Fachbereich II – Mathematik - Physik - Chemie

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

01/2020

Ulrike Grömping

Model-Agnostic Effects Plots for Interpreting Machine Learning Models

Modell-agnostische Effektdiagramme zur Interpretation maschinell gelernter Modelle (englischsprachig)

Reports in Mathematics, Physics and Chemistry

Berichte aus der Mathematik, Physik und Chemie

ISSN (print): 2190-3913

Reports in Mathematics, Physics and Chemistry

Berichte aus der Mathematik, Physik und Chemie

The reports are freely available via the Internet:
http://www1.beuth-hochschule.de/FB_II/reports/welcome.htm

01/2020, March 2020

Model-Agnostic Effects Plots for Interpreting Machine Learning Models
Modell-agnostische Effektdiagramme zur Interpretation maschinell gelernter Modelle (englischsprachig)

Editorial notice / Impressum

Published by / Herausgeber:
Fachbereich II
Beuth Hochschule für Technik Berlin
Luxemburger Str. 10
D-13353 Berlin
Internet: http://public.beuth-hochschule.de/FB_II/
E-Mail: fbiireports@beuth-hochschule.de

Responsibility for the content rests with the author(s) of the reports.
Die inhaltliche Verantwortung liegt bei den Autor/inn/en der Berichte.

ISSN (print): 2190-3913

Model-Agnostic Effects Plots for Interpreting Machine Learning Models

Ulrike Grömping

18 March 2020

Abstract

Partial dependence (PD) plots are an established tool for visualizing main effects from general black box models. In recent years they have been supplemented with individual conditional expectation plots (ICE plots). Furthermore, they have been fundamentally criticized as being invalid for correlated features, and average local effects plots (ALE plots) have been proposed as a remedy. This paper discusses the properties of PD plots, ICE plots and ALE plots both in terms of their estimands for linear models with interactions and in terms of their performance for nonparametric models that do not extrapolate well. A stratified PD plot is introduced, which is particularly useful for the visualization of interactions between correlated features. Recommendations for the use of model-agnostic effects plots are given, with special emphasis to nonparametric machine learning models.

1 Introduction

The use of black box machine learning models has generated a demand for tools that aid in achieving at least post-hoc interpretability. "Interpretable Machine Learning" (IML) is en vogue. A recent book by Molnar (2019) provides an overview over the tools that have been proposed for this purpose. Molnar's book focuses on *model-agnostic* methods, i.e. methods that can be applied to any model that can predict an outcome based on a vector of inputs.

This paper investigates model-agnostic effect plots (MAEPs). It focuses on MAEPs for tabular data: partial dependence (PD) plots (Friedman 2001) with individual conditional expectation (ICE) plots (Goldstein, Kapelner, Bleich and Pitkin 2015; these are sometimes also called Ceteris Paribus Profiles or What-If plots, see Biecek 2018) and average local effects (ALE) plots (Apley and Zhu 2019, initially proposed in earlier version of that report by Apley alone). Although they have only been published in a technical report, ALE plots are included, because they have been positively received in applied literature and software. For example, Molnar (2019) states in his Chapter 5.3.5: "All in all, in most situations I would prefer ALE plots over PDPs [PD plots, UG], because features are usually correlated to some extent." At this point, it should be mentioned that the machine learning community denotes as "features" what the statistical community calls "regressors" or "explanatory variables". This term will also be used in this paper.

There is a substantial body of literature on effect estimation and effects plots in linear models, generalized linear models or yet more general but still at least semi-parametric models. Lenth (2016) or Fox and Weisberg (2018) present modern implementations and overviews of historical developments. The MAEPs discussed in this paper cannot draw on assumptions regarding model structure, which implies that the established methods are not directly applicable. Besides exploring the properties of main effect MAEPs, this paper will also exploit the findings for proposing an interaction MAEP, called stratified PD plot, that is similar in spirit to the usual interaction plots for parametric models: it is based on stratified main effects ICE plots and PD plots, and it can be approximated by using ALE plots for resampled data, if computing resources are an issue. The proposed stratified PD plot is not only suitable as an interaction plot, but can also reduce distortions in the PD main effect plot, if the feature under investigation is strongly correlated to other features.

Why are PD plots criticized in case of correlated features? Calculation of a PD plot requires predictions from the Cartesian product of the feature values of the feature of interest with all feature value combinations from other features. Correlated features imply that parts of the resulting combinations belong to empty or very sparsely populated parts of the feature space. Molnar emphasizes the inadequacy of obtaining predictions for unlikely or even impossible feature combinations, like a person of height 2m with weight less than 50kg. Apley and Zhu (2019) voice a related (and from this author's point of view, more important) issue that may also occur for perfectly meaningful feature combinations: a nonparametric model that has been trained on heavily correlated data will have difficulties providing valid predictions for parts of the feature space that did not occur in the training data, i.e., it does not extrapolate well; most machine learning models are nonparametric and are thus affected by this weakness. (Semi-)Parametric models, on the other hand, can also draw on model structure and can thus extrapolate much more safely; nevertheless, as is e.g. known for linear models with multicollinear features, they can also be unable to distinguish between competing prediction models that may lead to quite different predictions in (almost) empty parts of the feature space. PD plots are averages; they may include some very poor predictions in such problematic situations. This and further aspects of method comparison will be discussed in more depth in this paper. The goal is to develop an understanding of the meanings of the different MAEPs in such situations, and to disentangle different aspects of usability for the MAEPs that have been somewhat mixed up in the discussion.

This paper discusses the MAEPs, using their estimands for the simplest possible interesting regression model, a normal linear model with main effects and interaction effect of two quantitative regressors, as a vehicle for demonstrating the conceptual properties of different MAEPs, when applied to models that extrapolate well (e.g. correctly specified "true" parametric models). With simulated data from the same simple model, it is also inspected how the MAEPs behave for nonparametric models that fit the true model well only locally – random forest models are used for that purpose. Attention is initially restricted to main effects plots, in full awareness that these are not the recommended type of visualization in the presence of an interaction. This approach has been chosen, because main effects plots are usually the first step of a more detailed analysis, and their results must therefore be sensible even in the presence of interactions; furthermore, main effects plots are the simplest cases for demonstrating generally valid properties of the MAEPs. The advantages and disadvantages of the different MAEPs will be conceptually discussed, and recommendations for adequate effect visualization will be given. This includes a new proposal for visualizing interactions between heavily correlated features using the afore-mentioned stratified PD curves, possibly accompanied by main effects ICE curves.

Section 2 presents notation and the linear model with interaction that will be used for inspecting the MAEPs. Section 3 presents the different MAEPs and provides their estimands both for the correctly-specified and a misspecified linear model; the section concludes with conceptual comparisons for the linear model case. Section 4 inspects MAEP performance for nonparametric (random forest) prediction models and introduces the stratified PD plot. Section 5 exemplifies the MAEPs for a data set on fuel consumption of different car models, and illustrates the findings of Sections 3 and 4. Section 6 discusses the roles of training data and MAEP generating data, which need not necessarily be the same; this section also introduces the afore-mentioned resampling-based ALE curve approximation for – possibly stratified – PD curves. The final discussion summarizes the findings and gives recommendations for use of MAEPs, especially with respect to nonparametric prediction models.

2 Notation, and the model for generating training data

We consider a true but unknown property of a phenomenon of interest, for example the expected value $f(X_1, \ldots, X_p) = E(Y|X_1, \ldots, X_p)$ of a response Y given the features $X_j, j = 1 \ldots p$. Furthermore, for a selected feature X_s of interest with the other $p-1$ features collected in $X_{C^{-s}}$, let $\hat{f}(X_s, X_{C^{-s}})$ denote the corresponding estimated prediction function. For simplicity, this paper focuses on a single quantitative feature X_s; general considerations remain valid without that constraint.

MAEPs will be inspected on data that have been generated from a linear model with two normally-distributed regressors and their interaction, i.e.

$$f(X_1, X_2) = E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2, \tag{1}$$

where the feature distribution is

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right) =: N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \tag{2}$$

and there is an independent normal additive error term so that $Y|X_1 = x_1, X_2 = x_2 \sim N(f(x_1, x_2), \sigma_\epsilon^2)$. The covariance element $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}\sigma_{22}}$ in (2) makes the problem easy or difficult: if $\sigma_{12} = 0$, the two regressors vary independently, and different methods lead to comparable results. The larger the correlation ρ_{12}, the larger the parts of the feature space for which there are no or very few data.

Throughout the paper, Model (1) with feature distribution (2) will be exemplified with parameters (3):

$$\begin{aligned} &\beta_0 = 200, \quad \beta_1 = 1, \quad \beta_2 = -0.5, \quad \beta_3 = 0.1 \\ &\mu_1 = 10, \quad \mu_2 = 50, \quad \sigma_{11} = 4, \quad \sigma_{22} = 25, \quad \rho_{12} \in \{-0.9, 0, 0.9\} \\ &\sigma_\epsilon^2 = 4. \end{aligned} \tag{3}$$

The true model $f(x_1, x_2)$ can be completely visualized: Figure 1 shows $f(x_1, x_2)$ with overlaid contours of the feature distributions for the three different correlations. Clearly, the information in the corners of the feature space is poorer than in the center; for the two correlated cases, there is a substantial lack of information on the relation between the response and the features in two diagonally opposite corners of the feature space. This will affect prediction quality of nonparametric prediction models (see Section 4).

