

---

# Uncertainty-aware Sensitivity Analysis Using Rényi Divergences (Supplementary material)

---

Topi Paananen<sup>1</sup>

Michael Riis Andersen<sup>2</sup>

Aki Vehtari<sup>1</sup>

<sup>1</sup>Aalto University, Department of Computer Science, Helsinki Institute for Information Technology

<sup>2</sup>Technical University of Denmark, Department of Applied Mathematics and Computer Science

## 1 R-SENS AND R-SENS<sub>2</sub> DERIVATIONS

### 1.1 R-SENS

$$\begin{aligned}
 & \left. \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}{(\partial x_d^{**})^2} \right|_{\mathbf{x}^{**}=\mathbf{x}^*} \\
 &= \left( \frac{\partial^2 \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial (x_d^*)^2} \right)^T \left( \frac{\mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}{\partial \boldsymbol{\lambda}^*(\mathbf{x}^{**})} \right) \Big|_{\mathbf{x}^{**}=\mathbf{x}^*} + \\
 & \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*} \right)^T \mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})} (\mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]) \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^{**})}{\partial x_d^{**}} \right) \Big|_{\mathbf{x}^{**}=\mathbf{x}^*}, \\
 &= \mathbf{0} + \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*} \right)^T \mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})} (\mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]) \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^{**})}{\partial x_d^{**}} \right) \Big|_{\mathbf{x}^{**}=\mathbf{x}^*}.
 \end{aligned} \tag{1}$$

### 1.2 R-SENS<sub>2</sub>

Here, we make the approximation that third and fourth derivatives of the Rényi divergence are zero. Let us start from the previous identity

$$\begin{aligned}
 & \left. \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}{(\partial x_d^{**})^2} \right|_{\mathbf{x}^{**}=\mathbf{x}^*} \\
 &= \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^*)}{\partial x_d^*} \right)^T \mathbf{H}_{\boldsymbol{\lambda}^*(\mathbf{x}^{**})} (\mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]) \left( \frac{\partial \boldsymbol{\lambda}^*(\mathbf{x}^{**})}{\partial x_d^{**}} \right) \Big|_{\mathbf{x}^{**}=\mathbf{x}^*} \\
 &= \sum_{k=1}^M \sum_{l=1}^M \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^*} \frac{\partial \lambda_k^*}{\partial x_d^*} \frac{\partial \lambda_l^*}{\partial x_d^{**}} \Big|_{\mathbf{x}^{**}=\mathbf{x}^*}.
 \end{aligned} \tag{2}$$

Then differentiating with respect to  $x_e$  gives the equality

$$\begin{aligned}
 & \left. \frac{\partial^3 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}{(\partial x_d^{**})^2 \partial x_e^{**}} \right|_{\mathbf{x}^{**}=\mathbf{x}^*} = \sum_{k=1}^M \sum_{l=1}^M \sum_{m=1}^M \frac{\partial^3 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^* \partial \lambda_m^*} \frac{\partial \lambda_k^*}{\partial x_d^*} \frac{\partial \lambda_l^*}{\partial x_d^*} \frac{\partial \lambda_m^*}{\partial x_d^{**}} \\
 & + \sum_{k=1}^M \sum_{l=1}^M \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^*} \left( \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial \lambda_l^*}{\partial x_d^*} + \frac{\partial \lambda_k^*}{\partial x_d^*} \frac{\partial^2 \lambda_l^*}{\partial x_d^* \partial x_e^*} \right) \\
 & = \sum_{k=1}^M \sum_{l=1}^M \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^*} \left( \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial \lambda_l^*}{\partial x_d^*} + \frac{\partial \lambda_k^*}{\partial x_d^*} \frac{\partial^2 \lambda_l^*}{\partial x_d^* \partial x_e^*} \right).
 \end{aligned} \tag{3}$$

Differentiating a second time gives

$$\begin{aligned} & \left. \frac{\partial^4 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^*)) || p(y^* | \boldsymbol{\lambda}^*(\mathbf{x}^{**}))]}{(\partial x_d^{**})^2 (\partial x_e^{**})^2} \right|_{\mathbf{x}^{**} = \mathbf{x}^*} \\ &= \sum_{k=1}^M \sum_{l=1}^M \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^*} \left( \frac{\partial^3 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial \lambda_l^*}{\partial x_d^*} + \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial^2 \lambda_l^*}{\partial x_d^* \partial x_e^*} + \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial^2 \lambda_l^*}{\partial x_d^* \partial x_e^*} + \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial^3 \lambda_l^*}{\partial x_d^* (\partial x_e^*)^2} \right). \end{aligned} \quad (4)$$

Dropping the third derivative terms and the factor 2 results in

$$\sum_{k=1}^M \sum_{l=1}^M \frac{\partial^2 \mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]}{\partial \lambda_k^* \partial \lambda_l^*} \frac{\partial^2 \lambda_k^*}{\partial x_d^* \partial x_e^*} \frac{\partial^2 \lambda_l^*}{\partial x_d^* \partial x_e^*} = \left( \frac{\partial^2 \boldsymbol{\lambda}^*}{\partial x_d^* \partial x_e^*} \right)^T \mathbf{H}(\mathcal{D}_\alpha [p(y^* | \boldsymbol{\lambda}^*) || p(y^* | \boldsymbol{\lambda}^*)]) \left( \frac{\partial^2 \boldsymbol{\lambda}^*}{\partial x_d^* \partial x_e^*} \right). \quad (5)$$

## 2 FINITE DIFFERENCE APPROXIMATION OF THE KULLBACK-LEIBLER DIVERGENCE

Consider two probability distributions,  $p(\cdot|\lambda^*)$  and  $p(\cdot|\lambda^{**})$  parameterised by vectors  $\lambda^*$  and  $\lambda^{**}$ , respectively. Keeping  $\lambda^*$  constant, let us make a second-order approximation of the Kullback-Leibler divergence between the distributions in the neighbourhood around  $\lambda^{**} = \lambda^*$ .

$$\begin{aligned} & \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**})) \\ &= \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**})) \Big|_{\lambda^{**}=\lambda^*} + \sum_{k=1}^M (\lambda_k^{**} - \lambda_k^*) \frac{\partial \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{\partial \lambda_k^{**}} \Big|_{\lambda^{**}=\lambda^*} \\ &+ \frac{1}{2} \sum_{k=1}^M \sum_{l=1}^M \left[ (\lambda_k^{**} - \lambda_k^*) (\lambda_l^{**} - \lambda_l^*) \frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{\partial \lambda_k^{**} \partial \lambda_l^{**}} \Big|_{\lambda^{**}=\lambda^*} \right] + \mathcal{O}(\|\lambda^{**} - \lambda^*\|^3). \end{aligned}$$

