Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-022-00841-4

Interpreting deep learning models with marginal attribution by conditioning on quantiles

Michael Merz[1] · Ronald Richman[2,3] · Andreas Tsanakas[4] · Mario V. Wüthrich[5]

Received: 22 March 2021 / Accepted: 18 April 2022

Abstract
A vast and growing literature on explaining deep learning models has emerged. This paper contributes to that literature by introducing a global gradient-based model-agnostic method, which we call Marginal Attribution by Conditioning on Quantiles (MACQ). Our approach is based on analyzing the marginal attribution of predictions (outputs) to individual features (inputs). Specifically, we consider variable importance by fixing (global) output levels, and explaining how features marginally contribute to these fixed global output levels. MACQ can be seen as a marginal attribution counterpart to approaches such as accumulated local effects, which study the sensitivities of outputs by perturbing inputs. Furthermore, MACQ allows us to separate marginal attribution of individual features from interaction effects and to visualize the 3-way relationship between marginal attribution, output level, and feature value.

Keywords Explainable AI (XAI) · Model-agnostic tools · Deep learning · Attribution · Accumulated local effects (ALE) · Partial dependence plot (PDP) ·

Responsible editor: Martin Atzmueller, Johannes Fürnkranz, Tomáš Kliegr and Ute Schmid.

✉ Mario V. Wüthrich
mario.wuethrich@math.ethz.ch

Michael Merz
michael.merz@uni-hamburg.de

Ronald Richman
ronaldrichman@gmail.com

Andreas Tsanakas
A.Tsanakas.1@city.ac.uk

1 Faculty of Business Administration, University of Hamburg, Hamburg, Germany

2 Old Mutual Insure, Johannesburg, South Africa

3 University of the Witwatersrand, Johannesburg, South Africa

4 Bayes Business School, City, University of London, London, UK

5 RiskLab, Department of Mathematics, ETH Zurich, Zürich, Switzerland

Published online: 18 June 2022

Locally interpretable model-agnostic explanation (LIME) · Variable importance · Post-hoc analysis · Interaction

1 Introduction

Deep learning models are typically trained to provide optimal predictive performance. Interpreting and explaining the results of deep learning models has, until recently, only played a subordinate role. With growing complexity of deep learning models, the need and requirement of being able to explain deep learning solutions has become increasingly important. This applies to many fields of application: deep learning findings in medical fields and health care need to make sense to patients, loan and mortgage evaluations and credit approvals need to be understandable to customers, insurance pricing must be explained to insurance policyholders, business processes and decisions need to be transparent to regulators, etc. These needs are even reinforced by the requirements of being able to prove that deep learning solutions do not discriminate w.r.t. protected features and are in line with data protection regulation; see, e.g., Lindholm et al. (2022) and the references therein. Thus, there is substantial social and political pressure to be able to explain, illustrate and verify deep learning solutions, in order to provide reassurance that these work as intended.

Recent research focuses on different methods for explaining deep learning decision making; an overview is given in Samek and Müller (2019). Some of these methods provide a post-hoc analysis, which aims at understanding global model behavior, by explaining individual outcomes and learned representations. Often this is done by analyzing representative examples. We will discuss some of these post-hoc analysis methods in the literature overview presented in the next section. Other methods aim at a wider interdisciplinary approach by more broadly examining how decision making is done in a social context, see, e.g., Miller (2019). All these approaches have in common that they try to "open up the black-box", to make model-driven decisions explainable to stakeholders.

Our paper contributes to this literature. We provide a novel gradient-based model-agnostic tool by analyzing marginal contributions to deep learning decisions in the spirit of salience methods, as described in Ancona et al. (2019). Salience methods are local model-agnostic tools that attribute marginal effects on outputs to different inputs, i.e., marginal attribution is understood in the sense that effects on outputs are allocated to individual components of the input features. The attributions we consider are motivated by sensitivity analysis tools in risk measurement, which aggregate local marginal attributions to a global picture at a given quantile level of the output variable, see Hong (2009) and Proposition 1 in Tsanakas and Millossovich (2016). We call this method Marginal Attribution by Conditioning on Quantiles (MACQ). Thus, our first contribution is that we provide a global model-agnostic tool that attributes output levels to input variables. In particular, this allows us to describe how the importance of inputs varies across different output levels. As a second contribution, we extend this view by including higher order derivatives beyond linear marginal contributions. This additional step allows us to analyze interactions, and it can be seen in the context of deep Taylor decompositions (DTD), similar to Montavon et al. (2017). A difficulty

in Taylor decompositions is that they depend on a reference point. By rearranging terms and taking advantage of our quantile view, we determine an optimal global reference point that allows us to quantify both variable importance and interaction strength in our MACQ approach. The third contribution is that we introduce graphic tools that illustrate the 3-way relationship between (i) marginal attributions, (ii) the response/output level and (iii) the feature value. Summarizing, our method allows us to simultaneously study variable importance and interaction strength at different output levels, i.e., it allows us to explain which inputs and interactions contribute to a high or a low output. This viewpoint is different from most other explainability tools, which are mostly based on perturbations of inputs. Moreover, our method can be applied at low computational cost, an important practical advantage.

Organization. The next section gives a literature overview that embeds our MACQ method into the present toolbox of model explainability. This literature overview is also used to introduce the relevant notation. Section 3 introduces our new proposal. This section is divided into three parts: in Sect. 3.1 we present our main idea of aggregating local marginal attributions to a quantile sensitivity analysis; Sect. 3.2 gives a higher order expansion that grounds the study of interaction strengths; and Sect. 3.3 discusses the choice of the reference point needed to calibrate our explainability tool. A synthetic data example is presented in Sect. 4, Sect. 5 gives a real data example, and in Sect. 6 we conclude. Appendix A revisits distortion risk measures, Appendix B gives additional analysis on the real data example, and in Appendix C we describe the data used.

2 Literature overview

We give a brief summary of recent developments in post-hoc interpretability and explainability tools for deep learning models. This summary also serves to introduce the relevant notation for this paper. Consider the following twice differentiable regression function

$$\mu : \mathbb{R}^q \to \mathbb{R}, \qquad \boldsymbol{x} \mapsto \mu(\boldsymbol{x}), \tag{2.1}$$

with features $\boldsymbol{x} = (x_1, \ldots, x_q)^\top \in \mathbb{R}^q$. The regression function μ describes the systematic effects of features $\boldsymbol{x}$ on the random variable Y via the (conditional) expectation

$$\mathbb{E}[Y|\boldsymbol{x}] = \mu(\boldsymbol{x}).$$

We assume twice differentiability in $\boldsymbol{x} \in \mathbb{R}^q$ of regression function (2.1) because our model-agnostic proposal will be gradient-based; discrete inputs, e.g., binary input components, will be embedded into $\mathbb{R}$. In our examples in Sects. 4 and 5, we will use a deep feed-forward neural network on tabular input data with the hyperbolic tangent activation function. This gives us a smooth regression function (and an interpolation for discrete input components), whose derivatives can be obtained in standard software such as TensorFlow/Keras and PyTorch.

2.1 Model-agnostic tools

Recent literature aims at understanding regression functions (2.1) coming from deep learning models. One approach is to analyze marginal plots. We select one component x_j of $\boldsymbol{x}$, and, by a slight abuse of notation, write $\boldsymbol{x} = (x_j, \boldsymbol{x}_{\setminus j})$, where $\boldsymbol{x}_{\setminus j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_q)^\top$ collects all components of $\boldsymbol{x}$ except x_j. We then study the regression function as a function of x_j by keeping the remaining components $\boldsymbol{x}_{\setminus j}$ of $\boldsymbol{x}$ fixed, that is,

$$x_j \in \mathbb{R} \;\mapsto\; \mu(x_j, \boldsymbol{x}_{\setminus j}).$$

This gives the method of individual conditional expectation (ICE) of Goldstein et al. (2015). If we have thousands or millions of instances $(Y, \boldsymbol{x})$, it might be advantageous to study ICE profiles on an aggregated level. This is the proposal of Friedman (2001) and Zhao and Hastie (2021), called partial dependence plots (PDPs). We introduce the feature distribution P which describes the statistical nature of all (potential) features $\boldsymbol{X} \sim P$. The PDP profile of component $1 \le j \le q$ is defined by

$$x_j \mapsto \mathbb{E}_P\left[\mu(x_j, \boldsymbol{X}_{\setminus j})\right] = \int \mu(x_j, \boldsymbol{x}_{\setminus j}) dP(\boldsymbol{x}_{\setminus j}). \tag{2.2}$$

The critical point in this approach is that it does not reflect the dependence structure between feature components X_j and $\boldsymbol{X}_{\setminus j}$, as described by feature distribution P, because we only integrate over the marginal distribution $P(\boldsymbol{x}_{\setminus j})$ of $\boldsymbol{X}_{\setminus j}$ in (2.2). The method of accumulated local effects (ALEs) introduced by Apley and Zhu (2020) aims at correctly incorporating the dependence structure of $\boldsymbol{X}$. The local effect of component x_j in instance $\boldsymbol{x}$ is given by the partial derivative

$$\mu_j(\boldsymbol{x}) = \frac{\partial \mu(\boldsymbol{x})}{\partial x_j}. \tag{2.3}$$

The average local effect of component $1 \le j \le q$ is obtained by

$$x_j \mapsto \Delta_j(x_j) = \mathbb{E}_P\left[\mu_j(\boldsymbol{X}) \middle| X_j = x_j\right] = \int \mu_j(x_j, \boldsymbol{x}_{\setminus j}) dP(\boldsymbol{x}_{\setminus j}|x_j), \tag{2.4}$$

where $P(\boldsymbol{x}_{\setminus j}|x_j)$ denotes the conditional distribution of $\boldsymbol{X}_{\setminus j}$, given $X_j = x_j$. ALEs integrate the average local effects $\Delta_j(\cdot)$ over their domain, thus, the ALE profile is defined by

$$x_j \mapsto \int_{x_{j0}}^{x_j} \Delta_j(z_j) dz_j = \int_{x_{j0}}^{x_j} \int \mu_j(z_j, \boldsymbol{x}_{\setminus j}) dP(\boldsymbol{x}_{\setminus j}|z_j) dz_j, \tag{2.5}$$

where x_{j0} is a given initialization point.

Remark 2.1 • The main difference between PDPs and ALEs is that the latter correctly consider the dependence structure between X_j and $\boldsymbol{X}_{\setminus j}$. The two profiles coincide if X_j and $\boldsymbol{X}_{\setminus j}$ are independent under P.

- Apley and Zhu (2020) provide a discretized version of the ALE profile, which can also be applied to non-differentiable regression functions $\mu(\cdot)$, such as those coming from regression trees and tree boosting methods. Basically, this can be received either by finite differences or by a local analysis in an environment of a selected feature value x_j.
- Local effect (2.3) allows us to consider a 1st order Taylor expansion. Denote by $\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x})$ the gradient of $\mu(\cdot)$ w.r.t. $\boldsymbol{x}$. We have

$$\mu(\boldsymbol{x}+\boldsymbol{\epsilon}) = \mu(\boldsymbol{x}) + (\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x}))^{\top}\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|), \tag{2.6}$$

  for $\boldsymbol{\epsilon} \in \mathbb{R}^q$ going to zero. This gives us a 1st order local approximation to $\mu(\cdot)$ in $\boldsymbol{x}$, which reflects the local (linear) behavior, similarly to the locally interpretable model-agnostic explanation (LIME) introduced by Ribeiro et al. (2016). That is, (2.6) fits a local linear regression model around $\mu(\boldsymbol{x})$ with regression parameters described by the components of the gradient $\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x})$. LIME then uses regularization, e.g., LASSO, to select the most relevant feature components in the neighborhood of $\mu(\boldsymbol{x})$.
- More generally, (2.6) defines a local surrogate model that can be used for a local sensitivity analysis by perturbing $\boldsymbol{x}$ within a small environment. "White-box" surrogate models are popular tools to explain complex regression functions; for instance, decision trees can be fit to network regression models for extracting the most relevant feature information.
- We can summarize the methods presented in this section as sensitivity tools that analyze the effects on outputs of changing inputs. Moreover, the presented tools do not study interaction effects of input components.

2.2 Gradient based model-agnostic tools

Gradient-based model-agnostic tools can be used to attribute outputs to (feature) inputs. Attribution denotes the process of assigning a relevance index to input components, in order to explain a certain output, see Efron (2020). Ancona et al. (2019) provide an overview of gradient-based attribution methods. In formula (2.3) of the previous subsection we have met a first attribution method, giving the sensitivity of the output $\mu(\boldsymbol{x})$ as a function of the input $\boldsymbol{x}$.

Marginal attribution is obtained by considering the directional derivative w.r.t. the features

$$x_j \;\mapsto\; x_j\mu_j(\boldsymbol{x}) \;=\; x_j\frac{\partial\mu(\boldsymbol{x})}{\partial x_j}. \tag{2.7}$$

This has first been discussed in the machine learning community by Shrikumar et al. (2016) who observed that this can make attribution more concise; these directional derivatives have been coined Gradient*Input in the machine learning literature, see

Ancona et al. (2019). Mathematically speaking, these marginal attributions can be understood as individual contributions to a certain value in a Taylor series sense (and relative to a reference point). For a linear regression model $\boldsymbol{x} \mapsto \beta_0 + \sum_{j=1}^{q} \beta_j x_j$, the marginal attributions give an additive decomposition of the regression function, and β_j can be considered as the relevance index of component j. In non-linear regression models, such a linear decomposition only holds true locally, see (2.6), and other methods such as the Shapley value (Shapley 1953) are used to quantify non-linear effects and interaction effects, see Lundberg and Lee (2017). We also mention (Sundararajan et al. 2017), who consider integrated gradients

$$x_j \;\mapsto\; x_j \int_0^1 \mu_j \left(\boldsymbol{x}_0 + z(\boldsymbol{x} - \boldsymbol{x}_0)\right) dz, \tag{2.8}$$

for a given reference point $\boldsymbol{x}_0$. This mitigates the problem of only being accurate locally. In practice, however, evaluation of (2.8) is computationally demanding, similarly to Shapley values.

There are other methods that are specific to deep networks. We mention layer-wise propagation (LRP) by Binder et al. (2016) and DeepLIFT (Deep Learning Important FeaTures) by Shrikumar et al. (2017). These methods use a backward pass from the output to the input. In this backward pass a relevance index (budget) is locally redistributed (recursively from layer to layer), resulting in a relevance index on the inputs (for the given output). Ancona et al. (2019) show in Propositions 1 and 2 that these two methods can be understood as averages over marginal attributions. We remark that these methods are mainly used for convolutional neural networks (CNNs), e.g., in image recognition, whereas our MACQ proposal is more suitable for tabular data, as we require differentiability w.r.t. the inputs $\boldsymbol{x}$. CNN architectures are often non-differentiable because of the use of max-pooling layers.

