Uncertainty-aware Sensitivity Analysis Using Rényi Divergences

Topi Paananen[1] **Michael Riis Andersen**[2] **Aki Vehtari**[1]

[1]Aalto University, Department of Computer Science, Helsinki Institute for Information Technology
[2]Technical University of Denmark, Department of Applied Mathematics and Computer Science

Preliminary version

Abstract

For nonlinear supervised learning models, assessing the importance of predictor variables or their interactions is not straightforward because importance can vary in the domain of the variables. Importance can be assessed locally with sensitivity analysis using general methods that rely on the model's predictions or their derivatives. In this work, we extend derivative based sensitivity analysis to a Bayesian setting by differentiating the Rényi divergence of a model's predictive distribution. By utilising the predictive distribution instead of a point prediction, the model uncertainty is taken into account in a principled way. Our empirical results on simulated and real data sets demonstrate accurate and reliable identification of important variables and interaction effects compared to alternative methods.

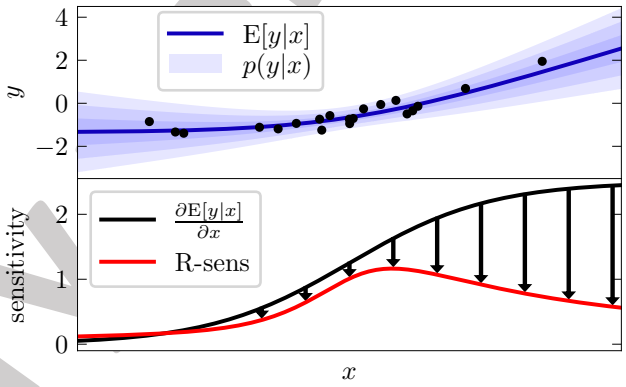

Figure 1: Top: Example of data and a probabilistic model with a Gaussian predictive distribution $p(y|x)$. The different shades of blue represent 1, 2, and 3 standard deviations of the predictive distribution. Bottom: The derivative of $\mathrm{E}[y|x]$ with respect to x (black) represents the naive sensitivity of the model's predictions to changes in x. The R-sens method proposed in this work (red) represents uncertainty-aware sensitivity as given by differentiating a Rényi divergence of predictive distributions, which adjusts the sensitivity according to uncertainty about y.

1 INTRODUCTION

Identifying important features and interactions from complex data sets and models remains a topic of active research. This is a fundamental problem with important applications in many scientific disciplines. Often the goal is to improve understanding of the model, but the identified features and interactions can also be used to build a simpler or more interpretable surrogate model.

For models that can capture nonlinear effects and interactions, the typical approach is to assess the contributions of individual predictors or interactions on the model's prediction at an individual observation. One approach is sensitivity analysis, which evaluates the change in predictions to small perturbations in the predictor values [Cacuci, 2003, Oakley and O'Hagan, 2004, Cacuci et al., 2005, Paananen et al., 2019]. For example, the partial derivative of the model's prediction with respect to the predictors can be a measure of importance [Guyon and Elisseeff, 2003]. Since the derivative can vary from positive to negative in the domain of the predictors, most approaches use absolute or squared derivatives averaged from the observations [Ruck et al., 1990, Dorizzi, 1996, Czernichow, 1996, Refenes and Zapranis, 1999, Leray and Gallinari, 1999, Sundararajan et al., 2017, Cui et al., 2020]. A similar approach is popular in image classification, where derivatives with respect to each pixel are called saliency maps [Simonyan et al., 2013, Zeiler and Fergus, 2014, Guidotti et al., 2018]. The average predictive comparison of Gelman and Pardoe [2007] uses the difference quotient of two predictions without taking the limit. By extending to cross-derivatives with respect to two predictors, one can also measure the interaction effect of predictors [Friedman et al., 2008, Cui et al., 2020]. A

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

closely related approach for sensitivity analysis is to directly estimate the contribution of predictor main effects or interactions to the variance of the target variable [Homma and Saltelli, 1996, Oakley and O'Hagan, 2004, Saltelli, 2002].

Recently, approaches that evaluate the differences of predictions in permuted training observations have gained popularity in machine learning. For example, Fisher et al. [2019] permute the observations of a single predictor, and examine the loss in predictive ability compared to the original data. Shapley values use permutations to assess the average marginal contribution of a predictor to a specific observation [Shapley, 1953, Štrumbelj and Kononenko, 2014]. Lundberg et al. [2018] extended this approach to evaluate second-order interactions based on the Shapley interaction index [Fujimoto et al., 2006]. Friedman et al. [2008] and Greenwell et al. [2018] use permutations and partial dependence functions [Friedman, 2001] to construct statistics that measure the strength of pairwise interactions. The individual conditional expectation plots of Goldstein et al. [2015] can also be used to identify interactions, but they rely on visualisation only.

This paper uses the curvature of Rényi divergence between predictive distributions to construct a uncertainty-aware sensitivity measure. Similar ideas have been used for measuring the sensitivity of Bayesian inference to the choice of prior [Al-Labadi et al., 2021]. Moreover, Dupuis et al. [2020] use Rényi divergence to measure the sensitivity of rare event probabilities.

The contributions of this work are summarised as follows. First, we present a novel method that generalises derivative and Hessian based sensitivity analysis to a Bayesian setting for models with a parametric predictive distribution or its approximation. Instead of using the first or second derivatives of the mean prediction of the model, we instead differentiate the Rényi divergence between two predictive distributions that coincide with each other, which takes into account the epistemic uncertainty of the predictions. Figure 1 gives an illustration of this method. Second, we show that our method is an analytical generalisation and extension of a previous finite difference method [Paananen et al., 2019]. Third, we show empirically that our proposed method can improve the accuracy of sensitivity analysis in situations where the used model has significant predictive uncertainty. Code for our method is freely available at https://github.com/topipa/rsens-paper.