The first two terms are zero, because the Kullback-Leibler divergence obtains a minimum value of zero at  $\lambda^{**} = \lambda^*$ . Dropping them and the third degree term, we are left with the approximation

$$\mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**})) \approx \frac{1}{2} \sum_{k=1}^M \sum_{l=1}^M \left[ (\lambda_k^{**} - \lambda_k^*) (\lambda_l^{**} - \lambda_l^*) \frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{\partial \lambda_k^{**} \partial \lambda_l^{**}} \Big|_{\lambda^{**}=\lambda^*} \right].$$

If the distributions  $p(\cdot|\lambda^*)$  and  $p(\cdot|\lambda^{**})$  are predictive distributions, then the parameters  $\lambda^*$  and  $\lambda^{**}$  depend on the predictor value  $\mathbf{x}$ , i.e.  $\lambda^{**} = \lambda(\mathbf{x}^{**})$ . When only one predictor variable,  $x_d$ , is varied, an infinitesimal change in the parameters can be written as

$$\lambda_k^{**} - \lambda_k^* = \frac{\partial \lambda_k^{**}}{\partial x_d^{**}} (x_d^{**} - x_d^*).$$

Thus we get

$$\mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**})) \approx \frac{1}{2} (x_d^{**} - x_d^*)^2 \sum_{k=1}^M \sum_{l=1}^M \left( \frac{\partial \lambda_k^{**}}{\partial x_d^{**}} \right) \left( \frac{\partial \lambda_l^{**}}{\partial x_d^{**}} \right) \frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{\partial \lambda_k^{**} \partial \lambda_l^{**}} \Big|_{\lambda^{**}=\lambda^*}.$$

Rearranging the terms gives the approximate equivalence

$$\begin{aligned} \frac{2\mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{(x_d^{**} - x_d^*)^2} &\approx \sum_{k=1}^M \sum_{l=1}^M \left( \frac{\partial \lambda_k^{**}}{\partial x_d^{**}} \right) \left( \frac{\partial \lambda_l^{**}}{\partial x_d^{**}} \right) \frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{\partial \lambda_k^{**} \partial \lambda_l^{**}} \Big|_{\lambda^{**}=\lambda^*} \\ &= \frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{(\partial x_d^{**})^2} \Big|_{\lambda^{**}=\lambda^*}. \end{aligned}$$

The last identity is based on the chain rule of differentiation. Finally, taking the square root gives the approximate equivalence

$$\frac{\sqrt{2\mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}}{|x_d^{**} - x_d^*|} \approx \sqrt{\frac{\partial^2 \mathcal{D}_{\text{KL}}(p(\cdot|\lambda^*)||p(\cdot|\lambda^{**}))}{(\partial x_d^{**})^2}} \Big|_{\lambda^{**}=\lambda^*},$$

where the left hand side is the finite difference KL method of Paananen et al. [2019] and the right hand side is the R-sens measure with  $\alpha = 1$ .

### 3 R-SENS<sub>2</sub> APPROXIMATION BENEFITS

In this section, we show an example of how the simplified R-sens<sub>2</sub> formula we use is better than the full formula that includes cross-derivative terms. With full formula we mean the fourth derivative of the Rényi divergence without dropping any terms. We replicate the simulation experiment from Section 3.2 of the main paper such that we compute interaction importance estimates using the simplified R-sens<sub>2</sub> formula and the full formula that is obtained with automatic differentiation. In Figure 1 we show the different interaction importances for a single simulation. The three annotated pairs are the true simulated interactions, whereas all the other interactions are irrelevant. The figure shows that the two formulas give almost equivalent importances for the true interactions, but the simplified formula gives much lower importance estimates for the irrelevant interactions, thus having significantly better ability to separate true interactions from nonexistent interactions.

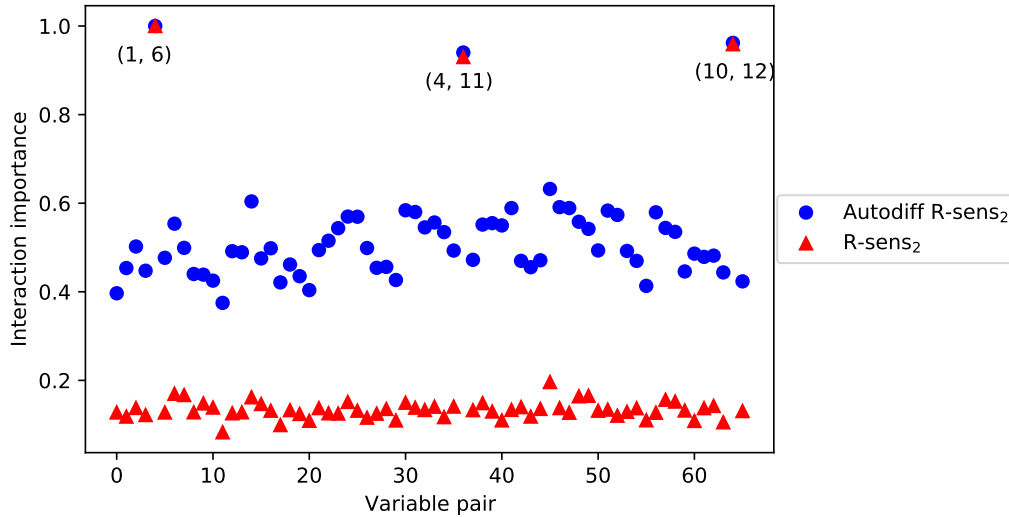


Figure 1: Comparison of interaction importance estimates for R-sens<sub>2</sub> (red) and the fourth derivative of the Rényi divergence (blue).

## 4 COMPUTATIONAL COST DETAILS

Here, we discuss the computational cost of the variable importance methods used in the main paper. Let us denote the number of observations with  $N$  and the number of predictor variables with  $D$ . Let us also denote the number of possible pairwise interactions with  $\frac{D(D+1)}{2} \equiv D_2$ . Let us denote the costs of making predictions from a regression model with a location-scale likelihood with

- $C_E$  cost of  $E[y]$ ,
- $C_V$  cost of  $\text{Var}[y]$ ,
- $\tilde{C}_E$  cost of  $\frac{\partial E[y]}{\partial x_d}$ ,
- $\tilde{C}_V$  cost of  $\frac{\partial \text{Var}[y]}{\partial x_d}$ ,
- $\hat{C}_E$  cost of  $\frac{\partial^2 E[y]}{\partial x_d^2}$ ,
- $\hat{C}_V$  cost of  $\frac{\partial^2 \text{Var}[y]}{\partial x_d^2}$ .

The computational cost of the variable importance methods can be tuned based on the amount of computational resources. We tried to tune the cost of each method roughly equal in order to make the comparison fair. The computational costs that were used in the experiment of Section 3.1 in the main paper are shown in Table 1, and the costs used in the Concrete data experiment of Section 3.3 are shown in Table 2.













