Understanding complex predictive models with ghost variables

Pedro Delicado[1,2*] and Daniel Peña[3,4]

[1*]Departament d'Estadística i Investigació Operativa.
[2*]IMTech. Institute of Mathematics of UPC-BarcelonaTech, Universitat Politècica de Catalunya·BarcelonaTech, Jordi Girona, 31, Barcelona, 08034, Spain.
[3]Departamento de Estadística.
[4]Big Data Institute, Universidad Carlos III de Madrid, Calle de Madrid, 126, Getafe (Madrid), 28903, Spain.

*Corresponding author(s). E-mail(s): pedro.delicado@upc.edu;
Contributing authors: daniel.pena@uc3m.es;

Abstract

Framed in the literature on Interpretable Machine Learning, we propose a new procedure to assign a measure of relevance to each explanatory variable in a complex predictive model. We assume that we have a training set to fit the model and a test set to check its out-of-sample performance. We propose to measure the individual relevance of each variable by comparing the predictions of the model in the test set with those obtained when the variable of interest is substituted (in the test set) by its ghost variable, defined as the prediction of this variable by using the rest of explanatory variables. In linear models it is shown that, on the one hand, the proposed measure gives similar results to leave-one-covariate-out (*loco*, with a lowest computational cost) and outperforms random permutations, and on the other hand, it is strongly related with the usual F-statistic measuring the significance of a variable. In non linear predictive models (as neural networks or random forests) the proposed measure shows the relevance of the variables in an efficient way, as shown by a simulation study comparing ghost variables with other alternative methods (including *loco* and random permutations, and also knockoff variables and estimated conditional distributions). Finally, we study the joint relevance of the variables by defining the relevance matrix as the covariance matrix of the vectors of effects on predictions when using every ghost variable. Our proposal is illustrated with simulated examples and the analysis of a large real data set.

Keywords: Explainable Artifical Inteligence, Estimated conditional distributions, Interpretable Machine Learning, Knockoffs,

Leave-one-covariate-out, Out-of-sample prediction, Partial correlation matrix, Random permutations

MSC Classification: 62R07 , 68T09 , 62G08

Acknowledgments

This research was supported by the Spanish Ministerio de Ciencia e Innovación grants MTM2017-88142-P, PID2020-116294GB-I00 (Pedro Delicado) and PID2019-109196GB-I00 (Daniel Peña). We are grateful to Alejandro Germán-Serrano for providing us with access to Idealista data.

Declarations

Competing Interests: Authors declare that they do not have financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

1 Introduction

In a stimulating and provocative paper, Breiman (2001) shook the statistical community by arguing that traditional Statistics was no longer the only way to reach conclusions from data. Breiman talked about the *two cultures* of data modeling, and he noted that, in addition to the *Data Modeling Culture* (traditional Statistics), was emerging an *Algorithmic Modeling Culture*. The algorithms proposed by the new culture (neural networks, support vector machines, random forests, etc.) have often better predictive accuracy than traditional statistical models (linear regression, logistic regression, etc.). On the other hand, statistical models explain better the relationship among response and input variables. In fact, these new-culture algorithms usually are called *black boxes.*

The need of understanding the effects of variables increases with prediction rules based on Big Data, that is, data sets with large number of variables, p, and observations, n, and even with $p > n$. The main reason is that in some models, as neural networks, the effect of a variable is a non linear function of different linear combinations with the rest of the variables, making its total effect difficult to see. Moreover, the effect of a variable strongly correlated with others cannot be well measured without considering their joint effects. The crucial importance of understanding the effect of the variables on algorithmic models has often been recognized. For instance, Ribeiro et al. (2016b) state that *a vital concern remains: if the users do not trust a model or a prediction, they will not use it.*

There is a powerful research line on *variable importance measures in algorithmic models*, whose main goal is to provide interpretability for these black box models. The interest on this field is rapidly growing. As a sample of it,

search queries in the Web of Science and in Scopus (performed in November 5th, 2021) looking for publications having as topic any of the terms "explainable artificial intelligence", "explainable machine learning" or "interpretable machine learning" found a total of 5673 publications in the Web of Science, from which the 51% were published in 2020 or later. In Scopus this percentage rose to 80% of the 7465 publications found. This vast literature has given rise to a considerable number of review papers (see for instance the good review in Barredo-Arrieta et al. 2020) and three monographs: Molnar (2019), Biecek and Burzykowski (2021) and Masís (2021).

Consider the general framework of a prediction problem involving the random vector (X, Z, Y), $X \in \mathbb{R}^p$, $Z \in \mathbb{R}$ and $Y \in \mathbb{R}$, where Y is the response variable that should be predicted from (X, Z). Assume that there exist a training sample of size n_1 (used to fit the prediction rule) and a test sample of size n_2 (used for checking the out-of-sample precision accuracy of the fitted prediction rule). A simple approach to define the importance (or relevance) of the variable Z consists in measuring its contribution to the forecast of the response. To compute this measure, the model is fitted twice: first including both X and Z, and then excluding Z. This approach, known as leave-one-covariate-out (*loco*), is often used to decide if the variable Z should be included in the model. However, other alternatives are possible. Instead of deleting the variable Z, Breiman (2001) proposed to randomly permute its values in the test sample to create a new variable Z', and compare the model predictions using the original Z variable and the randomized one Z'. This method has two main advantages over deleting the variable: (1) only one predictive model has to be fitted (the one using all the explanatory variables); and (2) the forecast comparison is made between two models with the same number of variables. A drawback is that Z' is unrelated to Y but also to X and the joint effect of the Z variable with the X will be lost. It is worth mentioning that both methods, deleting Z and replacing it by Z', are *model-agnostic* (Ribeiro et al. 2016a) because they can be applied to any predictive model, which is a desirable property.

Despite its popularity, using random permutations for interpreting black box prediction algorithms has received numerous criticisms, mainly when the explanatory variables present strong dependencies. See, for instance, Hooker et al. (2021), who explain why what they call *permute-and-predict* methods fail when covariates X and Z are dependent: X and Z' are independent by construction, then the support of (X, Z') is larger than that of (X, Z), which forces the prediction model to extrapolate. Hooker et al. (2021) propose two strategies to avoid this problem: to generate values Z' from the conditional distribution of Z given X, and to train again the model with covariates (X, Z') in the training set, which can be applied separately or jointly. When the conditional distribution of Z given X is unknown, they can be estimated as proposed by Tansey et al. (2022) in the context of *hold-out randomized tests*. Another possibility is to use knockoff variables (Barber and Candès 2015, Candès et al. 2018). Both approaches, knockoffs and hold-out randomized tests, appear in the variable

selection problem in models with a large number of possible explanatory variables with the objective of controlling the false discovery rate. Note that the problem of measuring relevance of variables is different to that of controlling the number of false positive results, although they are somehow related. The relevance variable problem appears when we have a non linear complex model where the effect of each variable is hard to see. Thus, we are interested in understanding the effects of each variable. On the other hand, the control of false discovery rate appears when we have thousands of noise variables, and we try to avoid including many of them in the model. In the first case the problem appears for the complexity of the model, whereas in the second one the model may be simple, as in linear regression, but the problem comes from the large number of possible irrelevant regressors. In Section 4 the use of these ideas for variable relevance will be discussed.

In this paper we propose a new approach to measure the effect of an explanatory variable, which combines the advantages of leaving-it-out or randomly permuting it, and at the same time avoids their drawbacks. First, we fit the model with all the original variables, X and Z, in the training sample. Then we apply the model in the test sample twice: first using the observed values of Z, and then using a new variable $\widehat{Z}$, which is uncorrelated to the response given the variables X (as in the randomized method) but reproduces the relationship between Z and X better than the randomized variable. We call this new variable a *ghost variable*, and it is computed as $\hat{Z} = \widehat{\mathbb{E}(Z \mid X)}$, the estimated expected value of Z given the other explanatory variables X. Finally, we expand our proposals to measure the joint relevance of groups of variables with the definition of the *relevance matrix*, as the covariance matrix of the individual effects, showing the importance of a joint analysis of the relationships between the individual effects of the variables to identify the most important groups of variables in the prediction.

Our proposal to use *ghost variables* has similarities and differences with the work of Hooker et al. (2021). (It is worth mentioning that both works were developed independently). The main similarity is that we use the conditional distribution of Z given X to generate pseudo-observations of Z, as in some of the proposals of Hooker et al. (2021). But, while we use just the conditional expectation of Z given X, Hooker et al. (2021) consider either taking conditional permutations of Z given X, or generating random data from this conditional distribution. Observe that modeling $\mathbb{E}(Z \mid X)$ is always simpler than modeling the complete conditional distribution of $(Z \mid X)$. Another difference with respect to Hooker et al. (2021) is that we avoid to learn the prediction model twice, which could be very demanding in time and resources (this is the main reason why we look for alternatives to *loco*). Finally, it is very important to take into account the joint relevance of sets of variables when they are correlated, and we introduce a relevance matrix to identify joint effects. This method represents a multivariate approach to relevance measures, while all the proposals of Hooker et al. (2021) are univariate.

The rest of the article is organized as follows. Section 2 presents the problem of measuring the relevance of a variable in a prediction problem at the population level, working with the joint distribution of random vectors. We briefly present leaving the co-variable out (*loco*), and substituting it by a random permutation. Then we introduce our proposal: replacing it by its ghost variable, defined as its expected value given the rest of the explanatory variables. Section 3 deals with the sample implementation of these three variable relevance measures. Other relevance measures, as those based on knockoffs or on conditional distributions, will be revised in Section 4, where their differences with ghost variables will be discussed. Section 5 shows the advantages of ghost variables versus random permutations in simple models, as multiple linear regression or additive models, proving that they work very well when exact results can be obtained and, in particular in the linear model, are asymptotically equivalent to *loco* and to the usual F-statistics for testing variables significance. This fact supports its application in more complex non linear models. Section 6 compares the practical performance of relevance based on ghost variables with alternative approaches in simulated examples. Section 7 introduces the joint relevance of groups of correlated variables and defines the relevance matrix. It is shown that, in linear regression, this matrix is closely related with the partial correlation matrix of the explanatory variables. The properties of the relevance matrix are illustrated in Section 8 in simulated data and in a real case, at which a neural network is fitted. Section 9 concludes. The proofs of the results in Sections 3 and 7 are deferred to the appendixes. An Online Supplement includes results on relevance by *loco* and by random permutations, additional outputs corresponding to the real data example, and the link to the R-scripts containing code to reproduce the computations and graphics in the paper.

2 Relevance for random variables

Let us consider the prediction problem involving the random vector (X, Z, Y), $X \in \mathbb{R}^p$, $Z \in \mathbb{R}$ and $Y \in \mathbb{R}$, where Y is the response variable that should be predicted from (X, Z). A *prediction function* $f : \mathbb{R}^{p+1} \to \mathbb{R}$ has *expected loss* $\mathbb{E}(L(f(X, Z), Y))$, where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a loss function measuring the *cost* associated with predicting Y by $f(X, Z)$. Probably the most widely used loss in practice is quadratic loss, for which the expected loss is the Mean Squared Prediction Error (MSPE):

$$\text{MSPE}(f) = \mathbb{E}((Y - f(X, Z))^2).$$

Even if we work here only with quadratic loss, we would like to stress that any other loss function could be used.

We consider the problem of measuring the relevance of Z given the prediction rule $f(X, Z)$. Relevance here is defined as the contribution of Z, or its importance, in the prediction of Y using $f(X, Z)$. We will compare in detail three approaches to measure the relevance of a single variable.

2.1 Relevance by leaving-one-covariate-out *(loco)*

The first method to measure the relevance is to delete the variable and compare the forecast with and without it. This procedure, leave-one-covariate-out or *loco*, has been often applied and it has a long tradition (for instance, it is the basic principle when testing $H_0 : \beta_Z = 0$ against $H_0 : \beta_Z \neq 0$ in a multiple linear regression model; see Lei et al. 2018 for a recent approach). The idea is to fit a new prediction rule that includes only the X variables and then compare the MSPE of both rules. Let $f_p(X)$ be the reduced version of f where the variable Z is not considered. Then the (population) relevance of Z by *loco* is usually measured in the literature as either the relative squared change in the prediction when it is left-out,

$$\mathbf{RV}_{\text{loco}}(Z) = \mathbb{E}\left((f(X,Z) - f_p(X))^2\right)/\text{MSPE}(f), \tag{1}$$

(see, for instance, Gregorutti et al. 2015 or Zhu et al. 2015) or the relative decrease in the MSPE when Z is removed from the predictive rule,

$$\mathbf{RV}^*_{\text{loco}}(Z) = (\text{MSPE}(f_p) - \text{MSPE}(f))/\text{MSPE}(f). \tag{2}$$

(see, for instance, Breiman 2001 or Hooker et al. 2021). Both measures coincide when $Y = f(X,Z) + \varepsilon$, where ε is a zero mean random variable such that $\mathbb{E}(\varepsilon f(X,Z)) = \mathbb{E}(\varepsilon f_p(X))$ (which happens, for instance, when (X,Z) and ε are independent; see also Theorem 2 in Hooker et al. 2021). Then,

$$\text{MSPE}(f_p) = \text{MSPE}(f) + \mathbb{E}\left((f(X,Z) - f_p(X))^2\right).$$

In this article we use both, but theoretical results are obtained for $\mathbf{RV}_{\text{loco}}(Z)$ because (1) it does not depend on the noisy part of the response variable Y, and (2) in linear models it is equivalent to the standard variable significance measures (see Theorem 1 bellow).