Our contribution builds on marginal attributions (2.7). Marginal attributions are, by definition, local explanations, and we will show how to integrate these local considerations into a global variable importance analysis. Samek and Müller (2019) call such an aggregation of individual explanations a *global meta-explanation*. As a consequence, our MACQ approach is the marginal attribution counterpart to ALEs by fixing (global) output levels and describing how features marginally contribute to these levels, whereas ALEs rather study the sensitivities of the outputs by perturbing the inputs. Thus, similar to LRP and DeepLIFT, we use a type of backward pass from the output to the input in our tool to explain the output.

3 Marginal attribution by conditioning on quantiles

3.1 First order attributions

We consider regression model (2.1) from a marginal attribution point of view, motivated by the risk sensitivity tools of Hong (2009) and Tsanakas and Millossovich (2016). Rather than considering average local effects (2.4), conditioned on event $\{X_j = x_j\}$, we try to understand how feature components contribute to a certain

response level $\mu(\boldsymbol{x})$. This allows us to study how the response levels are composed in different regions of the decision space, as this is of intrinsic interest, e.g., in financial applications.

Select a quantile level $\alpha \in (0, 1)$. The α-quantile of $\mu(\boldsymbol{X})$ is given by the left-continuous generalized inverse

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) = \inf\left\{y \in \mathbb{R};\ F_{\mu(\boldsymbol{X})}(y) \geq \alpha\right\},$$

where $F_{\mu(\boldsymbol{X})}(y) = P[\mu(\boldsymbol{X}) \leq y]$ describes the distribution function of $\mu(\boldsymbol{X})$.

The *1st order attributions* to components $1 \leq j \leq q$ at quantile level α are defined by

$$S_j(\mu; \alpha) \ = \ \mathbb{E}_P\left[X_j \mu_j(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right]. \tag{3.1}$$

These are the *Marginal Attributions by Conditioning on Quantiles* (MACQ).

Tsanakas and Millossovich (2016) show that (3.1) naturally arises via sensitivities of distortion risk measures. Choosing the α-Dirac distortion, which allocates a probability weight of size 1 to a given $\alpha \in (0, 1)$, we exactly receive (3.1), which corresponds to the sensitivities of the Value-at-Risk (VaR) risk measure at the given quantile level α. Thus, the sensitivities of the VaR risk measure can be described by the average of the marginal attributions $X_j \mu_j(\boldsymbol{X})$, conditioned on being the output at the corresponding quantile level. The interested reader is referred to Appendix A for a more detailed description of distortion risk measures.

Alternatively, we can describe 1st order attributions (3.1) by a 1st order Taylor expansion (2.6) in feature perturbation $\boldsymbol{\epsilon} = -\boldsymbol{x}$

$$\mu(\mathbf{0}) \ \approx \ \mu\left(\boldsymbol{x}\right) - \left(\nabla_{\boldsymbol{x}} \mu(\boldsymbol{x})\right)^{\top} \boldsymbol{x}. \tag{3.2}$$

This shows that the 1st order attributions (3.1) describe a 1st order Taylor approximation at the common *reference point* $\mathbf{0}$, and rearranging the terms we get the *1st order contributions* to a given response level

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) = \mathbb{E}_P\left[\mu\left(\boldsymbol{X}\right) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right] \ \approx \ \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q} S_j(\mu; \alpha). \tag{3.3}$$

Remark 3.1 • A 1st order Taylor expansion (2.6) gives a local model-agnostic description in the spirit of LIME. Explicit choice $\boldsymbol{\epsilon} = -\boldsymbol{x}$ provides (3.2), which can be viewed as a local description of $\mu(\mathbf{0})$ relative to $\boldsymbol{x}$. The 1st order contributions (3.3) combine all these local descriptions (3.1) w.r.t. a given quantile level to get the integrated MACQ view of $\mu(\mathbf{0})$, i.e.,

$$\begin{aligned}\mu(\mathbf{0}) \ &\approx \ \mathbb{E}\left[\mu\left(\boldsymbol{X}\right) - \left(\nabla_{\boldsymbol{x}} \mu(\boldsymbol{X})\right)^{\top} \boldsymbol{X} \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right] \\ &= F^{-1}_{\mu(\boldsymbol{X})}(\alpha) - \sum_{j=1}^{q} S_j(\mu; \alpha).\end{aligned}$$

This exactly corresponds to 1st order approximation (3.3). In the sequel it is less important that we can approximate $\mu(\mathbf{0})$ by this integrated view, but $\mu(\mathbf{0})$ plays the role of the *reference level* that calibrates our global meta-explanation. Thus, all explanations made need to be understood relative to this reference level $\mu(\mathbf{0})$.

- In (3.2)–(3.3) we implicitly assumed that $\mathbf{0}$ is a suitable reference point for calibrating our global meta-explanation. We further explore and improve this calibration in Sect. 3.3, below.
- Integrated gradients (2.8) integrate along a single path from a reference point $\boldsymbol{x}_0$ to $\boldsymbol{x}$ to make the 1st order Taylor approximation precise. We exchange the roles of the points, here, and we approximate the reference point by aggregating over all local descriptions in features $\boldsymbol{X}$.
- 1st order contributions (3.3) provide a 3-way description of the regression function, namely, they combine (i) marginal attribution $S_j(\mu;\alpha)$ as a function of $1 \leq j \leq q$, (ii) response level $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$ as a function of α, and (iii) feature values x_j. In our applications below we will illustrate the data from these different angles, each having its importance in explaining the response.
- 1st order attribution (3.1) combines marginal attributions $X_j\mu_j(\boldsymbol{X})$ by focusing on a common quantile level. A similar approach could also be done for other model-agnostic tools, such as the Shapley value.

Example 3.2 (linear regression) A linear regression model considers regression function

$$\boldsymbol{x} \;\mapsto\; \mu(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}, \tag{3.4}$$

with bias/intercept $\beta_0 \in \mathbb{R}$ and regression parameter $\boldsymbol{\beta} \in \mathbb{R}^q$. The 1st order contributions (3.3) are for $\alpha \in (0,1)$ given by

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) = \beta_0 + \sum_{j=1}^{q} \beta_j \mathbb{E}_P\left[X_j \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right] = \mu(\mathbf{0}) + \sum_{j=1}^{q} S_j(\mu;\alpha). \tag{3.5}$$

Thus, we weight the regression parameters β_j with the feature components X_j according to their contributions to quantile $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$; and the reference point $\mathbf{0}$ is given naturally providing initialization $\mu(\mathbf{0}) = \beta_0$.

This MACQ explanation (3.5) is rather different from the ALE profile (2.5). If we initialize $x_{j_0} = 0$, we receive ALE profile for the linear regression model

$$x_j \mapsto \int_0^{x_j} \Delta_j(z_j)dz_j = \beta_j x_j.$$

This is exactly marginal attribution (2.7) of component j in the linear regression model and it explains the change of the linear regression function if we change feature component x_j, whereas (3.5) describes the contribution of each feature component to an expected response level $\mu(\boldsymbol{x})$. Explanation (3.5) of the quantile level $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$ is exact in the linear regression case because there are no higher order terms in the Taylor expansion (2.6).

In general, the Taylor expansion (3.3) is accurate if the distance between $\mathbf{0}$ and $\boldsymbol{X}$ is small enough for all relevant $\boldsymbol{X}$, and if the regression function can be well described around $\mu(\boldsymbol{X})$ by a linear function. The former requires that the reference point is chosen somewhere "in the middle" of the feature distribution P. A useful consequence of our output-to-input view is that we can explicitly quantify the accuracy of the 1st order approximation by

$$\left| F^{-1}_{\mu(\boldsymbol{X})}(\alpha) - \mu\left(\mathbf{0}\right) - \sum_{j=1}^{q} S_j(\mu;\alpha) \right|. \tag{3.6}$$

In general, we want (3.6) to be small uniformly in quantile level α, for the given reference point $\mathbf{0}$. This then implies that the 1st order attributions give a good description on all quantile levels α. In the linear regression case this description is exact, see (3.5). In contrast to the Taylor decomposition in Montavon et al. (2017), the quantiles $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$ give us a natural anchor point for determining a suitable reference point, which is also computationally feasible. This idea will be developed in Sect. 3.3, below.

3.2 Second order attributions and interaction strength

Friedman and Popescu (2008) and Apley and Zhu (2020) have used higher order derivatives of $\mu(\cdot)$ to analyze interaction strength in systematic effects. This requires the study of higher order Taylor expansions. The 2nd order Taylor expansion is given by

$$\mu(\boldsymbol{x}+\boldsymbol{\epsilon}) = \mu(\boldsymbol{x}) + (\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x}))^{\top}\boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}^{\top}(\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{x}))\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|^2), \tag{3.7}$$

where $\nabla^2_{\boldsymbol{x}}\mu$ denotes the Hessian of μ w.r.t. $\boldsymbol{x}$. Setting $\boldsymbol{\epsilon} = -\boldsymbol{x}$ allows us, in complete analogy to (3.3), to study *2nd order contributions*

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \;\approx\; \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q} S_j(\mu;\alpha) - \frac{1}{2}\sum_{j,k=1}^{q} T_{j,k}(\mu;\alpha), \tag{3.8}$$

with *2nd order attributions*, for $1 \le j,k \le q$,

$$T_{j,k}(\mu;\alpha) = \mathbb{E}_P\left[X_j X_k \mu_{j,k}(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right]. \tag{3.9}$$

Slightly rearranging the terms in (3.7) allows us to study individual feature contributions and interaction terms separately, that is,

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \;\approx\; \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q}\left(S_j(\mu;\alpha) - \frac{1}{2}T_{j,j}(\mu;\alpha)\right) - \sum_{1\le j<k\le q} T_{j,k}(\mu;\alpha). \tag{3.10}$$

The last term quantifies all 2nd order contributions coming from interactions between X_j and X_k, $j \neq k$. We will show how interaction effects can be included in individual features' marginal attributions in Sect. 5.5, below.

Example 3.3 (quadratic regression function) A quadratic regression model considers the regression function

$$\boldsymbol{x} \;\mapsto\; \mu(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x} + \boldsymbol{x}^\top \mathbf{B} \boldsymbol{x}, \tag{3.11}$$

with parameters $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\mathbf{B} = (b_{j,k})_{j,k} \in \mathbb{R}^{q \times q}$. The 2nd order contributions are for $\alpha \in (0, 1)$ given by

$$\begin{aligned} F^{-1}_{\mu(\boldsymbol{X})}(\alpha) &= \beta_0 + \boldsymbol{\beta}^\top \mathbb{E}_P \left[\boldsymbol{X} \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right] \\ &\quad + \sum_{j,k=1}^{q} b_{j,k} \mathbb{E}_P \left[X_j X_k \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right] \\ &= \mu(\mathbf{0}) + \sum_{j=1}^{q} S_j(\mu; \alpha) - \frac{1}{2} \sum_{j,k=1}^{q} T_{j,k}(\mu; \alpha), \end{aligned}$$

with $\mu(\mathbf{0}) = \beta_0$. Thus, obviously, the explanation (3.10) of the quantile level $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$ is exact in the case of a quadratic regression function (3.11).

Remark 3.4 The motivation for studying 1st order attributions (3.1) has been given in terms of the risk sensitivity tools of Hong (2009) and Tsanakas and Millossovich (2016). These are obtained by calculating directional derivatives of distortion risk measures (using a Dirac distortion, see Appendix A). This argument does not carry forward to the 2nd order terms (3.9), as 2nd order directional derivatives of distortion risk measures turn out to be much more complicated, even in the linear case, see Property 1 in Gourieroux et al. (2000).

3.3 Choice of reference point

To obtain sufficient accuracy in 1st and 2nd order approximations, respectively, the reference point should lie somewhere "in the middle" of the feature distribution P. We elaborate on this in this section. Typically, we want to get the following expression small, uniformly in $\alpha \in (0, 1)$,

$$\left| F^{-1}_{\mu(\boldsymbol{X})}(\alpha) - \mu(\mathbf{0}) - \sum_{j=1}^{q} S_j(\mu; \alpha) + \frac{1}{2} \sum_{j,k=1}^{q} T_{j,k}(\mu; \alpha) \right|. \tag{3.12}$$

This expression is for reference point $\mathbf{0}$. However, we can select any other reference point $\boldsymbol{a} \in \mathbb{R}^q$, by exploring the 2nd order Taylor expansion (3.7) for $\boldsymbol{\epsilon} = \boldsymbol{a} - \boldsymbol{x}$. This latter reference point $\boldsymbol{a}$ then provides us with a 2nd order approximation

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \approx \mu(\boldsymbol{a}) - \mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top \nabla_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right] - \frac{1}{2}\mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top (\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{X}))(\boldsymbol{a}-\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right]. \tag{3.13}$$

The same can be received by translating the distribution P of the features by setting $\boldsymbol{X}^{\boldsymbol{a}} = \boldsymbol{X} - \boldsymbol{a}$ and letting $\mu^{\boldsymbol{a}}(\cdot) = \mu(\boldsymbol{a} + \cdot)$. The approximation (3.13) motivates us to look for a reference point $\boldsymbol{a} \in \mathbb{R}^q$ that makes the 2nd order approximation as accurate as possible for "all" quantile levels. Being a bit less ambitious, we select a discrete quantile grid $0 < \alpha_1 < \ldots < \alpha_L < 1$ on which we would like to have a good approximation capacity. Define the events $\mathcal{A}_l = \{\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha_l)\}$ for $1 \le l \le L$. Consider the objective function

$$\boldsymbol{a} \;\mapsto\; G(\boldsymbol{a};\mu) = \sum_{l=1}^{L}\left(F^{-1}_{\mu(\boldsymbol{X})}(\alpha_l) - \mu(\boldsymbol{a}) + \mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top \nabla_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mathcal{A}_l\right] + \frac{1}{2}\,\mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top (\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{X}))(\boldsymbol{a}-\boldsymbol{X})^\top \,\middle|\, \mathcal{A}_l\right]\right)^2. \tag{3.14}$$

Minimizing this objective function in $\boldsymbol{a}$ gives us an optimal reference point w.r.t. the quantile levels $(\alpha_l)_{1\le l\le L}$. Unfortunately, $\boldsymbol{a} \mapsto G(\boldsymbol{a};\mu)$ is not a convex function, and therefore numerical methods may only find local minima. These can be found by a plain vanilla gradient descent algorithm. We calculate the gradient of G w.r.t. $\boldsymbol{a}$

$$\begin{aligned}\nabla_{\boldsymbol{a}} G(\boldsymbol{a};\mu) = 2\sum_{l=1}^{L}\Bigg(&F^{-1}_{\mu(\boldsymbol{X})}(\alpha_l) - \mu(\boldsymbol{a}) + \mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top \nabla_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mathcal{A}_l\right] \\ &+ \frac{1}{2}\,\mathbb{E}_P\left[(\boldsymbol{a}-\boldsymbol{X})^\top (\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{X}))(\boldsymbol{a}-\boldsymbol{X})^\top \,\middle|\, \mathcal{A}_l\right]\Bigg) \\ &\times\Bigg(-\nabla_{\boldsymbol{a}}\mu(\boldsymbol{a}) + \mathbb{E}_P\left[\nabla_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mathcal{A}_l\right] \\ &- \mathbb{E}_P\left[\boldsymbol{X}^\top \nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mathcal{A}_l\right] + \frac{1}{2}\boldsymbol{a}^\top \mathbb{E}_P\left[\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{X}) \,\middle|\, \mathcal{A}_l\right]\Bigg).\end{aligned}$$

The gradient descent algorithm then provides for tempered learning rates $\varepsilon_{t+1} > 0$ updates at algorithmic time t

$$\boldsymbol{a}^{(t)} \;\mapsto\; \boldsymbol{a}^{(t+1)} = \boldsymbol{a}^{(t)} - \varepsilon_{t+1}\nabla_{\boldsymbol{a}} G(\boldsymbol{a}^{(t)};\mu). \tag{3.15}$$

Iteration step-wise locally decreases the objective function G.