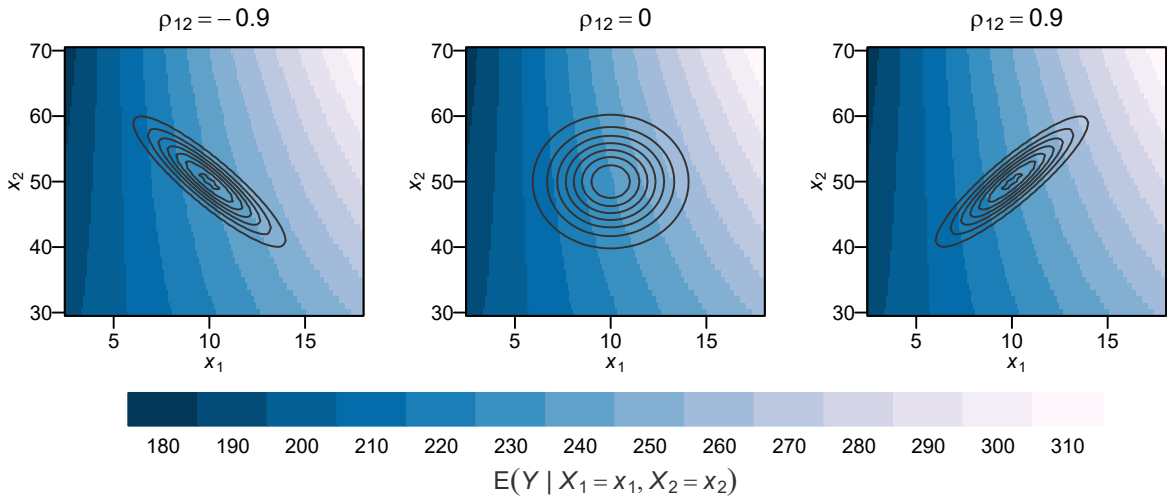

Figure 1: The true model (1) for $E(Y|X_1, X_2)$ with overplotted contours of the bivariate feature distribution (2) using parameters (3)

When fitting a linear model *without interaction* (i.e. with $\widetilde{\beta}_3 = 0$) to data from Model (1) with features from (2) and parameters (3), the "true" slopes for that misspecified model will be

$$\widetilde{\beta}_1 = \beta_1 + \beta_3\mu_2 = 6 \quad \text{and} \quad \widetilde{\beta}_2 = \beta_2 + \beta_3\mu_1 = 0.5, \tag{4}$$

with a correlation-dependent constant and an inflated error variance (see the appendix for the derivation of $\widetilde{\beta}_2$).

Using simulation parameters from (3), 2000 units each have been simulated for the true model depicted in Figure 1, by obtaining 2000 independent samples of the bivariate normal feature distribution (2) and obtaining realizations of Y by adding independent normal random errors with variance $\sigma_\epsilon^2 = 4$ to formula (1).

Figure 2 depicts a classical main effects plot for x_2 (averaged over x_1) for correctly-specified linear models in each of the three simulated data sets. Note that the plots have very narrow confidence bands, because

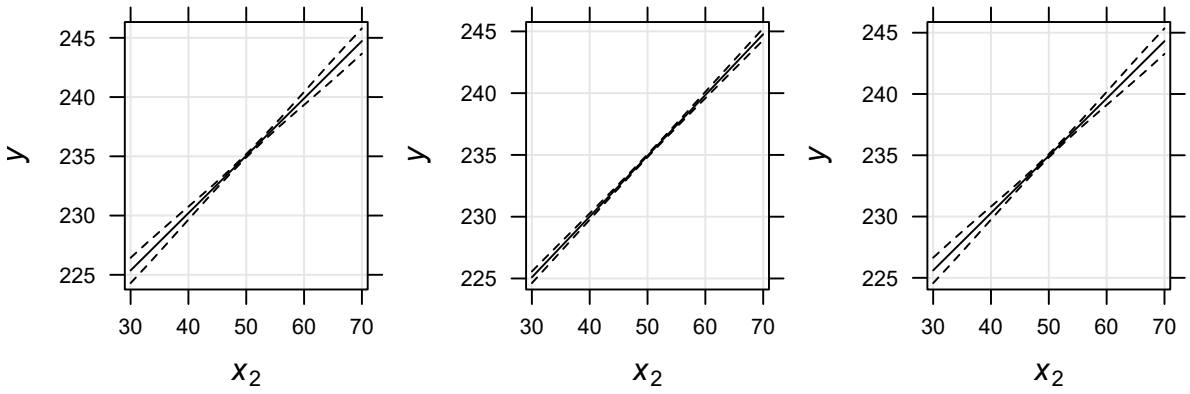

Figure 2: Classical main effects plot for x_2 for Model (1) with parameters from (2) (from left to right: $\rho_{12} = -0.9$, $\rho_{12} = 0$, $\rho_{12} = 0.9$). Dashed lines: Point wise 99 percent confidence bands. The figure has been produced with R package **effects** by Fox and Weisberg (2018).

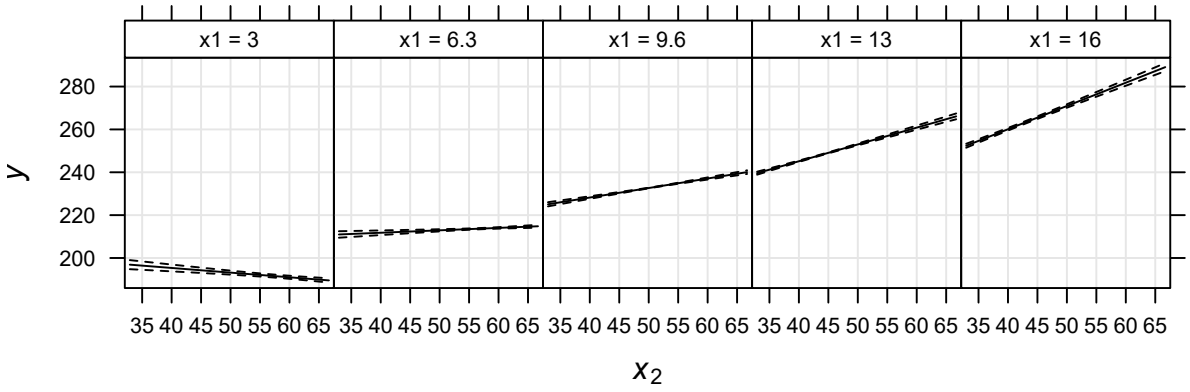

Figure 3: Classical predictor effects plot for x_2 for Model (1) with parameters from (2) ($\rho_{12} = -0.9$). Dashed lines: Point wise 99 percent confidence bands. The figure has been produced with R package **effects** by Fox and Weisberg (2018).

of the small error variance. Furthermore, note that the slopes for x_2 in Figure 2 estimate the slope $\widetilde{\beta}_2$ from (4). In the presence of interactions, it is well-known that main effects plots like those from Figure 2 yield an incomplete picture. Figure 3 provides what Fox and Weisberg (2018) call a predictor effects plot (shown for $\rho_{12} = -0.9$ only, the other two look similar). It gives a more realistic picture of the effect of x_2 on y for the left-hand side plot of Figure 2.

3 Model-agnostic effects plots and their estimands

This section explains each MAEP and inspects its main effect estimands for the correctly-specified linear model (1) and for the misspecified linear model that omits the interaction effect ($\widetilde{\beta}_3 = 0$ enforced by not including the interaction term, see (4)). All MAEPs are discussed in terms of their estimands for the feature X_2 (the weaker of the two features). The MAEPs are calculated and visualized based on the simulated training data. Section 6 will discuss the implications of using separate MAEP generating data instead of the training data.

The most important message from a main effect MAEP lies in the changes between different values of x_s; the overall level (i.e. the average response) is of less interest. ALE plots do not even provide a meaningful level. Therefore, it is not surprising that different software tools choose different overall levels for the MAEPs (approximately zero for both PD plot and ALE plot in package **ALEPlot** by Apley, level for PD curve taken as the average of the ICE curves in package **ICEbox** by Goldstein et al.; the latter coincides with the parametric main effects plot obtained from R package **effects** by Fox and Weisberg 2018). For supporting comparison of different curves in the same figure, this paper shifts the overall average of all MAEPs within the same figure either to the average response (which is very close to the average prediction for all models), or to the PD curve average, whenever ICE curves are involved.

3.1 The most naïve analysis approach, and M plots

For initial inspection, many researchers look at simple scatter plots of Y versus each individual feature. If a curve is fitted to those scatter plots, it models $E(Y|X_1 = x_1)$ for the y against x_1 plot and $E(Y|X_2 = x_2)$ for the y against x_2 plot, respectively. For calculating these conditional expectations, the conditional expectations of X_1 given $X_2 = x_2$ and vice versa are needed:

$$E(X_1|X_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \quad E(X_2|X_1 = x_1) \quad = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1 - \mu_1). \tag{5}$$

These formulae will also be important for obtaining the estimands of ALE plots.

Combining Equations (1), (2) and (5), the conditional expectation of Y given X_2 can be obtained as

$$\begin{aligned} M(x_2) = E(Y|X_2 = x_2) =& \beta_0 + \beta_1\mu_1 - \beta_1\frac{\sigma_{12}}{\sigma_{22}}\mu_2 \\ &+ (\beta_2 + \beta_1\frac{\sigma_{12}}{\sigma_{22}} + \beta_3\mu_1 - \beta_3\frac{\sigma_{12}}{\sigma_{22}}\mu_2)x_2 \\ &+ \beta_3\frac{\sigma_{12}}{\sigma_{22}}{x_2}^2. \end{aligned} \tag{6}$$

Figure 4 shows scatter plots of y against x_2 from the three simulated data sets for Model (1), with $M(x_2)$ added as line.

Plots like those in Figure 4, showing a curve based on the scatter of model predictions $\hat{f}(x_{i;s}, x_{i,C^{-s}})$ (instead of the y_i shown in Figure 4) versus $x_{i;s}$, have been called M plots by Apley and Zhu (2019). If $\hat{f}$ provides a consistent estimate for $E(Y|X_1, X_2)$ in Model (1), Equation (6) gives the estimand for the M plot for x_2. For the misspecified linear model without the interaction, the M plot estimand can also be obtained from (6), by using the modified slopes from (4) with $\widetilde{\beta}_3 = 0$ and an appropriately modified $\widetilde{\beta}_0$.

