi_{*,\Delta_j}}.$$

The measure of predictive relevance in equation (2) is then

$$r(i, j, \Delta) = \frac{\sqrt{2}}{\Delta} \sqrt{\pi_* \log \frac{\pi_*}{\pi_{*,\Delta_j}} + (1 - \pi_*) \log \frac{1 - \pi_*}{1 - \pi_{*,\Delta_j}}}.$$

Sensitivity of the KL Method to perturbation size Δ

We repeated the toy example from Section 4.1 and computed the KL relevance estimates with different values of the perturbation size Δ. All of the independent input variables have a uniform distribution $\mathrm{U}(-1, 1)$ and thus have a standard deviation of $1/\sqrt{3}$. Computed relevance estimates of the eight variables averaged from 50 data realizations are plotted in Figure 7. For reasonably small Δ values the results are identical. The results differ only when Δ is smaller than 10^{-7} or larger than 10^{-2}. $\Delta = 10^{-4}$ is a safe choice for most purposes unless the inputs have very small length-scale. In that case, one can make Δ smaller but should be cautious of numerical errors.

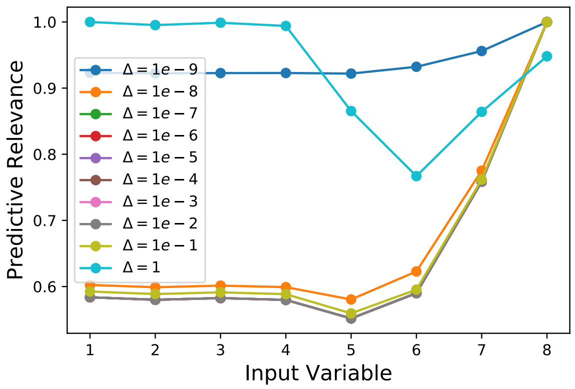

Figure 7: Relevance estimates given by the KL method for eight covariates in the toy example where each variable is equally relevant. The results are averaged over 50 data realizations and scaled so that the most relevant covariate has a relevance of one.

In-depth Look at Ranking Variability

To see the effect of ranking variability more clearly, we plotted markers for the variable ranks from each training split based on 50 training sets from the four regression data sets, and the results are presented in Figure 8. The markers are jittered horizontally to better illustrate the number of times each variable was assigned a specific relevance rank. The variables are ordered from left to right in terms of highest average relevance given by the KL method. A similar plot for the Pima Indians data set in shown in Figure 9.

For example, the plot of the Concrete data reveals the fact that the improved predictive performance in the chosen submodels is not only the result of being able to identify linear but relevant variables, but is also partly a result of less variation between different training sets. For example, the better performance in the submodel with six variables in Figure 3 is strictly the result of choosing variable 5 more often than variable 6, because all three methods always pick those two last, but ARD is more unsure about their order. The Housing data plot shows that while both the KL and VAR methods pick variable 5 as the most relevant in a majority of training samples, ARD is has more variability, choosing variables 12, 7, and 4 almost equiprobably.