Table 1: Computational costs of the variable importance methods used in the first experiment of the main paper.

| Method | Time complexity                       |
|--------|---------------------------------------|
| R-sens | $ND(C_V + \tilde{C}_E + \tilde{C}_V)$ |
| EAD    | $ND\tilde{C}_E$                       |
| AED    | $ND\tilde{C}_E$                       |
| APC    | $2ND\tilde{C}_E + D^3$                |
| SHAP   | $2ND\tilde{C}_E$                      |
| PD     | $9ND\tilde{C}_E$                      |
| PFI    | $N(D+1)(C_E + C_V)$                   |
| VAR    | $ND\tilde{C}_E + D^3$                 |

Table 2: Computational costs of the variable importance methods (for interactions) used in the Concrete experiment of the main paper.

| Method | Time complexity                     |
|--------|-------------------------------------|
| R-sens | $ND_2(C_V + \hat{C}_E + \hat{C}_V)$ |
| EAH    | $ND_2\hat{C}_E$                     |
| AEH    | $ND_2\hat{C}_E$                     |
| SHAP   | $4ND_2\hat{C}_E$                    |
| PD     | $9ND_2\hat{C}_E$                    |
| HS     | $N^2D_2\hat{C}_E$                   |

Table 3: Average error in rankings in the simulated example of the main paper. Each predictor has an independent standard normal distribution.

| Ground-truth models  |          |                                 |                                 |                                 |               |                                  |                |                                  |  |
|--|----------|---------------------------------|---------------------------------|---------------------------------|---------------|----------------------------------|----------------|----------------------------------|--|
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD                             | AED                             | APC                             | SHAP          | PD                               | PFI            | VAR                              |  |
|  $x$            | <b>0</b> | <b>0</b>                        | <b>0</b>                        | <b>0</b>                        | $0.9 \pm 0.1$ | $0.7 \pm 0.1$                    | $1.0 \pm 0.1$  | $0.4 \pm 0.1$                    |  |
|  $x^3$          | <b>0</b> | $0.0 \pm 0.1$                   | $0.0 \pm 0.1$                   | $5.9 \pm 0.4$                   | $2.8 \pm 0.2$ | $3.3 \pm 0.2$                    | $4.2 \pm 0.3$  | $2.0 \pm 0.2$                    |  |
|  $x + \cos(3x)$ | <b>0</b> | $0.0 \pm 0.0$                   | $3.9 \pm 0.2$                   | $8.0 \pm 0.3$                   | $0.8 \pm 0.1$ | <b><math>0.0 \pm 0.1</math></b>  | $0.7 \pm 0.1$  | <b><math>0.0 \pm 0.1</math></b>  |  |
|  $\sin(3x)$     | <b>0</b> | $0.0 \pm 0.0$                   | $21.0 \pm 0.6$                  | $10.8 \pm 0.3$                  | $0.4 \pm 0.1$ | $0.1 \pm 0.0$                    | $0.3 \pm 0.1$  | $3.2 \pm 0.2$                    |  |
|  $x \exp(-x)$   | <b>0</b> | $0.4 \pm 0.1$                   | $0.5 \pm 0.1$                   | $7.9 \pm 0.4$                   | $3.1 \pm 0.3$ | $1.6 \pm 0.2$                    | $6.0 \pm 0.3$  | $2.2 \pm 0.2$                    |  |
|  $\exp(-x^2)$   | 0        | $0.0 \pm 0.0$                   | $20.5 \pm 0.5$                  | $7.2 \pm 0.3$                   | $0.5 \pm 0.1$ | $0.4 \pm 0.1$                    | $0.4 \pm 0.1$  | <b><math>-0.1 \pm 0.0</math></b> |  |
| Imperfect models   |          |                                 |                                 |                                 |               |                                  |                |                                  |  |
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD                             | AED                             | APC                             | SHAP          | PD                               | PFI            | VAR                              |  |
|  $x$            | <b>0</b> | $0.2 \pm 0.1$                   | <b><math>0.1 \pm 0.2</math></b> | <b><math>0.2 \pm 0.2</math></b> | $1.3 \pm 0.2$ | <b><math>0.1 \pm 0.1</math></b>  | $1.9 \pm 0.2$  | $0.2 \pm 0.1$                    |  |
|  $x^3$          | <b>0</b> | <b><math>0.2 \pm 0.3</math></b> | <b><math>0.2 \pm 0.3</math></b> | $6.2 \pm 0.4$                   | $5.1 \pm 0.4$ | $2.0 \pm 0.3$                    | $7.5 \pm 0.5$  | $1.1 \pm 0.3$                    |  |
|  $x + \cos(3x)$ | <b>0</b> | $0.0 \pm 0.1$                   | $4.0 \pm 0.2$                   | $8.1 \pm 0.3$                   | $1.5 \pm 0.2$ | <b><math>-0.1 \pm 0.1</math></b> | $1.7 \pm 0.2$  | <b><math>0.0 \pm 0.1</math></b>  |  |
|  $\sin(3x)$     | <b>0</b> | $0.0 \pm 0.0$                   | $20.9 \pm 0.5$                  | $10.4 \pm 0.3$                  | $0.4 \pm 0.1$ | <b><math>0.1 \pm 0.1</math></b>  | $0.4 \pm 0.1$  | $3.2 \pm 0.2$                    |  |
|  $x \exp(-x)$   | <b>0</b> | $0.5 \pm 0.3$                   | $0.7 \pm 0.3$                   | $8.3 \pm 0.5$                   | $5.8 \pm 0.4$ | $2.8 \pm 0.4$                    | $10.0 \pm 0.5$ | $2.6 \pm 0.3$                    |  |
|  $\exp(-x^2)$   | 0        | $0.0 \pm 0.0$                   | $20.5 \pm 0.5$                  | $7.2 \pm 0.3$                   | $0.5 \pm 0.1$ | $0.5 \pm 0.1$                    | $0.4 \pm 0.1$  | <b><math>-0.1 \pm 0.0</math></b> |  |

## 5 SIMULATED INDIVIDUAL EFFECTS - ADDITIONAL RESULTS

In this Section, we show additional results for the simulated experiment of Section 3.1 in the main paper. Here, we show the results with different distributions for the predictor variables: 1) Independent Gaussians (Table 3), 2) mixtures of 2 Gaussians (Table 4), and 3) correlated Gaussians (Table 5).