2 UNCERTAINTY-AWARE SENSITIVITY

Consider a supervised learning model trained on data $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the design matrix and $\mathbf{y} \in \mathbb{R}^N$ is the vector of target observations. Let us denote the prediction function of the model for the target variable y as $f(\mathbf{x}^*)$. Derivative based sensitivity analysis can be used to assess the local sensitivity of f to the different predictors $(x_d)_{d=1}^D$. The sensitivity can be quantified by the partial derivative

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_d^*}.$$

Absolute values of local derivatives can be aggregated over the empirical distribution of $\mathbf{x}$ to obtain a global importance estimate for x_d, the expected absolute derivative [EAD; Leray and Gallinari, 1999, Cui et al., 2020]

$$\text{EAD}(x_d) = \text{E}_{p(\mathbf{x})}\left[\left|\frac{\partial f(\mathbf{x})}{\partial x_d}\right|\right]. \tag{1}$$

Similarly, absolute values of the elements of the Hessian matrix of f, that is, the second derivatives with respect to x_d and x_e, quantify the sensitivity to the joint interaction effect of x_d and x_e.

In this section, we present our proposed method, called R-sens, that extends derivative and Hessian based sensitivity analysis methods to a Bayesian setting where the evaluated model not only has a function for point predictions, but a *predictive distribution* $p(y^*)$. For now, we only consider predictive distributions that have some parametric form, which can be obtained exactly in closed form or it can be an approximation. Because the predictive distribution is obtained by integrating over uncertainty for the model parameters, it is important to utilise this uncertainty in sensitivity analysis as well. The predictive distribution does not need to be a posterior predictive distribution (i.e. conditioned on data), but in this work we only consider posterior predictive distributions.

To formulate a derivative based sensitivity measure for a model with a predictive distribution, we need a suitable functional of the predictive distribution, which to differentiate. We choose a family of statistical divergences called Rényi divergences due to their convenient properties, which we discuss later in this section. Rényi divergence of order α is defined for two probability mass functions $P = (p_1, ..., p_n)$ and $Q = (q_1, ..., q_n)$ as

$$\mathcal{D}_\alpha[P||Q] = \frac{1}{\alpha - 1} \log\left(\sum_{i=1}^{n} \frac{p_i^\alpha}{q_i^{\alpha-1}}\right)$$

when $0 < \alpha < 1$ or $1 < \alpha < \infty$ [Rényi et al., 1961, Van Erven and Harremos, 2014]. The definition generalises to continuous spaces by replacing the probabilities by densities and the sum by an integral. The divergences for values $\alpha = 0, 1$, and ∞ are obtained as limits. The most well-known Rényi divergence is the Kullback-Leibler divergence which is obtained in the limit $\alpha \to 1$ [Kullback and Leibler, 1951].

Let us consider a model with a predictive distribution parametrised by a vector $\boldsymbol{\lambda}^* = (\lambda_1^*, ..., \lambda_M^*)$, which depends on $\mathbf{x}^*$. Let us denote the predictive distribution for y

Preliminary version

conditional on predictor values $\mathbf{x}^*$ as

$$p(y^*) \equiv p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^*)).$$

Keeping $\mathbf{x}^*$ fixed, we denote the Rényi divergence of order α between two predictive distributions as a function of $\mathbf{x}^{**}$ as

$$\mathcal{D}_\alpha^p[\mathbf{x}^{**}] \equiv \mathcal{D}_\alpha[p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^*))||p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^{**}))].$$

We formalise the sensitivity of the predictive distribution to a change in a single predictor variable by differentiating the Rényi divergence in the limit when the distributions coincide, that is when $\mathbf{x}^{**} = \mathbf{x}^*$. However, because Rényi divergences obtain their minimum value when the two distributions coincide, the first derivative at this point is always zero. Hence, we formulate the uncertainty-aware sensitivity measure with respect to the predictor x_d using the second derivative

$$\left.\frac{\partial^2 \mathcal{D}_\alpha^p[\mathbf{x}^{**}]}{(\partial x_d^{**})^2}\right|_{\mathbf{x}^{**}=\mathbf{x}^*} = \left(\frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*}\right)^T \mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})}(\mathcal{D}_\alpha^p[\mathbf{x}^{**}]) \left(\frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^{**})}{\partial x_d^{**}}\right)\Bigg|_{\mathbf{x}^{**}=\mathbf{x}^*}, \tag{2}$$

where $\mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})}$ is the Hessian matrix of the Rényi divergence with second order derivatives with respect to $\boldsymbol{\lambda}^*(\mathbf{x}^{**})$.

The sensitivity measure in equation (2) has two kinds of partial derivatives: (i) second order derivatives of the Rényi divergence with respect to the parameters $\boldsymbol{\lambda}^*$ of the predictive distribution, and (ii) first order derivatives of the parameters $\boldsymbol{\lambda}^*$ with respect to the predictor x_d^*. These are obtained as follows:

(i) For sufficiently regular parametrisations, the second order Taylor approximation of the Kullback-Leibler divergence ($\alpha = 1$) gives an approximate equivalence between the Hessian of the divergence and the Fisher information matrix of $p(y^*)$ in the limit $\mathbf{x}^{**} - \mathbf{x}^* \to 0$ [Kullback, 1959, Van Erven and Harremos, 2014]. Haussler et al. [1997] state that this generalises to any Rényi divergence with $0 < \alpha < \infty$, leading to the relation

$$\mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})}(\mathcal{D}_\alpha^p[\mathbf{x}^{**}])|_{\mathbf{x}^{**}=\mathbf{x}^*} \approx \alpha \mathcal{I}(\boldsymbol{\lambda}^*(\mathbf{x}^*)), \tag{3}$$

where $\mathcal{I}(\boldsymbol{\lambda}^*(\mathbf{x}^*))$ is the Fisher information matrix of the distribution $p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^*))$.