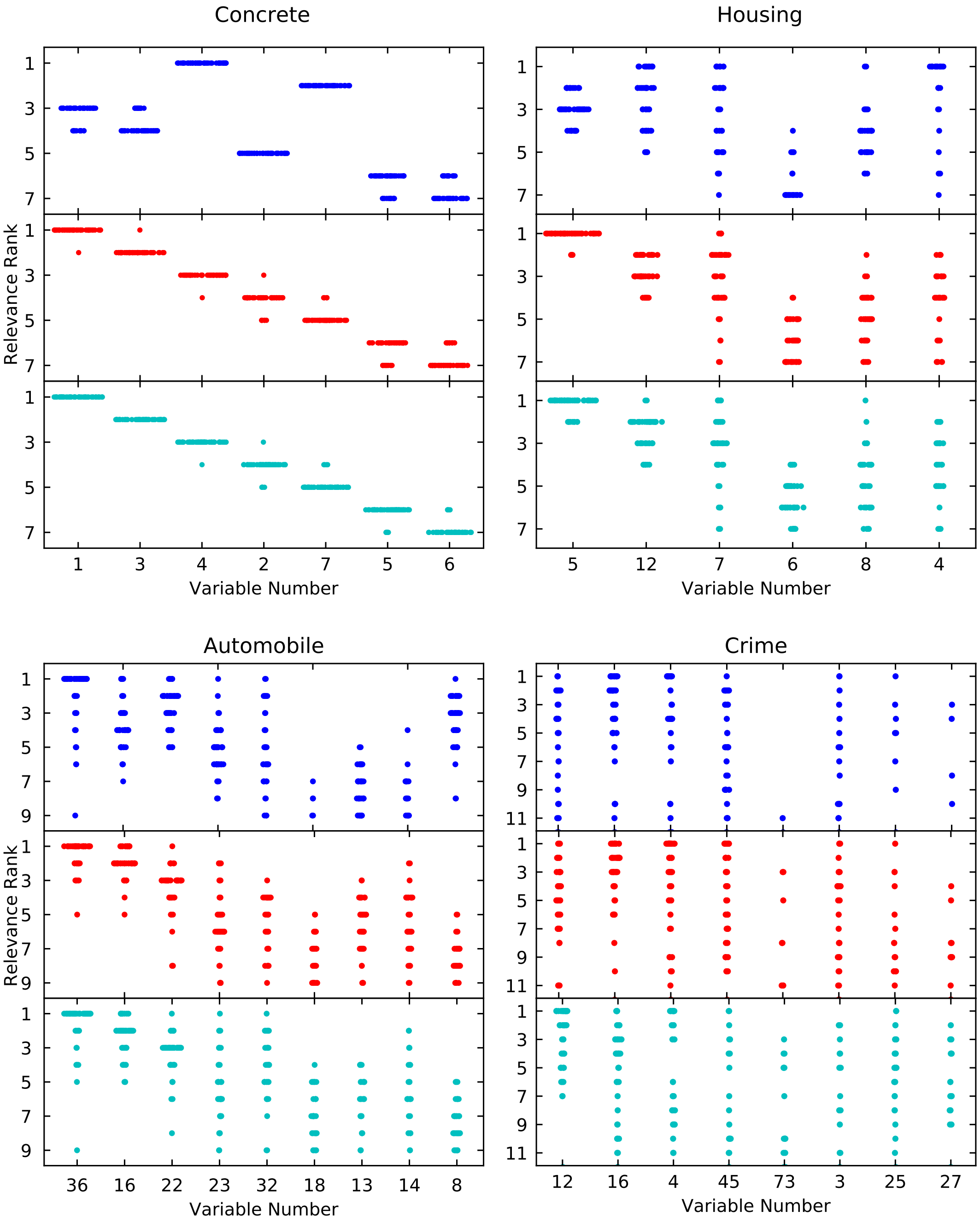

Figure 8: A plot representing the variability in relevance ranks between different training sets in the four regression data sets. Blue, red and cyan points represent ARD, KL and VAR ranking methods, respectively.

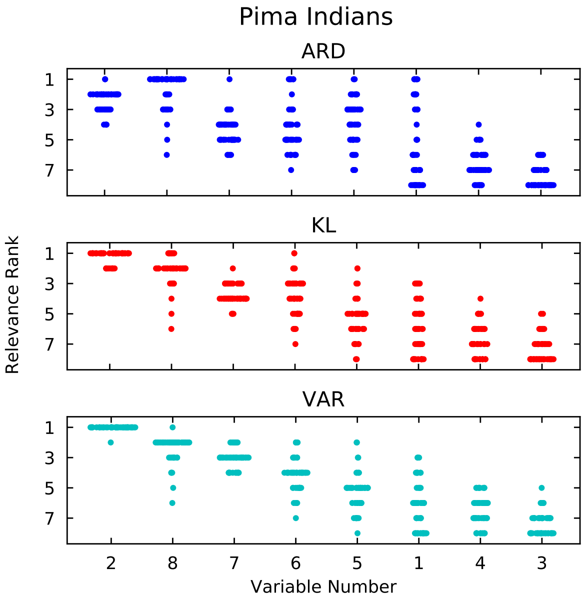

Figure 9: A plot representing the variability in relevance ranks between different training sets in the Pima Indians binary classification data set. Blue, red and cyan points represent ARD, KL and VAR ranking methods, respectively.

Toy Example With Irrelevant Variables

In the toy model presented in the paper, all input variables are equally relevant, thus it does not show how the methods treat irrelevant variables. We also tested an extension of the toy model with 50 variables, 42 of which had no impact on the target variable, and 8 equally relevant with each other. The 8 relevant variables range from linear to nonlinear similarly as in the original toy example in Section 4.1. The relevance values for the 50 variables are presented in Figure 10. The results show the same trend as the original toy example, namely that ARD overly prefers variables with a nonlinear response more than the KL and VAR methods.