Remark 3.5 • The above algorithm provides a global optimal reference point, thus, a calibration for a global 2nd order meta-explanation. In some cases this global calibration may not be satisfactory, in particular, if the reference point is far from

the feature values $\boldsymbol{X} = \boldsymbol{x}$ that mainly describe a given quantile level $F^{-1}_{\mu(\boldsymbol{X})}(\alpha)$ through the corresponding conditional probability $P[\cdot|\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)]$. In that case, one may be interested in different local reference points that are optimal for certain quantile levels, say, between 95 and 99%. In some sense, this will provide a more "honest" description (3.8) because we do not try to simultaneously describe all quantile levels. The downside of multiple reference points is that we lose comparability of marginal effects across the whole decision space.

- Our attribution method (starting from the quantile level) has the advantage that we can quantify the precision of our explanation through (3.12), and in the linear and quadratic regression cases of Examples 3.2 and 3.3 this description is exact.

4 Synthetic example

We start with a synthetic data example. A synthetic example has the advantage of a known (true) regression function μ. This allows us to verify that we draw the right conclusions.[1]

4.1 Generation of synthetic data

We choose $q = 7$ and generate i.i.d. features $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,7})^\top \sim P, 1 \le i \le n$, from a 7-dimensional standard Gaussian distribution having independent components. We generate $n = 10,000$ instances. For the regression function μ we set

$$\boldsymbol{x} \in \mathbb{R}^7 \quad \mapsto \quad \mu(\boldsymbol{x}) = \frac{1}{2}x_1^2 + \sin(x_2) + \frac{1}{2}x_3 \sin(x_4) - \frac{1}{2}x_5 x_6. \tag{4.1}$$

Thus, component x_7 does not enter regression function μ. Based on this regression function we generate for $1 \le i \le n$ independent Gaussian responses Y_i, given $\boldsymbol{x}_i$, that is

$$Y_i|_{\boldsymbol{x}_i} \sim \mathcal{N}(\mu(\boldsymbol{x}_i), 1). \tag{4.2}$$

This gives us data set $\mathcal{D} = \{(y_i, \boldsymbol{x}_i);\ 1 \le i \le n\}$. The resulting mean squared error (MSE) of the simulated samples is given by

$$\text{MSE}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu(\boldsymbol{x}_i))^2 = 1.0012. \tag{4.3}$$

This is an empirical approximation to the true variance of 1, see (4.2).

[1] The code of the synthetic example is available from https://github.com/RonRichman/MACQ.

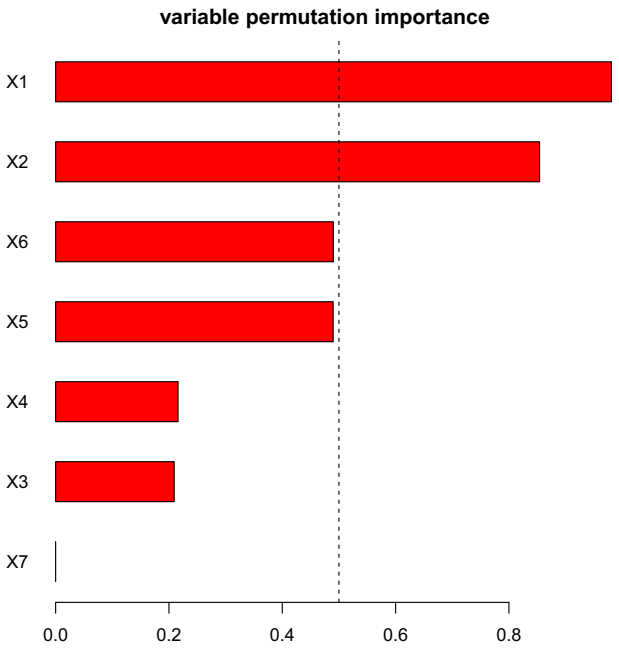

Fig. 1 Synthetic example: variable permutation importance

4.2 Variable permutation importance

We start by analyzing variable permutation importance (VPI) introduced by Breiman (2001). VPI is obtained by randomly permuting one component $1 \leq j \leq q$ of the features $\boldsymbol{x}_i \in \mathbb{R}^q$ at a time across all instances $1 \leq i \leq n$, and measuring the relative increase in MSE compared to (4.3). The bigger this relative increase in MSE the bigger the VPI of feature component x_j.

Figure 1 shows the results. From this we conclude that x_1 and x_2 are the most important feature components in this example, as permutation of these components increases the MSE by almost 100%. The permutation of feature components x_6 and x_5 leads to an increase of 50% and the permutation of x_4 and x_3 leads to an increase of 20%. The permutation of x_7 does not lead to any increase, showing that this component is unimportant. This gives us a first indication of variable importance.

4.3 MACQ explanation

We present the MACQ analysis for the regression function (4.1). In a first step, we need to find a suitable reference point $\boldsymbol{a} \in \mathbb{R}^q$ to obtain good accuracy in the 2nd order approximation (3.13). Therefore, we consider the objective function $G(\boldsymbol{a}; \mu)$ given in (3.14) and apply plain vanilla gradient descent updates (3.15) with learning rates $\varepsilon_{t+1} = 10^{-2}/\|\nabla_{\boldsymbol{a}} G(\boldsymbol{a}^{(t)}; \mu)\|$ to minimize this objective function. For the quantile grid we choose $\alpha_l \in \{1\%, \ldots, 99\%\}$. This optimization gives us reference point

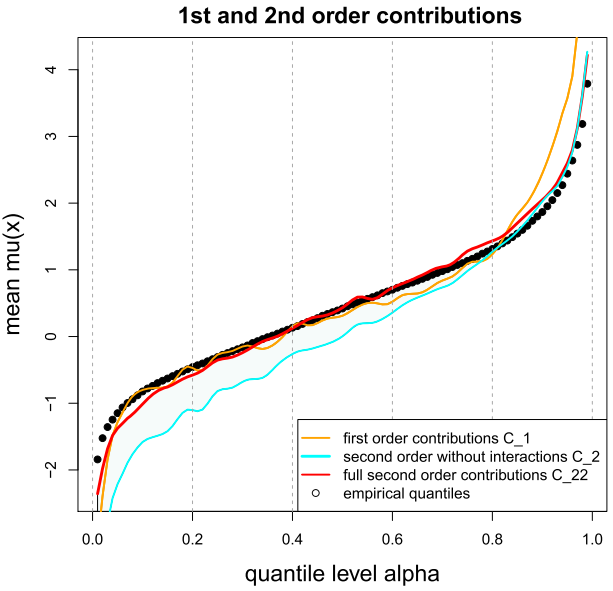

Fig. 2 1st and 2nd order contributions C_1, C_2 and $C_{2,2}$ compared to the empirical quantiles $\widehat{F}^{-1}_{\mu(\boldsymbol{X})}(\alpha_l)$, $1 \le l \le L$

$$\boldsymbol{a} = (0.000, -0.027, 1.187, -0.032, 0.891, 0.891, 0.000)^\top \in \mathbb{R}^7. \tag{4.4}$$

This reference point is centered except for the components a_3, a_5 and a_6, and $\boldsymbol{a}$ provides us with a base level of $\mu(\boldsymbol{a}) = -0.44$. We aim at studying the 1st and 2nd order attributions $S_j(\mu;\alpha)$ and $T_{j,k}(\mu;\alpha)$, respectively, relative to this base level. Working with the data $(Y_i, \boldsymbol{x}_i, \mu(\boldsymbol{x}_i))_{1\le i\le n}$, we need to estimate the conditional expectations on the events $\{\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\}$, $\alpha \in (0,1)$, to receive empirical versions of 1st and 2nd order attributions. We do this on a discrete grid by using a local smoother of degree 2. We use the R function `locfit`, see Loader et al. (2020), with parameters `deg=2` and `alpha=0.1` (the chosen bandwidth) applied to observations $x^{\boldsymbol{a}}_{i,j}\mu_j(\boldsymbol{x}_i)$ and $x^{\boldsymbol{a}}_{i,j}x^{\boldsymbol{a}}_{i,k}\mu_{j,k}(\boldsymbol{x}_i)$, $1 \le i \le n$, where we set $\boldsymbol{x}^{\boldsymbol{a}}_i = \boldsymbol{x}_i - \boldsymbol{a}$ for the chosen reference point (4.4). We then fit the local smoother to these observations, ordered w.r.t. the ranks of $\mu(\boldsymbol{x}_i)$. Thus, e.g., the $\boldsymbol{a}$-adjusted 1st order attributions $S_j(\mu;\alpha_l)$, $1 \le l \le L$, are estimated empirically by the pseudo code

$$\begin{aligned} &\texttt{predict(locfit(}x^{\boldsymbol{a}}_{i,j}\mu_j(\boldsymbol{x}_i) \sim \text{rank}(\mu(\boldsymbol{x}_i))/n, \\ &\texttt{alpha = 0.1, deg = 2), newdata = c(1 : 99)/100),} \end{aligned} \tag{4.5}$$

and correspondingly for the 2nd order attributions $T_{j,k}(\mu;\alpha_l)$, $1 \le l \le L$.

Figure 2 illustrates the results after optimizing for the reference point $\boldsymbol{a}$. The black dots show the empirical quantiles $\widehat{F}^{-1}_{\mu(\boldsymbol{X})}(\alpha_l)$, $1 \le l \le L$, obtained from the simulated data $\mu(\boldsymbol{x}_1), \ldots, \mu(\boldsymbol{x}_n)$. The orange line shows the total 1st order contributions $C_1 = \mu(\boldsymbol{a}) + \sum_{j=1}^q S_j(\mu;\alpha_l)$, see (3.3), received from approximations (4.5). The

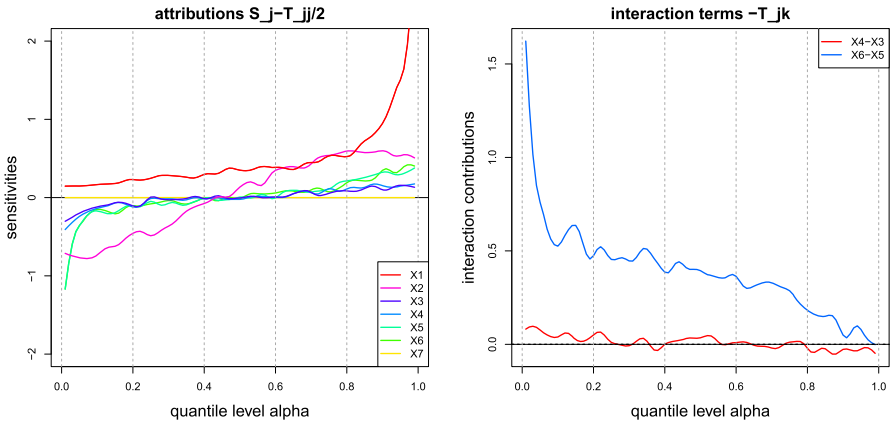

Fig. 3 (lhs) Attributions $S_j(\mu;\alpha) - \frac{1}{2}T_{j,j}(\mu;\alpha)$ excluding interaction terms, (rhs) interaction terms $-T_{j,k}(\mu;\alpha)$, $j \neq q$

cyan line shows the total 2nd order contributions without interaction terms $C_2 = \mu(\boldsymbol{a}) + \sum_{j=1}^{q}\left(S_j(\mu;\alpha_l) - \frac{1}{2}T_{j,j}(\mu;\alpha_l)\right)$, and the red line shows the full 2nd order contributions $C_{2,2} = \mu(\boldsymbol{a}) + \sum_{j=1}^{q}\left(S_j(\mu;\alpha_l) - \frac{1}{2}T_{j,j}(\mu;\alpha_l)\right) - \sum_{1\le j<k\le q} T_{j,k}(\mu;\alpha_l)$, see (3.8). Figure 2 is interpreted as follows. The full 2nd order contributions $C_{2,2}$ (red line) match the empirical quantiles $\widehat{F}^{-1}_{\mu(\boldsymbol{X})}(\alpha_l)$ (black dots) quite well; this shows that our choice of the reference point $\boldsymbol{a}$ leads to a small approximation error of the quantile function across all confidence levels. The shaded cyan area between C_2 (cyan line) and $C_{2,2}$ (red line) shows the attributions to the interaction terms $-T_{j,k}(\mu;\alpha)$, $j \neq k$. Since this area is comparably large, we conclude that interactions are relevant; we come back to this in Fig. 3 (rhs) below. Moreover, we observe that the orange line does not match the empirical quantiles as well as the red line, especially for large quantile levels α. This shows that a 1st order approximation is not sufficient to explain the (large) quantiles (note that we have a quadratic term $x_1^2/2$ in the regression function μ which is precisely the reason why we need to include 2nd order terms).