Table 4: Average error in rankings in the simulated example of the main paper. Each predictor is independently distributed with a mixture of 2 Gaussians.

























| Ground-truth models  |          |                                 |                                 |                                 |                |                                  |                |                                  |  |
|--|----------|---------------------------------|---------------------------------|---------------------------------|----------------|----------------------------------|----------------|----------------------------------|--|
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD                             | AED                             | APC                             | SHAP           | PD                               | PFI            | VAR                              |  |
|  $x$            | <b>0</b> | <b><math>0.0 \pm 0.0</math></b> | <b><math>0.0 \pm 0.0</math></b> | <b><math>0.0 \pm 0.0</math></b> | $0.5 \pm 0.1$  | $0.1 \pm 0.0$                    | $0.4 \pm 0.1$  | <b><math>0.0 \pm 0.0</math></b>  |  |
|  $x^3$          | 0        | <b><math>0.0 \pm 0.0</math></b> | $0.0 \pm 0.0$                   | $6.6 \pm 0.4$                   | $1.6 \pm 0.2$  | $1.4 \pm 0.1$                    | $2.0 \pm 0.2$  | $1.0 \pm 0.1$                    |  |
|  $x + \cos(3x)$ | 0        | $0.0 \pm 0.0$                   | $3.8 \pm 0.2$                   | $6.5 \pm 0.3$                   | $0.6 \pm 0.1$  | $0.1 \pm 0.1$                    | $0.5 \pm 0.1$  | <b><math>-0.1 \pm 0.0</math></b> |  |
|  $\sin(3x)$     | <b>0</b> | <b><math>0.0 \pm 0.0</math></b> | $3.6 \pm 0.2$                   | $8.1 \pm 0.3$                   | $0.5 \pm 0.1$  | <b><math>0.1 \pm 0.1</math></b>  | $0.3 \pm 0.1$  | $2.3 \pm 0.2$                    |  |
|  $x \exp(-x)$   | <b>0</b> | <b><math>0.1 \pm 0.1</math></b> | $0.2 \pm 0.1$                   | $8.2 \pm 0.4$                   | $1.2 \pm 0.2$  | $0.5 \pm 0.2$                    | $1.8 \pm 0.2$  | $2.5 \pm 0.2$                    |  |
|  $\exp(-x^2)$   | <b>0</b> | <b><math>0.0 \pm 0.0</math></b> | $20.4 \pm 0.5$                  | $9.8 \pm 0.4$                   | $0.7 \pm 0.1$  | $0.3 \pm 0.1$                    | $0.6 \pm 0.1$  | <b><math>0.0 \pm 0.0</math></b>  |  |
| Imperfect models   |          |                                 |                                 |                                 |                |                                  |                |                                  |  |
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD                             | AED                             | APC                             | SHAP           | PD                               | PFI            | VAR                              |  |
|  $x$            | 0        | $0.1 \pm 0.2$                   | $0.2 \pm 0.2$                   | $-0.6 \pm 0.1$                  | $-0.0 \pm 0.2$ | <b><math>-0.8 \pm 0.1</math></b> | $-0.2 \pm 0.1$ | $-0.7 \pm 0.1$                   |  |
|  $x^3$          | <b>0</b> | <b><math>0.1 \pm 0.2</math></b> | <b><math>0.1 \pm 0.2</math></b> | $6.0 \pm 0.4$                   | $1.4 \pm 0.2$  | $0.4 \pm 0.2$                    | $2.5 \pm 0.2$  | $1.3 \pm 0.2$                    |  |
|  $x + \cos(3x)$ | 0        | $0.0 \pm 0.1$                   | $3.9 \pm 0.2$                   | $6.6 \pm 0.3$                   | $0.7 \pm 0.1$  | $-0.1 \pm 0.1$                   | $0.7 \pm 0.1$  | <b><math>-0.3 \pm 0.1</math></b> |  |
|  $\sin(3x)$     | <b>0</b> | <b><math>0.0 \pm 0.0</math></b> | $3.7 \pm 0.2$                   | $7.7 \pm 0.3$                   | $0.5 \pm 0.1$  | <b><math>0.1 \pm 0.1</math></b>  | $0.4 \pm 0.1$  | $1.9 \pm 0.2$                    |  |
|  $x \exp(-x)$   | <b>0</b> | <b><math>0.1 \pm 0.2</math></b> | <b><math>0.2 \pm 0.2</math></b> | $7.8 \pm 0.4$                   | $1.1 \pm 0.3$  | <b><math>0.0 \pm 0.2</math></b>  | $2.6 \pm 0.3$  | $3.9 \pm 0.3$                    |  |
|  $\exp(-x^2)$   | <b>0</b> | <b><math>0.0 \pm 0.0</math></b> | $20.4 \pm 0.5$                  | $9.3 \pm 0.4$                   | $0.7 \pm 0.1$  | $0.3 \pm 0.1$                    | $0.6 \pm 0.1$  | <b><math>0.0 \pm 0.0</math></b>  |  |

Table 5: Average error in rankings in the simulated example of the main paper. The predictors have a multivariate Normal distribution with all correlations 0.8.

| <b>Ground-truth models</b>   |          |                  |                  |                  |           |                   |            |                   |  |
|--|----------|------------------|------------------|------------------|-----------|-------------------|------------|-------------------|--|
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD              | AED              | APC              | SHAP      | PD                | PFI        | VAR               |  |
|  $x$              | <b>0</b> | <b>0.0 ± 0.0</b> | <b>0.0 ± 0.0</b> | <b>0.0 ± 0.0</b> | 0.5 ± 0.1 | 0.1 ± 0.0         | 3.5 ± 0.2  | 0.4 ± 0.1         |  |
|  $x^3$            | <b>0</b> | <b>0.0 ± 0.0</b> | <b>0.0 ± 0.0</b> | 4.6 ± 0.4        | 2.6 ± 0.2 | 2.8 ± 0.2         | 5.2 ± 0.3  | 1.1 ± 0.1         |  |
|  $x + \cos(3x)$  | <b>0</b> | <b>0.0 ± 0.0</b> | 3.4 ± 0.2        | 6.1 ± 0.3        | 0.5 ± 0.1 | <b>-0.1 ± 0.1</b> | 2.5 ± 0.2  | <b>-0.1 ± 0.1</b> |  |
|  $\sin(3x)$     | <b>0</b> | <b>0.0 ± 0.0</b> | 20.8 ± 0.5       | 7.4 ± 0.3        | 0.5 ± 0.1 | 0.1 ± 0.0         | 1.0 ± 0.1  | <b>0.0 ± 0.0</b>  |  |
|  $x \exp(-x)$   | <b>0</b> | 0.2 ± 0.1        | 0.3 ± 0.1        | 4.4 ± 0.3        | 3.1 ± 0.2 | 1.6 ± 0.2         | 6.7 ± 0.3  | 1.4 ± 0.2         |  |
|  $\exp(-x^2)$   | <b>0</b> | <b>0.0 ± 0.0</b> | 18.9 ± 0.6       | 4.6 ± 0.3        | 0.4 ± 0.1 | <b>0.1 ± 0.1</b>  | 2.1 ± 0.2  | <b>0.0 ± 0.0</b>  |  |
| <b>Imperfect models</b>  |          |                  |                  |                  |           |                   |            |                   |  |
| Function $f_{\text{true},i}(x)$  | R-sens   | EAD              | AED              | APC              | SHAP      | PD                | PFI        | VAR               |  |
|  $x$            | 0        | 0.2 ± 0.1        | 0.2 ± 0.1        | 1.3 ± 0.2        | 0.7 ± 0.2 | <b>-0.4 ± 0.1</b> | 4.3 ± 0.3  | -0.0 ± 0.1        |  |
|  $x^3$          | <b>0</b> | <b>0.1 ± 0.3</b> | <b>0.1 ± 0.3</b> | 6.4 ± 0.4        | 3.8 ± 0.4 | 1.4 ± 0.3         | 9.5 ± 0.8  | 0.5 ± 0.3         |  |
|  $x + \cos(3x)$ | 0        | 0.0 ± 0.1        | 3.4 ± 0.2        | 6.6 ± 0.3        | 1.0 ± 0.1 | -0.2 ± 0.1        | 3.2 ± 0.2  | <b>-0.3 ± 0.1</b> |  |
|  $\sin(3x)$     | 0        | 0.0 ± 0.0        | 20.7 ± 0.5       | 7.5 ± 0.3        | 0.5 ± 0.1 | 0.0 ± 0.1         | 1.0 ± 0.1  | <b>-0.1 ± 0.0</b> |  |
|  $x \exp(-x)$   | <b>0</b> | 0.4 ± 0.3        | 0.5 ± 0.3        | 5.7 ± 0.4        | 4.8 ± 0.4 | 3.2 ± 0.5         | 11.1 ± 0.8 | 1.9 ± 0.3         |  |
|  $\exp(-x^2)$   | <b>0</b> | <b>0.0 ± 0.0</b> | 18.9 ± 0.6       | 4.6 ± 0.3        | 0.5 ± 0.1 | <b>0.1 ± 0.1</b>  | 2.1 ± 0.2  | <b>0.0 ± 0.0</b>  |  |