Friedman (2001) already criticized M plots (his formula (56)), because the effect of correlated features from $X_{C^{-s}}$ is reflected in $M(x_s)$: the summand $\beta_1\sigma_{12}/\sigma_{22}x_2$ in (6) indicates that the M plot slope depends on the slope of a correlated regressor, even if the model does not contain an interaction. This

author is not aware of any scientific recommendation in favor of the use of M plots. They are included here (and in Section 3.4) as a reference for an unsuitable tool. Nevertheless, it is of course adequate to inspect bivariate scatter plots with fitted lines, as long as one is aware of their limitations.

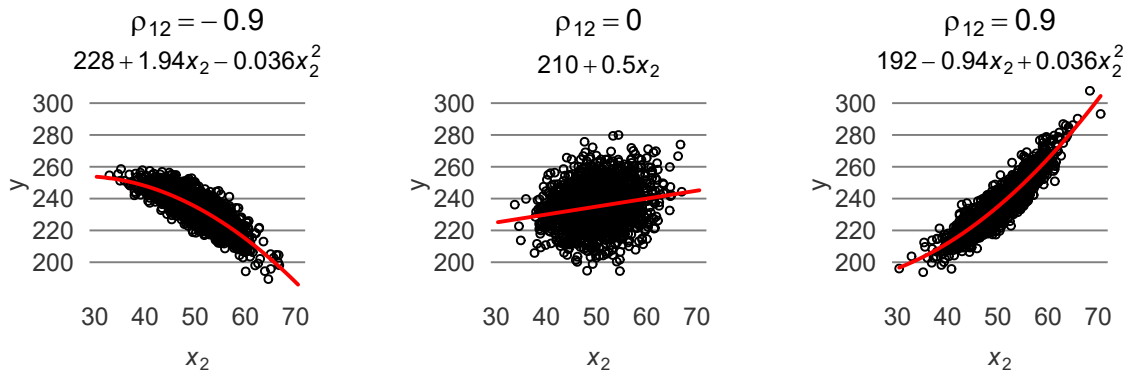

Figure 4: Model (1): Scatter plots of y_i versus x_{i2} with M plot estimands (line)

3.2 PD plots and ICE curves

PD plots were introduced in (2001) by Friedman, ICE curves add further detail to PD plots and were introduced by Goldstein et al. (2015). As PD plots can be derived as averages over ICE curves, ICE curves are discussed first. With the prediction function $\hat{f}$, the ICE curve for a selected feature X_s on a particular unit i is given as

$$ICE_i(x_s|X_{C^{-s}} = x_{i;C^{-s}}) = \hat{f}(x_s, x_{i;C^{-s}}), \tag{7}$$

where $x_{i;C^{-s}}$ indicates the realization of the remaining features for the ith unit. For Model (1), the expected individual ICE curves for x_2 are

$$ICE_i(x_2|X_1 = x_{i1}) = (\beta_0 + \beta_1 x_{i1}) + (\beta_2 + \beta_3 x_{i1})x_2, \tag{8}$$

and the actual curves use estimated coefficients $\hat{\beta}_j$ instead of unknown true ones. Note that ICE curves simply vary a specific feature (or in general a selection of two or even more features) over its entire range, fixing all other features at their actual value. As was mentioned in the introduction, this might create impossible combinations, like a 6-year-old child of height 193cm, which must of course be regarded with suspicion and is exactly Molnar's criticism of using PD curves in case of correlated features. This aspect is related to the extrapolation problem that was discussed in the introduction and will be further inspected in Section 4.3.

A PD plot is simply the average of all ICE curves. Its estimand is thus the expectation over $X_{C^{-s}}$ in formula (7) in general or (8) for Model (1), where

$$PD(x_2) = (\beta_0 + \beta_1\mu_1) + (\beta_2 + \beta_3\mu_1)x_2. \tag{9}$$

This estimand does not at all depend on the correlation between features. For the parameters of (3), (9) becomes $PD(x_2) = 210 + 0.5x_2$. Note that the slope coincides with that of the misspecified linear model ($\widetilde{\beta}_2$ of (4)), and the PD curve estimand is identical for the correct and the misspecified model.

Figure 5 shows PD curves with ICE curves for the correct and misspecified linear models on the simulated data with $\rho_{12} = -0.9$. ALE curves (see next section) are also included. Both PD curves and ALE curves coincide almost perfectly with their estimands.

3.3 ALE plots

Apley and Zhu (2019) introduced ALE plots, with the goal to avoid the perceived disadvantages of both PD plots and M plots: the idea is to work with conditioning on $X_s = x_s$ – like for M plots – but for the

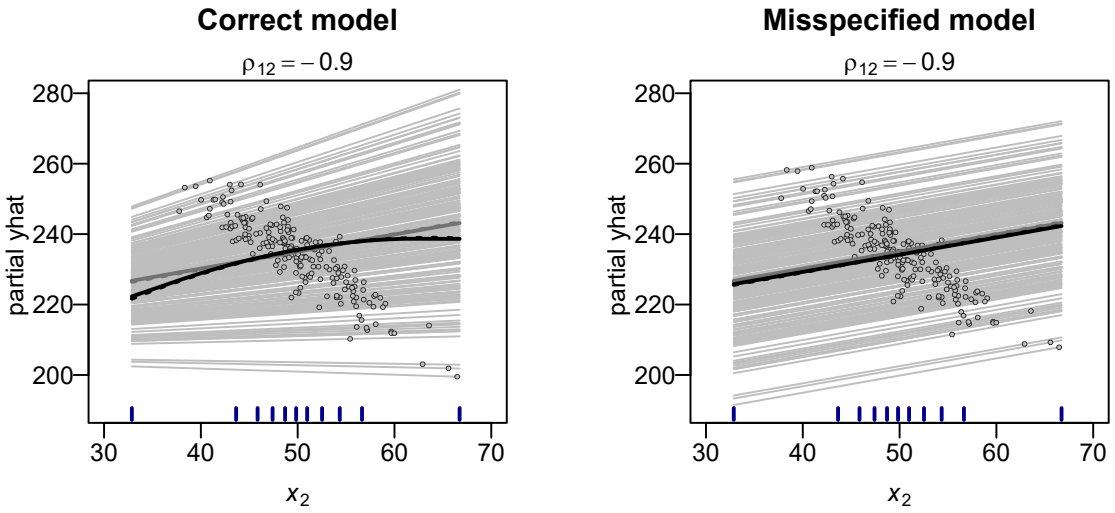

Figure 5: PD plots with random selection of ICE curves for the correct and misspecified linear models (data for $\rho = -0.9$). Grey: PD. Black: ALE. Solid: estimated curve, Dotted: estimand. Points: $(x_{i2}, \hat{f}(x_{i1}, x_{i2}))$ for the sampled ICE curves.

derivative of $\hat{f}$ instead of $\hat{f}$ itself. By subsequently integrating over the local changes measured by the derivative, one gets a change path over the entire range of x_s values, without having to obtain predictions from impossible combinations.

In formula terms, the derivative of the ALE function (with respect to its argument x_s) is given as

$$ALE'(x_s) = E_{X_{C^{-s}}|X_s=x_s}\left(\frac{\partial \hat{f}(X_s, X_{C^{-s}})}{\partial X_s}\middle| X_s = x_s\right). \tag{10}$$

This is then integrated (or, in estimation version, summed from left to right) for obtaining the ALE function itself. Note that this implies a limitation of ALE plots for nominal predictor variables: Categories must be ordered, in order to make "from left to right" meaningful.

For Model (1), the derivative with respect to $X_s = X_2$ is $\beta_2 + \beta_3 X_1$. The conditional expectation of this derivative given $X_s = X_2 = x_2 = x_s$ can be obtained in a straightforward way using Equation (5):

$$ALE'(x_2) = \beta_2 + \beta_3\mu_1 - \frac{\beta_3\sigma_{12}}{\sigma_{22}}\mu_2 + \frac{\beta_3\sigma_{12}}{\sigma_{22}}x_2. \tag{11}$$

Integrating this function yields the ALE estimand for x_2 in Model (1) as

$$ALE(x_2) = \text{const} + \left(\beta_2 + \beta_3\left(\mu_1 - \frac{\sigma_{12}}{\sigma_{22}}\mu_2\right)\right)x_2 + \beta_3\frac{\sigma_{12}}{2\sigma_{22}}{x_2}^2. \tag{12}$$

For the parameter settings (3), this yields the estimands $ALE(x_2) = \text{const} + 2.3x_2 - 0.018{x_2}^2$ for $\rho_{12} = -0.9$, $ALE(x_2) = \text{const} + 0.5x_2$ for $\rho_{12} = 0$ or $ALE(x_2) = \text{const} - 1.3x_2 + 0.018{x_2}^2$ for $\rho_{12} = 0.9$. An ALE plot for the correct linear model estimates these, as can be seen from Figure 5 for $\rho_{12} = -0.9$. The estimand for an ALE plot of the misspecified linear model can be obtained from (12) by replacing β_2 with $\widetilde{\beta}_2$ from (4) and β_3 with zero, i.e. the constant in the ALE estimand for the misspecified linear model can be chosen such that ALE and PD estimand coincide.

Equation (12) gives the estimand for the ALE curve, which has been obtained by integrating an expected derivative. Apley and Zhu (2019, and Apley's R package **ALEPlot**) estimate an ALE curve by

1. subdividing the range of X_s into intervals driven by percentiles,

2. for an interval $\mathcal{I} =]x_{lower}, x_{upper}]$, taking differences $\hat{f}(x_{upper}, x_{iC^{-s}}) - \hat{f}(x_{lower}, x_{iC^{-s}})$ for all units with $x_{is} \in \mathcal{I}$, and averaging these (averaging differences over the interval corresponds to estimating the slope within the interval as the average difference quotient),
3. obtaining provisional ALE curve values at interval borders as sums of all differences from left to right (starting with an arbitrary value at the minimum x_s value)
4. centering the provisional ALE curve values for obtaining a final version.