Next, we study the 1st and 2nd order attributions $S_j(\mu;\alpha) - \frac{1}{2}T_{j,j}(\mu;\alpha)$ for all feature components x_j, $1 \le j \le q$. Figure 3 (lhs) shows these attributions for the different chosen quantiles (and excluding interaction terms). This plot indicates feature importance at different quantile levels. First, we observe that component x_7 (yellow color) does not have any influence because $S_j(\mu;\alpha) - \frac{1}{2}T_{j,j}(\mu;\alpha) \equiv 0$ for $j = 7$. Second, the attributions that undergo the biggest changes from small to large quantiles are the ones of feature components x_1 and x_2. This shows that these two variables are the most important ones to explain the quantile levels of $\mu(\boldsymbol{x})$. This is in line with the VPI plot of Fig. 1, and, moreover, these two variables have a significant influence over the entire quantile range. The next important variables are x_5 and x_6, and, in particular, these variables are important to explain very small quantiles, noting that the term $-x_5x_6/2$ in μ is unbounded from below (having a quadratic behavior) which is the dominant term for small quantiles. The influence of the remaining variables x_3 and x_4 is smaller, though still clearly different from zero. Thus, we find a similar variable importance behavior for VPI and MACQ, but MACQ allows us to allocate importance to different quantile levels, specifically, from the VPI plot we observe that

x_1 and x_2 are similarly important, while MACQ additionally tells us that x_2 is more important for small quantile levels, whereas x_1 dominates high quantile levels. Thus, MACQ offers a more granular view.

Figure 3 (rhs) shows the interaction terms $-T_{j,k}(\mu;\alpha)$, $j \neq q$, that are different from zero. There are only two terms different from zero, and they are exactly the ones that have interactions in μ, see (4.1). Moreover, the interaction x_6–x_5 is clearly more important than the interaction x_4–x_3; the former is unbounded in both variables, while the latter is bounded for x_4 through the sine function in μ. We conclude that we can identify the important variables and interactions that mostly contribute to explain the different quantile levels of $\mu(\boldsymbol{X})$.

4.4 Contribution of individual instances

Finally, we analyze individual instances $\boldsymbol{x}_i^{\boldsymbol{a}} = \boldsymbol{x}_i - \boldsymbol{a}$ and study individual marginal contributions

$$\omega_{i,j} = (x_{i,j} - a_j)\mu_j(\boldsymbol{x}_i) - (x_{i,j} - a_j)^2 \mu_{j,j}(\boldsymbol{x}_i)/2$$

to the attribution $S_j(\mu;\alpha) - T_{j,j}(\mu;\alpha)/2$. For Fig. 4 we select at random 1000 different instances $\boldsymbol{x}_i$, and plot their individual marginal contributions $\omega_{i,j}$ to the attributions $S_j(\mu;\alpha) - T_{j,j}(\mu;\alpha)/2$ (black solid line) for selected feature components, $j = 1, 2, 3$. The ordering on the x-axis for the selected instances $\boldsymbol{x}_i$ is obtained by considering the empirical quantiles of the responses $\mu(\boldsymbol{x}_k)$ over all instances $1 \leq k \leq n$. The horizontal black line at 0 corresponds to the reference level. In addition to the attributions $S_j(\mu;\alpha) - T_{j,j}(\mu;\alpha)/2$ (black solid line), the plot is complemented by black dotted lines giving one (empirical) standard deviation

$$\mathrm{Var}_P\left((X_j - a_j)\mu_j(\boldsymbol{X}) - (X_j - a_j)^2 \mu_{j,j}(\boldsymbol{X})/2 \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right)^{1/2}. \tag{4.6}$$

The sizes of these standard deviations quantify the heterogeneity in the individual marginal contributions $\omega_{i,j}$. This can either be because of heterogeneity of the portfolio

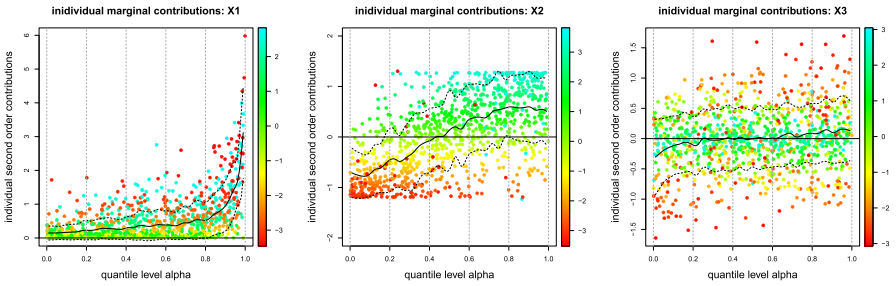

Fig. 4 Individual marginal contributions $\omega_{i,j}$ of 1,000 randomly selected instances $\boldsymbol{x}_i$ for (lhs) $j = 1$, (middle) $j = 2$, (rhs) $j = 3$; the black line shows attribution $S_j(\mu;\alpha) - T_{j,j}(\mu;\alpha)/2$ and the black dotted line gives one standard deviation; the colors illustrate the feature values x_j

$x_{i,j}$ at a certain quantile level, or because we have a rough regression surface implying heterogeneity in the derivatives $\mu_j(\boldsymbol{x}_i)$ and $\mu_{j,j}(\boldsymbol{x}_i)$.

We start by explaining Fig. 4 (lhs) which shows feature component x_1; this feature component enters the regression function μ as $x_1^2/2$, see (4.1). The color scale shows the feature values $x_{i,1}$ of the individual instances (indexed by i). From this plot we conclude that the feature values $x_{i,1}$ around zero (green, yellow color) give smaller contributions, while very negative values (red color) and very positive values (light blue color) give higher contributions also resulting in comparably larger values for $\mu(\boldsymbol{x})$. Of course, this makes perfect sense, given the quadratic term $x_1^2/2$ in μ.

Figure 4 (middle) basically shows one cycle of the sine function $\sin(x_2)$. Small values of $x_{i,2}$ (red color) explain small expected responses $\mu(\boldsymbol{x})$, whereas large values of $x_{i,2}$ (light blue color) explain bigger expected responses (for a linear function in x_2 we would receive a strictly horizontal coloring; an example is given in Fig. 16 in the appendix).

Finally, Fig. 4 (rhs) analyzes the feature component x_3. This term enters the regression function as $x_3 \sin(x_4)/2$. The sine function performs a (random) sine flip (X_3 and X_4 are independent), therefore, the sign of the contribution of component x_3 is not well-determined. This is clearly visible from Fig. 4 (rhs) because red and light-blue dots equally spread around the zero line, and only the interaction between x_3 and x_4 determines the explicit contribution in this case.

This finishes the synthetic data example; we will provide additional (different) graphs and analysis in the real data example presented in the next section.

5 Real data example: bike rentals

5.1 Model choice and model fitting

We consider the bike rental example of Fanaee-T and Gama (2014), which has also been studied in Apley and Zhu (2020). The data describes the bike sharing process over the years 2011 and 2012 of the Capital Bikesharing system in Washington DC. On an hourly time grid we have information about the proportion of casual bike rentals relative to all bike rentals (casual and registered users). This data set is complemented by explanatory variables such as weather conditions and seasonal variables. We provide a descriptive analysis of the data in Appendix C. On average, 17% of all bike rentals are by casual users and 83% by registered users. However, these proportions strongly fluctuate w.r.t. daytime, holidays, weather conditions, etc. This variability is illustrated in Fig. 19 in Appendix C. We design a neural network regression function to forecast the proportion of casual rentals. We denote the response variable (proportion) by Y, and we denote the features (explanatory variables) by $\boldsymbol{x} \in \mathbb{R}^q$.

We choose a fully-connected feed-forward neural network $\theta : \mathbb{R}^q \to \mathbb{R}$ of depth $d = 3$, having $(q_1, q_2, q_3) = (20, 15, 10)$ neurons in the three hidden layers. This gives the network regression function

$$\boldsymbol{x} \in \mathbb{R}^q \quad \mapsto \quad \mu(\boldsymbol{x}) = \sigma(\theta(\boldsymbol{x})) \ \in (0, 1), \tag{5.1}$$

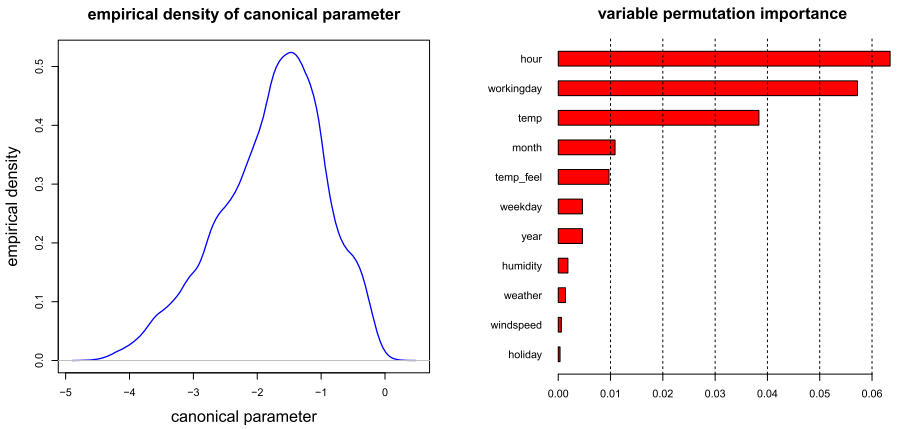

Fig. 5 (lhs) Empirical density of canonical parameter $(\theta(\boldsymbol{x}_i))_{1\le i\le n}$, (rhs) variable permutation importance

where σ is the sigmoid output activation and $\boldsymbol{x} \mapsto \theta(\boldsymbol{x})$ models the canonical parameter of a logistic regression model. In order to have a smooth network regression function we choose the hyperbolic tangent as activation function in the three hidden layers. We have implemented this network in TensorFlow, see Abadi et al. (2015), and in Keras, see Chollet et al. (2015); as mentioned above, this software allows us to formally calculate gradients and Hessians.

In all that follows we do not consider the attributions of the regression function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$ itself, but we directly focus on the corresponding attributions on the canonical scale $\boldsymbol{x} \mapsto \theta(\boldsymbol{x})$. This has the advantage that the results do not get distorted by the sigmoid output activation σ. Thus, we replace μ by θ in (3.8)–(3.10), resulting in studying 2nd order contributions

$$F^{-1}_{\theta(\boldsymbol{X})}(\alpha) \;\approx\; \theta(\boldsymbol{a}) + \sum_{j=1}^{q} \left(S_j(\theta;\alpha) - \frac{1}{2} T_{j,j}(\theta;\alpha) \right) - \sum_{1\le j<k\le q} T_{j,k}(\theta;\alpha), \tag{5.2}$$

for reference point $\boldsymbol{a} \in \mathbb{R}^q$. The network architecture is fitted to the available data using early stopping to prevent over-fitting. Importantly, we do not say anything here about the quality of the predictive model, but aim at understanding the fitted regression function $\boldsymbol{x} \mapsto \theta(\boldsymbol{x})$. This can be done regardless of whether the chosen model is suitable for the predictive task at hand. Figure 5 (lhs) shows the empirical density of the canonical parameters $\boldsymbol{x}_i \mapsto \theta(\boldsymbol{x}_i)$ of the fitted model over all instances $1 \le i \le n$. We have negative skewness in this empirical density.

5.2 Variable permutation importance and partial dependence plots

We start by presenting the classical explainability tools. As a first variable importance measure we again provide the VPI plot of Breiman (2001). As objective function we use the Bernoulli deviance loss which is proportional to the binary cross-entropy

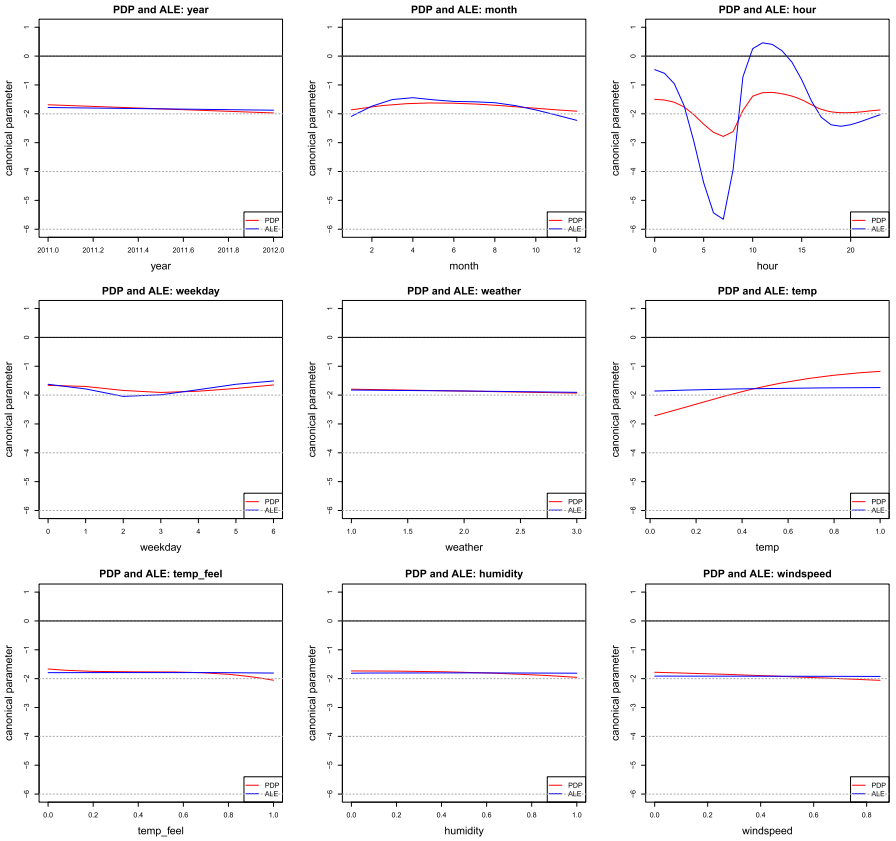

Fig. 6 PDPs and ALE profiles of the feature components having more than 2 levels

(also called log-loss). Figure 5 (rhs) shows the VPI. There are three variables (hour, working day and temperature) that highly dominate all others. Note that the VPI does not properly consider the dependence structure in $\boldsymbol{X}$, similarly to ICEs and PDPs, because permutation of x_j is done without impacting the remaining components $\boldsymbol{x}_{\setminus j}$.