(ii) The partial derivative of the parameter λ_k^* with respect to predictor variable x_d^* depends on the model where the predictive distribution is from.

We define R-sens, an uncertainty-aware sensitivity measure for predictor x_d at $\mathbf{x}^*$ as

$$\text{R-sens}(\mathbf{x}^*, x_d, \alpha) \equiv \sqrt{\alpha \left(\frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*}\right)^T \mathcal{I}(\boldsymbol{\lambda}^*(\mathbf{x}^*)) \left(\frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*}\right)}. \tag{4}$$

In a similar fashion as above, we generalise the Hessian based sensitivity to a Bayesian predictive distribution by differentiating the Rényi divergence four times, i.e. twice with respect to two predictors. However, the full fourth derivative contains cross-derivative terms, which we drop for two reasons. First, based on our experiments we concluded that the simplified formula we use is better at identifying interactions, meaning that the dropped terms do not contain useful information about the interaction effect between x_d and x_e. Second, the simplified formula is similar to the R-sens measure and is thus more easily interpretable and computationally cheaper. We define R-sens$_2$, the uncertainty-aware sensitivity measure for the interaction effect between variables x_d and x_e as

$$\text{R-sens}_2(\mathbf{x}^*, (x_d, x_e), \alpha) \equiv \sqrt{\alpha \left(\frac{\partial^2 \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^* \partial x_e^*}\right)^T \mathcal{I}(\boldsymbol{\lambda}^*(\mathbf{x}^*)) \left(\frac{\partial^2 \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^* \partial x_e^*}\right)}. \tag{5}$$

In the supplementary material, we show the full equation and an illustration of the benefit of equation (5) compared to the full fourth derivative.

2.1 ORDER OF RÉNYI DIVERGENCE

In the R-sens and R-sens$_2$ equations, the order of Rényi divergence α is only a prefactor. Thus, when using the uncertainty-aware sensitivity analysis for comparing observations or predictors to each other, the value of α often does not make a difference in practice. In all experiments, we use the value $\alpha = 1$. Other families of divergences may provide different results and serve as an interesting direction for future research.

2.2 LOCATION-SCALE FAMILY

For distributions in the location-scale family ($p(y|\lambda_1, \lambda_2) = g((y - \lambda_1)/\lambda_2)/\lambda_2$), the Fisher information of the location parameter λ_1 depends only on the scale parameter λ_2, but not the location parameter itself [Shao, 2006, Ch.3, Ex. 20]. This has two implications. First, if the predictive distribution is in the location-scale family, the R-sens and R-sens$_2$ measures are direct extensions to the absolute derivative or absolute Hessian of the mean prediction. The extension is twofold, as they introduce the derivative of the scale parameter and possible auxiliary parameters, and also multiplication with the Fisher information matrix. The uncertainty-aware measures can be also viewed as the Mahalanobis norm of the differentiated parameters of the predictive distribution instead of a simple Euclidean norm. Second, for a sufficiently regular model, where the posterior uncertainty vanishes in the asymptotic regime due to the Bernstein-von Mises theorem [Walker, 1969] such that the predictive distribution converges to a limiting distribution in the location-scale family, R-sens tends to the absolute derivative of the

Preliminary version

mean prediction. Here, we have to assume that the Fisher information exists and converges to the Fisher information of the limiting distribution.

2.3 ILLUSTRATIVE EXAMPLE

To illustrate the effects of the different components in equation (4), we analyse a Bayesian linear regression model as an example. We use the standard Gaussian likelihood with noise variance σ^2 and denote the regression coefficients as $\boldsymbol{\beta}$. Using an improper uniform prior on $(\boldsymbol{\beta}, \log\sigma)$, integrating over the uncertainty about all the parameters makes the posterior predictive distribution at any point $\mathbf{x}^*$ (treated as a row vector)

$$
\begin{aligned}
p(y^*|\mathbf{X},\mathbf{y}) &= \text{Student-}t(\mathrm{E}[y^*], \mathrm{Var}[y^*], \nu), \text{ where}\\
\mathrm{E}[y^*] &= \mathbf{x}^*\widehat{\boldsymbol{\beta}},\\
\mathrm{Var}[y^*] &= s^2(1+\mathbf{x}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{*T}),\\
\nu &= N-D,\\
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},\\
s^2 &= \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{N-D}.
\end{aligned}
$$

Here, ν represents the degrees of freedom, N and D are the number of observations and predictor variables, respectively, and $\widehat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of the regression coefficients. The derivative of ν with respect to x_d^* is zero, and the derivatives of the other two parameters of $p(y^*|\mathbf{X},\mathbf{y})$ are

$$
\begin{aligned}
\frac{\partial \mathrm{E}[y^*]}{\partial x_d^*} &= \widehat{\beta}_d,\\
\frac{\partial \mathrm{Var}[y^*]}{\partial x_d^*} &= 2s^2[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{*T}]_d \equiv 2s^2 V_d.
\end{aligned}
$$

Multiplying these with the Fisher information matrix of the Student-t predictive distribution, the R-sens sensitivity measure for the predictor x_d from equation (4) evaluates to

$$
\text{R-sens}(\mathbf{x}^*, x_d, \alpha=1) = \sqrt{\frac{(\nu+1)\widehat{\beta}_d^2 + \frac{2\nu s^4 V_d^2}{\mathrm{Var}[y^*]}}{(\nu+3)\mathrm{Var}[y^*]}}. \quad (6)
$$