Consider the case of f being additive in X and Z: $f(X,Z) = \beta_0 + s_1(X) + s_2(Z)$, with $\mathbb{E}(s_1(Z)) = \mathbb{E}(s_2(Z)) = 0$. Then $f_p(X) = \mathbb{E}(Y \mid X) = \beta_0 + s_1(X) + \mathbb{E}(s_2(Z) \mid X) = \beta_0 + s_1^*(X)$, and the relevance of Z by *loco* is

$$\mathbf{RV}_{\text{loco}}(Z) = \eta\mathbb{E}((s_2(Z) - \mathbb{E}(s_2(Z) \mid X))^2) = \eta\mathbb{E}(\text{Var}(s_2(Z) \mid X)), \tag{3}$$

where $\eta = (\text{MSPE}(f))^{-1}$. Assuming additional linearity, $Y = \beta_0 + X^T\beta_X + Z\beta_Z + \varepsilon$, we obtain that

$$\mathbf{RV}_{\text{loco}}(Z) = \eta\mathbb{E}(\text{Var}(Z\beta_Z \mid X)) = \eta\beta_Z^2\mathbb{E}(\text{Var}(Z \mid X)).$$

2.2 Relevance by random permutation

A second approach, proposed by Breiman (2001), is to replace the variable Z in the prediction rule by an independent copy Z': a random variable with the

same marginal distribution as Z but independent from (X, Y) and Z. Then, define the (population) relevance by random permutations by

$$\mathbf{RV}_{\text{rp}}(Z) = \mathbb{E}((f(X, Z) - f(X, Z'))^2)/\text{MSPE}(f).$$

This procedure has the advantage that does not require the fitting of the reduced model, $f_p(X)$, but has some limitations. Consider the previous additive case. Then,

$$\mathbf{RV}_{\text{rp}}(Z) = \eta\mathbb{E}\left(\{(\beta_0 + s_1(X) + s_2(Z)) - (\beta_0 + s_1(X) + s_2(Z'))\}^2\right) =$$

$$2\eta\text{Var}(s_2(Z)).$$

where, as before, $\eta = (\text{MSPE}(f))^{-1}$. If additional linearity happens, $s_2(Z) = Z\beta_z$, then this relevance is $2\eta\beta_z^2\text{Var}(Z)$. These results are similar to those obtained by Gregorutti et al. (2017, Propositions 1 and 2) or Hooker et al. (2021, Theorem 1). At a first glance this relevance measure seems to be suitable, but:

(1) The method of random permutations assigns to Z the same relevance when it is independent from X or when it is strongly related with X. Clearly Z is more relevant in the first case than in the second one.
(2) The replacement of Z by an independent copy Z' adds some noise to the prediction function $f(X, Z)$. For instance, in the linear predictor $f(X, Z) = \beta_0 + X^T\beta_x + Z\beta_z$ replacing Z by Z' is equivalent to using the following reduced version of f:

$$f_p(X) = \beta_0' + X^T\beta_x + \nu,$$

where $\beta_0' = \beta_0 + \beta_z\mathbb{E}(Z)$ and $\nu = \beta_z(Z' - \mathbb{E}(Z))$, a zero mean random variable independent from (X, Y) that adds noise to the prediction of Y. A better alternative would be to use the reduced version of f given just by $\beta_0' + X^T\beta_x$, which is equivalent to replacing Z by $\mathbb{E}(Z)$ in $f(X, Z)$.

2.3 Relevance by a ghost variable

We propose in this article a new procedure that combines the conceptual simplicity of the *loco* approach and the computational advantage of the random permutation approach. The idea is to replace Z by its ghost variable that is defined as a variable independent of the response that is as close as possible as the one we want to replace. Using quadratic loss the ghost variable of Z is $\mathbb{E}(Z \mid X)$, the best prediction of Z given X. Note that if the variable is independent of the others, this will imply replacing the varible by a constant, $\mathbb{E}(Z)$, and the procedure will be similar to deleting the variable. However, if there is dependency between X and Z, $|Z - \mathbb{E}(Z \mid X)|$ is expected to be lower than $|Z - \mathbb{E}(Z)|$, so $f(X, \mathbb{E}(Z \mid X))$ is expected to be closer to $f(X, Z)$ than $f(X, \mathbb{E}(Z))$. Therefore, when Z is not available, replacing it by $\mathbb{E}(Z \mid X)$ allows

X to contribute a little bit more in the prediction of Y than replacing Z by $\mathbb{E}(Z)$. The greater this additional contribution of X, the lower the relevance of Z in the prediction of Y. The (population) relevance of Z with ghost variables is measured by

$$\mathbf{RV}_{\text{gh}}(Z) = \mathbb{E}\left((f(X,Z) - f(X,\mathbb{E}(Z \mid X)))^2\right)/\text{MSPE}(f).$$

Observe that replacing Z by $\mathbb{E}(Z \mid X)$ is equivalent to consider $f_p(x) = f(x,\mathbb{E}(Z \mid X = x))$ as the reduced version of $f(X,Z)$. It follows that $f(X,\mathbb{E}(Z \mid X)) = \mathbb{E}(f(X,Z) \mid X)$ for linear functions $f(X,Z)$, leading to the well known result that removing Z or replacing it by $\mathbb{E}(Z \mid X)$ are equivalent in multiple linear regression. When f is additive in X and Z, and calling as before $\eta = (\text{MSPE}(f))^{-1}$, we have

$$\mathbf{RV}_{\text{gh}}(Z) = \eta\mathbb{E}\left((s_2(Z) - s_2(\mathbb{E}(Z \mid X)))^2\right).$$

If, additionally, $s_2(Z) = Z\beta_z$,

$$\mathbf{RV}_{\text{gh}}(Z) = \eta\beta_z^2\mathbb{E}\left((Z - \mathbb{E}(Z \mid X))^2\right) = \eta\beta_z^2\mathbb{E}(\text{Var}(Z \mid X)), \tag{4}$$

which coincides with $\mathbf{RV}_{\text{loco}}(Z)$. Observe that $\eta\beta_z^2\mathbb{E}(\text{Var}(Z \mid X))$ coincides with $\eta\beta_z^2\text{Var}(Z)$ when X and Z are independent, but otherwise the first one would be preferred to the second one as relevance measure of Z. So we conclude that measuring variable relevance by *loco* or by ghost variables is preferred to using random permutations in the important case of multiple linear regression models. We expect that this better performance will be maintained when these relevance measures are applied to more complex predictive models.

We conclude that for defining the relevance of a random variable Z in a non linear model the ghost variable approach has several advantages over the other two alternative procedures. First, in contrast to the *loco* approach, we do not need to fit two complex predictive models for the response, although we need to compute $\mathbb{E}(Z \mid X)$, which can be approximated by a linear or additive function. Second, in contrast to the random permutation approach, the relevance by ghost variable is sensible to changes in the degree of dependence between Z and the other explanatory variables.

3 Relevance in data sets

Now we consider the practical implementation of the approaches introduced so far for measuring the relevance of a variable. Calling $f(x,z) = \mathbb{E}(Y \mid X = x, Z = z)$ to the *regression function* of Y on (X,Z), the best prediction function for Y under quadratic loss, any prediction function of Y is also an estimator $\hat{f}(x,z)$ of the regression function $f(x,z)$, and vice versa. So, from now on, we will talk indistinctly of prediction functions or regression function estimators.

Consider a training sample of n_1 independent realizations of (X,Z,Y) and we assume, to simplify the notation and without loss of generality, that all

the variables have sample mean equal to zero. Let $(\mathbf{X}_1, \mathbf{z}_1, \mathbf{y}_1)$ be the matrix representation of the training sample, which is used to estimate the regression function $\hat{f}(x, z)$, and let $(\mathbf{X}_2, \mathbf{z}_2, \mathbf{y}_2)$ be the test sample in matrix format and $\mathbf{z}_2' \in \mathbb{R}^{n_2 \times 1}$ be a random permutation of the elements of the column vector $\mathbf{z}_2$. The estimation of $\text{MSPE}(\hat{f})$ from the test sample is

$$\widehat{\text{MSPE}}(\hat{f}) = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_{2.i} - \hat{f}(x_{2.i}, z_{2.i}))^2.$$

The test sample is used to obtain the sampling versions of the three variable relevance measures introduced in Section 2. Calling $\text{RV}_v(Z)$ to the sample estimate of the relevance of Z by method $v \in \{\text{loco}, \text{rp}, \text{gh}\}$, this statistic is computed by

$$\text{RV}_v(Z) = \frac{1}{n_2} \sum_{i=1}^{n_2} (\hat{f}(x_{2.i}, z_{2.i}) - \hat{f}_v(i))^2 / \widehat{\text{MSPE}}(\hat{f}),$$

where for $v = \text{loco}$, $\hat{f}_v(i) = \hat{f}_p(x_{2.i})$, for $v = \text{rp}$, $\hat{f}_v(i) = \hat{f}(x_{2.i}, z'_{2.i})$, and for $v = \text{gh}$ $\hat{f}_v(i) = \hat{f}(x_{2.i}, \widehat{z}_{2.i})$ with $\widehat{z}_{2.i} = \hat{\mathbb{E}}(Z \mid X = x_{2.i})$.

This analysis of the relevance of individual variables can be applied to any prediction rule. When using *loco*, we need to compute two prediction rules: $\hat{f}(x_{2.i}, z_{2.i})$ and $\hat{f}_p(x_{2.i})$. For instance in a neural network we have to run the estimation process twice in the training sample. By random permutation we only estimate the model once, and then use the random permutation of the variable in the test sample to compute the predictions. With ghost variables we also estimate the model in the training sample only once, and then in the test sample we compute the ghost variable (the prediction function of Z given the X variables) and use it in the prediction function computed in the training set. The method with the least computer time is random permutation, which can be used directly on the estimated prediction rule. The second fastest is, in general, the ghost variable approach, which requires computing the ghost values of Z to be used in the prediction rule in the test sample. The process can be sped up by using linear regression to estimate the conditional expectation of Z given X. In general, the most computationally time-intensive is the *loco* approach, which requires the estimation of two (usually complex) predictive models with the training sample.

4 Other possible relevance measures based on perturbations

Random permutations and ghost variables methods for computing relevance of an explanatory variable Z follow a general scheme: to replace the values of Z in the test set by *"perturbed"* values of them, which are independent of the response variable Y, given the other explanatory variables X.

We revise now other possibilities of *"perturbation"* of Z which have been considered recently in the literature. We will see that these proposals (which essentially draw random values from a conditional distribution, known or estimated) are more difficult to apply than ghost variables (which consists of estimating just a conditional expectation).

Hooker et al. (2021) propose to replace z_i by a random value coming from the conditional distribution of $(Z \mid X = x_i)$, which is usually known for simulated data. For realistic settings, where the conditional distribution is unknown, Hooker et al. (2021) propose to use the Model-X (MX) knockoff framework proposed by Candès et al. (2018) to generate values of Z. In particular, they sample second-order multivariate Gaussian knockoff variables as implemented in the R package `knockoff` (Patterson and Sesia 2022).

Knockoff variables were defined by Barber and Candès (2015) for linear regression models as variables unrelated to the response and that jointly have the same distribution as the original ones, but being as different as possible from them. This idea was extended by Candès et al. (2018) assuming that the explanatory variables are random variables with some joint distribution. Then the set of model-X (MX) knockoff variables, $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$, are unrelated to the response, and the joint distribution of $(X, \tilde{X})$ does not change if we swap or interchange any set of original variables and their knockoff variables. From the definition of MX knockoffs it follows that X and $\tilde{X}$ have the same distribution.

Any realization of the random variable $(\tilde{X} \mid X = x)$ can be used as valid values of the knockoff variables (Candès et al. 2018). In particular, a vector of random observations of $(\tilde{X}_j \mid X = x_{2.i})$, $i = 1, \ldots, n_2$ can be used as a perturbation of the variable X_j at the test set, as done in Hooker et al. (2021).

For instance, for normal zero mean variables the joint distribution only depends on the covariance matrix and the swapping rule will be true if this covariance matrix is of the form

$$\mathrm{Var}(X, \tilde{X}) = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \mathbf{D} \\ \mathbf{\Sigma} - \mathbf{D} & \mathbf{\Sigma} \end{pmatrix}$$

where $\mathbf{D}$ is a diagonal matrix so that $\mathrm{Var}(X, \tilde{X})$ is positive definite which is equivalent to say that the matrix $2\mathbf{\Sigma} - \mathbf{D}$ has this property.