Figure 6 gives the PDPs (red color) and the ALE profiles (blue color) of all feature components that have more than 2 levels; the y-scale is the same in all plots. As explained in Sect. 2.1, these techniques analyze the sensitivities of $\mu(\boldsymbol{x})$ in the individual feature components of $\boldsymbol{x}$. PDPs do not respect the dependence structure within $\boldsymbol{x}$, whereas ALE profiles do. From Fig. 6 we observe that these dependence structures may play an important role, e.g., in the most important variable `hour` (daytime) the dependence structure significantly influences the graph. In fact, `hour` and `temp` are highly dependent (nighttime is colder than daytime), and this implies that we do not observe a temperature of 30 degree Celsius during nighttime. ALE profiles respect this, but PDPs do not. In general, these sensitivity plots reflect the empirical marginal plots of Fig. 19 in the appendix, but beyond that they do not provide much further insight, e.g., concerning interactions of feature components (theoretically, it would be possible to produce two-dimensional PDPs and ALEs, but in high-dimensional problems this is not feasible).

5.3 1st and 2nd order contributions

We now turn our attention to the MACQ approach to interpret the fitted network of the bike rental data. The accuracy of the 2nd order contributions (5.2) will depend on the choice of the reference point $\boldsymbol{a} \in \mathbb{R}^q$. For network gradient descent fitting we have normalized the feature components to be centered and have unit variance, i.e., $\mathbb{E}_P[\boldsymbol{X}] = \mathbf{0}$ and $\mathrm{Var}_P(X_j) = 1$ for all $1 \leq j \leq q$. This pre-processing is needed to efficiently apply stochastic gradient descent network fitting, and all subsequent interpretations should be understood in terms of the scaled feature components. Of course, by a simple back-transformation we get back to interpretations on the original feature scale. We then translate these feature components by choosing a reference point $\boldsymbol{a}$ such that the objective function $G(\boldsymbol{a}; \theta)$ is minimized, see (3.14); this is done as in Sect. 4 with the same learning rates $\varepsilon_{t+1} = 10^{-2}/\|\nabla_{\boldsymbol{a}} G(\boldsymbol{a}^{(t)}; \theta)\|$. The resulting decrease in the objective function $G(\cdot; \theta)$ is plotted in Fig. 7 (lhs).

The chosen reference point is given by

$$\boldsymbol{a} = (-0.27, 0.01, -0.18, 0.59, -0.18, -0.58, 0.15, -0.46, -0.48, 0.13, 0.15)^\top \in \mathbb{R}^{11}.$$

This gives canonical parameter of $\theta(\boldsymbol{a}) = -1.12$ and logistic probability $\mu(\boldsymbol{a}) = \sigma(\theta(\boldsymbol{a})) = 24\%$. Thus, the reference point lies at a higher probability level than the overall empirical probability of 17%.

Next we determine the 1st and 2nd order contributions in complete analogy to Sect. 4 using the local smoother (4.5). Figure 7 (rhs) gives the results after optimizing for the reference point $\boldsymbol{a}$. The orange line shows the 1st order contributions $C_1 = \theta(\boldsymbol{a}) + \sum_{j=1}^q S_j(\theta; \alpha)$, see (3.3), the cyan line the 2nd order contributions without interaction terms $C_2 = \theta(\boldsymbol{a}) + \sum_{j=1}^q (S_j(\theta; \alpha) - \frac{1}{2} T_{j,j}(\theta; \alpha))$, and the red line the full 2nd order contributions $C_{2,2} = \theta(\boldsymbol{a}) + \sum_{j=1}^q (S_j(\theta; \alpha) - \frac{1}{2} T_{j,j}(\theta; \alpha)) - \sum_{1 \leq j < k \leq q} T_{j,k}(\theta; \alpha)$, see (3.8)–(3.10). Figure 7 (rhs) tells us that the full 2nd order contributions $C_{2,2}$

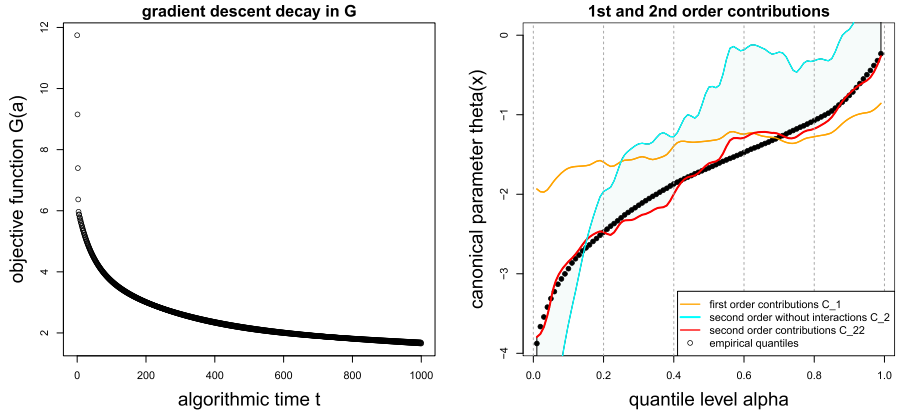

Fig. 7 (lhs) Gradient descent loss decay for determining reference point $\boldsymbol{a}$, (rhs) 1st and 2nd order contributions C_1, C_2 and $C_{2,2}$ compared to the empirical quantiles $\widehat{F}^{-1}_{\theta(\boldsymbol{X})}(\alpha_l)$, $1 \leq l \leq L$

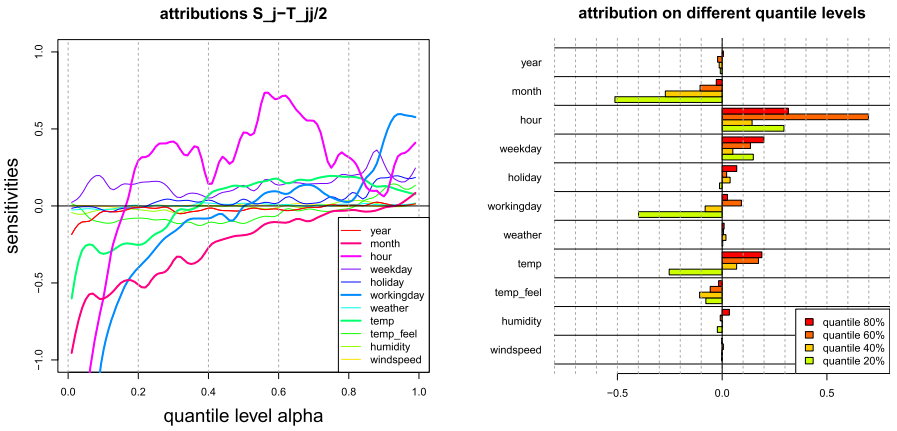

Fig. 8 (lhs) Attributions $S_j(\theta;\alpha) - \frac{1}{2}T_{j,j}(\theta;\alpha)$ excluding interaction terms, see (5.2), (rhs) attributions $S_j(\theta;\alpha) - \frac{1}{2}T_{j,j}(\theta;\alpha)$ for selected quantile levels $\alpha \in \{20\%, 40\%, 60\%, 80\%\}$

match the empirical quantiles $\widehat{F}^{-1}_{\theta(\boldsymbol{X})}(\alpha_l)$ (black dots) rather well. The shaded cyan area between C_2 (cyan line) and $C_{2,2}$ (red line) shows the significant influence of the interaction terms $T_{j,k}(\theta;\alpha)$, $j \neq k$. This implies that a simple generalized additive model (GAM) will not be able to model these data accurately.

In Fig. 8 (lhs) we show the attributions $S_j(\theta;\alpha) - \frac{1}{2}T_{j,j}(\theta;\alpha)$. These attributions show the differences relative to the canonical parameter in the reference point $\theta(\boldsymbol{a})$; when aggregating over $1 \leq j \leq q$ this results in the cyan line of Fig. 7 (rhs). Fig. 8 (lhs) shows substantial sensitivities in the variables `hour`, `workingday`, `temp` and `month`. From this we conclude that these are the most important variables in the regression model to explain the systematic effects in responses Y. In contrast to the VPI plot of Fig. 5 (rhs) and the PDPs of Fig. 6, this assessment correctly considers the dependence structure within the features $\boldsymbol{X}$; note that the variable `month` receives a higher importance in MACQ than in the VPI plot of Fig. 5 (rhs). (In Fig. 12, below, we will see that `month` has important interaction effects with other variables which may partly explain the differences between VPI and this MACQ assessment.) Figure 8 now allows us to analyze variable importance at different quantile levels by considering vertical slices. We consider such vertical slices in Fig. 8 (rhs) for four selected quantile levels $\alpha \in \{20\%, 40\%, 60\%, 80\%\}$. We observe that the variables `month`, `hour`, `workingday` and `temp` undergo the biggest changes when moving from small quantiles to big ones. The quantile level at 20% can be explained by the three features `temp`, `month` and `workingday`, whereas the quantile level at 60% has `hour` (daytime) as an important variable. Note that this is not the full MACQ picture, yet, as we do not consider interactions in these vertical slices; the importance of interactions is indicated by the cyan shaded area in Fig. 7 (rhs) for different quantile levels.

The 1st and 2nd order contribution results given in Figs. 7 and 8 have been obtained by one single network that has been fitted to the data. Since network fitting lacks a certain degree of robustness, due to the fact that stochastic gradient descent may

explore different local minimums of the loss surface, we verify in Appendix B.1 that the proposed MACQ analysis gives similar results also for other networks that have been fitted to the same data.

Remark 5.1 The feature components of $\boldsymbol{x}$ need pre-processing in order to be suitable for gradient descent fitting. Continuous and binary variables have been centered and normalized so that their gradients live on a similar range. This makes gradient descent fitting more efficient because all partial derivatives in the gradient are directly comparable. Our example does not have categorical feature components. Categorical feature components can be treated in different ways. For our MACQ proposal we envisage two different treatments. Firstly, dummy coding could be used. This requires the choice of a reference level, and considers all other levels relative to this reference level. The resulting marginal attributions should then be interpreted as differences to the reference level. Secondly, one can use embedding layers for categorical variables, see Bengio et al. (2003) and Guo and Berkhahn (2016). In that case the attribution analysis can directly be done on these learned embeddings of categorical levels, in complete analogy to the continuous variables.

5.4 Contribution of individual instances

Next, we focus on individual instances $\boldsymbol{x}_i^{\boldsymbol{a}} = \boldsymbol{x}_i - \boldsymbol{a}$ and study individual marginal contributions $\omega_{i,j} = (x_{i,j} - a_j)\theta_j(\boldsymbol{x}_i) - (x_{i,j} - a_j)^2\theta_{j,j}(\boldsymbol{x}_i)/2$ to attribution $S_j(\theta;\alpha) - T_{j,j}(\theta;\alpha)/2$. This is in analogy to Fig. 4, but again we study the contributions on the canonical scale θ.

For Fig. 9 we select at random 1,000 different instances $\boldsymbol{x}_i$, and plot their individual marginal contributions $\omega_{i,j}$ (colored dots) to the attributions $S_j(\theta;\alpha) - T_{j,j}(\theta;\alpha)/2$ (black solid line). The ordering on the x-axis is w.r.t. the quantile levels $\alpha \in (0, 1)$, the black solid line shows the attributions and the black dotted line gives one empirical standard deviation, see (4.6). We start with Fig. 9 (bottom-right), which shows the binary variable `workingday`. This variable clearly differentiates low from high quantiles $F^{-1}_{\theta(\boldsymbol{X})}(\alpha)$, showing that the casual rental proportion Y is in average bigger for non-working days (red dots). Moreover, for low quantiles levels the working day variable clearly lowers expected response $\theta(\boldsymbol{x})$ compared to the reference level $\theta(\boldsymbol{a})$, as the cyan dots are below the horizontal black line at 0, which corresponds to the reference level.

Next, we study the variable `temp` of Fig. 9 (bottom-left). In this plot we see a clear positive dependence between quantile levels and temperature, showing that casual rentals are generally low for low temperatures, which can either be the calendar season or bad weather conditions. We have clearly more heterogeneity in features (and resulting derivatives $\theta_j(\boldsymbol{x}_i)$ and $\theta_{j,j}(\boldsymbol{x}_i)$) contributing to low quantile levels than to higher ones. The variable `temp` is highly correlated with calendar month, and the calendar month plot in Fig. 9 (top-left) looks similar, showing that casual rental proportions Y are negatively impacted by winter seasons. There are some low proportions, though, also for summer months, these need to be explained by other variables, e.g., they may correspond to a rainy day or to a specific daytime. The interpretation of the variable `hour` in Fig. 9 (top-right) is slightly more complicated since we do not have

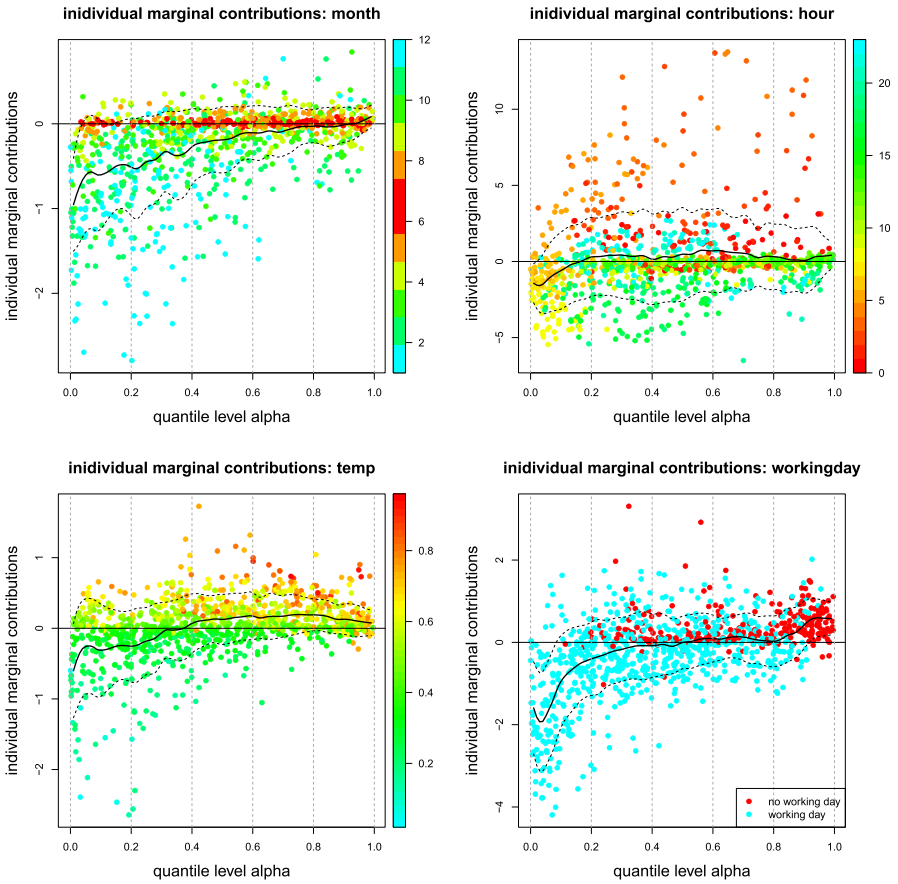

Fig. 9 Individual marginal contributions $\omega_{i,j}$ of 1,000 randomly selected instances $\boldsymbol{x}_i$ for (top-left) $j = \texttt{month}$, (top-right) $j = \texttt{hour}$, (bottom-left) $j = \texttt{temp}$ and (bottom-right) $j = \texttt{workingday}$; the black line shows attribution $S_j(\theta;\alpha) - T_{j,j}(\theta;\alpha)/2$ and the black dotted line gives one standard deviation; the y-scales differs in the plots and the colors illustrate the feature values x_j

monotonicity of $\theta(\boldsymbol{x})$ in this variable, see also Fig. 19 in the appendix. Nevertheless we also see a separation between working and leisure times (for the time-being ignoring interactions with holidays and weekends).