The two summands have the following interpretations: In the absence of the second term, the measure would be proportional to $|\widehat{\beta}_d|$ divided by the standard deviation of the predictive distribution. The first term thus measures the absolute derivative of the mean prediction, but predictions with high uncertainty are given less weight. Also the second term in equation (6) quantifies the amount of uncertainty in the predictive distribution, but in a different way. Even if $\widehat{\beta}_d$ would be exactly zero, the second summand is nonzero as long as there is uncertainty about the model parameters that causes the predictive uncertainty to vary with respect to $\mathbf{x}^*$. There are thus two separate mechanisms that include epistemic uncertainty in the sensitivity analysis [O'Hagan, 2004, Kendall and Gal, 2017]. As N (and hence also ν) approaches infinity and the posterior uncertainty vanishes, the R-sens measures for both variables approach a constant proportional to $|\widehat{\beta}_d|$.

To visualise the effects of the two terms in equation (6), we simulated 10 observations from a linear model with two predictor variables x_1 and x_2 whose true regression coefficients are $\beta_1 = 1$ and $\beta_2 = 0$. The predictor variables are independent and normally distributed with zero mean and standard deviation one, and the simulated noise standard deviation is $\sigma_{\text{true}} = 0.5$. The R-sens sensitivities for both variables are shown in the bottom part of Figure 2 with solid lines. The dashed lines represent the contribution of the $\partial\mathrm{E}[y^*]/\partial x_d^*$ term, i.e. R-sens if $\partial\mathrm{Var}[y^*]/\partial x_d^*$ were zero, whereas the dotted line represents R-sens if $\partial\mathrm{E}[y^*]/\partial x_d^*$ were zero. The red color depicts the predictive distribution $p(y|x_1, x_2 = 0)$ (top) and R-sens (bottom) for x_1. The R-sens value is dominated by the contribution from the first summand in equation (6), where the Fisher information weighs down the sensitivity at the edges of the data because of the larger uncertainty. The blue color shows the predictive distribution $p(y|x_2, x_1 = 0)$ and R-sens for x_2. Now the first summand in equation (6) is almost zero because $\widehat{\beta}_2$ is small. In this case, the second term dominates because there is still a significant amount of epistemic uncertainty in the model. Comparing R-sens for x_1 and x_2 illustrates the two different ways that R-sens takes uncertainty into account.

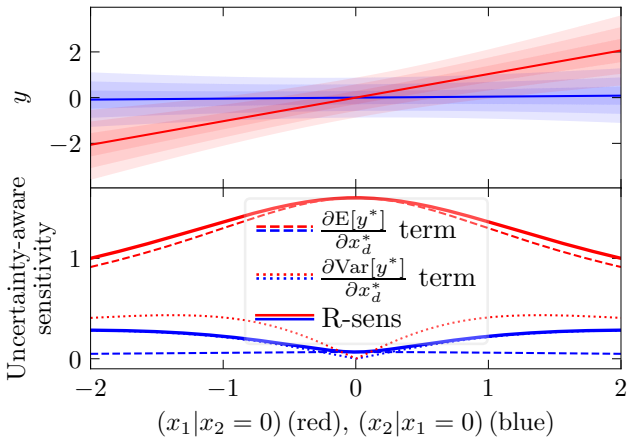

Figure 2: Top: Predictive distributions $p(y|x_1, x_2 = 0)$ (red) and $p(y|x_2, x_1 = 0)$ (blue) for the linear regression model in Section 2.3. Bottom: R-sens uncertainty-aware sensitivity measure for x_1 (red) and x_2 (blue). The dashed and dotted lines show the contributions of the two summands in equation (6).

Note that R-sens is model-agnostic in the sense that it does not take into account the fact that the prediction function is constrained to be linear in the example above. In addition, the different terms in equation (4) may not have such clear

Preliminary version

interpretations for other likelihoods or models. For example, in a binary classification task, the posterior predictive distribution can be considered a Bernoulli distribution which has only a single parameter. Nevertheless, the principle of taking into account the uncertainty through the predictive distribution still holds.

2.4 APPLICABILITY

The requirements for using the proposed R-sens and R-sens$_2$ methods are that we must have an analytical representation of the predictive distribution conditioned on the predictor variables, and that the derivatives of its parameters with respect to the predictors must be available. This somewhat restricts their applicability, but computational tools such as automatic differentiation make practical implementation easier [Baydin et al., 2018]. There are no restrictions to the support of the predictive distribution, and the methods are thus applicable to many learning tasks. However, because the methods adjust sensitivity based on the model uncertainty, the results may be misleading if the model lacks proper uncertainty quantification. We thus recommend using the methods for probabilistic models where the uncertainty is properly taken into consideration. In practice, the methods are most useful when the number of observations is relatively small and there is a lot of uncertainty about the parameters of the model.

Because the proposed methods measure the importance of predictor variables locally, they are most useful for nonlinear and complex models. For example, they can be useful for sensitivity analysis with Gaussian process models, which can represent flexible nonlinear functions with interactions, but still have good uncertainty quantification [O'Hagan, 1978, MacKay, 1998, Neal, 1998, Rasmussen and Williams, 2006]. Moreover, the predictive distribution is available in a parametric form, although for certain likelihoods some approximations are required. In the supplementary material we show the derivatives required for the R-sens and R-sens$_2$ methods for Gaussian processes and commonly used likelihoods.

The added computational expense of the R-sens and R-sens$_2$ methods compared to just differentiating the mean prediction depends on the used model. For many models, the cost is not significant compared to the cost of inference.

2.4.1 Global Measures

For assessing the global importance of predictor variables or pairs of variables, the local R-sens and R-sens$_2$ sensitivity measures can be aggregated over the empirical distribution of the predictors similarly to EAD in equation (1). By using the global measures, the predictor variables or pairwise interactions can be ranked by global importance. This approach is also taken in the experiments of Section 3.