Observe that in the definitions of knockoff variables in Barber and Candès (2015) and Candès et al. (2018) the idea of building new variables associated to each of the explanatory variables that are the conditional expectation of the variable given the others does not appear at all. Even though, it is worth remarking here that ghost variables and knockoff variables are different concepts.

Consider the p-dimensional random variable $X = (X_1, \ldots, X_p)$ from which a realization is $x = (x_1, \ldots, x_p)$. The corresponding vector of ghost variable

values is $X^g = (X_1^g, \ldots, X_p^g)$ where

$$X_j^g = E(X_j \mid X_{-j} = x_{-j}),\ j = 1, \ldots, p.$$

From the definition of MX knockoffs it follows that X_j and $\tilde{X}_j$ have the same marginal distribution. It is obvious that X_j and the ghost variable $X_j^g = E(X_j \mid X_{-j})$ have different distributions. Therefore ghost variables are not knockoffs.

To fix ideas, assume that X is a bivariate normal distribution with zero means, unit variances and correlation $\rho > 0$:

$$X \sim N_2\left(\mathbf{0}_2, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{5}$$

Then, from Candès et al. (2018), the joint distribution of $(X, \tilde{X})$ is

$$\begin{pmatrix} X \\ \tilde{X} \end{pmatrix} \sim N_4\left(\mathbf{0}_4, \begin{pmatrix} \Sigma & \Sigma - \mathrm{diag}(s) \\ \Sigma - \mathrm{diag}(s) & \Sigma \end{pmatrix}\right)$$

with s chosen as large as possible respecting that the joint covariance matrix is positive semidefinite: $s = \min\{2(1-\rho), 1\}$. Therefore, for $x = (x_1, x_2)^T$ the values of the knockoff variables, given that $X = x$, are generated from

$$(\tilde{X} \mid X = x) \sim N_2\left((\mathbf{I}_2 - s\Sigma^{-1})x, 2s\mathbf{I}_2 - s^2\Sigma^{-1}\right).$$

On the other hand, the ghost variables in this case are

$$X_1^g = E(X_1 \mid X_2) = \rho X_2,\ X_2^g = E(X_2 \mid X_1) = \rho X_1.$$

Calling $X^g = (X_1^g, X_2^g)^T$, the joint distribution of X and X^g is

$$\begin{pmatrix} X \\ X^g \end{pmatrix} \sim N_4\left(\mathbf{0}_4, \begin{pmatrix} \Sigma & \rho\Sigma \\ \rho\Sigma & \rho^2\Sigma \end{pmatrix}\right),$$

that does not verify the definition of knockoff variables. In particular, the distribution of X^g does not coincide with that of X because $\mathrm{Var}(X^g) \neq \mathrm{Var}(X)$.

Both concepts, ghost variables are knockoffs, are related with the conditional distributions used by Hooker et al. (2021) when they are known. To see that the three concepts are different, we use again the previous example of X following the bivariate normal model (5). Let X_1^c be a random variable distributed as $(X_1 \mid X_2)$:

$$X_1^c = E(X_1 \mid X_2) + \varepsilon_1 = \rho X_2 + \varepsilon_1,$$

where $\varepsilon_1 \sim N(0, \text{Var}(X_1 \mid X_2) = 1 - \rho^2)$ is independent of X. Analogously, let

$$X_2^c = E(X_2 \mid X_1) + \varepsilon_2 = \rho X_1 + \varepsilon_2,$$

where $\varepsilon_2 \sim N(0, \text{Var}(X_2 \mid X_1) = 1 - \rho^2)$ independent of X and ε_1. Let $X^c = (X_1^c, X_2^c)^T$. The joint distribution of X and X^c is

$$\begin{pmatrix} X \\ X^c \end{pmatrix} \sim N_4 \left(\mathbf{0}_4, \begin{pmatrix} \Sigma & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \right),$$

where

$$\mathbf{A} = \text{Corr}(X^c) = \begin{pmatrix} 1 & \rho^3 \\ \rho^3 & 1 \end{pmatrix} \text{ and } \mathbf{B} = \text{Corr}(X, X^c) = \begin{pmatrix} \rho^2 & \rho \\ \rho & \rho^2 \end{pmatrix}.$$

Observe that the variables following the conditional distributions are not knockoffs and that they are different from ghost variables.

Finally, following the classical regression formulae, the value of the bivariate conditional variable, given that $X = x$, is

$$(X^c \mid X = x) \sim N_2 \left(\rho \check{x}, (1 - \rho^2) \mathbf{I}_2 \right),$$

where $\check{x} = (x_2, x_1)^T$. Observe that there is still randomness in the distribution of $(X^c \mid X = x)$.

When the conditional distribution $(Z \mid X = x_i)$ is unknown, it can be estimated following the proposals of Tansey et al. (2022) when they describe the general Holdout Randomized Test (HRT). Then random samples from the estimated conditional distribution can be drawn. This last step can be done just once, or it can be repeated several times and then record the average results (this option implies an additional computational cost).

It is worth to say that the estimation of the conditional distribution models in HRT is a complicated task requiring a considerable computing effort. Tansey et al. (2022) model the conditional distribution of $X_j \mid X_{-j} = x_{-j}$ as a mixture of univariate Gaussian distributions. They fix the number of components in the mixture at 5. Then there are $5+5+(5-1) = 14$ conditional parameters to be estimated as functions of the $(p-1)$ values of x_{-j}. Following the proposal of Bishop (1994) on *mixture density networks*, Tansey et al. (2022) use a neural network with 14 neurons in the output layer (one for each parameter), instead of having just one output neuron as it happens when the goal is to estimate simply the conditional expectation $E(X_j \mid X_{-j} = x_{-j})$.

On the contrary, ghost variables requires only to estimate the conditional expectation using the regression model preferred by the user. For instance, linear or additive models (or their generalized versions, if the nature of X_j requires it) can be used. If there are many variables, it may be better to use lasso type estimation.

5 Relevance measures in linear regression

As we have stressed, the procedures for finding the relevance of a variable can be applied in any prediction model. However, in the general case exact results to compare the procedures cannot be found. Therefore, we will compare them in the linear case and, when a closed solution can be found, in the additive model.

5.1 Relevance by *loco* in linear regression

The estimated linear regressions are now $\hat{\mathbf{y}}_{1.X.z} = \mathbf{X}_1\hat{\beta}_x + \mathbf{z}_1\hat{\beta}_z$, and $\hat{\mathbf{y}}_{1.X} = \mathbf{X}_1\hat{\beta}_0$. Let $\text{se}(\hat{\beta}_z)$ be the estimated standard error of $\hat{\beta}_z$ and let $t_z = \hat{\beta}_z/\text{se}(\hat{\beta}_z)$ the standard t-test statistic for the null hypothesis $H_0 : \beta_z = 0$. Let $F_z = t_z^2$ be the F-statistic for testing the same null hypothesis. The relevance by leaving the covariate Z out evaluated in the training sample is

$$\text{RV}_{\text{loco}}^{\text{Train}}(Z) = \frac{1}{n_1}\sum_{i=1}^{n_1}(\hat{f}(x_{1.i}, z_{1.i}) - \hat{f}_p(x_{1.i}))^2/\widehat{\text{MSPE}}^{\text{Train}}(\hat{f}),$$

where $\widehat{\text{MSPE}}^{\text{Train}}(\hat{f})$ is the usual estimator of the residual variance, $\hat{\sigma}^2_{n_1}$. Standard computations in the linear model (see, e.g., Seber and Lee 2003, or the supplemental materials, where we have included them for the sake of completeness) lead to

$$\text{RV}_{\text{loco}}^{\text{Train}}(Z) = F_z/n_1,$$

It follows that evaluating the relevance of a variable by *loco* in the training sample is equivalent to computing the statistic for testing its significance and that we can compute the relevance by dividing the squared t statistic of the variable by the sample size. The same computations indicate that $\text{RV}_{\text{loco}}^{\text{Train}}(Z) = \hat{\beta}_z^2\hat{\sigma}^2_{z.x,n_1}/\widehat{\text{MSPE}}^{\text{Train}}(\hat{f})$, where $\hat{\sigma}^2_{z.x,n_1}$ is a consistent estimator of $\sigma^2_{z.x}$, the residual variance in the model $Z = X^T\alpha + \varepsilon_z$, computed from the training sample. This is a sampling version of the expression (3) obtained in Section 2.1.

When the relevance by *loco* is computed in the test sample (as we advocate) similar results are obtained. In this case the vectors of predicted values are $\hat{\mathbf{y}}_{2.X.z} = \mathbf{X}_2\hat{\beta}_x + \mathbf{z}_2\hat{\beta}_z$ and $\hat{\mathbf{y}}_{2.X} = \mathbf{X}_2\hat{\beta}_0$, and the relevance by *loco* of the variable Z is

$$\text{RV}_{\text{loco}}(Z) = \frac{(\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X})^T(\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X})}{(\mathbf{y}_2 - \hat{\mathbf{y}}_{2.X.z})^T(\mathbf{y}_2 - \hat{\mathbf{y}}_{2.X.z})}.$$

Then, it can be proved (detailed computation can be found in the supplemental materials; see also Theorem 3 in Hooker et al. 2021) that

$$\text{RV}_{\text{loco}}(Z) = \frac{F_z}{n_1}\frac{\hat{\sigma}^2_{z.x,n_1,n_2}}{\hat{\sigma}^2_{z.x,n_1}} = \frac{F_z}{n_1}\left(1 + O_p\left(\min\{n_1, n_2\}^{-1/2}\right)\right),$$

and

$$\mathrm{RV}_{\mathrm{loco}}(Z) = \hat{\beta}_z^2 \hat{\sigma}^2_{z.x,n_1,n_2} / \widehat{\mathrm{MSPE}}(\hat{f}),$$

where $\hat{\sigma}^2_{z.x,n_1,n_2}$ and $\hat{\sigma}^2_{z.x,n_1}$ are consistent estimators of the same parameter $\sigma^2_{z.x}$, the residual variance in the linear regression model $Z = X^T\alpha + \varepsilon_z$. This is another sampling version of the equation (3). It follows that, for large values of n_1 and n_2, $\hat{\sigma}^2_{z.x,n_1,n_2}/\hat{\sigma}^2_{z.x,n_1} \approx 1$, and then

$$\mathrm{RV}_{\mathrm{loco}}(Z) \approx F_z/n_1,$$

approximately the same relationship we have found when computing the relevance by *loco* in the training sample.

5.2 Relevance by random permutations in linear and additive models

We will analyze directly the additive model and show the results for linear regression as a particular case. Assume that an *additive model* is fitted in the training sample

$$\hat{f}(x,z) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{s}_j(x_j) + \hat{s}_{p+1}(z),$$

$\hat{\beta}_0 = \sum_{i=1}^{n_1} y_{1.i}/n_1$, $\sum_{i=1}^{n_1} \hat{s}_j(x_{1.i})/n_1 = 0$, $j = 1,\ldots,p$, and for identifiability reasons $\sum_{i=1}^{n_1} \hat{s}_{p+1}(z_{1.i})/n_1 = 0$. These identities are only approximately true when taking averages at the test sample. Observe that

$$\widehat{\mathrm{MSPE}}(\hat{f})\mathrm{RV}_{\mathrm{rp}}(Z) = \frac{1}{n_2}\sum_{i=1}^{n_2}(\hat{s}_{p+1}(z_{2.i}) - \hat{s}_{p+1}(z'_{2.i}))^2 =$$

$$2\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{s}_{p+1}(z_{2.i})^2 - 2\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{s}_{p+1}(z_{2.i})\hat{s}_{p+1}(z'_{2.i}) \approx$$

$$2\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{s}_{p+1}(z_{2.i})^2 = 2\widehat{\mathrm{Var}}(\hat{s}_{p+1}(Z)).$$

The approximation follows from the fact that $\sum_{i=1}^{n_2} \hat{s}_{p+1}(z_{2.i})\hat{s}_{p+1}(z'_{2.i})/n_2$ has expected value over random permutations equal to $\{\sum_{i=1}^{n_2} \hat{s}_{p+1}(z_{2.i})/n_2\}^2 \approx 0$ and variance of order $O(1/n_2)$.