For better understanding, Fig. 9 should be compared to the corresponding plots of a generalized linear model (GLM). We provide a GLM in Appendix B.2, and the crucial property of the GLM plot is a horizontal layering of the colors due to linearity after applying the link function, see Fig. 16 in Appendix B.2.

In Fig. 9 we have plotted the individual marginal contributions $\omega_{i,j}$ on the y-axis against the quantiles $\alpha \in (0, 1)$ on the x-axis to explain how the features $\boldsymbol{x}_i$ enter the quantile levels $F^{-1}_{\theta(\boldsymbol{X})}(\alpha)$. This is the 3-way analysis mentioned above, where the third dimension is highlighted by using different colors in Fig. 9. Alternatively, we can also try to understand how this third dimension of different feature values x_j contributes to the individual marginal contributions $\omega_{i,j}$. Figure 10 plots the individual marginal

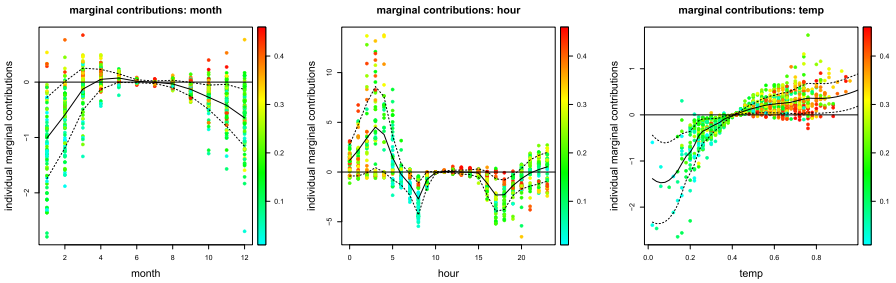

Fig. 10 Individual marginal contributions $\omega_{i,j}$ of 1,000 randomly selected instances $\boldsymbol{x}_i$ for (lhs) $j = \texttt{month}$, (middle) $j = \texttt{hour}$ and (rhs) $j = \texttt{temp}$; the black line shows the empirical average; the colors show the expected responses $\mu(\boldsymbol{x}_i) \in (0, 1)$ (casual rental proportions)

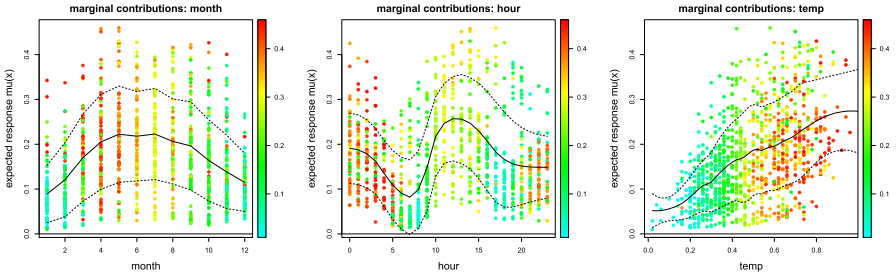

Fig. 11 Expected responses $\mu(\boldsymbol{x}_i)$ of 1,000 randomly selected instances $\boldsymbol{x}_i$ for (lhs) $j = \texttt{month}$, (middle) $j = \texttt{hour}$ and (rhs) $j = \texttt{temp}$; the black line shows the empirical average; the colors show the the individual marginal contributions $\omega_{i,j}$

contributions $\omega_{i,j}$ on the y-axis against the feature values x_j on the x-axis. The black line shows the averages of $\omega_{i,j}$ over all instances, and the colored dots show the 1,000 randomly selected instances $\boldsymbol{x}_i$ with the colors illustrating the expected responses, i.e., the expected casual rental proportions $\mu(\boldsymbol{x}_i) = \sigma(\theta(\boldsymbol{x}_i)) \in (0, 1)$. The general shape of the black lines in these graphs reflects well the PDPs and ALE profiles in Fig. 6. However, the detailed structure slightly differs in these plots as they do not exactly show the same quantity, Fig. 6 shows marginal empirical graphs, whereas Fig. 10 quantifies individual marginal contributions to expected responses $\theta(\boldsymbol{x})$ in an additive way (on the canonical scale). Figure 10 (rhs) shows a clear monotone plot which also results in a separation of the colors, whereas the colors in Fig. 10 (lhs, middle) can only be fully understood by also studying contributions and interactions with other components $x_{i,k}$, $k \neq j$.

Figure 11 shows the same data as Fig. 10, but it exchanges the role of the individual marginal contributions $\omega_{i,j}$ and the expected responses $\mu(\boldsymbol{x}_i)$. In Fig. 11 we show the responses on the y-axis and the color scale is chosen w.r.t. the individual marginal contributions $\omega_{i,j}$. We observe a strong positive correlation between the individual marginal contributions $\omega_{i,j}$ and the expected responses $\mu(\boldsymbol{x}_i)$, which is better visible here than in Fig. 10. Of course, this makes perfect sense as the individual marginal contributions are 2nd order approximations to the expected responses (subject to the interactions we are going to study in the next section).

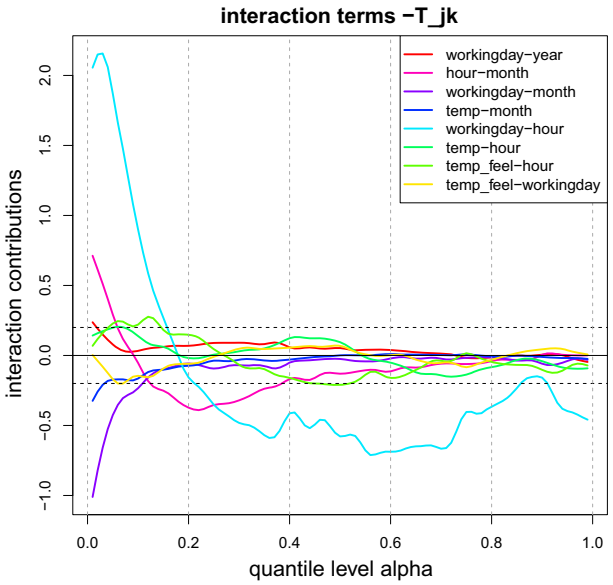

Fig. 12 Off-diagonal terms $-T_{j,k}(\theta;\alpha)$ giving the interactions

5.5 Importance of interaction terms

There remains the analysis of the interaction terms $-T_{j,k}(\theta;\alpha)$, $j \neq k$, which account for the cyan shaded are in Fig. 7 (rhs). These interaction terms are shown in Fig. 12.

To not overload Fig. 12 we only show the interaction terms $T_{j,k}$ for which $\max_\alpha |T_{j,k}(\theta;\alpha)| > 0.2$. We identify three major interaction terms: `workingday-hour`, `workingday-month` and `hour-month`. Of course, these interactions are very reasonable to understand casual rental proportions. For small quantiles also interactions `temp-month` and `temp-hour` are important. Interestingly, we also find an interaction `workingday-year`: in the data there is a positive trend of registered rental bike users (in absolute terms) which interacts differently on working and non-working days because casual rentals are more frequent on non-working days. Identifying the importance of these interactions highlights that it will not be sufficient to work within a GLM or a GAM unless we add explicit interaction terms to them.

In the final step we combine the attributions $S_j(\theta;\alpha) - T_{j,j}(\theta;\alpha)/2$ with the interaction terms $T_{j,k}(\theta;\alpha)$, $k \leq j$. A natural way is to just allocate half of the interaction terms $T_{j,k}(\theta;\alpha)$ to each component j and k. This then provides allocated 2nd order attributions to components $1 \leq j \leq q$

$$\begin{aligned} V_j(\theta;\alpha) &= S_j(\theta;\alpha) - T_{j,j}(\theta;\alpha)/2 - \sum_{j\neq k} T_{j,k}(\theta;\alpha)/2 \\ &= S_j(\theta;\alpha) - \sum_{k=1}^{q} T_{j,k}(\theta;\alpha)/2. \end{aligned}$$

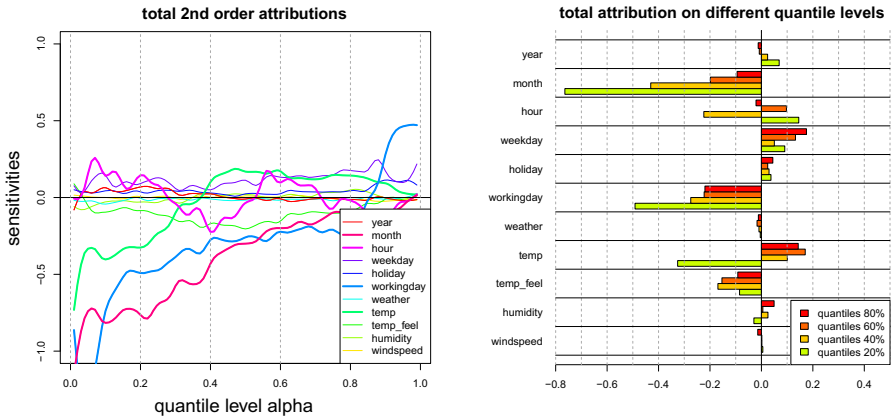

Fig. 13 (lhs) 2nd order attributions $V_j(\theta;\alpha)$ including interaction terms, and (rhs) $V_j(\theta;\alpha)$ for selected quantile levels $\alpha \in \{20\%, 40\%, 60\%, 80\%\}$

Adding the reference level $\theta(\boldsymbol{a})$, we again receive the full 2nd order contributions $C_{2,2} = \theta(\boldsymbol{a}) + \sum_{j=1}^{q} V_j(\theta;\alpha)$ illustrated by the red line in Fig. 7 (rhs). In Fig. 13 we provide these attributions $V_j(\theta;\alpha)$ for quantiles $\alpha \in (0, 1)$. These plots differ from Fig. 8 only by the inclusion of the 2nd order off-diagonal (interaction) terms. Comparing the right-hand sides of these two plots we observe that firstly the level is shifted, which is explained by the shaded cyan area in Fig. 7 (rhs). Secondly, interactions impact mainly the small quantiles in our example, as is made clear from Fig. 12.

6 Conclusions

This article proposes a novel gradient-based global model-agnostic tool that can be calculated efficiently for differentiable deep learning models and produces informative visualizations. This tool studies marginal attribution to feature components at a given response level. Marginal attributions allow us to separate marginal effects of individual feature components from interaction effects, and they allow us to study the resulting variable importance plots in different regions of the decision space, characterized by different response levels. Variable importance is measured w.r.t. a reference point that is calibrated on the entire space for our explanation. Finding a good reference point has been efficiently performed by a simple gradient descent search. A main outcome of our model-agnostic tool is a 3-way relationship between marginal attribution, output level and feature value, which can be illustrated in different ways.

Our method complements commonly used response sensitivity analyses, such as variable permutation importance or accumulated local effects, by an additional marginal attribution view. It should be preferred over these alternative methods, when variable importance varies across response levels and if interactions play an important role in the systematic effects on responses. In particular, in our bike rental example we have been able to explicitly (at low computational cost) extract granular feature information at different response levels also providing insight into systematic effects coming from variable interactions.

Our method is most suitable if the true regression surface is comparably smooth, and if it can be characterized by second order terms. In fact, our method is exact in the quadratic regression case, and in the general case we can explicitly quantify the approximation error due to the conditioning on the quantile level considered. If this approximation error is too big, one can still consider a local description, local in terms of response levels by choosing different reference points for different response intervals, but then one loses the global model-agnostic view, as the decision space becomes partitioned w.r.t. response levels.