2.5 FINITE DIFFERENCE APPROXIMATION

Paananen et al. [2019] propose a sensitivity analysis method abbreviated KL, which is a finite difference like method that evaluates the Kullback-Leibler divergence of predictive distributions when the predictor variables are perturbed. If we set $\alpha = 1$ (where Rényi divergence equals the Kullback-Leibler divergence), we show that the KL method is approximately equivalent to the second order Taylor approximation of R-sens

$$\text{R-sens}(\mathbf{x}^*, x_d, \alpha = 1) \approx \frac{\sqrt{2\mathcal{D}_1[p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^*))||p(y^*|\boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}}{|x_d^{**} - x_d^*|}.$$

Here, $\mathbf{x}^{**}$ is equivalent to $\mathbf{x}^*$ with predictor x_d perturbed, and $\mathcal{D}_1$ denotes the Kullback-Leibler divergence. We show the full derivation in the supplementary material. The benefit of the finite difference approximation is its generality, as it requires only an analytical representation of the predictive distribution but not the derivatives of its parameters with respect to the predictors. However, using R-sens avoids numerical errors related to finite difference and is easier because the selection of the perturbation size is avoided. For an appropriately chosen perturbation, the two equations produce practically identical results up to a small numerical error.

Our proposed R-sens$_2$ measure is not directly approximable with finite differences in the same way as R-sens. This is because it would require second-order finite differences, but the first-order finite difference using Kullback-Leibler divergence already reduces the predictive distribution into a single number. Riihimäki et al. [2010] perturb two predictor variables at a time with a unit length perturbation and measure the change in predictions by Kullback-Leibler divergence. For an infinitesimal perturbation this would be equivalent to a directional derivative instead of the cross-derivative in the R-sens$_2$ method that is required to properly assess interaction effects.

3 EXPERIMENTS

In this section, we demonstrate the practical utility of the methods discussed in Section 2 for identifying important predictor variables and interactions in nonlinear models. First, we evaluate different variable importance methods on simulated data using a hypothetical predictive function. This way we can control the quality of the model fit and limit the comparison strictly to the variable importance methods. Second, we will use Gaussian process models to evaluate ranking of main effects and interactions on both simulated and real data.

We compare the R-sens and R-sens$_2$ measures to several alternative variable importance methods: 1) Expected absolute derivative (EAD) or expected absolute Hessian (EAH),

Preliminary version

which correspond to R-sens and R-sens$_2$ without predictive uncertainty [Cui et al., 2020], 2) Absolute expected derivative (AED) or absolute expected Hessian (AEH) that take the expectation over $\mathbf{x}$ inside the absolute value [Cui et al., 2020], 3) Average predictive comparison (APC) [Gelman and Pardoe, 2007], 4) Shapley values [Shapley, 1953, Štrumbelj and Kononenko, 2014, Lundberg et al., 2018], 5) Partial dependence based importance (PD) [Greenwell et al., 2018], 6) Permutation feature importance (PFI) [Fisher et al., 2019], 7) Variance of the predictive mean (VAR) [Paananen et al., 2019], and 8) H-statistic [Friedman et al., 2008]. We omit comparison to the KL method of Paananen et al. [2019] because it is equivalent to R-sens up to numerical accuracy. We still show the results of their VAR method, which has no direct connection to R-sens. We have used the methods such that their computational cost is approximately equivalent. In the supplementary material, we detail the practical computational cost of the compared methods.

3.1 SIMULATED INDIVIDUAL EFFECTS

In the first experiment we compare different methods for ranking individual predictors based on their importance. We simulate 200 observations from 10 predictors, and construct the target variable y as a sum of 10 effects with added Gaussian noise

$$x_i \sim p_{x_i}(x_i),\ i = 1, \ldots, 10,$$
$$y = f_{\text{true}}(\mathbf{x}) = \sum_{i=1}^{10} A_i f_{\text{true},i}(x_i) + \varepsilon.$$

The shape of each effect $f_{\text{true},i}(x_i)$ is the same for all i, but they have different strengths varying from $A_1 = 1$ to $A_{10} = 10$. We consider 6 different experiments with different function shapes. By considering only a single effect shape for each experiment, we can unambiguously define the true importance of each predictor. We also repeat the experiment with 4 different distributions for the predictors.

When evaluating the ranking methods, we first use the true data generating function $f_{\text{true}}(\mathbf{x})$ as the mean prediction of the model. To simulate the uncertainty of the prediction model, we set the predictive distribution as Gaussian whose variance increases quadratically as distance from the mean of the data increases. Using the true data generating function allows us to strictly compare the ranking methods without being obfuscated by a non-optimal model fit. Because all of the compared methods use the mean prediction for ranking the predictors, using the true data generating function does not favour any single method over the others.

To consider the effect of taking uncertainty into account in the ranking, we also consider an imperfect version of the ground-truth model, where each term $A_i f_{\text{true},i}(x_i)$ is multiplied with a term $(|A_{\text{bias},i}||x_i|^3 + 1)$ where $A_{\text{bias},i}$ is drawn from a normal distribution with mean 0 and standard deviation 0.02. This simulates a situation where the model is correct where the uncertainty is small, but is biased at the edges of the data where the uncertainty is larger.

In Table 1 we show the results of different ranking methods for 6 different functions $f_{\text{true}}(\mathbf{x})$. Here, the distributions p_{x_i} of the predictors are independent Student-t_3 distributions. In the supplementary material, we show the results of the experiment with three alternative distributions p_{x_i}. The results are generated from 500 independent repetitions. In each repetition, the 10 predictors are ranked in importance from 1 to 10. For each data realisation, we compute the average error in the ranks across the predictors with respect to the true ranking, and compare that error to the ranking error of R-sens. A negative number thus means that the error is on average smaller than for R-sens. Table 1 reports the mean and 95% uncertainty invervals of the comparative ranking errors across the 500 independent data realisations.