## 6 R-SENS FOR GAUSSIAN PROCESS MODELS

Gaussian process models are widely used in supervised learning, where the task is to predict an output  $y$  from a  $D$ -dimensional input  $\mathbf{x}$ . The type of functions the Gaussian process can represent are determined by its covariance function, which is a key decision made during modelling. The covariance function  $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$  defines the covariance between the function values at the input points  $\mathbf{x}^{(i)}$  and  $\mathbf{x}^{(j)}$ . We assume the Gaussian process has a zero mean, in which case the joint distribution of the latent output values  $f$  at the training points is

$$p(f(\mathbf{X})) = p(\mathbf{f}) = \text{Normal}(\mathbf{f} \mid 0, \mathbf{K}),$$

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  $\mathbf{K}_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ .

In this work, we use the exponentiated quadratic covariance function

$$k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{k=1}^D \frac{(x_k^{(i)} - x_k^{(j)})^2}{l_k^2}\right). \quad (6)$$

Here, the hyperparameter  $\sigma_f$  determines the overall variability of the functions, and  $(l_1, \dots, l_D)$  are the length-scales of each input dimension. By defining an observation model that links the observations to the latent values of the Gaussian process, the model can be used for inference and predictions in many supervised learning tasks.

For example, in regression with an assumption of Gaussian noise, the posterior distribution of latent values for a new input point  $\mathbf{x}^*$  is a univariate normal distribution with mean and variance

$$\begin{aligned} \mathbb{E}[f^* | \mathbf{x}^*, \mathbf{y}] &= k(\mathbf{x}^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \\ \text{Var}[f^* | \mathbf{x}^*, \mathbf{y}] &= k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{X}, \mathbf{x}^*), \end{aligned} \quad (7)$$

where  $\sigma^2$  is the noise variance,  $\mathbf{I}$  is the identity matrix, and  $\mathbf{y}$  is the vector of training outputs. For many other observation models, the posterior of latent values is not Gaussian, but it is commonplace to approximate it with a Gaussian distribution during inference, and many methods have been developed for doing so [Williams and Barber, 1998, Opper and Winther, 2000, Minka, 2001, Rasmussen and Williams, 2006]. The variable importance assessment thus depends implicitly on the posterior approximation, as does any general method that uses the model's predictions.

### 6.1 DIFFERENTIATING GAUSSIAN PROCESSES

We assume that the posterior distribution of latent values is Gaussian. Because differentiation is a linear operation, the derivatives of the parameters of a Gaussian process posterior distribution with respect to predictor variables are available in closed form [Solak et al., 2003, Rasmussen, 2003]. For example, for the Gaussian observation model, the derivatives of the mean and variance of the predictive distribution in equation (7) with respect to the predictor variable  $x_d$  at point  $\mathbf{x}^*$  are given as

$$\begin{aligned} \frac{\partial \mathbb{E}[f^* | \mathbf{x}^*, \mathbf{y}]}{\partial x_d^*} &= \frac{\partial k(\mathbf{x}^*, \mathbf{X})}{\partial x_d^*} (k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \\ \frac{\partial \text{Var}[f^* | \mathbf{x}^*, \mathbf{y}]}{\partial x_d^*} &= \frac{\partial k(\mathbf{x}^*, \mathbf{x}^*)}{\partial x_d^*} - \frac{\partial k(\mathbf{x}^*, \mathbf{X})}{\partial x_d^*} (k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{X}, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X}) (k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \frac{\partial k(\mathbf{X}, \mathbf{x}^*)}{\partial x_d^*}. \end{aligned}$$

For the exponentiated quadratic covariance function in equation (6), the partial derivatives with respect to the predictor variable  $x_d$  are

$$\begin{aligned} \frac{\partial k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\partial x_d^{(i)}} &= \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{k=1}^D \frac{(x_k^{(i)} - x_k^{(j)})^2}{l_k^2}\right) \left(-\frac{x_d^{(i)} - x_d^{(j)}}{l_d^2}\right), \\ \frac{\partial k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\partial x_d^{(j)}} &= -\frac{\partial k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\partial x_d^{(i)}}. \end{aligned}$$

The second derivatives are

$$\frac{\partial^2 k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\partial x_d^{(i)} \partial x_e^{(i)}} = \frac{\partial^2 k_{\text{EQ}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\partial x_d^{(j)} \partial x_e^{(j)}} = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{k=1}^D \frac{(x_k^{(i)} - x_k^{(j)})^2}{l_k^2}\right) \left(\frac{x_d^{(i)} - x_d^{(j)}}{l_d^2}\right) \left(\frac{x_e^{(i)} - x_e^{(j)}}{l_e^2}\right).$$



For the R-sens measure, we need derivatives with respect to the parameters of the predictive distribution and not the posterior of the latent values. However, for many observation models these are obtained as a function of the derivatives of the latent values. In this section, we derive the equations for some commonly used observation models.