Apley and Zhu (2019, Chapter 3) describe the process more formally. With this process in mind, Figure 5 illustrates the reason behind the curvature of the ALE plot in the model with interaction: the local slopes for units with smaller x_2 values are steeper than those for units with larger x_2 values; as the ALE plot integrates (sums from left to right) over those local slopes, the outcome is a curve. It will be discussed in Section 3.4, whether or not this curvature is desirable behavior for a main effect MAEP.

3.4 Summary of conceptual comparisons

The discussion refers to equations (6) for the M plot estimand (as a reference for an unsuitable tool), (9) for the PD plot estimand and (12) for the ALE plot estimand.

Correlation with unplotted features that do not interact with plotted features: For M plots, the estimand heavily reflects influences from correlated features (see Equation (6)), and only some of these disappear in the absence of interactions ($\beta_3 = 0$). This is the reason, why M plots are universally rejected as tools for effect assessment (by Friedman 2001, formula (56), not yet called M plots; and also by Apley and Zhu 2019 or Molnar 2019). For both PD plots and ALE plots, non-interacting uncorrelated features do not affect the estimand: Equations (9) and (12) yield slope β_2 in the absence of the interaction ($\beta_3 = 0$).

Interactions with unplotted uncorrelated features: PD plots and ALE plots (and even M plots) show the same reasonable behavior, as can be seen by eliminating all summands involving σ_{12} from Equations (6), (12) and (9): they average over the uncorrelated interacting variable. As a result, the main effect MAEPs are linear with modified slope: β_2 is modified into $\widetilde{\beta}_2 = \beta_2 + \beta_3\mu_1$, i.e. exactly into the respective slope from a misspecified linear model without the intercept (see (4)).

Interactions with unplotted correlated features: PD plot estimands behave exactly like for interactions with unplotted *uncorrelated* features: the PD plot estimand does not contain any expression that depends on the correlation. Interaction with an unplotted correlated feature affects M plots similarly to ALE plots (in addition to the M plots' aforementioned inadequate behavior for unplotted correlated features in general that is unrelated to interactions): both receive an identical contribution $-\beta_3\sigma_{12}x_2/\sigma_{22}$ for the linear part, and both receive a quadratic contribution, with that for ALE being half that for M ($\beta_3\sigma_{12}{x_2}^2/(2\sigma_{22})$ for ALE). The ALE plot main effect estimand thus contains correlation-driven portions of the interaction term, which is quite counter-intuitive and in the same league as the heavily criticized M plot behavior. Apley and Zhu (2019) already note this behavior and defend it, stating that one should not expect interaction with an unplotted *correlated* feature to stay out of the main effect. This author considers this reasoning flawed; it could be applied with the same right to M plots and their way of incorporating main effect contributions of an unplotted *correlated* feature. Thus, ALE plots have the same conceptual problems for the estimation of local changes that M plots have for the estimation of the function itself. Hence, whenever both correlations and interactions must be considered somewhat important, ALE plots are conceptually flawed.

Apley and Zhu (2019) and Molnar (2019) emphasize that ALE plots are "unbiased". Only Apley and Zhu (2019) detail what is meant by this claim: For additive $\hat{f}$ regardless of correlation, and for multiplicative $\hat{f}$ with uncorrelated features, the decomposition of $\hat{f}$ into contributions from individual features is correctly recovered. Apley references Hastie, Tibshirani and Friedman (2009, chapter 10.13), according to whom this property holds even more generally for PD curves, which have the corresponding unbiasedness property even for purely multiplicative $\hat{f}$ regardless of correlation. This result is responsible for main effects PD plots acting reasonably on a linear model $\hat{f}$ that includes an interaction with a correlated feature (while the ALE plot estimand contains an inadequate quadratic portion that arises from the interaction term). The unbiasedness claim is correct for a purely additive or a purely multiplicative *estimated* $\hat{f}$.

Because of the above-mentioned conceptual flaws of M plots and ALE plots, PD plots are the only *conceptually* convincing MAEP. Remember that all considerations in this section worked under data

generated from a linear model that were also modeled as such. If a nonparametric model is used for which predictions outside the region of high feature density do not follow the underlying model well, MAEPs may behave quite differently. In particular, the above-mentioned unbiasedness claims may go wrong, even if they would apply for the true underlying model. The next section will study this case.

4 MAEP performance for nonparametric prediction models

MAEPs – as opposed to classical effects plots – are applied mainly for nonparametric prediction models. This section will inspect their behavior for two random forest models in comparison to the parametric models that were already considered. In the parametric models, the empirical MAEPs almost perfectly coincided with their estimands (formulae (6), (9) and (12)), because the estimated prediction function $\hat{f}$ was very close to the true f. For the nonparametric models of this section, differences between the actual MAEPs for the simulated data and the reference estimands arise from deviations between the nonparametric $\hat{f}$ and the true f. As the paper deals with describing the fitted prediction models and not with assessing prediction accuracy, and as there has been now extensive model building, it is sufficient to discuss goodness of fit in terms of the R^2 values obtained from the training data.

4.1 The four prediction models

The following prediction models are considered:

1. The correctly specified linear model (Equation (1)) with coefficient estimands as given in (3) yields the prediction function $\hat{f}_{\text{lin}}$. For the simulated data, the model has R^2 values 96.2% for $\rho_{12} = -0.9$, 97.5% for $\rho_{12} = 0$ and 98.2% for $\rho_{12} = 0.9$.
2. A linear model without interaction effect yields the prediction function $\hat{f}_{\text{lin;misspec}}$. For the simulated data, the model has R^2 values 94.4% for $\rho_{12} = -0.9$, 96.9% for $\rho_{12} = 0$ and 97.2% for $\rho_{12} = 0.9$. Although the interaction is significant in the correctly specified model, its omission only slightly reduces explained variance, because modification of main effects parameters (see Formula (4) for coefficient estimands) can largely compensate for the missing interaction term.
3. A random forest with default $m_{\text{try}} = 1$ (1000 trees) yields the prediction function $\hat{f}_{\text{rf1}}$. For the simulated data, this forest has R^2 values 95.3% for $\rho_{12} = -0.9$, 96.7% for $\rho_{12} = 0$ and 97.7% for $\rho_{12} = 0.9$. For $m_{\text{try}} = 1$, individual trees in a forest cannot model interaction, which is why one can expect the MAEPs from such a forest to be closer to those of the misspecified linear model than to those of the correctly specified one.
4. A random forest with $m_{\text{try}} = 2$ (i.e. no random feature selection) yields the prediction function $\hat{f}_{\text{rf2}}$. For the simulated data, this forest has R^2 values 95.3% for $\rho_{12} = -0.9$, 96.8% for $\rho_{12} = 0$ and 97.7% for $\rho_{12} = 0.9$, i.e., R^2 values are almost identical to those of $\hat{f}_{\text{rf1}}$.

Not surprisingly, the R^2 values from the correctly-specified linear model are better than those from the other three models, but the difference is not very large.

Figure 6 shows the contours of the predictions overlaid with the contours of the empirical feature distributions. We see that the correctly-specified linear model, by means of its structure, recovers the true model of Figure 1 well in the entire feature space, regardless of feature correlation. The misspecified linear model yields visually relatively similar predictions, but does of course not capture the interaction effect. The random forests appear to handle densely populated areas better than the empty corners; for uncorrelated features, the corners and margins are not fitted as well as the center areas of the plot, but the functional structure is recovered reasonably well. For the correlated features, the forest with $m_{\text{try}} = 1$ tends to create discrimination along the principal axis of the elliptical bivariate density, i.e. the direction with most information gets split into different predictions. For the positively correlated features, because of the nature of the target function, this still works OK. For the negatively correlated features, however, the prediction function is very poor, since the most informative direction of the data is very different from the most interesting direction of the underlying model. The forest with $m_{\text{try}} = 2$ captures the underlying data-generating model much better, but again misses its details in the empty corners of the feature space.

The substantial difference between prediction models in terms of their ability to extrapolate to unusual (x_1, x_2) pairs is not reflected by the reported R^2 values, since these are taken from the actual data values