The top section of Table 1 shows the ranking errors when using the ground-truth predictions function. R-sens and EAD are almost equivalent in many cases, but R-sens is significantly better for functions that have a large derivative in the tails of the data (x^3 and $x\exp(-x)$). Both R-sens and EAD outperform the other methods in most cases. AED does well for function that are monotonic, but it fails badly for non-monotonic functions. This is expected because the derivative of these functions varies from positive to negative.

In the bottom section of Table 1 when the model's predictions are imperfect, the difference in the ranking errors of R-sens and EAD is significantly larger in favour of R-sens. R-sens is also consistently better than the alternative methods with just a few exceptions. This shows that the uncertainty-aware sensitivity can be more reliable when there is a significant amount of uncertainty. In the supplementary material, we repeat the experiment with three alternative distributions p_{x_i}, including independent and correlated normal distributions. These results have similar conclusions: R-sens is mostly similar to EAD, but better in specific situations.

3.2 SIMULATED INDIVIDUAL AND PAIRWISE EFFECTS

In the second experiment, we study how accurately different methods evaluate interactions when the model has both main effects and interaction effects. We simulate 12 predictors and 8 main effects with different shapes and strengths, and three equally important pairwise interaction effects which are simply the product of the two predictors, i.e. $x_d x_e$. These are chosen such that the predictors of the first interaction do not have main effects, one of the predictors in the second interaction has a main effect, and both predictors of the third interaction have a main effect. To study how many observations the different methods require to reliably detect the true interactions in the data, we generate data with different

Preliminary version

Table 1: Average relative errors in rankings compared to R-sens and 95% uncertainty intervals from 500 data realisations in the simulated example of Section 3.1. A negative value indicates better ranking than R-sens.

Ground-truth models

Function $f_{\text{true},i}(x)$		R-sens	EAD	AED	APC	SHAP	PD	PFI	VAR
	x	**0**	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	2.3 ± 0.2	1.1 ± 0.1	3.6 ± 0.2	3.9 ± 0.2
	x^3	**0**	2.0 ± 0.2	2.0 ± 0.2	9.5 ± 0.5	10.0 ± 0.4	1.2 ± 0.3	15.8 ± 0.5	5.7 ± 0.3
	$x + \cos(3x)$	0	$\mathbf{-0.1 \pm 0.0}$	4.0 ± 0.2	5.9 ± 0.3	1.7 ± 0.2	1.8 ± 0.2	2.8 ± 0.2	3.0 ± 0.2
	$\sin(3x)$	**0**	$\mathbf{0.0 \pm 0.0}$	20.3 ± 0.6	11.5 ± 0.4	0.4 ± 0.1	0.3 ± 0.1	0.4 ± 0.1	8.3 ± 0.5
	$x \exp(-x)$	0	0.6 ± 0.1	0.6 ± 0.1	0.7 ± 0.3	1.1 ± 0.2	$\mathbf{-16.5 \pm 0.7}$	5.6 ± 0.7	-4.6 ± 0.7
	$\exp(-x^2)$	**0**	$\mathbf{0.0 \pm 0.0}$	20.5 ± 0.6	9.3 ± 0.3	0.3 ± 0.1	$\mathbf{-0.1 \pm 0.1}$	0.3 ± 0.1	$\mathbf{0.0 \pm 0.1}$

Imperfect models

Function $f_{\text{true},i}(x)$		R-sens	EAD	AED	APC	SHAP	PD	PFI	VAR
	x	**0**	2.6 ± 0.5	2.7 ± 0.5	10.8 ± 0.7	20.1 ± 0.7	22.4 ± 0.9	21.3 ± 0.7	1.1 ± 0.4
	x^3	0	1.4 ± 0.5	1.7 ± 0.5	6.7 ± 0.7	9.7 ± 0.8	12.3 ± 0.9	9.7 ± 0.8	$\mathbf{-4.4 \pm 0.5}$
	$x + \cos(3x)$	**0**	2.6 ± 0.4	6.7 ± 0.5	14.0 ± 0.6	23.4 ± 0.7	25.1 ± 0.9	24.8 ± 0.6	4.0 ± 0.4
	$\sin(3x)$	**0**	2.1 ± 0.3	18.9 ± 0.6	10.5 ± 0.5	14.8 ± 0.6	5.3 ± 0.9	20.8 ± 0.7	1.0 ± 0.3
	$x \exp(-x)$	0	0.2 ± 0.3	0.5 ± 0.3	4.3 ± 0.8	3.9 ± 0.8	5.2 ± 1.0	4.0 ± 0.8	$\mathbf{-2.7 \pm 0.8}$
	$\exp(-x^2)$	**0**	$\mathbf{0.0 \pm 0.0}$	20.6 ± 0.6	9.1 ± 0.3	0.3 ± 0.1	$\mathbf{0.0 \pm 0.1}$	0.3 ± 0.1	$\mathbf{0.0 \pm 0.1}$

numbers of observations ranging from 50 to 300.

In Figure 3, we plot the importance values averaged from 50 simulations for the three interacting variable pairs as well as three variable pairs without an interaction effect. The solid lines represent pairs with a true interaction, and the dotted lines are pairs without an interaction. In the left plot, both variables in the pairs have a main effect. In the middle plot, only one variable in the pairs has a main effect, and in the right plot neither variable has a main effect. For each of the 50 simulations, the interaction importance values are scaled so that the maximum given by each method is one. Thus, the ideal value is 1 for the solid lines and 0 for the dotted lines.