Rank One Update of Cholesky Decomposition

This section presents the method for obtaining the Cholesky decomposition of a submatrix with one row and one column removed. This is done by updating the Cholesky decomposition of the full matrix with a rank-one update (Hager, 1989). Denote the full matrix and its Cholesky decomposition as $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\mathsf{T} \in \mathbb{R}^{p \times p}$. The goal is to obtain the Cholesky decomposition of the submatrix $\mathbf{\Sigma}_{-j,-j} = \mathbf{L}_{-j,-j}\mathbf{L}^\mathsf{T}_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$, where the row j and column j are removed from the full matrix $\mathbf{\Sigma}$. A direct Cholesky decomposition of the submatrix has a computational complexity of $\mathcal{O}(p^3)$, but a rank one update has only $\mathcal{O}(p^2)$. If the parts of the lower triangular matrix $\mathbf{L}$ are denoted as

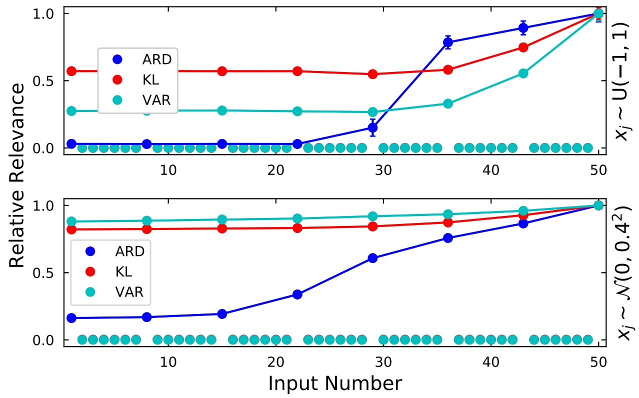

Figure 10: Relevance estimates for 50 covariates in the toy model with 8 equally relevant covariates and 42 irrelevant covariates. The estimates are computed with ARD (blue), KL (red), VAR (cyan) methods. The 8 relevant covariates are joined with a line, and range from linear (variable 1) to nonlinear (variable 50). The results are averaged over 50 data realizations and scaled so that the most relevant covariate has a relevance of one.

$$\mathbf{L} = \begin{array}{c c} & \begin{array}{ccc} <j & j & >j \end{array} \\ \begin{array}{c} <j \\ j \\ >j \end{array} & \begin{pmatrix} \mathbf{L}_A & \mathbf{0} & \mathbf{0} \\ \mathbf{l}_B^\mathsf{T} & l_{j,j} & \mathbf{0}^\mathsf{T} \\ \mathbf{L}_C & \mathbf{l}_D & \mathbf{L}_E \end{pmatrix} \end{array} \in \mathbb{R}^{p \times p}, \qquad (7)$$

The corresponding triangular matrix of the submatrix $\mathbf{\Sigma}_{-j,-j}$ is obtained as

$$\begin{aligned} \mathbf{L}_{-j,-j} &= \begin{pmatrix} \mathbf{L}_A & \mathbf{0} \\ \mathbf{L}_C & \tilde{\mathbf{L}}_E \end{pmatrix} \in \mathbb{R}^{(p-1)\times(p-1)}, \\ \tilde{\mathbf{L}}_E\tilde{\mathbf{L}}_E^\mathsf{T} &= \mathbf{L}_E\mathbf{L}_E^\mathsf{T} + \mathbf{l}_D\mathbf{l}_D^\mathsf{T}. \end{aligned} \qquad (8)$$

Because $\mathbf{l}_D$ is a vector, the modification to the Cholesky decomposition in equation (8) is a rank-one update.

Additional Predictive Performance Utilities for the Real World Data Sets

This section shows the predictive performance of chosen submodels in the real world data sets using different performance utilities. Figure 11 is the same as Figure 3, but shows mean squared error instead of mean log predictive density. Figure 12 is the same as Figure 4, but shows classification accuracy, precision, recall, and the F1 score instead of the mean log predictive density.