In the special case of linear regression, $\hat{s}_j(x_j) = \hat{\beta}_j(x_j - \bar{x}_{1.j})$ for all $j = 1,\ldots,p$, and $\hat{s}_{p+1}(z) = \hat{\beta}_z(z - \bar{z}_1)$, where $\bar{x}_{1.j} = \sum_{i=1}^{n_1} x_{1.ij}/n_1$ and $\bar{z}_1 = \sum_{i=1}^{n_1} z_{1.i}/n_1$. Then, in this case

$$\widehat{\mathrm{MSPE}}(\hat{f})\mathrm{RV}_{\mathrm{rp}}(Z) \approx 2\hat{\beta}_z^2\frac{1}{n_2}\sum_{i=1}^{n_2}(z_{2.i} - \bar{z}_1)^2 \approx 2\hat{\beta}_z^2\frac{1}{n_2}\sum_{i=1}^{n_2}(z_{2.i} - \bar{z}_2)^2,$$

and

$$\mathrm{RV}_{\mathrm{rp}}(Z) \approx 2\hat{\beta}_z^2\widehat{\mathrm{Var}}(Z)/\widehat{\mathrm{MSPE}}(\hat{f}),$$

a sampling version of the expression found for random variables in Section 2.2 (see also Theorem 1 in Hooker et al. 2021). Observe that $\mathrm{RV}_{\mathrm{rp}}(Z)$ is not a multiple of the statistic F_z used to test the null hypothesis $H_0 : \beta_z = 0$, because the variance of $\hat{\beta}_z$, the OLS estimator of β_z, is not a multiple of $\widehat{\mathrm{MSPE}}(\hat{f})/\widehat{\mathrm{Var}}(Z)$, except in the particular case that $\mathbf{X}_1$ and $\mathbf{z}_1$ are uncorrelated.

5.3 Relevance by ghost variables in linear regression

To get a model-agnostic proposal, the training sample $\mathcal{S}_1$ should be used only through the estimated prediction function. Therefore, we propose to estimate $\mathbb{E}(Z \mid X)$ using the data in the test sample $\mathcal{S}_2$. Let us assume that the regression function of Y given (X, Z) is linear and that it is estimated by OLS in the training sample $(\mathbf{X}_1, \mathbf{z}_1, \mathbf{y}_1) \in \mathbb{R}^{n_1 \times (p+2)}$. Let $\hat{\beta}_x$ and $\hat{\beta}_z$ be the estimated coefficients, and let $\hat{\sigma}^2_{n_1}$ be the estimated residual variance. Let $(\mathbf{X}_2, \mathbf{z}_2, \mathbf{y}_2) \in \mathbb{R}^{n_2 \times (p+2)}$ be the test sample in matrix format. We fit the linear model $Z = X\alpha + \varepsilon_z$ by OLS in the test sample to obtain the ghost values for $\mathbf{z}_2$:

$$\hat{\mathbf{z}}_{2.2} = \mathbf{X}_2 \hat{\alpha}_2,$$

with $\hat{\alpha}_2 = (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T \mathbf{z}_2$. Then,

$$\hat{\mathbf{y}}_{2.X.z} = \mathbf{X}_2 \hat{\beta}_x + \mathbf{z}_2 \hat{\beta}_z,$$

when using (X, Z) as predictors, and

$$\hat{\mathbf{y}}_{2.X.\hat{z}} = \mathbf{X}_2 \hat{\beta}_x + \hat{\mathbf{z}}_{2.2} \hat{\beta}_z,$$

when using $(X, \hat{Z}_X)$, that is replacing Z by the ghost variable. Let us define

$$\hat{\sigma}^2_{n_1,n_2} = \widehat{\mathrm{MSPE}}(\hat{f}) = \frac{1}{n_2} (\mathbf{y}_2 - \hat{\mathbf{y}}_{2.X.z})^T (\mathbf{y}_2 - \hat{\mathbf{y}}_{2.X.z}),$$

which is an estimator of the residual variance depending on both, the training and the test samples. Therefore, the *relevance by a ghost variable* of the variable Z is

$$\mathrm{RV}_{\mathrm{gh}}(Z) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2} (\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X.\hat{z}})^T (\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X.\hat{z}}).$$

The following result states a sampling version of equation (4) in Section 2.2, and connects the relevance measure based on ghost variables with the F-test for β_z. The proof can be found in Appendix A.2.

Theorem 1 *Assume that the regression function of Y over (X, Z) is linear, that it is estimated by OLS, and that the ghost variable for Z is also estimated by OLS. Then*

$$\mathrm{RV}_{gh}(Z) = \frac{F_z}{n_1} \frac{\hat{\sigma}^2_{z.x,n_2}}{\hat{\sigma}^2_{z.x,n_1}} \frac{\hat{\sigma}^2_{n_1,n_2}}{\hat{\sigma}^2_{n_1}} = \frac{F_z}{n_1} \left(1 + O_p\left(\min\{n_1, n_2\}^{-1/2}\right)\right),$$

and

$$\mathrm{RV}_{gh}(Z) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \hat{\beta}_z^2 \hat{\sigma}^2_{z.x,n_2},$$

where $\hat{\sigma}^2_{z.x,n_2}$ *and* $\hat{\sigma}^2_{z.x,n_1}$ *are consistent estimators of the same parameter* $\sigma^2_{z.x}$ *(the residual variance in the linear regression model* $Z = X^T\alpha + \varepsilon_z$*), the first one depending on the test sample, and the second one on the training sample.*

The properties of the relevance measure based on ghost variables established in Theorem 1 are similar to those stated in Hooker et al. (2021, Theorem 3) for variable importance measures based on sampling conditional distributions and/or re-learning (fitting twice the predictive model). It is worth noting that using ghost variables is easier than implementing the proposals of Hooker et al. (2021) because ghost variables only require modeling and estimating conditional expectations, instead of having to deal with the whole conditional distribution. Additionally, ghost variables do not require re-learning. A simulation study comparing ghost variables, conditional distributions, knockoff variables, as well as *loco* and random permutations, is presented in Section 6.

In the linear regression model, Theorem 1 establishes a parallelism between deleting the variable Z and replacing it by a ghost variable. Nevertheless, using the ghost variable has two clear advantages: first, only one model has to be fitted in the training sample (the model with all the explanatory variables), and second, the estimated prediction function is the only element we have to save from the training sample (and, consequently, our proposal is model-agnostic).

As a last remark, we would like to emphasize that the relevance by a ghost variable allows us to approximate a very relevant statistic in the linear regression model, namely the F-statistic for testing the null hypothesis $H_0 : \beta_z = 0$. This approximation only requires the estimated prediction function from a training sample and a test sample. Therefore, the relevance measure by a ghost variable allows us to compute a pseudo F-statistic for algorithmic predictive model as $\mathrm{RV}_{\mathrm{gh}}(Z)n_1$. In fact, the examples provided in Sections 6 and 8 include fitting neural networks, random forest or additive models, as well as linear models estimated by OLS or by lasso.

6 Comparison of relevance measures by simulation

We introduce two examples of synthetic data (of small and medium sizes, respectively) for which several prediction models are fitted (neural networks, random forest, and linear models estimated by OLS or lasso). For each fitted model, the relevance of the predictive variables is computed using different approaches. Data from both examples are repeatedly simulated with three main goals: to calibrate the computational efficiency of the relevance methods, to check the validity of their results under different scenarios, and to evaluate their adaptability when the model characteristics change.

6.1 Example 1. A model with 10 explanatory variables

We present a simulated example following the simple design proposed by Hooker et al. (2021), Section 2: a multiple linear regression model with 10 explanatory variables uniformly distributed on $[0, 1]$, all independent except perhaps the first two of them, which could be possibly correlated through a Gaussian copula with $\rho = 0$ or $\rho = 0.9$. Data are generated from the model

$$Y = x_1 + x_2 + x_3 + x_4 + x_5 + 0x_6 + 0.5x_7 + 0.8x_8 + 1.2x_9 + 1.5x_{10} + \varepsilon,$$

where $\varepsilon \sim N(0, 0.1^2)$. We have repeated 50 times the generation of a training set of size 2000, plus a test set of size 1000.

We start comparing the computational performance of our ghost variables proposal with perturbing with estimated conditional distributions, for using which we have taken the Python code provided by Tansey et al. (2022), available at https://github.com/tansey/hrt. For this comparison we have implemented our proposal in Python as the estimation procedure for conditional distributions used by Tansey et al. (2022) is not easily transferable to R because it relies on a neural network to fit a conditional mixture of normal distributions, as proposed by Bishop (1994).

We have fitted a linear model by OLS, and then a random forest (using the function `RandomForestRegressor` from `sklearn.ensemble`, Pedregosa et al. 2011, with default parameter values). Variable importance is computed in three different ways: by *loco*, by ghost variables, and by replacing an explanatory variable in the test set with a random draw from the estimated conditional distribution.

To facilitate comparisons with Hooker et al. (2021), the relevance has been computed based on relative differences in MSPE ($\mathbf{RV}^*$, as defined in equation (2) for *loco*) when a feature is removed from the model (in *loco*) or replaced by its perturbed version. Additionally, for ghost variables the relevance has also been computed based on differences in predictions ($\mathbf{RV}$, as in equation (1)).

Table 1 summarizes the computing times (in seconds) required by each combination of fitting model and variable relevance computation. It can be seen that measuring the feature relevance using ghost variables is extremely much faster than using random values from the estimated conditional distributions (even if only one randomization is done: the really time consuming task is the estimation of the conditional mixture). Additionally, relevance by *loco* is faster than ghost variables when OLS is used to fit a linear model, but it is much slower when a random forest has to be fitted for each explanatory variable when it is left out.

Regarding the importance relative rankings for *loco*, ghost variables and random data from the estimated conditional distribution, they are quite similar in the simulated examples, as it can be seen in the following Figure 1, which shows the average importance rank (lowest to highest) of each of the 10 explanatory variables according to the different relevance measures. We use

Table 1 Computation times (in seconds) of different relevance measures applied to two regression models. Implementation done in Python.

	Relevance measures		
Models	*loco*	ghost variables	estimated conditional distribution
Linear model (OLS)	0.23	0.34	5102.78
Random forest	1341.62	29.83	12771.48

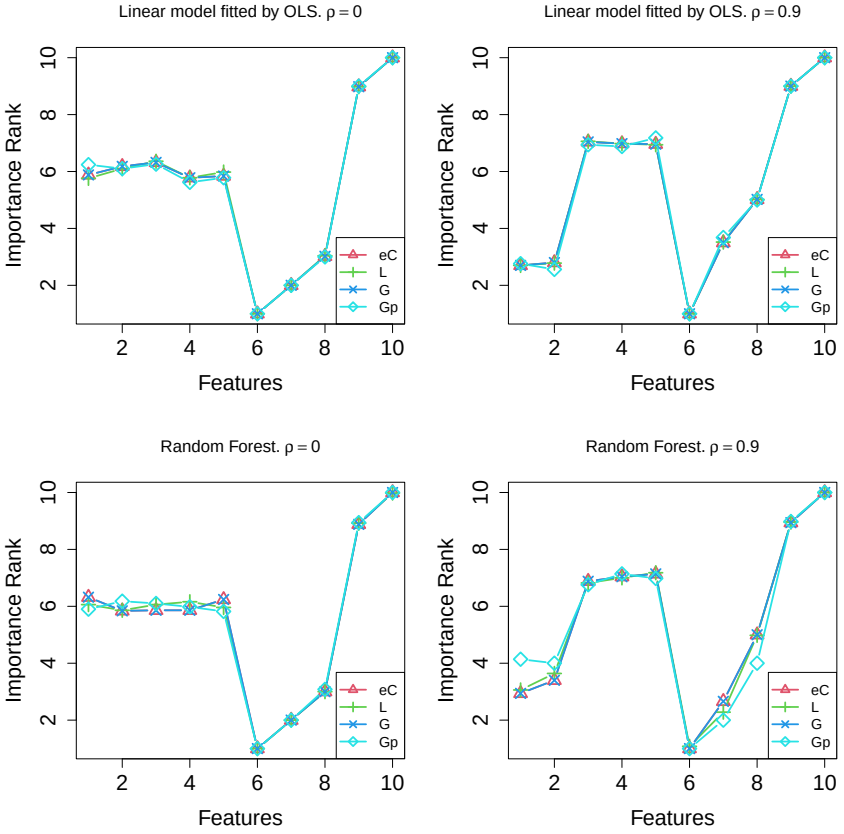

Fig. 1 Relative rankings of the explanatory variables according to different relevance measures applied two regression models. Implementation done in Python.

labels `eC` for estimated conditional distribution, `L` for *loco*, `G` for ghost variables using differences in MSPE, and `Gp` for ghost variables using differences in predictions. These results are similar to those reported by Hooker et al. (2021).

The relevance scores obtained by ghost variables are similar to those obtained by *loco* (graphics not included here), and they are in general lower than (in median, the 70% of) those obtained when using random data from

Table 2 Computation times (in seconds) of different relevance measures applied three regression models. Implementation done in R.

	loco	ghost variables	true conditional distribution	random permutations	knockoffs
Time (in seconds)	4267.84	49.78	42.93	43.07	46.76

the estimated conditional distribution. Nevertheless the three methods lead to similar importance relative rankings, as Figure 1 shows.