Figure 3 shows that even when increasing the number of observations, the HS method over-emphasizes the variable pair where neither variable has a main effect (right plot), whereas the PD method over-emphasizes the variable pair where both variables have a main effect (left plot). The other methods correctly identify the interactions as equally relevant when increasing the number of observations. For the true interactions (solid lines), EAH and R-sens$_2$ are almost indistinguishable, but R-sens$_2$ gives higher importance to the nonexistent interactions (dotted lines) when there is significant uncertainty because the number of observations is small.

3.3 BENCHMARK DATA SETS

In real data experiments, we focus on assessing the performance of the pairwise interaction method R-sens$_2$, because the experiments of Paananen et al. [2019] already demonstrate the effectiveness of (the finite difference approximation of) R-sens empirically. We use two publicly available data sets. The first is the Concrete Slump data set where the compressive strength of concrete is predicted based on the amount of different components included [Yeh, 2007]. The second is a Bike sharing data set, where the target variable is the hourly number of bike uses from a bicycle rental system [Fanaee-T and Gama, 2014]. The Concrete data has 103 observations and 7 predictors. From the Bike sharing data, we picked observations from February across two years, resulting in 1339 observations and 6 predictors. We model the problems using Gaussian process models with an exponentiated quadratic covariance function and either Gaussian (Concrete) or Poisson (Bike) likelihood. With the Poisson likelihood, we use the Laplace approximation for the latent values, and thus the resulting predictive distribution is an approximation but has an analytical solution. The details of the models and the derivatives needed for R-sens$_2$ are presented in the supplementary material.

To evaluate the plausibility of the interactions identified by the different methods, we compare the out-of-sample predictive performance of models with explicit interaction terms chosen based on interactions identified by each method. We compare the performance of the models using cross-validation with 50 random splits into training and test sets, and log predictive density as the utility function. The number of training observations used is 80 in the Concrete data and 500 in the Bike sharing data. For each training set, the Gaussian process model with full interactions is fitted, and the pairwise interactions are identified with R-sens$_2$ and 5 other methods. Based on these, models with only 0 to 5 pairwise interaction terms are fitted again, and their predictive performance is evaluated on the test data.

Preliminary version

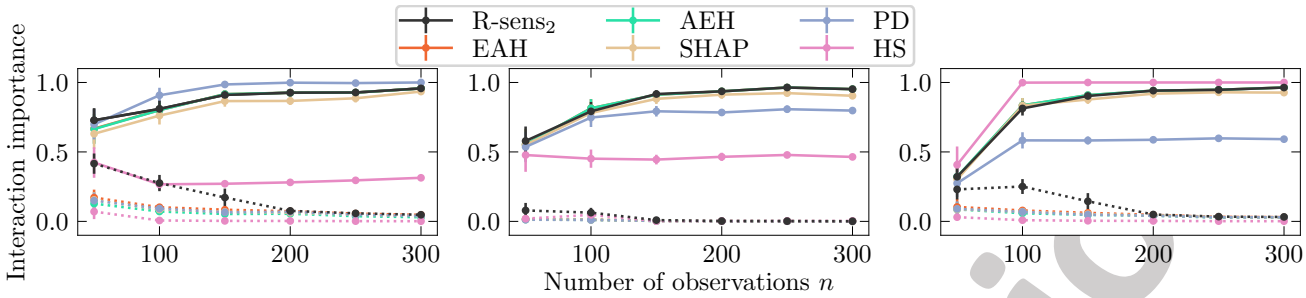

Figure 3: The interaction importance values given to six variable pairs in data sets with different numbers of observations. The solid lines represent pairs with a true interaction, and dotted lines are pairs without an interaction. The left plot depicts two pairs where both variables in the pairs have a main effect. In the middle plot, only one variable in the pairs has a main effect, and in the right plot neither variable has a main effect. In all plots, the ideal values would be 1 and 0 for the solid and dotted lines, respectively. The error bars represent 95% uncertainty intervals for the means from 50 simulated data sets.

The mean log predictive densities (MLPDs) across different test splits as well as 95% uncertainty intervals of the means are shown in Figure 4. The figure shows that in the Bike

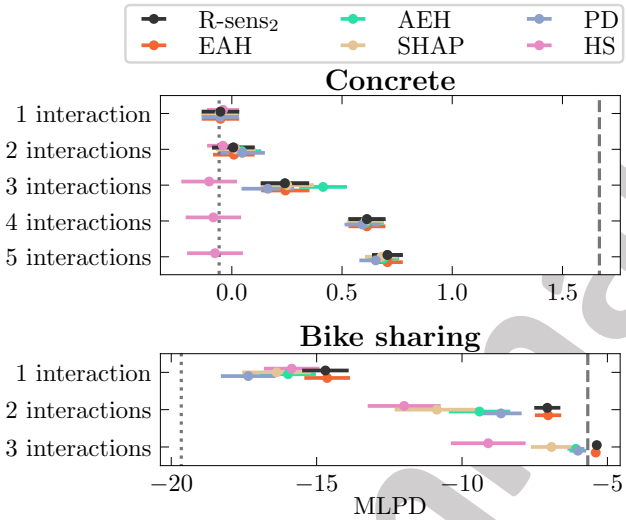

Figure 4: Mean log predictive densities (MLPDs) on independent test sets from the Concrete and Bike sharing data sets for Gaussian process models with different numbers of interactions. With each method, the interactions were identified from each training data set using the model with all interactions. The error bars represent 95% uncertainty intervals for the means from 50 different train-test splits. The dotted and dashed lines represent the MLPD for models with no interactions and all interactions, respectively.

sharing data, modelling only the three strongest pairwise interactions increases the out-of-sample predictive performance to the level of the model with all possible interactions. R-sens$_2$ does equally well compared to EAH, which both identify more important interactions on average than the competing methods. In the Concrete data set, adding even 5 pairwise interactions does not reach the performance of the model with all interactions. In this data there are no significant differences between the methods, except for HS which is clearly worse than the rest.