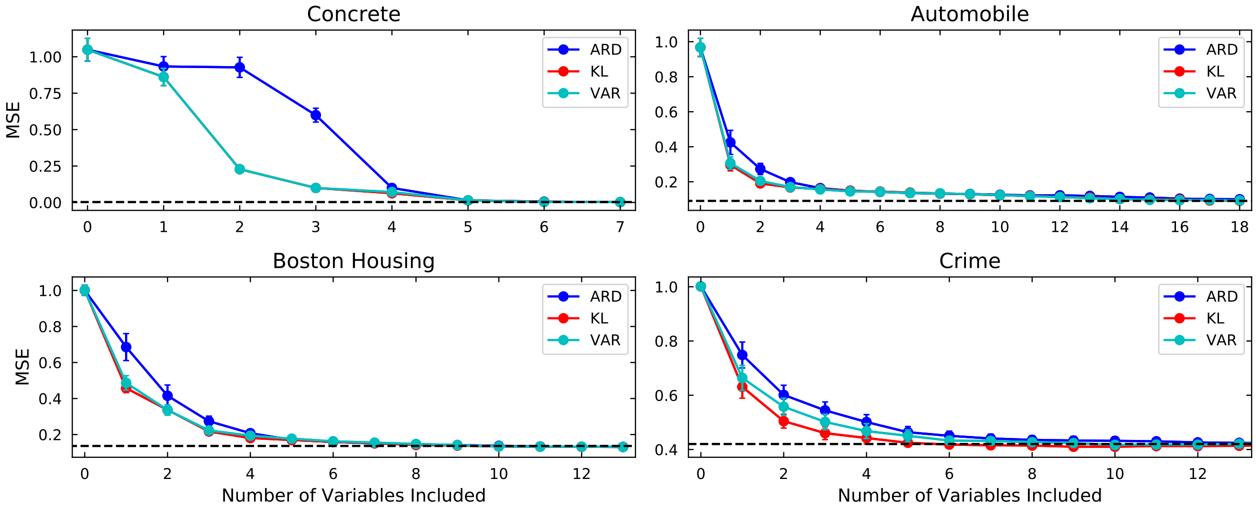

Figure 11: Mean squared errors (MSEs) of the test sets with 95% confidence intervals for submodels as a function of variables included in the submodel. Blue depicts variables sorted using ARD, red and cyan depict the KL and VAR methods, respectively. The dashed horizontal line depicts the MSE of the full model with hyperparameters sampled using the Hamiltonian Monte Carlo algorithm.

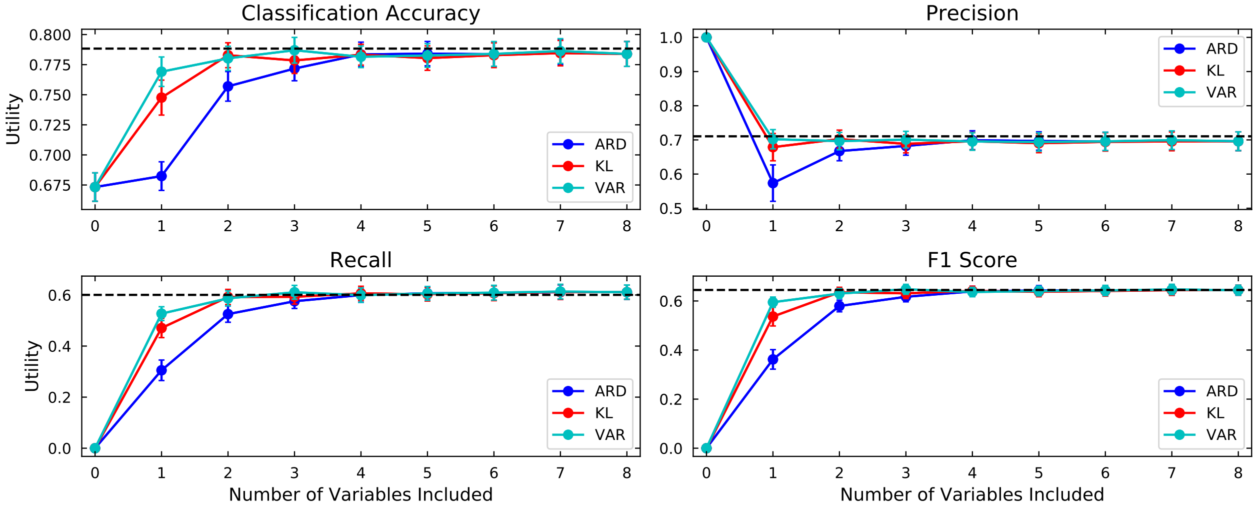

Figure 12: classification accuracy, precision, recall, and the F1 score of the test sets of the Pima indians data set with 95% confidence intervals for submodels as a function of the number of variables included in the submodel. The dashed horizontal line depicts the utilities of the full model with hyperparameters sampled using the Hamiltonian Monte Carlo algorithm.