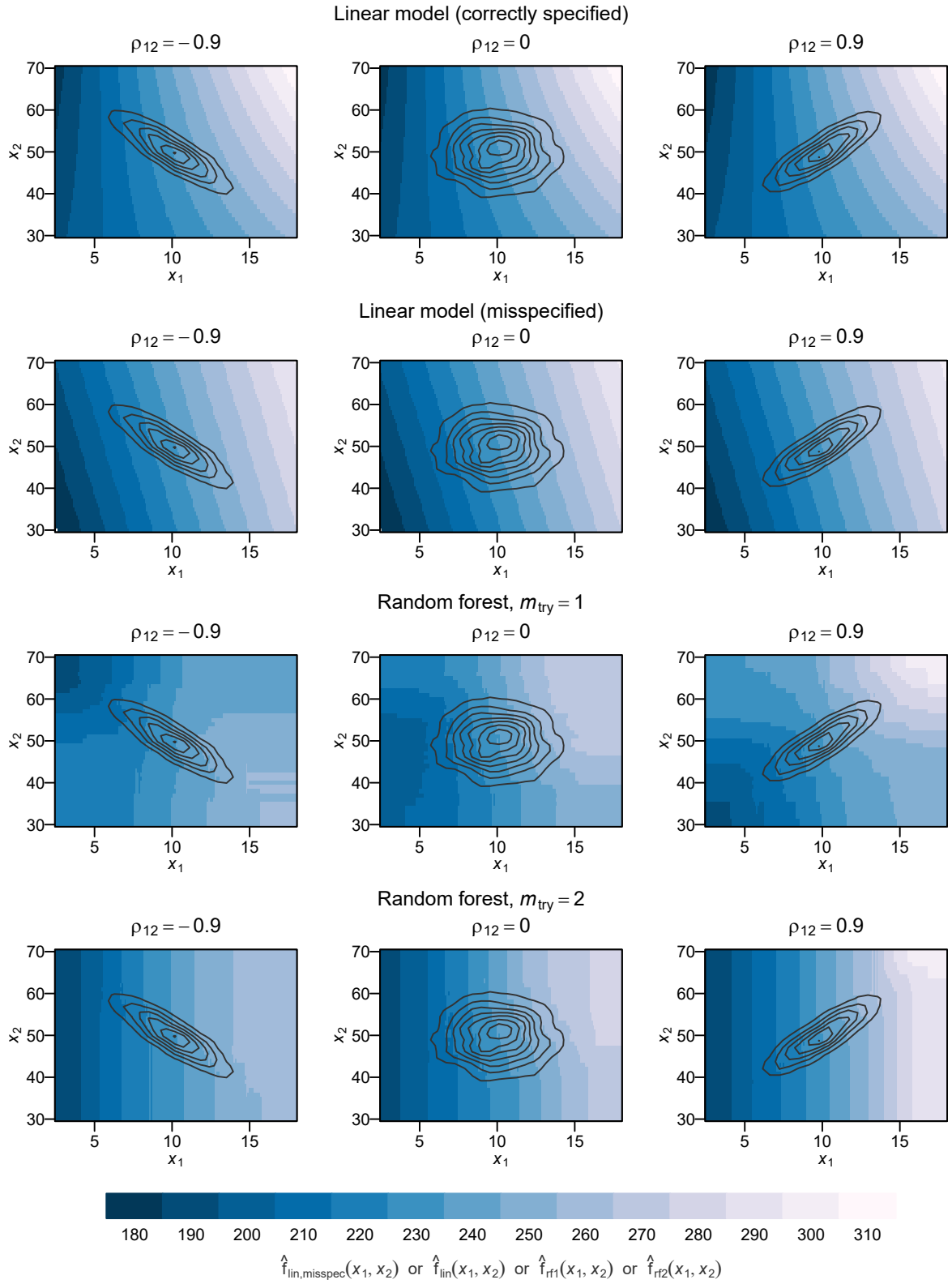

Figure 6: Prediction functions for Y with overplotted contours of the empirical bivariate feature distribution

Table 1: Mean square errors on a grid of feature values

	$\rho_{12} = -0.9$	$\rho_{12} = 0$	$\rho_{12} = 0.9$
Linear model (correct)	0.044	0.037	0.448
Linear model (misspecified)	29.316	28.477	29.548
Random forest $m_{\text{try}} = 1$	336.391	84.123	175.230
Random forest $m_{\text{try}} = 2$	133.447	31.897	78.229

only, which do not occur in the "would-be-badly-predicted" empty corners. Table 1 quantifies the lack of prediction quality over the grid of feature values that was used for producing Figure 1, by providing mean squares of deviations between true and fitted values. We see that the parametric prediction models fit the data generating model reasonably well everywhere, because they are correctly specified or almost so. The forests are much worse, especially for $m_{\text{try}} = 1$. This is due to the forests' nonparametric nature, which leads to poor extrapolation abilities: they fit the data well in the regions where training data feature pairs occur with high probability (see the elliptical contours of the feature distribution in Figure 1 or 6), but reflect the data generating model (1) much worse outside of those regions. Of course, such inspections can only be made in a simulation, where the true underlying model is known. Both forests fit the training data approximately equally well; the information in the training data is thus not sufficient to distinguish between them, although their extrapolation properties are quite different.

4.2 Performance of the MAEPs for the nonparametric models

Figure 7 shows PD and ALE curves with their corresponding estimands from the correctly-specified linear model; the PD plot estimand is simultaneously the estimand from the misspecified linear model for both PD and ALE curve. M curves are not shown, since they are clearly not adequate for describing the contribution of x_2 in a model that also contains x_1 as a predictor. For the uncorrelated case, all estimands coincide, and both PD and ALE plots from both forests match these quite well; only for x_2 values close to the margins of the x_2 range, there are small deviations. For $m_{\text{try}} = 1$, PD and ALE curve for the forest with positively correlated features are still similar and seem to achieve a compromise between PD and ALE estimands from the correct linear model. For the other correlated cases, PD plot and ALE plot are quite different from each other, and also from their target curves, in particular for the negative correlation.

We now focus on the particularly messy random forest with negative correlation and $m_{\text{try}} = 1$. Its prediction function is depicted in more detail in Figure 8. That figure also shows the PD plot with ICE curves and an overlaid ALE curve (constant for ALE curve modified such that the mean is the overall mean of predictions). From Figure 7 we know that neither PD plots nor ALE plots approximate their true estimands from the correct linear model, which is due to the fact that the prediction model is poor outside the ellipsoidal region where the training data are most concentrated. With the help of Figure 8, we can understand the different strategies by which PD and ALE plot depict the role of x_2 in the prediction model:

- The actual data points tend to occur on slightly increasing bits of the ICE curves, i.e. locally around these locations, there is an increase with respect to x_2. Outside of this increasing band, the ICE curves decrease.
- The PD curve averages over the ICE curves vertically, i.e. for fixed x_2 values. This almost completely eliminates the increasing portions, because these slide through the x_2 range and are always a minority versus decreasing parts on other ICE curves. The overall PD curve is therefore almost flat, with decreases at the outer areas and a very slight increase in the middle.
- The ALE curve adds up all the local increases in the vicinity of the data points and thus produces a relatively strong increase in the center area of the x_2 range; this increase roughly corresponds to the PD or ALE curve slope of the underlying data model. Thus, in that center area, the ALE curve manages to aggregate the model's local behavior in the vicinity of the data points into something that resembles the true PD or ALE curve. Outside of that center area, the ALE plot shows a slight decrease, presumably because there are not enough data points for the ALE plot to maintain its behavior.

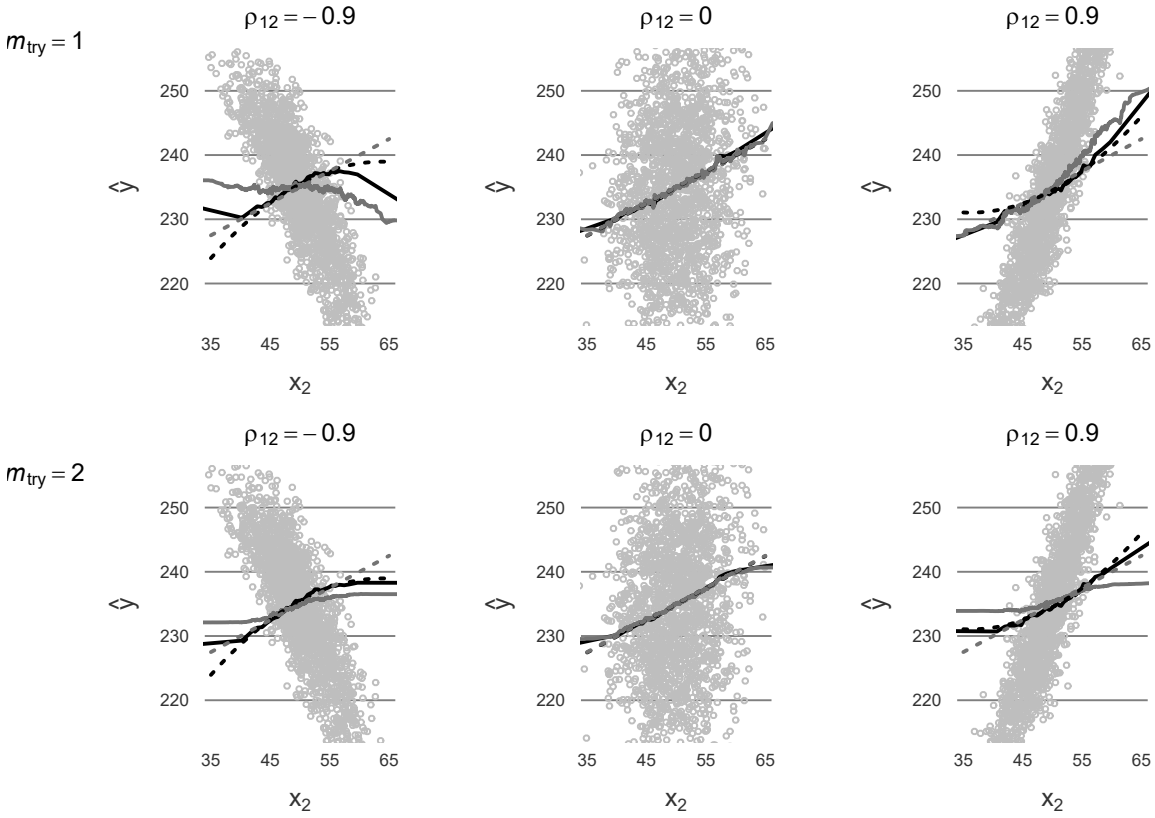

Figure 7: MAEPs for x_2 for the two random forests, applied to data from Model (1). Grey: PD, black: ALE. Solid: estimated, dotted: estimand under correctly-specified linear model. Background: scatter of predictions (would extend beyond vertical axis limits).