Our method is based on aggregating local Taylor expansions, conditioned on a given a response level. A similar concept could also be applied to other local model-agnostic decomposition tools such as SHAP. In this sense, our proposal can be seen as a more general concept that can be applied w.r.t. different local attribution mechanisms.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

A Sensitivities in distortion risk measures

The purpose of this appendix is to briefly explain distortion risk measures and how they relate to marginal attribution. For this discussion we impose stronger assumptions than we have needed above, i.e., these more restrictive assumptions are only made for the explanation here. Assume the expected response $\mu(\boldsymbol{X})$ has a continuous distribution function $F_{\mu(\boldsymbol{X})}$. It follows that $U_{\mu(\boldsymbol{X})} = F_{\mu(\boldsymbol{X})}(\mu(\boldsymbol{X}))$ is uniformly distributed on [0, 1]. Choose a density ζ on [0, 1]. We can interpret $\zeta(U_{\mu(\boldsymbol{X})})$ as a probability distortion (probability re-weighting scheme inducing a change of probability measure) because we have

$$\mathbb{E}_P\left[\zeta(U_{\mu(\boldsymbol{X})})\right] = \int_0^1 \zeta(u)du = 1.$$

The distorted expected response can then be defined by

$$\varrho(\mu(\boldsymbol{X}); \zeta) = \mathbb{E}_P\left[\mu(\boldsymbol{X})\zeta(U_{\mu(\boldsymbol{X})})\right].$$

The functional $\varrho(\mu(\boldsymbol{X}); \zeta)$ describes a *distortion risk measure*, see Wang (1996) and Acerbi (2002). It can be interpreted via a Radon–Nikodým derivative $dP_\zeta(\boldsymbol{X} = \boldsymbol{x}) = \zeta(U_{\mu(\boldsymbol{x})})dP(\boldsymbol{X} = \boldsymbol{x})$. We study the sensitivities of this distortion risk measure w.r.t. the components of $\boldsymbol{X}$. Assume that the following directional derivatives exist in zero for

all $1 \leq j \leq q$

$$S_j(\mu;\zeta) = \frac{\partial}{\partial \varepsilon}\, \varrho\left(\mu\left((X_1,\ldots,X_{j-1},X_j(1+\varepsilon),X_{j+1},\ldots X_q)^\top\right);\zeta\right)\Big|_{\varepsilon=0}.$$

Then, $S_j(\mu;\zeta)$ can be interpreted as the sensitivity of $\boldsymbol{X} \mapsto \mu(\boldsymbol{X})$ in feature component X_j. Hong (2009) and Tsanakas and Millossovich (2016) prove under different sets of assumptions that these sensitivities satisfy

$$S_j(\mu;\zeta) = \mathbb{E}_P\left[X_j \mu_j(\boldsymbol{X}) \zeta(U_{\mu(\boldsymbol{X})})\right].$$

Observe that this exactly uses marginal attribution (2.7). We still have the freedom of choosing the density ζ on $[0,1]$. If we choose the uniform distribution $\zeta \equiv 1$ on $[0,1]$ we receive the average expected response and its average marginal attribution

$$\varrho(\mu(\boldsymbol{X});\zeta\equiv 1) = \mathbb{E}_P[\mu(\boldsymbol{X})] \qquad \text{and} \qquad S_j(\mu;\zeta\equiv 1) = \mathbb{E}_P[X_j\mu_j(\boldsymbol{X})].$$

If we choose for density ζ the Dirac measure δ_α in $\alpha \in (0,1)$, which allocates probability weight 1 to α, this gives us the α-quantile

$$\varrho(\mu(\boldsymbol{X});\zeta=\delta_\alpha) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha).$$

For its sensitivities we receive for $1 \leq j \leq q$

$$S_j(\mu;\zeta=\delta_\alpha) \;=\; \mathbb{E}_P\left[X_j\mu_j(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right],$$

which exactly corresponds to 1st order attribution (3.1). We could choose any other density ζ on $[0,1]$ to obtain sensitivities of other distortion risk measures. Such other choices may also have interesting counterparts in interpreting smooth deep learning models, by directing attention to different areas of the prediction space.

B Further analysis of the real data example

This appendix provides further analysis of the real data example of Sect. 5. First, we verify the robustness of the MACQ approach by fitting multiple networks to the same data. Second, we compare the fully-connected feed-forward neural network regression function given in (5.1) to a generalized linear model (GLM) regression. Finally, we illustrate what we can learn from the MACQ analysis about representation learning in different network layers.

B.1 Robustness of 1st and 2nd order contributions

In Fig. 14 we analyze the robustness of the attribution results. We do this by considering different networks $\boldsymbol{x} \mapsto \theta(\boldsymbol{x})$ for predicting the response variable Y.

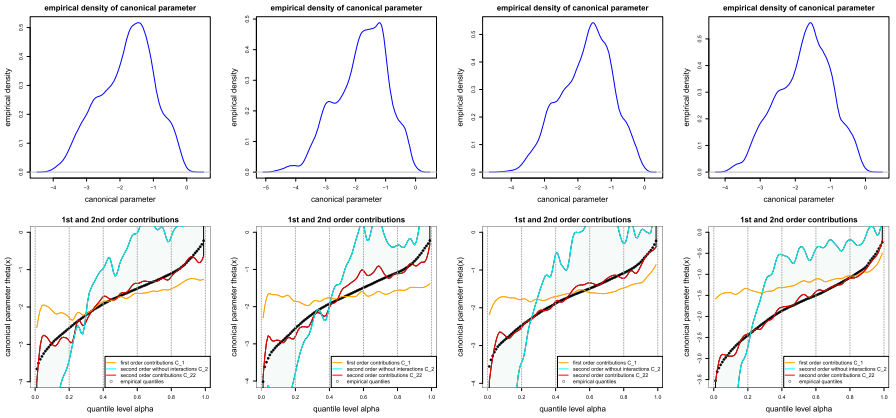

Fig. 14 Robustness of 1st and 2nd order contributions across 4 different networks: (top row) empirical densities of canonical parameters $(\theta(\boldsymbol{x}_i))_{1\le i\le n}$, (bottom row) 1st and 2nd order contributions (5.2)

Network regression models lack a certain degree of robustness as gradient descent network fitting explores different (local) minima of the objective function; note that, in general, neural network fitting is not a convex minimization problem. This issue of non-uniqueness of good predictive models has been widely discussed in the literature, and ensembling may be one mitigation strategy; we refer to Dietterich (2000a, b), Zhou et al. (2002), Zhou (2012), Richman and Wüthrich (2020) and Wüthrich and Merz (2021). The top row of Fig. 14 shows the empirical distributions of the canonical parameter $(\theta(\boldsymbol{x}_i))_{1\le i\le n}$ for 4 different networks; we observe that there are some differences in these empirical densities. The bottom row shows the corresponding 1st and 2nd order contributions (5.2), split by 1st order contributions C_1, 2nd order contributions without interactions C_2 and the full 2nd order contributions $C_{2,2}$. At this level, we judge the attributions made to be rather robust over the different models, as the general shapes of these graphs are similar, and the interaction terms $C_{2,2} - C_2$ show a similar structure and magnitude across the 4 different network models.

From Fig. 14 we also observe that the 1st order contributions C_1 intersect the quantiles $F^{-1}_{\theta(\boldsymbol{X})}(\alpha)$ at different levels for the 4 different calibrations. This indicates that the optimal reference point $\boldsymbol{a}$ is chosen differently in the different networks. Figure 15 shows the chosen reference points $\boldsymbol{a}$ of the 4 different networks in relation to the centered and normalized features $(\boldsymbol{x}_i)_{1\le i\le n}$. Some feature components have a very skewed distribution as can be seen from the thicker horizontal boxplot lines showing the median of each feature component $(x_{i,j})_{1\le i\le n}$, $1 \le j \le q$. The reference point mostly lies within the interquartile range (IQR).

B.2 Generalized linear model

To better understand the individual marginal contribution plots of Fig. 9, we also fit a logistic GLM to the rental bike data. A GLM has a linear regression structure on the canonical scale θ. We fit this GLM to the rental bike data, and in this fitted GLM

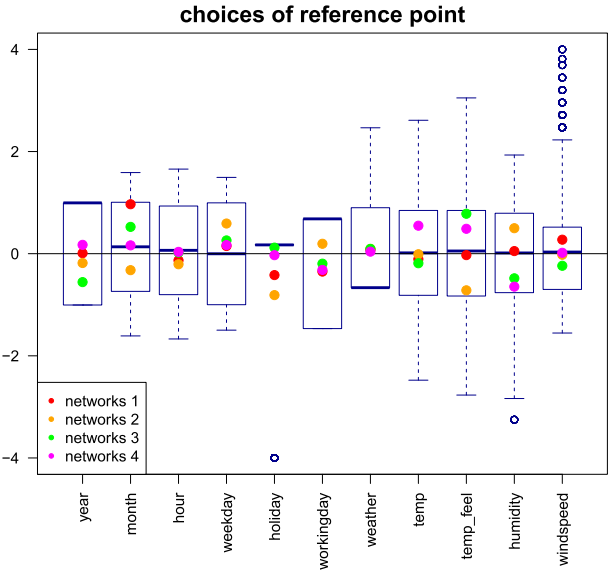

Fig. 15 Choice of reference point $\boldsymbol{a}$ across 4 different networks illustrated for all feature components $1 \leq j \leq q$

the variable `temp_feel` receives the largest regression parameter $\beta_j = 0.3197$. We remark that the biggest regression parameter is not attained by the most important variables, here, because these most important variables enter the regression function non-monotonically (e.g., cyclically in case of `hour`) and a GLM cannot properly cope with such non-monotonicity.

Figure 16 shows the individual marginal contributions $\omega_{i,j}$ in a GLM for variable `temp_feel`. A GLM is linear in the feature values on the canonical scale (for canonical link), and this can clearly be seen from Fig. 16, as the resulting coloring has a (strict) horizontal structure. Moreover, the regression parameter β_j is positive which results in an increasing slope w.r.t. the quantile levels. If we compare Figs. 9 and 16, we conclude that the former plots do not have a strict horizontal structure in the colors which says that none of these variables can be modeled by a GLM term.

B.3 Scrolling through the network layers

Our MACQ proposal is also useful to understand representation learning of neural networks. A deep feed-forward neural network $\theta : \mathbb{R}^q \to \mathbb{R}$ is a composition of d hidden neural network layers $z^{(k)} : \mathbb{R}^{q_{k-1}} \to \mathbb{R}^{q_k}$, $1 \leq k \leq d$; we initialize input dimension $q_0 = q$. Define the composition $\boldsymbol{x} \mapsto z^{(d:1)}(\boldsymbol{x}) = (z^{(d)} \circ \ldots \circ z^{(1)})(\boldsymbol{x})$ which maps input $\boldsymbol{x} \in \mathbb{R}^q$ to the last hidden network layer having dimension q_d. Network (5.1) with logistic output can then be written as

$$\boldsymbol{x} \in \mathbb{R}^q \quad \mapsto \quad \mu(\boldsymbol{x}) = \sigma(\theta(\boldsymbol{x})) = \sigma\left(\beta_0 + \boldsymbol{\beta}^\top z^{(d:1)}(\boldsymbol{x})\right),$$

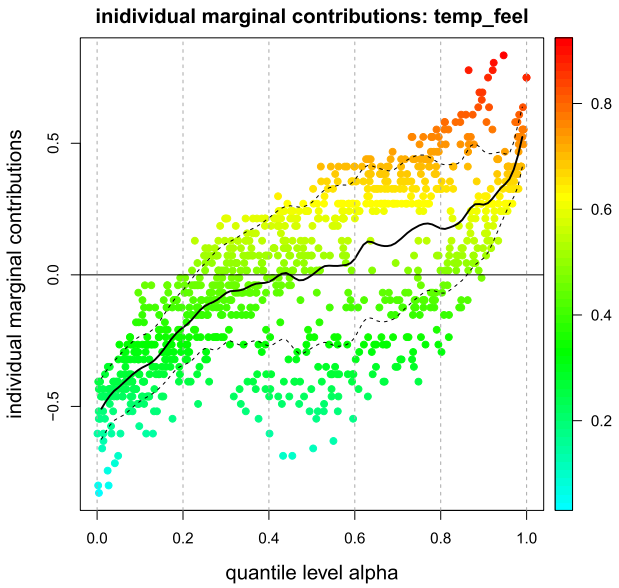

Fig. 16 Individual marginal contributions $\omega_{i,j}$ of 1,000 randomly selected instances $\boldsymbol{x}_i$ in the logistic GLM for $j = \texttt{temp\_feel}$; the black line shows attribution $S_j(\theta;\alpha)$ and the black dotted line gives one standard deviation; colors illustrate the feature values x_j

with bias/intercept $\beta_0 \in \mathbb{R}$ and regression parameter/weight $\boldsymbol{\beta} \in \mathbb{R}^{q_d}$. This should be compared to linear regression (3.4).

Each hidden layer learns a new representation of the inputs $\boldsymbol{x}_i$, that is, the representations learned in layer k are given by $\boldsymbol{x}_i^{(k:1)} := (\boldsymbol{z}^{(k)} \circ \ldots \circ \boldsymbol{z}^{(1)})(\boldsymbol{x}_i)$, for $1 \le i \le n$, we also refer to Section 7.1 in Wüthrich and Merz (2021). We can view these learned representations as new inputs to the remaining network after hidden layer k

$$\boldsymbol{x} \in \mathbb{R}^{q_k} \quad \mapsto \quad \sigma\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{z}^{(d:k+1)}(\boldsymbol{x})\right) = \sigma\left(\beta_0 + \boldsymbol{\beta}^\top (\boldsymbol{z}^{(d)} \circ \ldots \circ \boldsymbol{z}^{(k+1)})(\boldsymbol{x})\right).$$

In the following analysis we consider the instances $(Y_i, \boldsymbol{x}_i^{(k:1)})$ with these learned features $\boldsymbol{x}_i^{(k:1)}$ as inputs to the remaining network $\boldsymbol{z}^{(d:k+1)}$ after layer k, and we perform the same MACQ analysis as above in these reduced setups.

Figure 17 provides the 1st and 2nd order contributions (5.2) of the original inputs (lhs), the learned representations $\boldsymbol{x}_i^{(1:1)}$ in the first hidden layer (middle), and the learned representations $\boldsymbol{x}_i^{(2:1)}$ in the second hidden layer (rhs) on the corresponding remaining networks $\boldsymbol{z}^{(3:k+1)}$. We interpret these MACQ results as follows. The first hidden layer (middle graph) has mainly a smoothing effect in recomposing the inputs $\boldsymbol{x}_i$ suitably. The second layer takes care of the interaction effects diminishing the cyan shaded area in Fig. 17 (rhs). Of course, this makes perfect sense as the output layer considers a linear function with weight $\boldsymbol{\beta} \in \mathbb{R}^{q_d}$ which no longer allows for

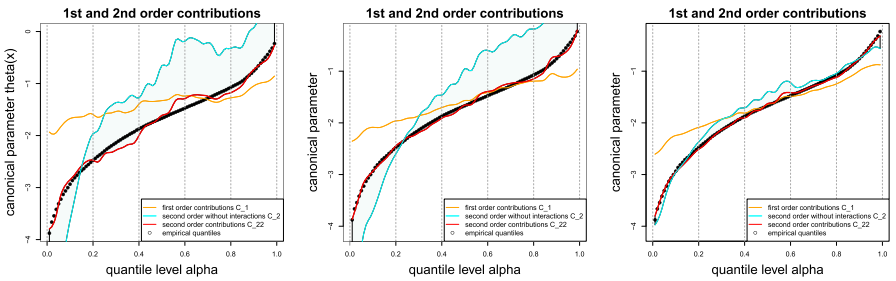

Fig. 17 1st and 2nd order contributions (5.2) of the (learned) representations: (lhs) original inputs $\boldsymbol{x}_i$, (middle) learned representations $\boldsymbol{x}_i^{(1:1)}$, and (rhs) learned representations $\boldsymbol{x}_i^{(2:1)}$

interactions. Therefore, interactions need to be learned in the previous layers. The same applies to non-linear structures (on the canonical scale).