So, we conclude that using ghost variables to assess the relevance of the explanatory variables gives in this case similar results to those obtained when estimating conditional distributions, but is much more efficient from the computational point of view. A similar statement applies when comparing ghost variables with *loco*, if the prediction model is hard to be fitted (as random forest).

Given the very large computing time required by the estimation of conditional distributions, we have decided to remove this method from further simulation studies. This decision allows us to work exclusively in R. This way we can use part of the R code provided as electronic supplemental material by Hooker et al. (2021). Now, we simulate in R data following the previous linear model and we compare 5 ways to assess relevance of features (the last 4 of them following the general scheme of perturbing the features in the test set): *loco*, ghost variables, random data from the true conditional distribution which is known in this simulated example, random permutations, and using knockoffs (Model-X second-order multivariate Gaussian knockoff variables as implemented in the R package `knockoff`, Patterson and Sesia 2022). Three predicting models have been fitted: a linear model fitted by OLS, a random forest (using the function `randomForest` from the R package `randomForest`, Liaw and Wiener 2002, with default parameter values), and a 1-hidden layer neural network (using `nnet` from package `nnet`, Venables and Ripley 2002, with 20 neurons in the hidden layer; we have proceed as in Hooker et al. (2021) to choose the best neural network among 10 randomly initialized fits).

In our implementation in R, for each simulated data set the four sets of perturbed values of the explanatory variables are generated only once, and the three models are fitted to the same data set. Therefore there is no way to assign computing times to each specific combination of fitting model and variable relevance computation. The computing times (in seconds) required by each method of variable relevance computation are shown in Table 2. It follows that, in terms of computing time, using ghost variables is comparable with generating data from the known conditional distributions (which is not feasible in real settings), using random permutations, or using knockoffs, and that they are much faster than using *loco*.

Regarding the quality of the computed relevance measures, Figure 2 shows the results. The main conclusions are the following:

- The random permutation method is giving bad results when there are some inter-dependent features, as expected from our arguments as well as those given by Hooker et al. (2021).
- Ghost variables and knockoffs perform similarly to using random data from the true conditional distributions, with the advantage that the former are feasible in a real setting while the latter is not.
- Ghost variables and knockoffs perform similar to *loco* (except perhaps when fitting neural networks), with the advantage that the former are much faster that the latter.

It follows that using ghost variables or using knockoffs to compute relevance of features are comparable strategies regarding the quality of the resulting relevance measures, as well as regarding computational efficiency, and that both are preferred to other alternatives considered in our simulation study.

Let us make a final remark on comparing ghost variables and knockoffs. When using ghost variables the practitioner has to propose regression models of each explanatory variable over the others, and then fit these models. This is a routine process which is easily implemented in any standard platform (R or Python, for instance), even if the linearity assumption is not fulfilled by our data. On the other hand, generating knockoffs variables is difficult even in the most standard settings. Moreover, when the data are far from well mimicked with Model-X Gaussian knockoffs there is no easy way to generate knockoffs.

To explore the effect of nonlinear relation between explanatory variables on both, ghost variables and knockoffs, we modify the previous multiple linear regression model introducing a nonlinear dependence between the first two explanatory variables (X_1, X_2). First we generate data (θ, R) uniformly in the set $\{[0, \pi/2] \cup [\pi, 3\pi/2]\} \times [0.9, 1]$. Then we define

$$X_1 = (R\cos(\theta) + 1)/2, \;\; X_2 = (R\sin(\theta) + 1)/2.$$

This way X_1 and X_2 are both in $[0, 1]$ and they present a non-linear dependence pattern. We proceed as before with the only difference that now the ghost variables are fitted using a generalized additive model using the function `gam` in the R package `mgcv` (Wood 2017). Figure 3 shows the results for the random forest and the neural network (results for the linear model fitted by OLS are similar to those of the neural network). It can be seen that the performance of ghost variables is similar to that of using the true conditional distribution (which is unknown in real settings) and close to *loco* (mainly for random forests). Random permutations gives very poor results: the reported relevance for X_1 and X_2 is larger than that of X_3, X_4 and X_5, even if the there is a strong dependence between X_1 and X_2. Using knockoffs gives results midway between *loco* and random permutations, because the MX Gaussian knockoff method is not able to reproduce the joint distribution of X_1 and X_2, which is very far from joint normality.

We conclude that the simplicity and flexibility of the ghost variable procedure are a clear advantage with respect to using knockoffs.

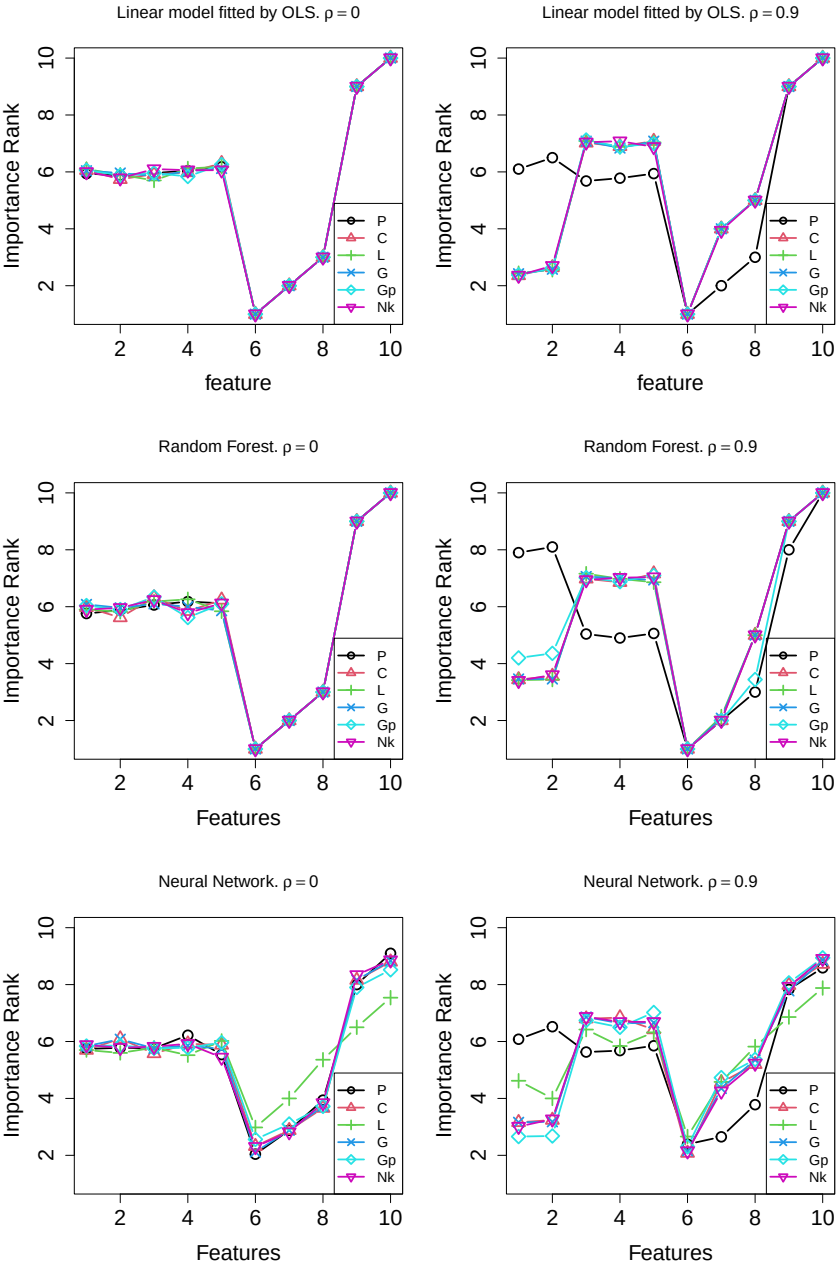

Fig. 2 Relative rankings of the explanatory variables according to different relevance measures applied three regression models. Implementation done in R.

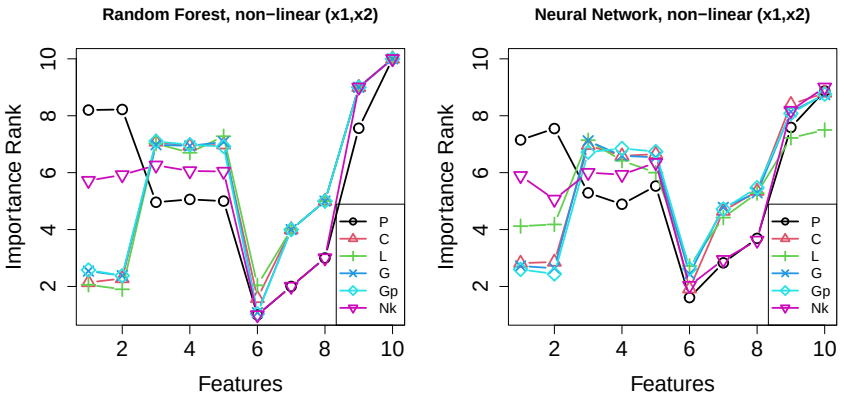

Fig. 3 Relative rankings of the explanatory variables according to different relevance measures applied three regression models. A nonlinear dependence pattern has been simulated between the first two explanatory variables.

6.2 Example 2. A large model with 100 features

We simulate now data following a linear model with 100 explanatory Gaussian variables, grouped into three subsets with 5, 45 and 50 variables each, respectively. The 5 variables in the first group are independent standard normal. In the second group, the 45 variables are marginally standard normal but they are correlated to each other with correlation coefficient $\rho_2 = 0.95$. The 50 variables in the third group are independent normal with zero mean and standard deviation $\sigma_3 = 2$. Variables in different groups are independent from each other. For each observed set of explanatory variables, $x_1, \ldots, x_{100}$, the response variable Y is generated from the linear model

$$Y = \sum_{j=1}^{100} \beta_j x_j + \varepsilon,$$

where ε follows a $N(0,1)$ and $\beta_j = \gamma_1 = 0.5$, for $j = 1, \ldots, 5$, $\beta_j = \gamma_2 = 1$, for $j = 6, \ldots, 50$, and $\beta_j = \gamma_3 = 0.1$, for $j = 51, \ldots, 100$.

The parameters that control the simulated example, ρ_2, σ_3, γ_1, γ_2, γ_3, have been fixed at the previously indicated values with the objective of having similar relevance (measured by ghost variables) for variables in groups 2 and 3, and higher relevance for variables in the first group (see Figure 4). We look for such a configuration because later (in Section 8) we use this Example to illustrate the use of the relevance matrix, defined in Section 7. There we will see that using the relevance matrix allows to distinguish between the groups of variables 2 and 3, which can not be differentiated using only individual relevance measures.

The following procedure has been repeated 100 times. We generate a training set of size 1000, and a test set of size 500. The training set is used to fit the

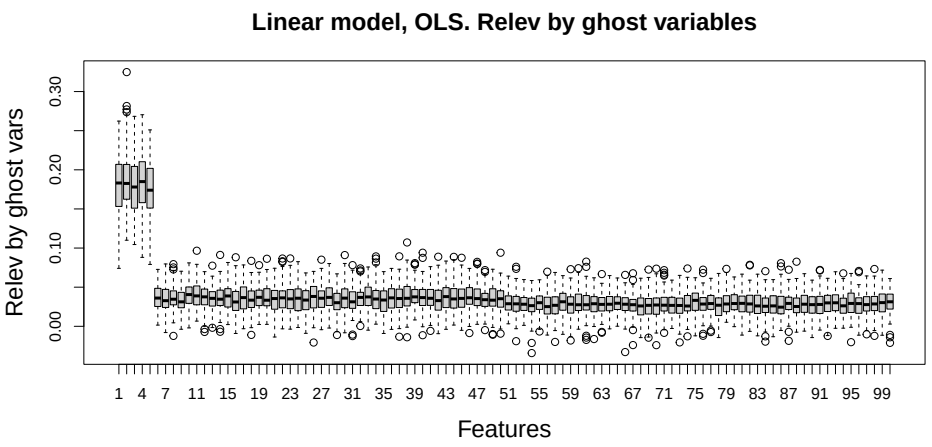

Fig. 4 Relevance of the explanatory variables in the linear model with 100 features of Example 2, estimated by OLS. The relevance is measured as the relative increment in MSPE when each feature is replaced by its ghost variables in the test sample. The boxplots show the relevance values measured in 100 simulated data sets (training set of size 1000, test set of size 500).

linear model, first by OLS and then using lasso (as implemented in the R package `glmnet`, Friedman et al. 2010, using 10-fold cross-validation for choosing the penalty parameter). The test set is used to measure relevance of variables by different methods: *loco* (`L`, with a cost of 1327.08 seconds in computing time), ghost variables in its both versions (`Gp` compares predictions, as **RV** in equation (1), and `G` compares MSPEs, as $\mathbf{RV}^*$ in equation 2; the required time was 220.26 seconds in total), knockoffs perturbations (`Nk`, 53.78 seconds), and random permutations (`P`, 35.51 seconds).