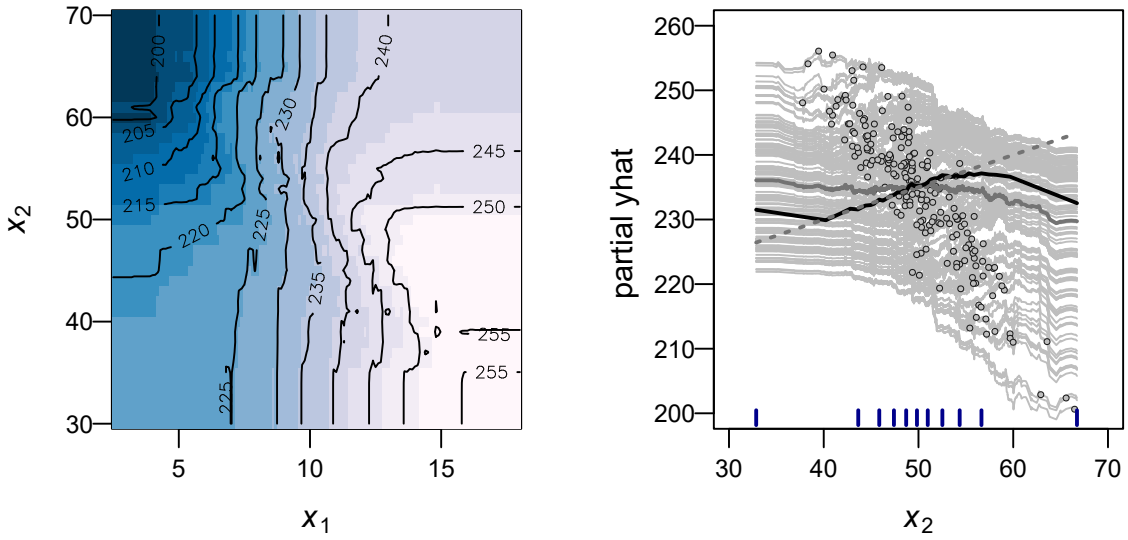

Figure 8: Left-hand side: Prediction function of the random forest with $m_{\text{try}} = 1$ and $\rho_{12} = -0.9$. Right-hand side: lighter grey=10% sample of ICE curves, grey=PD curve, black=ALE curve, dotted=PD estimand.

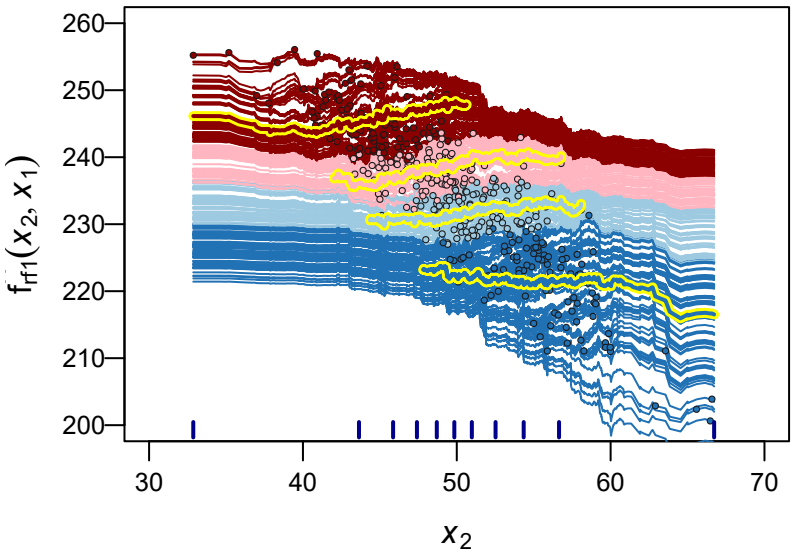

Figure 9: Stratified PD plot for the random forest with $m_{\text{try}} = 1$ and $\rho_{12} = -0.9$, with ICE curves: darker blue=lowest x_1 quarter, lighter blue=second x_1 quarter, lighter red=third x_1 quarter, darker red=highest x_1 quarter.

4.3 Summary of MAEP assessment regarding the model's extrapolation behavior

It is well-known, but nevertheless once more emphasized, that nonparametric models cannot be expected to work well on unseen feature combinations, even though many models will provide a prediction uncomplainingly. Because of this problem, ICE curves from nonparametric models for a feature that strongly correlates with other features will often be meaningless in much of the feature's range, because parts of the x_s values are not compatible with the unit's $x_{i,C^{-s}}$ (or have at least never been encountered in combination in the training data). Whenever this is the case, methods that rely on averaging over ICE curves will yield distorted results (likely usually towards less pronounced effects). Apley's and Zhu's (2019) reason for proposing ALE plots was that they identified a problematic behavior when applying PD plots to nonparametric prediction models for feature distributions like (2) with strong positive or negative correlation for which substantial areas of the Cartesian product of the domain of X_s with the domain of $X_{C^{-s}}$ are (almost) empty.

It has already been discussed that M plots are easily estimable, because they condition on the variable of interest and can be estimated well whenever the model produces good predictions for the training data. The price is that they are conceptually seriously flawed, by including into the effect of a particular feature influences from all correlated features. We will thus not discuss M plots any further.

Main effect MAEPs are supposed to provide insight into how a specific feature drives the model behavior over and above what other model features contribute. PD curves average over the distribution of $X_{C^{-s}}$, which is known to provide the full picture only in the absence of strong interactions, even for perfectly extrapolating models. One might think of limiting the range of x_s values to look at, depending on $x_{i;C^{-s}}$; simple averaging of this kind of restricted ICE curves will modify the average curve from a PD curve towards an M curve; thus, this route cannot be recommended (but see the next sub section). Apley and Zhu (2019) moved the restricted averaging from the curve itself to its derivative; in the presence of both interaction and correlation, this also yields a very problematic behavior conceptually (see Section 3.4). For predictions from nonparametric prediction models, such conceptually flawed behavior of ALE plots appears to be less pronounced, presumably because the nonparametric prediction model tends to miss out on the interaction behavior in empty extreme corners (it tends to be flat there). Nevertheless, this author lacks the confidence that ALE curves always estimate something reasonable.

Table 2: Correlation among quantitative features of auto mpg data ('mpg' is the response)

	mpg	cyl	displ	hp	wt	accel	MY
mpg	1.0000	-0.7776	-0.8051	-0.7784	-0.8322	0.4233	0.5805
cyl	-0.7776	1.0000	0.9508	0.8430	0.8975	-0.5047	-0.3456
displ	-0.8051	0.9508	1.0000	0.8973	0.9330	-0.5438	-0.3699
hp	-0.7784	0.8430	0.8973	1.0000	0.8645	-0.6892	-0.4164
wt	-0.8322	0.8975	0.9330	0.8645	1.0000	-0.4168	-0.3091
accel	0.4233	-0.5047	-0.5438	-0.6892	-0.4168	1.0000	0.2903
MY	0.5805	-0.3456	-0.3699	-0.4164	-0.3091	0.2903	1.0000

4.4 Stratified PD curves

ICE curves condition on the entire feature vector of each unit, except for x_s, which is varied. Classical effect plot tools for generalized linear models often condition on features that interact with the feature under investigation, and average over the other features (see e.g. Figure 3); a related approach can be implemented for MAEPs by averaging over subsets of ICE curves which are determined via strata w.r.t. a correlated – and possibly interacting – feature. Figure 9 shows such a stratified PD plot, with PD curves for x_2 created separately for four different ranges of x_1 values. The PD curves are drawn only over the x_2 range that occurs in the respective stratum. Such an approach may be useful whenever two features are strongly correlated, and will be particularly interesting, if there is a relevant amount of interaction between strongly correlated features. In Figure 9, we see an increase for increasing x_2 values, where x_1 is in any of the top three quarters, and a slight decrease for x_1 in the bottom quarter.

5 Application to a real world example

We now consider an automotive data set from the UCI Machine Learning repository (Dua and Graeff 2019; https://archive.ics.uci.edu/ml/datasets/auto+mpg). The response variable is fuel economy in miles per gallon (`mpg`). There are 392 observations without missing values (six observations with missings have been omitted). The data set has seven features for explaining `mpg`, four of which are quantitative continuous, two quantitative discrete (number of cylinders and model year) and one nominal with three categories (origin). There is substantial correlation between quantitative features (see Table 2). Consequently, variance inflation factors (VIFs) in a linear model with all main effects would be quite large for features `cyl` to `wt` (about 10 to 11), and particularly for the feature `displ` (about 23).

The response `mpg` is modeled with a random forest (1000 trees are built, each using $m_{\text{try}} = 2$ features). In the absence of knowledge about the true model, target curves for PD plots and ALE plots are of course unknown. A pair of variables for inspection has been picked from feature importance and interaction importance outcomes that were calculated with functions `featureImp$New` and `Interaction$new` of R package **iml**: weight, displacement, model year and horsepower are the most important variables, with displacement, horsepower and model year featuring strongest in terms of interactions. Displacement and horsepower is the most interesting pair, as these two are heavily correlated with a strong interaction. For comparison, a linear model has been hand-crafted, including all main effects plus a few plausible interaction effects (displacement by horsepower, weight by horsepower, acceleration by model year). The R^2 values are 87.93% for the random forest and 87.55% for the linear model, i.e. the two models have similar ability to explain the variability in the fuel economy responses (the linear model has been chosen to include the displacement by horsepower interaction, even though a different linear model would have produced a better R^2; remember the large VIFs that indicate that the information in the data is not sufficient for distinguishing different linear prediction model variants). Figure 10 shows the linear model interaction plot for displacement by horsepower, which indicates decreasing fuel economy with higher displacement for low horsepower and slightly increasing fuel economy with higher displacement for high horsepower (with a lot of variability around the prediction line). Figure 11 shows the MAEPs for both the linear model and the random forest. We see that ALE and PD plots are relatively similar, except for horsepower in the linear model. A likely reason for the larger difference there is that the random forest ICE curves are flat for extreme value combinations (like, e.g., large horsepower with small displacement),

while linear model ICE curves reflect large contributions from interactions; predictions for such extreme value combinations do affect PD plots but do not affect ALE plots, so that the linear model PD curve differs from the other three curves.