C Descriptive analysis of bike rental example

This appendix gives a brief descriptive analysis of the data, which helps us interpret the network regression models. The data comprises the number of casual and registered bike rentals every hour from 2011/01/01 until 2012/12/31. This data set has originally been studied in Fanaee-T and Gama (2014) and Apley and Zhu (2020), and can be downloaded from https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. Listing 1 gives a short excerpt of the data.

Listing 1 Excerpt of bike rental data.

```
'data.frame':   17379 obs. of  13 variables:
 $ date       : Date, format: "2011-01-01" "2011-01-01" "2011-01-01" ...
 $ year       : num  2011 2011 2011 2011 2011 ...
 $ month      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ hour       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ weekday    : int  6 6 6 6 6 6 6 6 6 6 ...
 $ holiday    : Factor w/ 2 levels "holiday","no-holiday": 2 2 2 2 2 2 2 2 2 2 ...
 $ workingday: Factor w/ 2 levels "no-working","workingday": 1 1 1 1 1 1 1 1 1 1 ...
 $ weather    : num  1 1 1 1 1 2 1 1 1 1 ...
 $ temp       : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
 $ temp_feel  : num  0.288 0.273 0.273 0.288 0.288 ...
 $ humidity   : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
 $ windspeed  : num  0 0 0 0 0 0.0896 0 0 0 0 ...
 $ casual     : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
 $ count      : int  16 40 32 13 1 1 2 3 8 14 ...
```

As response variable we consider the proportion of casual rentals relative to all rentals, thus, we set the response $Y = \texttt{casual/count} \in [0, 1]$ on an hourly grid over the entire observation period. These are $n = 17,379$ hours from 2011/01/01 until 2012/12/31, see line 1 of Listing 1. We note that $\texttt{count} \geq 1$ for all observations, which makes Y well-defined throughout the whole observation period. The goal is to

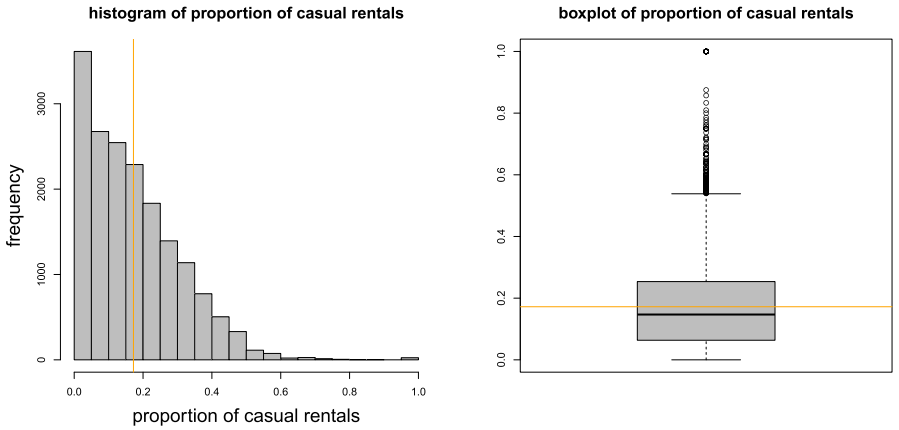

Fig. 18 (lhs) Histogram and (rhs) boxplot of (hourly) responses $Y = \texttt{casual/count} \in [0, 1]$ over the entire observation period; the orange line shows the empirical mean of 17%

predict this response variable Y based on the available feature information $\boldsymbol{x}$, which is provided on lines 3-13 of Listing 1. These are the `year`, `month` and `hour` of the observations Y. The `weekday` (with 0 for Sunday), `holiday` (yes/no for public holiday), `workingday` (yes/no, the former neither being a public holiday nor a weekend), `weather` (1,2 and 3 for clear, cloudy and rain/snow), temperature `temp`, the felt temperature `temp_feel`, `humidity` and `windspeed`. Note that all these features are continuous or binary, thus, we can directly use this feature encoding for regression modeling.

We illustrate this data. Figure 18 shows the observed responses $Y = \texttt{casual/count}$ over the entire observation period. In average the casual rentals make 17% of all rentals, and the empirical density of Y is strongly skewed.

In Fig. 19 we provide the marginal observed responses for each level of all features. The top-left shows the average response for each calendar week from 2011/01/01 until 2012/12/31. This depicts a strong seasonal pattern of the casual rental proportion. Moreover, daytime, weekdays, working days/holidays and weather conditions such as temperature give important information for predicting the proportion of casual rentals. Only wind speed does not seem to be very relevant. From the top-middle we also observe that the proportion of casual rentals slightly decreases over time which can be explained by increasing regular rental subscriptions from 2011 to 2012.

For many of the feature components it is clear that they are highly correlated. In Fig. 20 we plot temperature, humidity and wind speed against calendar month (top row), daytime (middle row) and weather conditions (bottom row). These plots clearly show this dependence. Moreover, humidity is negatively correlated with wind speed and positively correlated with temperature (at least up to moderate temperatures).

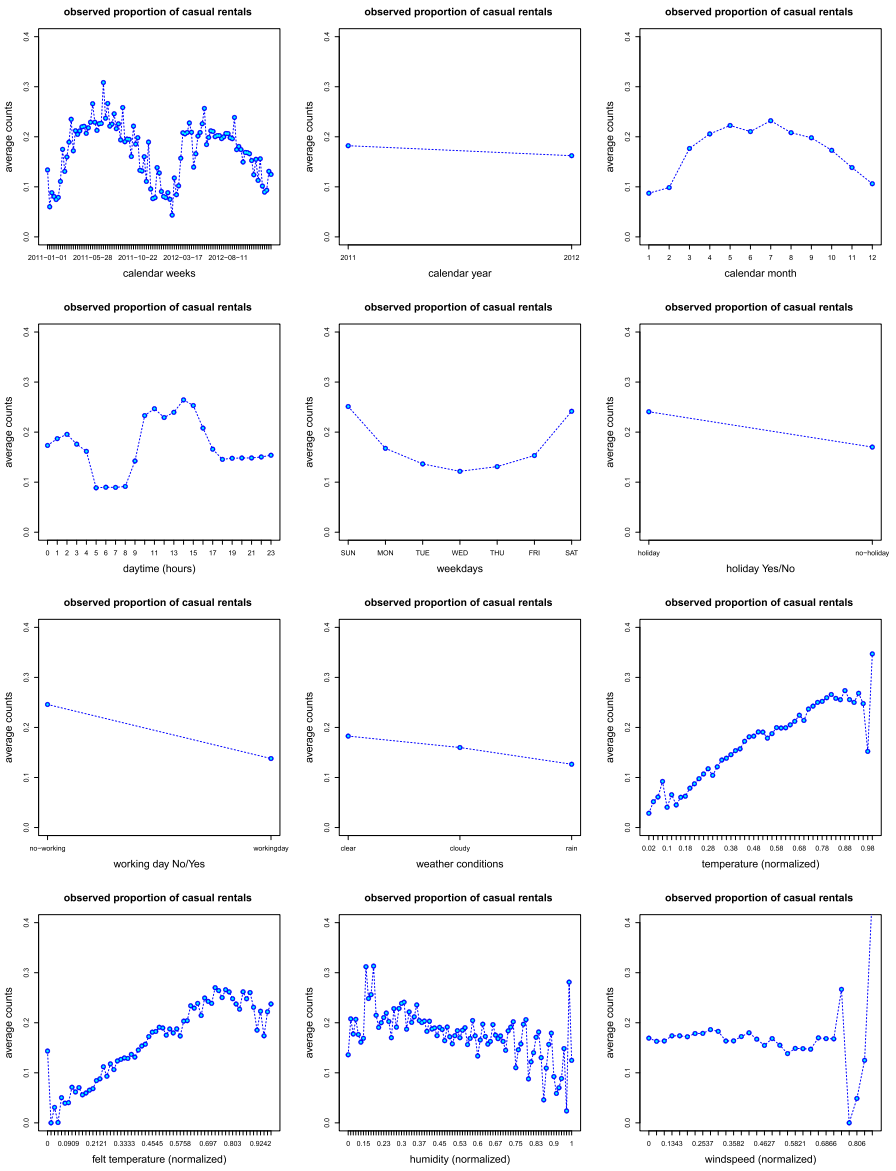

Fig. 19 Average response Y for each level of all features `date` (in weekly units), `year`, `month`, `hour`, `weekday`, `holiday`, `workingday`, `weather`, `temp`, `temp_feel`, `humidity` and `windspeed`

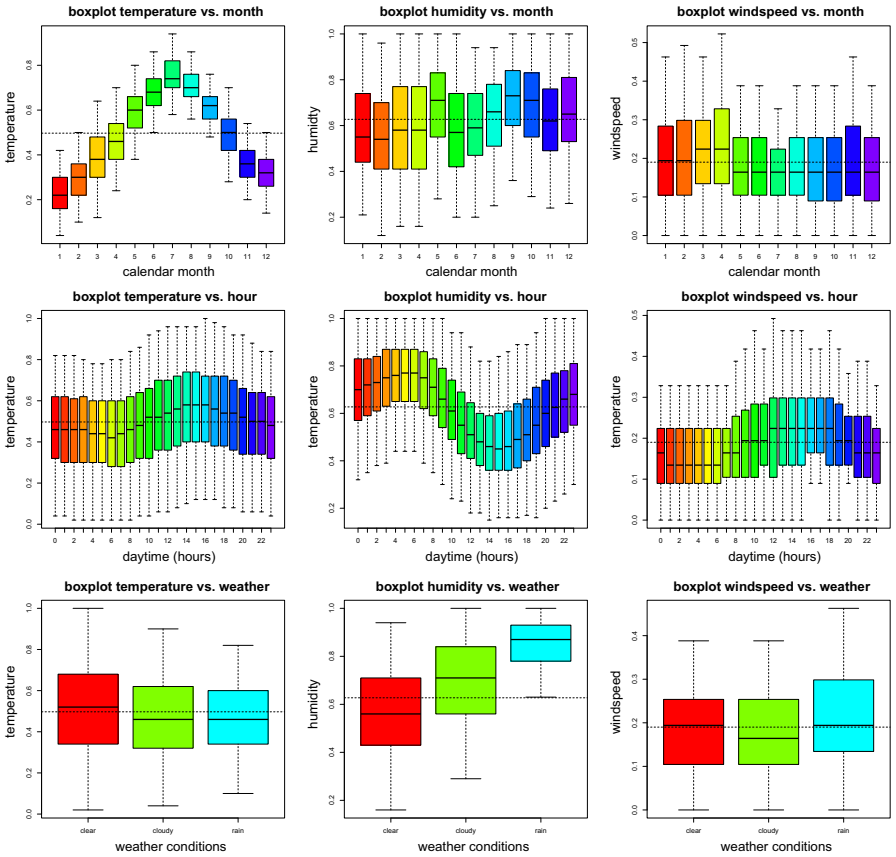

Fig. 20 Dependence between feature components: (top) temperature, humidity and wind speed against calendar month, (middle) temperature, humidity and wind speed against daytime, (bottom) temperature, humidity and wind speed against weather conditions

References

Abadi M et al (2015) TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/

Acerbi C (2002) Spectral measures of risk: a coherent representation of subjective risk aversion. J Bank Finance 7:1505–1518

Ancona M, Ceolini E, Öztireli C, Gross M (2019) Gradient-based attribution methods. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (eds) Explainable AI: interpreting, explaining and visualizing deep learning, lecture notes in artificial intelligence 11700. Springer, pp 168–191

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B 82(4):1059–1086

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Binder A, Bach S, Montavon G, Müller K-R, Samek W (2016) Layer-wise relevance propagation for deep neural network architectures. In: Kim K, Joukov N (eds) Information science and applications (ICISA), lecture notes in electrical engineering 376. Springer

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chollet F et al (2015) Keras. https://github.com/fchollet/keras

Dietterich TG (2000a) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40(2):139–157
Dietterich TG (2000b) Ensemble methods in machine learning. In: Kittel J, Roli F (eds) Multiple classifier systems, lecture notes in computer science, 1857. Springer, pp 1–15
Efron B (2020) Prediction, estimation and attribution. Int Stat Rev 88(S1):S28–S59
Fanaee-T H, Gama J (2014) Event labeling combining ensemble detectors and background knowledge. Prog Artif Intell 2:113–127
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2(3):916–954
Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65
Gourieroux C, Laurent JP, Scaillet O (2000) Sensitivity analysis of values at risk. J Empir Finance 7:225–245
Guo C, Berkhahn F (2016) Entity embeddings of categorical variables. arXiv:1604.06737
Hong LJ (2009) Estimating quantile sensitivities. Oper Res 57(1):118–130
Lindholm M, Richman R, Tsanakas A, Wüthrich MV (2022) Discrimination-free insurance pricing. ASTIN Bull 52(1):55–89
Loader C, Sun J, Technologies Lucent, Liaw A (2020) locfit: local regression, likelihood and density estimation. R package version 1.5-9.4
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30. Curran Associates, Montreal, pp 4765–74
Miller T (2019) Explanation in artificial intelligence: insights form social sciences. Artif Intell 267:1–38
Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 65:211–222
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD'16). Association for Computing Machinery, New York, pp 1135–1144
Richman R, Wüthrich MV (2020) Nagging predictors. Risks 8/3, article 83
Samek W, Müller K-R (2019) Toward explainable artificial intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (eds) Explainable AI: interpreting, explaining and visualizing deep learning, lecture notes in artificial intelligence 11700. Springer, pp 5–23
Shapley LS (1953) A value for n-Person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games (AM-28), vol II. Princeton University Press, Princeton, pp 307–318
Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: learning important features through propagating activation differences. arXiv:1605.01713
Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: Proceedings of the 34th international conference on machine learning, proceedings of machine learning research, PMLR, vol 70. International Convention Centre, Sydney, Australia, pp 3145–3153
Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th international conference on machine learning, proceedings of machine learning research (PMLR), vol 70. International Convention Centre, Sydney, Australia, pp 3319–3328
Tsanakas A, Millossovich P (2016) Sensitivity analysis using risk measures. Risk Anal 36(1):30–48
Wang S (1996) Premium calculation by transforming the layer premium density. ASTIN Bull 26(1):71–92
Wüthrich MV, Merz M (2021) Statistical foundations of actuarial learning and its applications. SSRN Manuscript ID 3822407
Zhao Q, Hastie T (2021) Causal interpretations of black-box models. J Bus Econ Stat 39(1):272–281
Zhou Z-H (2012) Ensemble methods: foundations and algorithms. Chapman & Hall/CRC, London
Zhou Z-H, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1–2):239–263