Figure 4 shows, for each explanatory feature, the boxplot of its relevance values in the 100 simulated data sets, when ghost variables is used to measure relevance. It can be seen that the relevance distributions are similar for variables in groups 2 and 3, as desired.

Figure 5 show the relative ranks, according to different relevance measures, of the 100 explanatory averaged over the 100 simulations. In the left hand side graphic, corresponding to OLS estimation of the linear model, it can be seen that *loco* (`L`) gives similar results to ghost variables in any of its both versions, comparing predictions (`Gp`) or comparing MSPEs (`G`). These three relevance measures are approximately equivalent to compute the F (or t) statistics for testing that the coefficients in the linear model are equal to zero. Figure 5 also shows that random permutations (`P`) and knockoffs (`Nk`) do not agree with the other relevance measures. Using knockoffs allows to detect variables in group 1 as the most relevant but this method switch the relevance order of variables in groups 2 and 3.

When the lasso estimation is analyzed (right hand side of Figure 5) we can see that the results are qualitative similar to those of the OLS estimation,

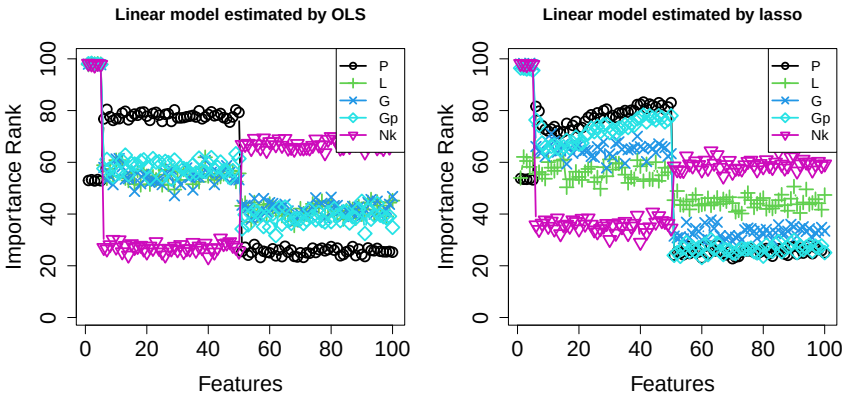

Fig. 5 Average relative rankings, according to different relevance measures, of the 100 explanatory variables in the large linear model of Example 2, estimated by OLS and by lasso.

with certain particularities that we list now. Loco (L) assigns similar average relative ranks to variables in groups 1 and 2, these being only slightly larger than those it assigns to variables in group 3. Now ghost variables results are no longer so similar to those of *loco* for variables in groups 2 and 3, but its version comparing MSPEs is still the closest method to *loco*.

We conclude that measuring relevance based on ghost variables has given results as least as good as those of *loco*, with a much lower computational cost. Using knockoffs, a method that is faster than ghost variables in this example, has given unsatisfactory relevance results. The same applies to using random permutations, whose unsatisfactory performance has already been reported in this article and others as Hooker et al. (2021).

7 Understanding sets of variables: The relevance matrix

In complex models with many possibly grouped variables, the measure of relevance could be more important for groups of variables than for individual variables. The ideas presented in the previous sections can be generalized easily when Z is a set of variables, instead of a single one. However, with many variables a key problem is to find the relevant sets of variables with grouped effects. Finding relevant groups of variables is similar to the detection of groups of influential cases and outliers in regression, where a standard approach (similar to *loco*) is to remove each data in turns and compute the effect of these deletions on the values predicted by the model (or on the estimates of the parameters). Peña and Yohai (1995) introduced the *influence matrix* by first looking at the influence vectors that measure the effect of deleting each observation on the vector of forecasts for the whole data set and then computing the $n \times n$ symmetric matrix that has the scalar products between these vectors. Thus the influence matrix has in the diagonals Cook's statistics and outside the diagonal the scalar products of these effects. These authors showed that the eigen-structure of the influence matrix contains useful information to detect

influential subsets or multiple outliers avoiding masking effects. We propose a similar idea by computing a *case-variable* relevance matrix (denoted by $\mathbf{A}$ below).

We will present the case when variables are replaced by their ghost variables and, in the supplemental material, their replacement by random permutations. Observe that similar definitions could be done for other relevance measures based on perturbing variables, as those based on conditional distributions or on knockoffs.

We consider again the prediction of Y from the p components of X through the regression function $f(x) = \mathbb{E}(Y \mid X = x)$ estimated in a training sample $(\mathbf{X}_1, \mathbf{y}_1) \in \mathbb{R}^{n_1 \times (p+1)}$ that will be tested in the test sample $(\mathbf{X}_2, \mathbf{y}_2) \in \mathbb{R}^{n_2 \times (p+1)}$.

Let $\mathbf{x}_{2.1}, \ldots, \mathbf{x}_{2.p}$ be the columns of $\mathbf{X}_2$. For $j = 1, \ldots, p$, let $\mathbf{X}_{2.[j]} = (\mathbf{x}_{2.1}, \ldots, \mathbf{x}_{2.(j-1)}, \mathbf{x}_{2.(j+1)}, \ldots, \mathbf{x}_{2.p})$ be the matrix $\mathbf{X}_2$ without the j-th column, and $\mathbf{H}_{2.[j]} = \mathbf{X}_{2.[j]}(\mathbf{X}_{2.[j]}\mathbf{X}_{2.[j]})^{-1}\mathbf{X}_{2.[j]}^{\mathrm{T}}$ be the projection matrix on the column space of $\mathbf{X}_{2.[j]}$. Let $\hat{\mathbf{x}}_{2.j} = \mathbf{H}_{2.[j]}\mathbf{x}_{2.j}$ be projection of $\mathbf{x}_{2.j}$ over the column space of the other columns of $\mathbf{X}_2$. We will take $\hat{\mathbf{x}}_{2.j}$ as the j-th ghost variable. Note that alternative regression models (additive models, for instance, or non linear models) could be also used to define the ghost variable. Let

$$\mathbf{X}_{2.\hat{j}} = (\mathbf{x}_{2.1}, \ldots, \mathbf{x}_{2.j-1}, \hat{\mathbf{x}}_{2.j}, \mathbf{x}_{2.j+1}, \ldots, x_{2.p})$$

be the regressor matrix in the test sample where the j-th variable has been replaced by the j-th ghost variable. We use $\hat{\mathbf{Y}}_2 = \hat{f}(\mathbf{X}_2)$ to denote the n_2-dimensional column vector of forecasts with all the variables and $\hat{\mathbf{Y}}_{2.\hat{j}} = \hat{f}(\mathbf{X}_{2.\hat{j}})$ to the vector of forecast with the j-th ghost variable. Define the $n_2 \times p$ matrix of forecast changes as

$$\mathbf{A} = (\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_{2.\hat{1}}, \ldots, \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_{2.\hat{p}}),$$

where the element a_{ij} of $\mathbf{A}$, $i = 1, \ldots, n_2$, $j = 1, \ldots, p$, measures the change in the response prediction for the i-th case in the test sample, when the j-th variable has been replaced by its ghost variable. Finally, we define the *relevance matrix* as the $p \times p$ matrix

$$\mathbf{V} = \frac{1}{\widehat{\mathrm{MSPE}}(\hat{f})} \frac{1}{n_2} \mathbf{A}^{\mathrm{T}}\mathbf{A}.$$

Then, the element (j, k) of $\mathbf{V}$ is

$$v_{jk} = \frac{1}{\widehat{\mathrm{MSPE}}(\hat{f})} \frac{1}{n_2} \sum_{i=1}^{n_2} (\hat{f}(x_{2.j.i}) - \hat{f}(x_{2.\hat{j}.i}))(\hat{f}(x_{2.k.i}) - \hat{f}(x_{2.\hat{k}.i})),$$

where $x_{2.j.i}$ and $x_{2.\hat{j}.i}$ are, respectively, the i-th element of $\mathbf{x}_{2.j}$ and $\hat{\mathbf{x}}_{2.j}$. In particular, the diagonal of the relevance matrix $\mathbf{V}$ has elements

$$v_{jj} = \text{RV}_{\text{gh}}(X_j),\ j = 1, \ldots, p.$$

The advantage of working with the matrix $\mathbf{V}$, instead of just computing univariate relevance measures, is that $\mathbf{V}$ contains additional information in its out-of-the-diagonal elements, which we are exploring through the examination of its eigen-structure. If, for $j = 1 \ldots, p$, $(1/n_2)\sum_i \hat{f}(x_{2.j.i}) = (1/n_2)\sum_i \hat{m}(x_{2.\hat{j}.i})$ then $\mathbf{V}$ is proportional to the variance-covariance matrix of $\mathbf{A}$.

7.1 The relevance matrix for linear regression

Consider the particular case that $\hat{m}(x)$ is the OLS estimator of a multiple linear regression. Then

$$\hat{f}(x) = x^{\text{T}}\hat{\beta},\ \text{with}\ \hat{\beta} = (\mathbf{X}_1^{\text{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\text{T}}\mathbf{Y}_1.$$

Therefore, the vector of predicted values in the test sample is $\hat{\mathbf{Y}}_2 = \mathbf{X}_2\hat{\beta}$, and

$$\hat{\sigma}^2_{n_1,n_2} = \widehat{\text{MSPE}}(\hat{f}) = \frac{1}{n_2}(\mathbf{Y}_2 - \hat{\mathbf{Y}}_2)^{\text{T}}(\mathbf{Y}_2 - \hat{\mathbf{Y}}_2),$$

which is an estimator of the residual variance in the regression of Y over X. Writing $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\text{T}}$, the predicted values when using the j-th ghost variable is

$$\hat{\mathbf{Y}}_{2.\hat{j}} = \mathbf{X}_{2.\hat{j}}\hat{\beta} = \mathbf{X}_2\hat{\beta} - (\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})\hat{\beta}_j = \hat{\mathbf{Y}}_2 - (\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})\hat{\beta}_j,$$

the matrix $\mathbf{A}$ is

$$\mathbf{A} = (\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_{2.\hat{1}}, \ldots, \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_{2.\hat{p}}) = (\mathbf{X}_2 - \hat{\mathbf{X}}_2)\text{diag}(\hat{\beta}),$$

where $\hat{\mathbf{X}}_2$ is the matrix with each variable replace by its ghost variable. The relevance matrix is

$$\mathbf{V} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}}\frac{1}{n_2}\text{diag}(\hat{\beta})(\mathbf{X}_2 - \hat{\mathbf{X}}_2)^{\text{T}}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)\text{diag}(\hat{\beta}) =$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}}\text{diag}(\hat{\beta})\mathbf{G}\text{diag}(\hat{\beta}),$$

where $\mathbf{G} = (1/n_2)(\mathbf{X}_2 - \hat{\mathbf{X}}_2)^{\text{T}}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)$. The elements (j, k) of $\mathbf{G}$ and $\mathbf{V}$ are, respectively,

$$g_{jk} = \frac{1}{n_2}(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})^{\text{T}}(\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k}),\ \text{and}\ v_{jk} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}}\hat{\beta}_j\hat{\beta}_k g_{jk}.$$

Observe that, in the regression of $\mathbf{x}_{2.j}$ over $\mathbf{X}_{2.[j]}$, $\hat{\sigma}^2_{[j]} = g_{jj}$ is the residual variance estimation that uses n_2 as denominator. The following result summarizes the properties of the relevance matrix $\mathbf{V}$ and, in particular, its relationship with the partial correlation matrix. The proof can be found in Appendix A.3.

Theorem 2 *Let* $\mathbf{P}$ *be the matrix that contains the partial correlation coefficients in the test sample as non-diagonal elements and has* -1 *in the diagonal. Then*

$$\mathbf{G} = \frac{1}{n_2}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)^\top(\mathbf{X}_2 - \hat{\mathbf{X}}_2) = -\mathit{diag}(\hat{\sigma}_{[1]}, \ldots, \hat{\sigma}_{[p]})\, \mathbf{P}\, \mathit{diag}(\hat{\sigma}_{[1]}, \ldots, \hat{\sigma}_{[p]}),$$

and consequently

$$\mathbf{V} = -\frac{1}{\hat{\sigma}^2_{n_1,n_2}}\, \mathit{diag}(\hat{\beta})\, \mathit{diag}(\hat{\sigma}_{[1]}, \ldots, \hat{\sigma}_{[p]})\, \mathbf{P}\, \mathit{diag}(\hat{\sigma}_{[1]}, \ldots, \hat{\sigma}_{[p]})\, \mathit{diag}(\hat{\beta}).$$

Therefore $RV_{gh}(X_j)$*, the* j*-th element of the diagonal of* $\mathbf{V}$*, admits the alternative expression*

$$RV_{gh}(X_j) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \hat{\beta}^2_j \hat{\sigma}^2_{[j]},$$

and the partial correlation coefficient between variables j *and* k *when controlling by the rest of variables can be computed as*

$$\hat{\rho}_{jk.R} = -\frac{g_{jk}}{\sqrt{g_{jj} g_{kk}}} = -\frac{v_{jk}}{\sqrt{v_{jj} v_{kk}}}.$$

The expressions for the partial correlation coefficient appearing in Theorem 2 reminds the well known formula

$$\hat{\rho}_{jk.R} = -\frac{s^{jk}}{\sqrt{s^{jj} s^{kk}}},$$

where the s^{jk} is the element (j, k) of $\mathbf{S}_2^{-1}$, the inverse of the covariance matrix of the test sample $\mathbf{X}_2$, $\mathbf{S}_2$. This coincidence, and the observation that s^{jj} is the inverse of $(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})^{\mathrm{T}}(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})$ (a consequence of the inverse formula for a block matrix; see Appendix A.1), imply the next Corollary.