## 6.2 REGRESSION WITH GAUSSIAN OBSERVATION MODEL

In regression problems, it is commonly assumed that the noise has a Gaussian distribution. For a Gaussian observation model, the predictive distribution for a new observation  $y^*$  at a single predictor value  $\mathbf{x}^*$  is a normal distribution, which we will denote

$$p(y^*|\mathbf{x}^*, \mathbf{y}) = \text{Normal}(y^*|E[y^*], \text{Var}[y^*]) = \text{Normal}(y^*|E[f^*], \text{Var}[f^*] + \sigma^2),$$

where  $E[f^*]$  and  $\text{Var}[f^*]$  are the mean and variance of the posterior distribution of latent values at  $\mathbf{x}^*$ , and  $\sigma^2$  is the noise variance. Now, the derivatives of  $E[y^*]$  and  $\text{Var}[y^*]$  with respect to predictor variable  $x_d^*$  are simply

$$\begin{aligned} \frac{\partial E[y^*]}{\partial x_d^*} &= \frac{\partial E[f^*]}{\partial x_d^*} \\ \frac{\partial \text{Var}[y^*]}{\partial x_d^*} &= \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \\ \frac{\partial^2 E[y^*]}{\partial x_d^* \partial x_e^*} &= \frac{\partial^2 E[f^*]}{\partial x_d^* \partial x_e^*} \\ \frac{\partial^2 \text{Var}[y^*]}{\partial x_d^* \partial x_e^*} &= \frac{\partial^2 \text{Var}[f^*]}{\partial x_d^* \partial x_e^*}. \end{aligned}$$

The Fisher information elements of the normal distribution are

$$\begin{aligned} \mathcal{I}_N(E[y^*]) &= \frac{1}{\text{Var}[y^*]} \\ \mathcal{I}_N(\text{Var}[y^*]) &= \frac{1}{2(\text{Var}[y^*])^2}. \end{aligned}$$

Thus, the R-sens measure takes the form

$$\text{R-sens}(\mathbf{x}^*, x_d, \alpha = 1) = \sqrt{\frac{1}{\text{Var}[y^*]} \left( \frac{\partial E[f^*]}{\partial x_d^*} \right)^2 + \frac{1}{2(\text{Var}[y^*])^2} \left( \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right)^2}.$$

Here, the first term is proportional to the slope of the mean prediction scaled by the predictive uncertainty, as with the linear regression model discussed in Section 2.1 of the main paper. The R-sens<sub>2</sub> measure evaluates to

$$\text{R-sens}_2(\mathbf{x}^*, (x_d, x_e), \alpha = 1) = \sqrt{\frac{1}{\text{Var}[y^*]} \left( \frac{\partial^2 E[f^*]}{\partial x_d^* \partial x_e^*} \right)^2 + \frac{1}{2(\text{Var}[y^*])^2} \left( \frac{\partial^2 \text{Var}[f^*]}{\partial x_d^* \partial x_e^*} \right)^2}.$$

## 6.3 BINARY CLASSIFICATION

For binary classification problems, the predictive distribution is a Bernoulli distribution with only one parameter, the probability of positive classification. This is obtained by squashing the latent Gaussian process function through a link function and integrating over the posterior of the latent function values. Two commonly used link functions for Gaussian process classification are the logit and probit. The Probit link function has the benefit that the predictive distribution has an analytical formula when the posterior distribution of latent values is approximated with a Gaussian. Using a Probit link function, the predictive probability has thus an approximate analytical form

$$\pi^* = p(y = 1|\mathbf{x}^*, \mathbf{y}) = \Phi \left( \frac{E[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right),$$

where  $\Phi$  is the cumulative distribution of the standard normal distribution.

Now, the derivatives of  $\pi^*$  with respect to  $x_d^*$  are

$$\begin{aligned} \frac{\partial \pi^*}{\partial x_d^*} &= \text{Normal} \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \left[ \frac{1}{\sqrt{1 + \text{Var}[f^*]}} \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} - \frac{\mathbb{E}[f^*]}{2(1 + \text{Var}[f^*])^{3/2}} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right], \\ \frac{\partial^2 \pi^*}{\partial x_d^* \partial x_e^*} &= \text{Normal} \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \left[ \frac{1}{\sqrt{1 + \text{Var}[f^*]}} \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} - \frac{\mathbb{E}[f^*]}{2(1 + \text{Var}[f^*])^{3/2}} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right] \times \\ &\quad \left[ \frac{1}{\sqrt{1 + \text{Var}[f^*]}} \frac{\partial \mathbb{E}[f^*]}{\partial x_e^*} - \frac{\mathbb{E}[f^*]}{2(1 + \text{Var}[f^*])^{3/2}} \frac{\partial \text{Var}[f^*]}{\partial x_e^*} \right] + \\ &\quad \text{Normal} \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \left[ \frac{1}{\sqrt{1 + \text{Var}[f^*]}} \frac{\partial^2 \mathbb{E}[f^*]}{\partial x_d^* \partial x_e^*} - \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} \frac{1}{2(1 + \text{Var}[f^*])^{3/2}} \frac{\partial \text{Var}[f^*]}{\partial x_e^*} \right. \\ &\quad \left. - \frac{\partial^2 \text{Var}[f^*]}{\partial x_d^* \partial x_e^*} \frac{\mathbb{E}[f^*]}{2(1 + \text{Var}[f^*])^{3/2}} - \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_e^*} \frac{1}{2(1 + \text{Var}[f^*])^{3/2}} - \frac{3\mathbb{E}[f^*]}{4(1 + \text{Var}[f^*])^{3/2}} \frac{\partial \text{Var}[f^*]}{\partial x_e^*} \right) \right]. \end{aligned}$$

The Fisher information of the Bernoulli distribution is

$$\mathcal{I}_{\text{Bern}}(\pi^*) = \frac{1}{\pi^*(1 - \pi^*)} = \left( \Phi \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \right)^{-1} \left( 1 - \Phi \left( \frac{\mathbb{E}[f^*]}{\sqrt{1 + \text{Var}[f^*]}} \right) \right)^{-1}.$$

The R-sens and R-sens<sub>2</sub> measures take the form

$$\begin{aligned} \text{R-sens}(\mathbf{x}^*, x_d, \alpha = 1) &= \sqrt{\mathcal{I}_{\text{Bern}}(\pi^*) \left( \frac{\partial \pi^*}{\partial x_d^*} \right)^2}, \\ \text{R-sens}_2(\mathbf{x}^*, (x_d, x_e), \alpha = 1) &= \sqrt{\mathcal{I}_{\text{Bern}}(\pi^*) \left( \frac{\partial^2 \pi^*}{\partial x_d^* \partial x_e^*} \right)^2}. \end{aligned}$$

## 6.4 POISSON OBSERVATION MODEL

For modelling count data with Gaussian processes, it is common to use a combination of a count observation model with a link function that transforms the positively constrained parameters to unconstrained scale where the Gaussian Process prior is placed. Here, we derive the equations needed for the R-sens method for the case of Poisson likelihood and exponential link function.