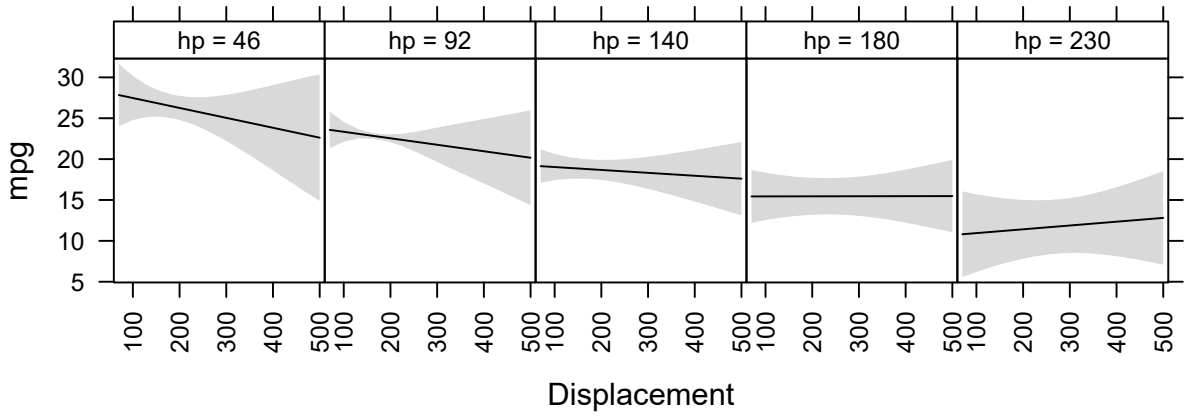

Figure 10: Classical interaction plot for miles per gallon from a hand-crafted linear model.

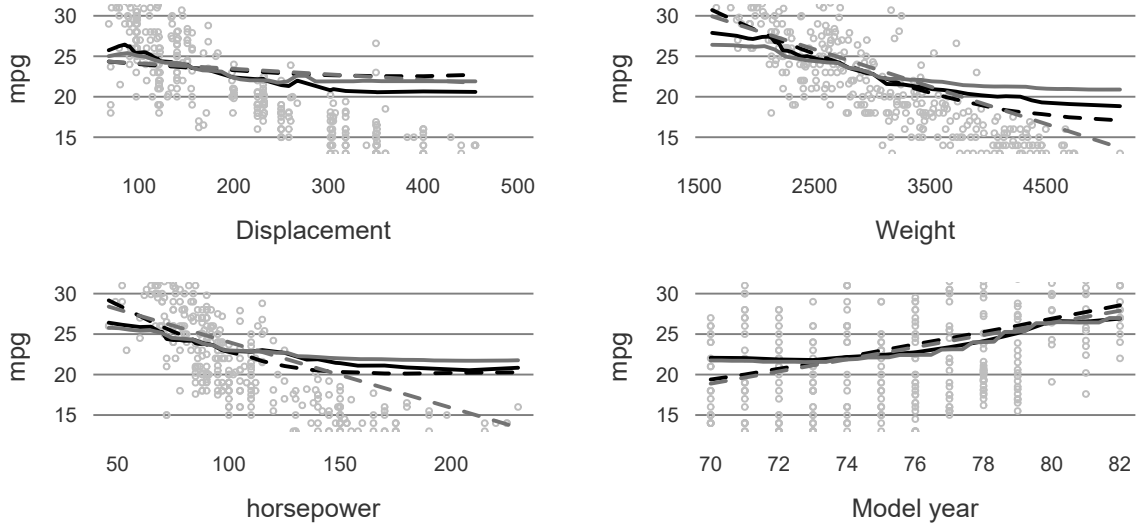

Figure 11: ALE and PD curves from linear model and random forest model. Grey: PD, black: ALE. Solid: random forest, dashed: linear model. Back ground: scatter of data points (would extend beyond vertical axis limits).

Figure 12 applies the stratified PD plot (introduced with Figure 9) to this real data example: ICE curves for displacement are stratified by horsepower quintiles, and the bold lines surrounded by yellow space show the separately calculated PD curves. The figure reflects a decrease in fuel economy with increasing displacement, which is steepest for lowest horsepower (dark blue) and quite flat for highest horsepower (the initial stronger decrease for highest horsepower is not really supported by data). Part (b) of the figure will be considered in Section 6.2.

6 The role of the data

There are two data sets involved in obtaining a MAEP:

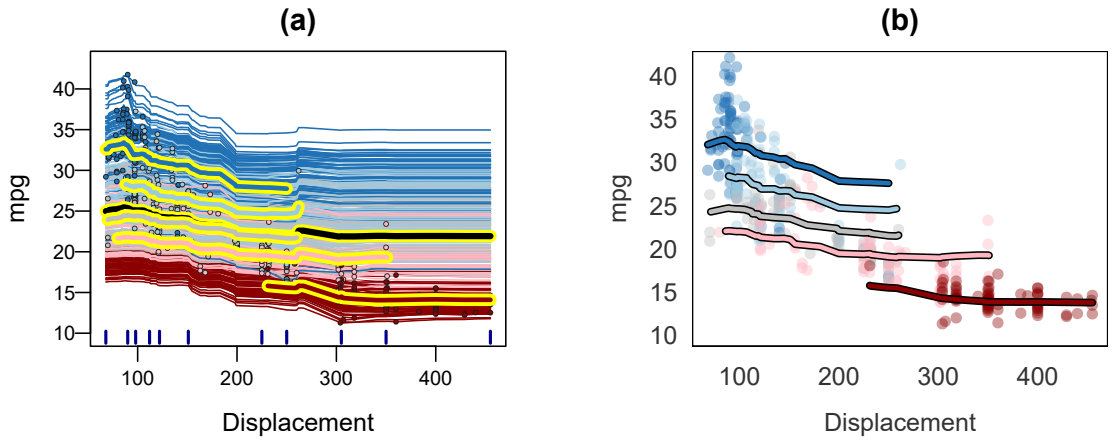

Figure 12: Stratified PD curves (a) calculated by averaging ICE curves or (b) approximated as ALE curves from resampled data. Legend: Overall PD curve (black) and separate PD curves for horsepower quintiles (darkest blue: lowest horsepower, darkest red: highest horsepower, grey: middle quintile). The separate PD curves are limited to the range of displacement values for the horsepower stratum. Vertical locations of ALE plot approximations in (b) adjusted to average stratum prediction. Left-hand side: points show predictions for respective ICE curves. Right-hand side: points show all training data predictions (partly transparent colors support visibility of overlap).

- The training data are used for creating the prediction function $\hat{f}$ (which is supposed to estimate Formula (1) in the simple linear model example of Section 2). If the training data contain highly correlated features, this affects the quality of that function: multicollinear data imply that quite different $\hat{f}$ have almost the same predictions for the training data and thus are indistinguishable based on training data information. This problem is relevant for linear models and is even more severe for nonparametric models. Hence, predictions in – sometimes large – parts of the Cartesian product of the domain of X_s with the set of realizations $\{x_{i,C^{-s}}, i = 1, \ldots, n\}$ can be very poor (often worse so for nonparametric than for parametric models).
- For creating the MAEP from $\hat{f}$, a different data set can be chosen, even though this paper so far always chose the original training data. These MAEP generating data are responsible for estimating the feature distribution, which in the simple linear model example of Section 2 amounts to estimating the conditional expectations (5) and the implied estimands in Equations (6), (9) or (12).

6.1 Different MAEP generating data

If the MAEP generating data are chosen differently from the training data, these data are used for estimating the influence of features on $\hat{f}$; the training data feature distribution ((2) for the simple example) is only indirectly involved by the influence it had on creating $\hat{f}$. If the estimated model was a linear model with two features and an interaction and the new feature distribution can also be described by (2) with modified parameters $\dot{\boldsymbol{\mu}}$ and $\dot{\boldsymbol{\Sigma}}$, the estimands for the MAEPs can be obtained from Equations (6), (9) or (12), by replacing β_j with the estimates $\hat{\beta}_j$ for the prediction model, and $\boldsymbol{\mu}$ with $\dot{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ with $\dot{\boldsymbol{\Sigma}}$ for the feature distribution.

For the choice of MAEP generating data, two aspects are to be considered:

1. What is the feature distribution for which $\hat{f}$ is to be applied?
2. What is the feature distribution under which $\hat{f}$ yields valid answers (in terms of an underlying truth)?

Of course, if the feature distribution for 1. is more general than that for 2., there is a problem, since application of $\hat{f}$ will not yield valid answers for all use cases. For the second question, it was already discussed that a nonparametric machine learning model generally yields $\hat{f}$ that is only applicable to sections of the feature space for which there was a sufficient amount of training data. This has direct

ethical implications (which are well-known and trivial, but nevertheless occasionally overlooked in naïve enthusiasm about the possibilities opened by machine learning): it can never be expected that a nonparametric machine learning model arrives at appropriate predictions for units that are exceptional relative to the body of data on which the algorithm was trained.

6.2 Approximating PD plots by ALE plots from uncorrelated data

Whenever obtaining a prediction requires some effort, calculation of PD plots is much slower than calculation of ALE plots, because more predictions have to be calculated. ALE plots can be used for approximating a PD plot with reduced calculation effort: the trick is to use the – possibly replicated – training data, but with x_s values resampled so that existing correlations between X_s and X_{C-s} are eliminated. This works, because

- the PD plot calculation algorithm for x_s gives each combination of x_s with $x_{i,C-s}$ to the prediction model, which emulates independence between X_s and X_{C-s}, even if the training data exhibit strong correlation.
- the resulting ALE curve does not suffer from the conceptual problem inherent in ALE plots of including interaction effects between x_s and its correlated features into the main effect of x_s, because all such correlations are eliminated by the resampling.