Corollary 1 Let $\mathbf{S}_2$ be the covariance matrix of the test sample $\mathbf{X}_2$. $\mathbf{G}$ and $\mathbf{S}_2^{-1}$ verify that

$$\mathbf{G} = \frac{n_2 - 1}{n_2}\, \mathrm{diag}(\hat{\sigma}^2_{[1]}, \ldots, \hat{\sigma}^2_{[p]})\, \mathbf{S}_2^{-1}\, \mathrm{diag}(\hat{\sigma}^2_{[1]}, \ldots, \hat{\sigma}^2_{[p]}).$$

The relevance matrix when random permutations are used (say $\tilde{\mathbf{V}}$) is analyzed in the supplemental material in the case of multiple linear regression. There it is proved that

$$\tilde{\mathbf{V}} \approx \frac{1}{\hat{\sigma}^2_{n_1,n_2}} 2\, \mathrm{diag}(\hat{\beta})\, \mathrm{diag}(S_1, \ldots, S_p)\, \mathbf{R}\, \mathrm{diag}(S_1, \ldots, S_p)\, \mathrm{diag}(\hat{\beta}),$$

where S_j^2 is the sample variance (computed dividing by n_2) of $\mathbf{x}_j$, and $\mathbf{R}$ is the correlation matrix of the test sample $\mathbf{X}_2$. This expression suggests that the eigen-structure of $\tilde{\mathbf{V}}$ will probably be related with the principal component analysis of the test sample explanatory matrix $\mathbf{X}_2$.

On the contrary, the eigen-structure of $\mathbf{V}$ differs from the principal components of the explanatory variables. This leads us to expect that the study of $\mathbf{V}$ reveals relevant knowledge that would be hidden if we limit ourselves to analyzing the covariance structure of the explanatory variables.

8 Relevance matrix in action

We analyze now the performance of the relevance matrix in practice. We use synthetic data in Section 8.1 and 8.2, while real data are used in Section 8.3.

8.1 Analyzing one case from Example 1

We consider a data set generated as in Example 1 (Section 6.1), following the linear model with 10 explanatory variables proposed in Hooker et al. (2021), with the first two variables strongly related ($\rho = 0.9$). We have fitted three models to these data in R: a linear model fitted by OLS, a random forest and a one-hidden-layer neural network, as described in Section 6.1. Linear model and neural network give very good fits, both with values of the multiple R-squared around 0.99. Random forest multiple R-squared is 0.88.

We compute the variable relevance, as well as the relevance matrix, for the neural network model using ghost variables (the results for the linear model are very similar). Figure 6 summarizes our findings. The relevance of each variable is represented in the upper left plot, and they are according to the design of the generating linear model. Variables X_{10} and X_9 are the most relevant (in this order), followed by X_5, X_3 and X_4, almost tied, then X_8 and X_7, and finally variables X_1 and X_2 have a similar smaller relevance. Variable X_6 has almost null relevance. Note that this relevance information, which is not included in the standard output of the fitting neural networks routines, is as informative about the statistical significance of each explanatory variable as the t-values reported in the standard output of a linear model.

The relevance matrix $\mathbf{V}$ provides information on the joint effect that variables have on the response. The upper middle plot in Figure 6 shows the eigenvalues of $\mathbf{V}$, and the other plots represent the components of each eigenvector. This eigen-structure reveals the following facts. The first and second largest eigenvalue are linked to eigenvectors defined by the columns of the case-variable relevance matrix $\mathbf{A}$ corresponding to variables X_{10} and X_9, respectively. The eigenvectors 3, 4 and 5 are jointly related to the variables X_3, X_4 and X_5, and they correspond to 3 eigenvalues with similar values. Therefore, these three eigenvectors expand a quite spherical 3-dimensional subspace. For a different sample, probably the components of these eigenvectors in the columns of matrix $\mathbf{A}$ corresponding to variables X_3, X_4 and X_5 would be quite different, even if the spanned subspace would remain approximately equal.

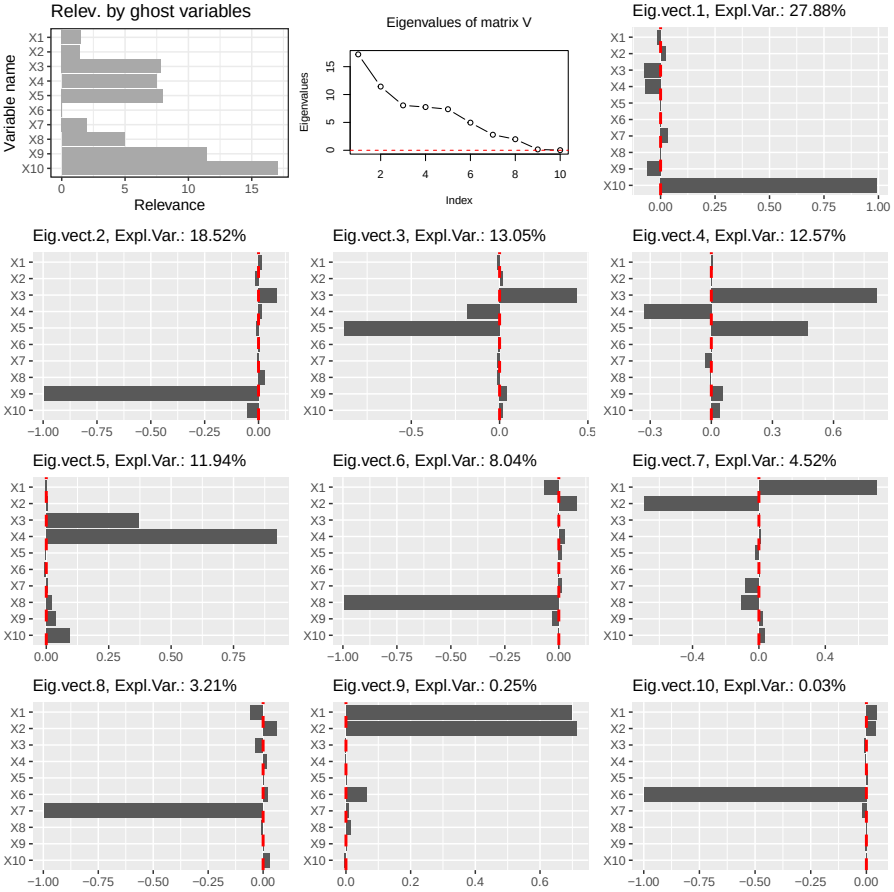

Fig. 6 Ghost variables relevance matrix analysis in one data set generated according to Example 1 in Section 6.1.

Eigenvectors 6 and 8 are given by variables X_8 and X_7, respectively. The eigenvectors 7 and 9 jointly correspond to the strongly related variables X_1 and X_2, and they are associated with 2 quite different eigenvalues. This indicates that the subspace spanned by the first two columns of the case-variable relevance matrix $\mathbf{A}$ is elliptic, with high eccentricity, a consequence of the strong linear dependence between the columns in $\mathbf{A}$ corresponding to variables X_1 and X_2. The last eigenvector, corresponding to an almost null eigenvalue, is associated with X_6.

8.2 Analyzing one case from Example 2

We consider a data set generated as in Example 2 (Section 6.2), following the linear model with 100 explanatory variables. We fit a linear model, first using OLS and then using lasso. For both fitted models, the relevance matrix using ghost variables (with differences in predictions, as in equation (1)) have

been computed. The results are summarized in Figure 7 (the large number of explanatory variables forces to use a format different from that of Figure 6). The upper graph shows the variable relevance ranks for both, OLS and lasso. The first 5 variables are clearly the most relevant in both fits. When using lasso estimation, the second group of 45 correlated variables are detected as most relevant than the third group, integrated by 50 uncorrelated variables. Nevertheless, variables in both groups have similar relevance when using OLS as fitting procedure. In this last case, we are seeing below that the relevance matrix reveals the different nature of variables in groups 2 or 3.

Let $\mathbf{V}$ be the relevance matrix computed as proposed in Section 7. The second graph in Figure 7 is a scree plot showing the eigenvalues of $\mathbf{V}$ in decreasing order. To select which eigenvectors we should explore further, we follow the usual recommendations in Principal Component Analysis (see, for instance, Section 8.3 of Johnson et al. 2002): eigenvectors associated with the largest eigenvalues are relevant; an eigenvector corresponding to an eigenvalue near zero may indicate an almost exact linear dependency in the columns of the case-variable relevance matrix $\mathbf{A}$, pointing out a possible group of explanatory variables having a joint effect on the model predictions; large steps in the scree plot occurs between consecutive eigenvalues with potentially different behavior, and it might be interesting to explore the eigenvectors associated to both.

In this example, just looking for large steps in the scree plot (we select those that are marked as outliers in the boxplot of the step sizes, in logarithmic scale) results on the identification of 9 potentially interesting eigenvectors: the seven largest ones plus the two smallest ones. The 9 small plots in the lower part of Figure 7 represent the coefficients of each column of $\mathbf{A}$ in the definition of these 9 eigenvectors. It follows that eigenvectors 1 to 5 are associated with the first 5 most relevant variables. Eigenvectors 6, 7 and 99 seem not to be of special interest. Finally, the last eigenvector identifies the variables in group 2, from X_6 to X_{50}, as a set of variables having a joint effect over the model predictions. This is something that we new in advance (because the simulated model was designed this way) but that was overlooked by the individual relevance measures. This is a clear example of the added value that calculating the relevance matrix can imply.

8.3 A real data example: Rent housing prices

We present a real data example on rental housing, coming from the Spanish real estate search portal Idealista (www.idelista.com) which allows customers to search for housing based on several criteria (geography, number of rooms, price, and various other filters) among the offers posted by other customers. We started working from the data set downloaded from www.idelista.com by Alejandro Germán-Serrano on February 27th 2018 (available at https://github.com/seralexger/idealista-data; accessed April 12th 2019). This data set contained 67201 rows (posted rental offers actives at the download date)

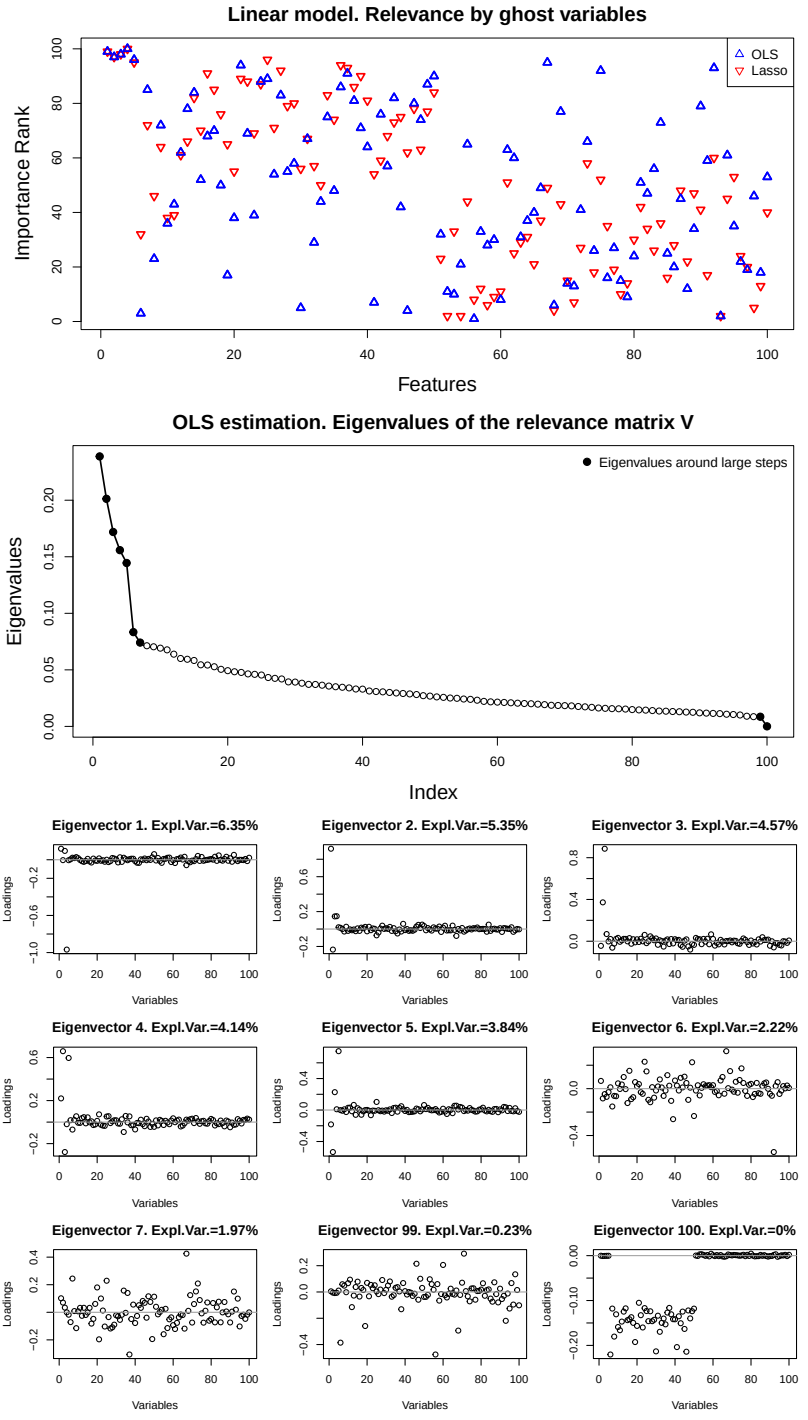

Fig. 7 Ghost variables relevance matrix analysis in one data set generated according to Example 2 in Section 6.2.