The likelihood is

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^n \text{Poisson}(y_i | \lambda_i(f_i)) = \prod_{i=1}^n \text{Poisson}(y_i | \exp(f_i)).$$

Now, the Gaussian process prior is placed on the unconstrained latent values. If one uses a Gaussian approximation to the posterior of the latent values, then the transformed  $\lambda$ 's have a log-normal distribution. The intensity  $\lambda$  at any input point is given by integrating over the approximate posterior  $q(f^*|\mathbf{y}, \mathbf{x}^*)$

$$\lambda^* = \int \exp(f^*) q(f^*|\mathbf{y}, \mathbf{x}^*) df^*.$$

This evaluates to the mean of the log-normal distribution

$$\lambda^* = \mathbb{E}[\text{Lognormal}(\mathbb{E}[f^*], \text{Var}[f^*])] = \exp(\mathbb{E}[f^*] + \text{Var}[f^*]/2).$$

The derivatives of this with respect to the predictor variables are

$$\begin{aligned} \frac{\partial \lambda^*}{\partial x_d^*} &= \exp(\mathbb{E}[f^*] + \text{Var}[f^*]/2) \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right), \\ \frac{\partial^2 \lambda^*}{\partial x_d^* \partial x_e^*} &= \exp(\mathbb{E}[f^*] + \text{Var}[f^*]/2) \times \\ &\quad \left[ \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right) \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_e^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_e^*} \right) + \left( \frac{\partial^2 \mathbb{E}[f^*]}{\partial x_d^* \partial x_e^*} + \frac{1}{2} \frac{\partial^2 \text{Var}[f^*]}{\partial x_d^* \partial x_e^*} \right) \right]. \end{aligned}$$

The Fisher information of the Poisson distribution is

$$\mathcal{I}_{\text{Pois}}(\lambda^*) = \frac{1}{\lambda^*} = \frac{1}{\exp(\mathbb{E}[f^*] + \text{Var}[f^*]/2)}.$$

Thus, the R-sens and R-sens<sub>2</sub> measures take the form

$$\text{R-sens}(\mathbf{x}^*, x_d, \alpha = 1) = \sqrt{\exp\left(\mathbb{E}[f^*] + \frac{\text{Var}[f^*]}{2}\right) \left| \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right|},$$

$$\text{R-sens}_2(\mathbf{x}^*, (x_d, x_e), \alpha = 1)$$

$$= \sqrt{\exp\left(\mathbb{E}[f^*] + \frac{\text{Var}[f^*]}{2}\right) \left| \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_d^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_d^*} \right) \left( \frac{\partial \mathbb{E}[f^*]}{\partial x_e^*} + \frac{1}{2} \frac{\partial \text{Var}[f^*]}{\partial x_e^*} \right) + \left( \frac{\partial^2 \mathbb{E}[f^*]}{\partial x_d^* \partial x_e^*} + \frac{1}{2} \frac{\partial^2 \text{Var}[f^*]}{\partial x_d^* \partial x_e^*} \right) \right|}.$$

## 7 ILLUSTRATIVE EXAMPLE - LINEAR REGRESSION

This section shows an extended version of Figure 2 in the main paper, where the variables  $x_1$  and  $x_2$  range from  $-20$  to  $20$ .

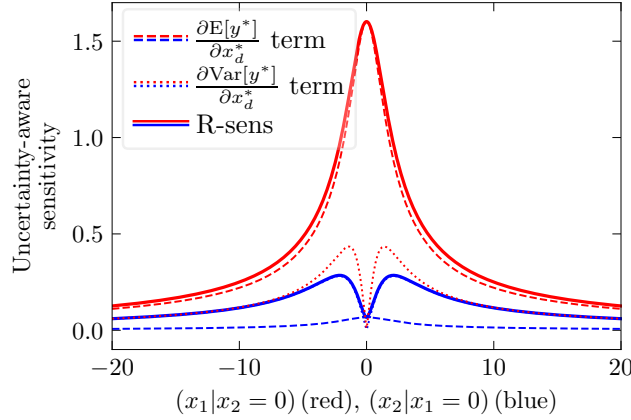


Figure 2: R-sens uncertainty-aware sensitivity measure for  $x_1$  (red) and  $x_2$  (blue) for the linear regression model from Section 2.2 in the main paper.

## 8 ILLUSTRATIVE EXAMPLE - LOGISTIC REGRESSION

In this section, we show an illustrative example similar to the main paper, but with a logistic regression model where the target variable  $y$  is binary. As the inverse link function, we use the cumulative Normal distribution. We consider a simple multivariate Gaussian prior on the regression coefficients. Contrary to the linear regression example, the posterior distribution has no closed form. We will use the Laplace approximation to get a Gaussian approximation to the posterior. The approximate posterior is

$$p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \approx \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{H}^{-1}),$$

where  $\hat{\boldsymbol{\beta}}$  is the maximum a posteriori estimate of the regression coefficients. When using the inverse cumulative Normal distribution as the link function, the predictive distribution at a new point  $\mathbf{x}^*$  has a closed form equation.

$$p(y^* = 1|\mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \pi^*(\mathbf{x}^*) = \Phi\left(\frac{\mathbf{x}^* \hat{\boldsymbol{\beta}}}{\sqrt{1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T}}}\right).$$

The derivative of the success probability is

$$\frac{\partial \pi^*(\mathbf{x}^*)}{\partial x_d^*} = \mathcal{N}\left(\frac{\mathbf{x}^* \hat{\boldsymbol{\beta}}}{\sqrt{1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T}}}\right) \left[ \frac{\hat{\beta}_d}{\sqrt{1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T}}} - \frac{\mathbf{x}^* \hat{\boldsymbol{\beta}} [(\mathbf{H})^{-1} \mathbf{x}^{*T}]_d}{(1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T})^{3/2}} \right].$$

The Fisher information of the Bernoulli distribution is

$$\mathcal{I}_{\text{Ber}}(\pi^*) = \frac{1}{\pi^*(1 - \pi^*)}.$$

The R-sens sensitivity measure for variable  $x_d^*$  thus evaluates to

$$\sqrt{\frac{1}{\pi^*(1 - \pi^*)}} \mathcal{N}\left(\frac{\mathbf{x}^* \hat{\boldsymbol{\beta}}}{\sqrt{1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T}}}\right) \left| \frac{\hat{\beta}_d}{\sqrt{1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T}}} - \frac{\mathbf{x}^* \hat{\boldsymbol{\beta}} [(\mathbf{H})^{-1} \mathbf{x}^{*T}]_d}{(1 + \mathbf{x}^* (\mathbf{H})^{-1} \mathbf{x}^{*T})^{3/2}} \right|. \quad (8)$$

To illustrate the R-sens measure in equation (8), we simulated 1000 observations from a logistic regression 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. The R-sens sensitivities for both variables given by equation (8) are shown in Figure 3. The dashed line shows the derivative of the prediction function without the Fisher information term. Because of the link function, this derivative is not constant and is much larger close to the decision boundary. The Fisher information term does not remove this effect, but gives a bit more weight to points further away.