We also evaluate the stability of the interaction rankings by computing the variability in the rankings across 100 Bootstrap samples of the data. Table 2 shows the entropy in the rankings of each method across the Bootstrap samples. In both data sets, R-sens$_2$ and EAH have smaller entropies than the competing methods, meaning that their rankings are more stable.

Table 2: Variability in rankings across different Bootstrap samples of the benchmark data sets.

Data	R-sens$_2$	EAH	AEH	PD	HS	SHAP
Concrete	1.86	**1.80**	1.99	2.04	2.26	1.90
Bike	**1.61**	1.63	2.28	2.16	2.61	2.58

4 CONCLUSION

In this work, we presented an uncertainty-aware sensitivity analysis method that is based on differentiating Rényi divergences of predictive distributions. We showed that the method takes model uncertainty into account in a principled way and generalises to different likelihoods. For likelihoods in the location-scale family, the method is a direct extension to the absolute derivative or absolute Hessian of the mean prediction which are non-Bayesian sensitivity measures. Even though the method generalises to different predictive distributions, we recommend using it for models that have well calibrated uncertainty. The proposed method requires an analytical representation of the predictive distribution of the model, which is not available for all models. This could be generalised further, which is a possible direction for future research.

We demonstrated empirically that the method can reliably identify main effects as well as interactions in nonlinear models for complex data sets. In a controlled simulation

Preliminary version

setting, we showed that using uncertainty-aware sensitivity is beneficial in the presence of uncertainty when the used model may be wrong. Moreover, the proposed methods were equally good or better than previous derivative based sensitivity analysis methods in all of the tested cases. We can thus recommend using uncertainty-aware sensitivity analysis in modelling situations with little data and/or lots of uncertainty. We also demonstrated with two real data sets that our proposed method identifies pairwise interactions in nonlinear models that, when added to a model, improve its predictive performance. In addition, the ranking of the interactions between different Bootstrap samples of the data has less variation compared to many alternative variable importance methods.

Acknowledgements

We thank Alejandro Catalina and Kunal Ghosh for their helpful comments, and Mostafa Abdelrahman for GPyTorch and autodiff implementations. We thank the anonymous reviewers for helpful comments to improve the manuscript. We also acknowledge the computational resources provided by the Aalto Science-IT project and support by the Academy of Finland Flagship programme: Finnish Center for Artificial Intelligence, FCAI.

References

Luai Al-Labadi, Forough Fazeli Asl, and Ce Wang. Measuring Bayesian robustness using Rényi divergence. *Stats*, 4 (2):251–268, 2021.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153), 2018.

Dan G Cacuci. *Sensitivity and Uncertainty Analysis, volume I: Theory*. Boca Raton: Chapman and Hall/CRC, 2003.

Dan G Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. *Sensitivity and uncertainty analysis, volume II: applications to large-scale systems*. CRC press, 2005.

Tianyu Cui, Pekka Marttinen, and Samuel Kaski. Learning global pairwise interactions with Bayesian neural networks. *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

Thomas Czernichow. Architecture selection through statistical sensitivity analysis. In *International Conference on Artificial Neural Networks*, pages 179–184. Springer, 1996.

B Dorizzi. Variable selection using generalized RBF networks: Application to the forecast of the French T-bonds. *Proceedings of IEEE-IMACS'96, Lille, France*, 1996.

Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, Luc Rey-Bellet, et al. Sensitivity analysis for rare events based on Rényi divergence. *Annals of Applied Probability*, 30 (4):1507–1533, 2020.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman, Bogdan E Popescu, et al. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

Andrew Gelman and Iain Pardoe. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37(1): 23–51, 2007.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

David Haussler, Manfred Opper, et al. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996.

Preliminary Version

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5580–5590, 2017.

S Kullback. Statistics and information theory. *J. Wiley and Sons, New York*, 1959.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951.

Philippe Leray and Patrick Gallinari. Feature selection with neural networks. *Behaviormetrika*, 26(1):145–166, 1999.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

David JC MacKay. Introduction to Gaussian processes. In J Bishop, editor, *Neural Networks and Machine Learning*, pages 133–166. Springer Verlag, 1998.

Radford Neal. Regression and classification using Gaussian process priors (with discussion). In J Bernardo, J Berger, A Dawid, and A Smith, editors, *Bayesian statistics*, volume 6, pages 475–501. Oxford University Press, 1998.

Jeremy E Oakley and Anthony O'Hagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

Anthony O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

Tony O'Hagan. Dicing with the unknown. *Significance*, 1 (3):132–133, 2004.

Topi Paananen, Juho Piironen, Michael Riis Andersen, and Aki Vehtari. Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1743–1752. PMLR, 2019.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

A-PN Refenes and AD Zapranis. Neural model identification, variable selection and model adequacy. *Journal of Forecasting*, 18(5):299–332, 1999.

Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Jaakko Riihimäki, Reijo Sund, and Aki Vehtari. Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach. *Health care management science*, 13(2):170–181, 2010.

Dennis W Ruck, Steven K Rogers, and Matthew Kabrisky. Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, 2(2):40–48, 1990.

Andrea Saltelli. Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590, 2002.

Jun Shao. *Mathematical statistics: exercises and solutions*. Springer Science & Business Media, 2006.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3): 647–665, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Andrew M Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1):80–88, 1969.

I-Cheng Yeh. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and concrete composites*, 29(6):474–480, 2007.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Preliminary version