Where features are correlated in the real world, the ALE plot approximation for the PD curve will of course have to use predictions from less likely feature combinations for which the prediction model may yield poor predictions; this is unavoidable for approximating a PD curve.

ALE plot approximation can also be applied to stratified PD plots. For the example data, an ALE plot approximated stratified PD plot for displacement, obtained by separate resampling within strata of the strongly correlated feature horsepower, is shown in part (b) of Figure 12. The approximation apparently works quite well for the fuel economy random forest. This may be related to the fact that stratification on the correlated variable horsepower, because of the strong correlations within an entire group of features, implicitly includes also a (weaker) stratification on correlated features like cylinders and weight.

7 Discussion

PD plots are an established tool for depicting main effects in nonparametric prediction models, included e.g. in Hastie et al. (2009). Although they are conceptually more sound than ALE plots for "everywhere" correctly estimated prediction functions, they suffer from a severe extrapolation problem for nonparametric models with correlated features, as was highlighted by Apley and Zhu (2019), who proposed ALE plots as a remedy.

ALE plots have recently received praise in applied literature (e.g. Molnar 2019) and social media and are implemented in several packages in both R and Python (to name a few, without judgment and without attempting completeness: R: **ALEPlot**, **iml**, **DALEX** (via auxiliary packages); Python: **ALEPython**, **Pytalite**). Their advantages are calculation speed and avoiding extrapolation; their disadvantage is that they attribute part of the interaction effect to the main effect if there are interactions between correlated features. Consequently, ALE interaction plots (which were not discussed in this paper) miss out on those parts of the interaction effect that were already attributed to main effects in case of correlated features. ALE plots must be considered as defective, because of the severity of their conceptual problem.

It is recommended to use PD plots in combination with ICE curves, including also points for actual predictions (see e.g. right-hand side graph in Figure 8). The points help to understand whether there is a severe extrapolation problem, and whether there are systematic differences between ICE curves that are likely due to an interaction effect. In case of strong correlations and/or interactions, stratifying PD curves w.r.t. strongly correlated and/or interacting features can be considered (see Figure 9 for the small simulated example, and 12 for the fuel economy example). If computing resources for obtaining PD curves are a limiting factor, Section 6 pointed out how ALE curves can be used as approximations for PD curves, by emulating uncorrelatedness between x_s and the other features through resampling (overall or stratified). Where one resorts to ALE curves from resampled data as approximation for PD curves, ICE curves are no longer available; plotting points of $(x_{i;s}, \hat{f}(x_{i;s}, x_{C-s}))$ still provides the connection

to the raw predictions (see Figure 12). In the author's opinion, where feature correlation is high, the proposed stratified PD plots are preferable to the usual PD interaction plots (which were not discussed in this paper), because they show much more clearly where curves are supported by data. Even if there are no interactions, stratification with respect to a correlated feature may pay off, because the feature may represent a group of correlated features, so that the extrapolation problem of PD curves may be reduced by the stratification (this has likely been the case in the fuel economy example). However, the cautioning comments of the following paragraph are also applicable.

Hastie et al. (2009) use very careful wording in advertising PD curves: they justly claim that PD plots "can help to provide a qualitative description of [the] properties" of $\hat{f}$. Furthermore, they caution that insights can only be expected for x_s consisting of at most three features (this author considers three as already very ambitious), and only for those that involve highly relevant features. Furthermore, they state that the plots will be most revealing for prediction functions that are additive or multiplicative in nature (see the unbiasedness discussion in Section 3.4). All these are relevant comments. In addition, users of PD plots should always make themselves aware of the very local nature of nonparametric prediction models, particularly in case of correlated features; this awareness can be supported by the proposals of the previous paragraph.

Finally, it is once more emphasized that nonparametric prediction functions must be used with great care (or not at all), whenever one needs predictions for objects / situations / people that are exceptional relative to the body of data on which the model was trained. This well-known and trivial fact is not always treated with the priority that it deserves.

References

Apley, D. and Zhu, J. (2019). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. https://arxiv.org/abs/1612.08468v2.

Apley, D. (2018). ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots. R package version 1.1.

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research* **19**(84), 1-5.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fox, J. and Weisberg, S. (2018). Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals. *Journal of Statistical Software* **87**(9), 1-27.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5), 1189-1232.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* **24**(1), 44-65. doi:10.1080/10618600.2014.907095.

Grömping, U. (2019). South German Credit Data: Correcting a Widely Used Data Set. Report 4/2019, *Reports in Mathematics, Physics and Chemistry*, Beuth University of Applied Sciences Berlin.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, New York.

Isserlis, L. (1918). On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables. *Biometrika* **12**, 134–139. https://doi.org/10.1093/biomet/12.1-2.134

Lenth, R.V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software* **69**(1), 1-33.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* **2**(3), 18-22. URL = https://CRAN.R-project.org/doc/Rnews/.

Molnar, C. (2019). *Interpretable Machine Learning*. Christoph Molnar, Munich. ISBN 9780244768522. Most recent online version at https://christophm.github.io/interpretable-ml-book/.

Riedel, K.S. (1992). A Sherman–Morrison–Woodbury Identity for Rank Augmenting Matrices with Application to Centering, SIAM Journal on Matrix Analysis and Applications **13**, 659-662, https://doi.org/10.1137/0613040.

Appendix: Proof for Equation (4)

We consider matrices $\mathbf{X}_\mathrm{I} = (\mathbf{1}_n \quad \mathbf{X}_1 \quad \mathbf{X}_2)$ and $\mathbf{X}_\mathrm{II} = (\mathbf{X}_1 \cdot \mathbf{X}_2)$, where $\mathbf{X}_j$ denotes the one-column matrix of n realizations of the random variable X_j, and $\cdot$ the element-wise product. Given the actual matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, the true parameters in the misspecified model without interaction are: $(\beta_0, \beta_1, \beta_2)^\top + (\mathbf{X}_\mathrm{I}{}^\top \mathbf{X}_\mathrm{I})^{-1} \mathbf{X}_\mathrm{I}{}^\top \mathbf{X}_\mathrm{II} \beta_3$.

$$\frac{1}{n} \mathbf{X}_\mathrm{I}{}^\top \mathbf{X}_\mathrm{II} = \frac{1}{n} \left(\sum_{i=1}^{n} x_{i1} x_{i2}, \quad \sum_{i=1}^{n} x_{i1}{}^2 x_{i2}, \quad \sum_{i=1}^{n} x_{i1} x_{i2}{}^2 \right)^\top$$

converges against

$$\big(E(X_1 X_2), E(X_1{}^2 X_2), E(X_1 X_2{}^2)\big)^\top = \big(\mu_1\mu_2 + \sigma_{12}, \quad \mu_1{}^2\mu_2 + \sigma_{11}\mu_2 + 2\sigma_{12}\mu_1, \quad \mu_1\mu_2{}^2 + \sigma_{22}\mu_1 + 2\sigma_{12}\mu_2\big)^\top$$

(using a generalization of a result from Isserlis 1918).

$\frac{1}{n} \mathbf{X}_\mathrm{I}{}^\top \mathbf{X}_\mathrm{I}$ is consistent for a matrix $\mathbf{\Omega}$ of non-central first and second moments, which can be written as the sum of the outer product of the vector $\mathbf{u} = (1, \mu_1, \mu_2)^\top$ with a singular 3×3 matrix $\mathbf{A}$ whose lower right block is the covariance matrix $\mathbf{\Sigma}$ from (2) (and all other elements are zeroes). Choosing vectors $\mathbf{v}_1 = \mathbf{v}_2 = (0, \mu_1, \mu_2)^\top$ and $\mathbf{w}_1 = \mathbf{w}_2 = (1, 0, 0)^\top$, the limit can be written as $\mathbf{\Omega} = \mathbf{A} + (\mathbf{v}_1 + \mathbf{w}_1)(\mathbf{v}_2 + \mathbf{w}_2)^\top$. With Theorem 3 of Riedel (1992), its inverse becomes

$$\mathbf{A}^+ - \frac{\mathbf{w}_2 \mathbf{v}_2{}^\top \mathbf{A}^+}{|\mathbf{w}_2|^2} - \frac{\mathbf{A}^+ \mathbf{v}_1 \mathbf{w}_1{}^\top}{|\mathbf{w}_1|^2} + \big(1 + \mathbf{v}_2{}^\top \mathbf{A}^+ \mathbf{v}_1\big) \frac{\mathbf{w}_2 \mathbf{w}_1{}^\top}{|\mathbf{w}_1|^2 |\mathbf{w}_2|^2} = \begin{pmatrix} 1 + (\mu_1 \;\; \mu_2)\mathbf{\Sigma}^{-1} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & -(\mu_1 \;\; \mu_2)\mathbf{\Sigma}^{-1} \\ -\mathbf{\Sigma}^{-1} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & \mathbf{\Sigma}^{-1} \end{pmatrix}.$$

Thus, the coefficient vector of the misspecified linear model is consistent for

$$\begin{pmatrix} \widetilde{\beta}_0 \\ \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 + (\mu_1 \;\; \mu_2)\mathbf{\Sigma}^{-1} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & -(\mu_1 \;\; \mu_2)\mathbf{\Sigma}^{-1} \\ -\mathbf{\Sigma}^{-1} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & \mathbf{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \mu_1\mu_2 + \sigma_{12} \\ \mu_1{}^2\mu_2 + \sigma_{11}\mu_2 + 2\sigma_{12}\mu_1 \\ \mu_1\mu_2{}^2 + \sigma_{22}\mu_1 + 2\sigma_{12}\mu_2 \end{pmatrix} \beta_3. \qquad (13)$$

Now, simple but slightly tedious calculations yield the equalities $\widetilde{\beta}_1 = \beta_1 + \mu_2\beta_3$ and $\widetilde{\beta}_2 = \beta_2 + \mu_1\beta_3$.