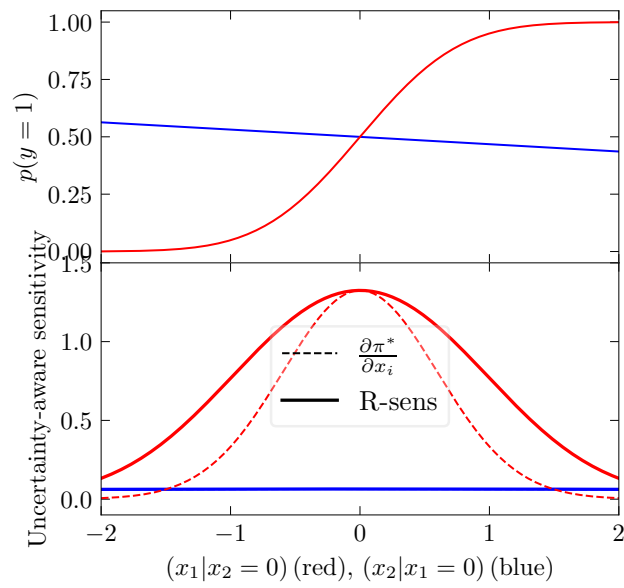


Figure 3: Top: Predictive distributions  $p(y = 1|x_1, x_2 = 0)$  (red) and  $p(y|x_2, x_1 = 0)$  (blue) for the logistic regression model. Bottom: R-sens uncertainty-aware sensitivity measure given by equation (8) for  $x_1$  (red) and  $x_2$  (blue).

## 9 ASYMPTOTIC RESULTS FOR GENERALISED LINEAR MODELS

In Section 2.1 in the main paper, we discussed the behaviour of the posterior predictive distribution for a Bayesian linear regression model as the number of observations goes to infinity. In this case, the predictive uncertainty tends to a constant, and the R-sens local sensitivity measure for each predictor is proportional the absolute value of the maximum likelihood estimate of the regression coefficient,  $|\widehat{\beta}_d|$ . In this section we discuss the asymptotic results of the logistic and Poisson regression models, which are nonlinear models that can be used to model binary or integer data.

### 9.1 LOGISTIC REGRESSION MODEL

#### 9.1.1 Logit Link Function

The predictive distribution of a logistic regression model is a Bernoulli distribution. In the asymptotic limit, the posterior of the regression coefficients concentrates to a point  $\widehat{\beta}$ , and the “success probability” parameter as a function of the predictor variables is the logistic function

$$p(y^* = 1 | \mathbf{X}, \mathbf{y}, \mathbf{x}^*) := \pi^*(\mathbf{x}^*) = \frac{\exp(\mathbf{x}^* \widehat{\beta})}{1 + \exp(\mathbf{x}^* \widehat{\beta})}.$$

The derivative of  $\pi^*$  with respect to  $x_d^*$  is

$$\frac{\partial \pi^*}{\partial x_d^*} = \widehat{\beta}_d \pi^* (1 - \pi^*).$$

The Fisher information of the Bernoulli distribution

$$\mathcal{I}_{\text{bern}}(\pi^*) = \frac{1}{\pi^*(1 - \pi^*)}.$$

In the limit when the number of observations goes to infinity, the R-sens measure thus evaluates to

$$\sqrt{\mathcal{I}_{\text{bern}}(\pi^*) \left( \frac{\partial \pi^*}{\partial x_d^*} \right)^2} = |\widehat{\beta}_d| \sqrt{\pi^*(1 - \pi^*)}.$$

The R-sens importance measure for the logistic regression model is proportional to the absolute value of the regression coefficient. In addition, due to the logistic (inverse) link function, the local importance measure is higher for points close to the decision boundary  $p(y^* = 1) = 0.5$  compared to points further away. Because the term  $\sqrt{\pi^*(1 - \pi^*)}$  is the same for each predictor variable, ranking the variables with R-sens is equivalent to ranking with the absolute regression coefficients  $|\widehat{\beta}_d|$  in the limit of infinite data.

Contrary to the linear regression example in the main paper, in logistic regression, the R-sens measure gives more importance to observations further from the decision boundary. It can be interpreted in the sense that the derivative of the logistic prediction function,  $\frac{\partial \pi^*}{\partial x_d^*} = \widehat{\beta}_d \pi^* (1 - \pi^*)$  gives too much emphasis to points near the decision boundary, and the R-sens measure makes the sensitivity measure more even.

#### 9.1.2 Inverse Normal Link Function

Now, the “success probability” parameter as a function of the predictor variables is the cumulative Normal distribution function

$$p(y^* = 1 | \mathbf{X}, \mathbf{y}, \mathbf{x}^*) := \pi^*(\mathbf{x}^*) = \Phi(\mathbf{x}^* \widehat{\beta}).$$

The derivative of  $\pi^*$  with respect to  $x_d^*$  is

$$\frac{\partial \pi^*}{\partial x_d^*} = \widehat{\beta}_d \mathcal{N}(\mathbf{x}^* \widehat{\beta}).$$

The Fisher information of the Bernoulli distribution

$$\mathcal{I}_{\text{bern}}(\pi^*) = \frac{1}{\pi^*(1 - \pi^*)}.$$

In the limit when the number of observations goes to infinity, the R-sens measure thus evaluates to

$$\sqrt{\mathcal{I}_{\text{bern}}(\pi^*) \left( \frac{\partial \pi^*}{\partial x_d^*} \right)^2} = |\hat{\beta}_d| \frac{\mathcal{N}(\mathbf{x}^* \hat{\beta})}{\sqrt{\pi^* (1 - \pi^*)}}.$$

## 9.2 POISSON REGRESSION MODEL

In a Poisson regression model, the predictive distribution is a Poisson distribution. Here, we consider the commonly used logarithmic link function, where the mean of the predictive distribution is

$$\mathbb{E}[y^*] = \exp(\mathbf{x}^* \beta).$$

In the asymptotic limit of infinite data, the posterior of the regression coefficients concentrates to a point  $\hat{\beta}$ , and the mean of the predictive distribution is given by the exponential function

$$\mathbb{E}[y^*] = \exp(\mathbf{x}^* \hat{\beta}).$$

The derivative of  $\mathbb{E}[y^*]$  with respect to  $x_d^*$  is

$$\frac{\partial \mathbb{E}[y^*]}{\partial x_d^*} = \hat{\beta}_d \exp(\mathbf{x}^* \hat{\beta}).$$

The Fisher information of the Poisson distribution is

$$\frac{1}{\mathbb{E}[y^*]}.$$

In the limit when the number of observations goes to infinity, the R-sens measure thus evaluates to

$$|\hat{\beta}_d| \exp\left(\frac{1}{2} \mathbf{x}^* \hat{\beta}\right).$$

## References

- Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.
- Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11): 2655–2684, 2000.
- 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.
- CE Rasmussen. Gaussian processes to speed up hybrid monte carlo for expensive Bayesian integrals. *Bayesian statistics*, 7: 651–659, 2003.
- Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl E Rasmussen. Derivative observations in Gaussian process models of dynamic systems. In *Advances in neural information processing systems*, pages 1057–1064, 2003.
- Christopher KI Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.