Table 3 List of variables used in the rent housing prices example. The response variable is `log.price`, and other are the explanatory variables. The district price level indicator `categ.distr` has been computed in Barcelona and Madrid separately. For each district (there are $N_B = 10$ districts in Barcelona and $N_M = 21$ in Madrid) the third quartile of `price` is computed. Then these N values (where $N = N_B$ or N_M) are classified by their own quartiles, and values -3, -1, 1 and 3 are assigned to them accordingly. Finally this district value is used to define `categ.distr` for all the houses in each district.

`log.price`	Monthly rental price, the response variable (in logarithms).
`Barcelona`	1 for houses in Barcelona, 0 for those in Madrid.
`categ.distr`	An indication of the district price level, taking the values -3, -1, 1 and 3. See the caption for details.
`type.chalet`	These 4 variables are the binarization of the
`type.duplex`	original variable `type` with 5 levels:
`type.penthouse`	flat (the most frequent), chalet, duplex,
`type.studio`	penthouse and studio.
`floor`	Floor where the house is located.
`hasLift`	1 if the house has lift, 0 otherwise.
`floorLift`	`abs(floor)*(1-hasLift)`
`log.size`	Surface, in squared meters (in logarithms).
`exterior`	1 if the house is exterior, 0 otherwise.
`rooms`	Number of bedrooms.
`bathrooms`	Number of bathrooms.
`hasParkingSpace`	1 if the house has a parking space, 0 otherwise.
`ParkingInPrice`	1 if the parking space is included in the price, 0 otherwise.
`log_activation`	logarithm of the number of days since the first activation of the post.

and 19 attributes, corresponding to all cities in Spain. Some offers where activated for the first time in 2010.

We have selected posts corresponding to Madrid and Barcelona (16480 rows) and finally work with 17 variables (some of them already in the original data set, other calculated from the original attributes) listed in Table 3.

In order to predict the logarithm of prices as a function of the other 16 explanatory variables, we have fitted three predictive models: a linear regression, an additive model, and a neural network. For each model, the variables relevance has been computed by ghost variables (a 70% of the data are used as training set, and the rest as test set). The standard outputs of the linear and the additive models offer a rich information about the statistical significance of each explanatory variable. Then the relevance analysis represents a complementary information, which in most cases confirms the standard one, although matrix relevance can add new lights. The situation is different for the neural network model: in this case the relevance analysis will provide genuine new insights on the importance of the explanatory variables, or groups of them.

The variable relevance results for the three models are broadly consistent. We show below those corresponding to the neural network. The results for the linear model and for the additive model are accessible as supplemental materials.

Table 4 Rent housing prices: Output of the neural network model.

```
# > nnet.logprice
#
# a 16-10-1 network with 181 weights
# inputs: Barcelona categ.distr type.chalet type.duplex type.penthouse
# type.studio floor hasLift floorLift log.size exterior rooms bathrooms
# hasParkingSpace ParkingInPrice log_activation
#
# output(s): log.price
#
# options were - linear output units  decay=0.5
#
# > 1-mean(nnet.logprice$residuals^2)
# [1] 0.8009131
```

A one-hidden-layer neural network has been fitted using the `nnet` function from the R package `nnet` (Venables and Ripley 2002). The response and the explanatory variables were centered and scaled before the fitting. Tuning parameters, `size` (number of neurons in the hidden layer) and `decay` parameter, are chosen using `caret` (Kuhn 2018) by 10-fold cross validation (in the training set). The candidates values for `size` were 10, 15, and 20, and they were 0, 0.1, 0.3, and 0.5 for `decay`. Finally the chosen values were `size=10` and `decay=0.5`. With these values, the whole training sample was used to fit the neural network and the results were stored in the object `nnet.logprice`.

Table 4 shows the little information obtained when printing the output of the `nnet` function. Additionally, the multiple R-squared has been computed and printed, with a value of 0.80 (the variance of the response variable is 1 in the test sample because the data has been previously standardized). It is a little bit larger than those for the linear and the additive models (not reported here; see the supplemental materials). You can see that the output in Table 4 does not provide any insight about which explanatory variables are more responsible for that satisfactory fit. The equation that defines the neural network is not helpful for this, as it explains the predictions as a non linear combination of 10 (the number of hidden nodes) variables which are linear combinations of the original ones. As indicated in the output, 181 parameters have been fitted. Relevance measures will be of help in this respect.

Results on relevance by ghost variables for the fitted neural network are shown in Figure 8. We can see (first row, first column plot) that `log.size` is the most relevant variable, followed by `categ.distr`, `type.chalet`, `Barcelona`, `bathrooms`, and `log_activation`, in this order. The relevance of `rooms`, `floor` and `type.studio` is lower.

Regarding the analysis of the relevance matrix $\mathbf{V}$, only the 10 eigenvectors explaining more than 1% of the total relevance are plotted. The first eigenvector accounts for the 43% of total relevance, and it is associated with the size of houses (mainly with `log.size`, and to a lesser extent with `bathrooms` and `rooms`). The second eigenvector (18% of total relevance) is mostly related

to the district price level (`categ.distr`) the third one (10%) to `type.chalet` and the fourth one (7%) to `Barcelona`. The variables `bathrooms` and `rooms` appear together in several eigenvectors (with larger influence in eigenvectors 5th and 9th), always accompanying other variables. This fact indicates that this pair of variables probably have a joint effect on model predictions. Similar joint behavior present `Barcelona` and `log_activation`. Variables referring to other types of houses appear at eigenvectors 5th, 6th, 8th, and 10th. The 7th eigenvector is mainly related with `floor`. Eigenvectors 11th to 16th (not shown here) are related with the 7 less relevant variables, some of which appear only in one eigenvector (as `exterior`) and other appear in pairs in several ones (as `hasLift` and `floorLift`, which are closely related).

9 Conclusions

We have defined the relevance of a variable in a complex model by its contribution to out-of-sample prediction and proposed a new way to measure this contribution: to compare the predictions of the model in the test set with those obtained when the variable of interest is replaced (in the test set) by its ghost variable, which is defined as a prediction of the variable by using the other explanatory variables. We have also shown that this approach has advantages over other approaches: ghost variables require much less computing time than leaving-one-covariate-out or using estimated conditional distributions, they give better results than random permutations when the covariates are dependent, and they involve a much more flexible modeling stage than using knockoffs. We have proved that in linear regression this approach is related to the F-statistic used to check the significance of the variable and, therefore, the computation of the relevance by ghost variables in a complex predictive model is an extension of the significance concept to other models in which this concept is not usually considered. With many dependent variables, the relevance of a each variable separately is less useful than considering the joint contribution of sets of variables. Taken that into account, we have introduced the relevance matrix as a way to explore joint effects in out-of-sample prediction. In the linear model, we have proved the relationship between the relevance matrix and the matrix of partial correlations of the explanatory variables.

Finally, we would like to emphasize the strength that research in Interpretable Machine Learning has taken in recent years. Paraphrasing Breiman (2001), we want to alert statisticians to be aware that traditional Statistics is no longer the only way to reach *understandable* conclusions from data.

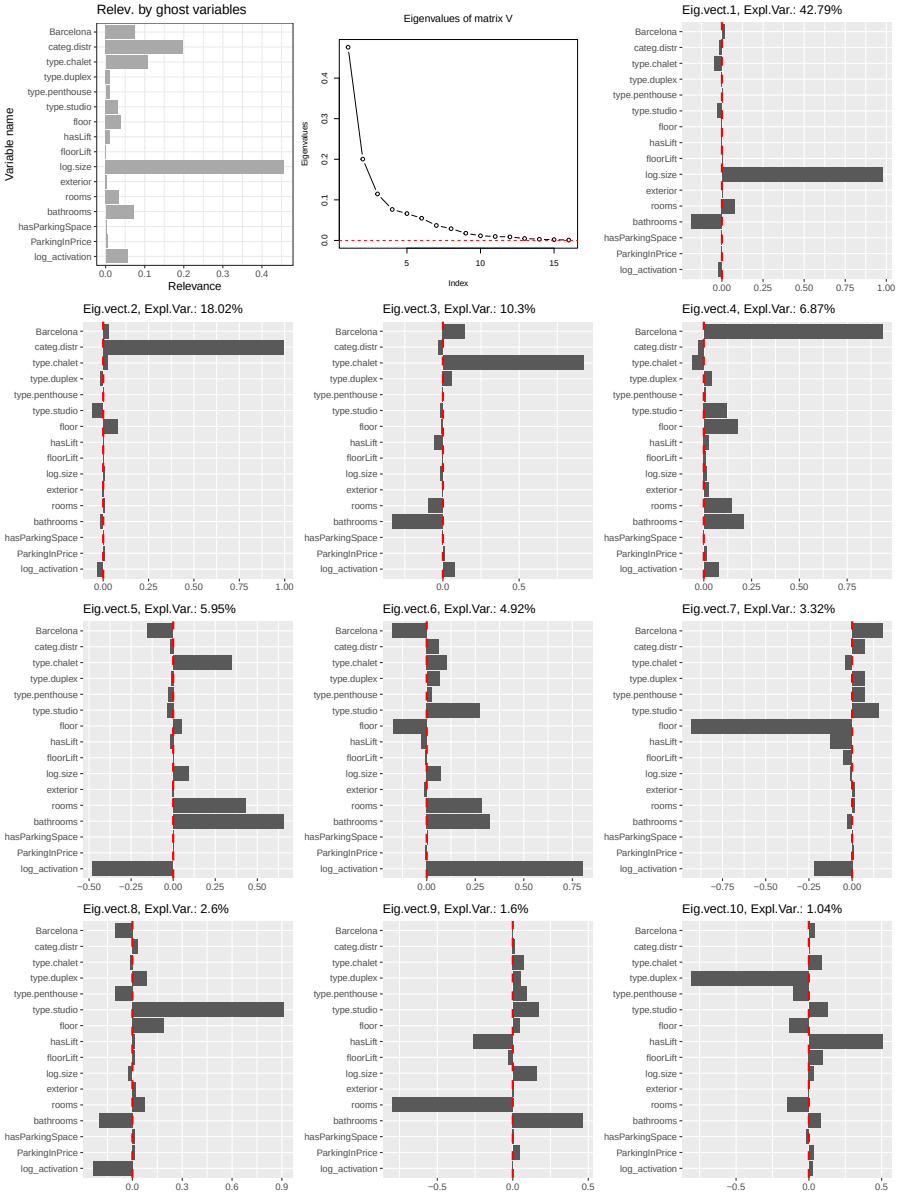

Fig. 8 Rent housing prices: Relevance by ghost variables for the neural network model.

A Proof of the results

A.1 Updating formula for the OLS estimator

Let $\mathbf{b} \in \mathbb{R}^p$, $c \in \mathbb{R}$, and $\mathbf{A} \in \mathbb{R}^{p \times p}$, an invertible matrix. The expression of the inverse of an invertible block matrix is as follows:

$$\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{k}\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{A}^{-1} & -\frac{1}{k}\mathbf{A}^{-1}\mathbf{b} \\ -\frac{1}{k}\mathbf{b}^T\mathbf{A}^{-1} & \frac{1}{k} \end{pmatrix},$$

where $k = c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$.

Consider the linear regression with responses in vector $\mathbf{y} \in \mathbb{R}^n$ and regression matrix $(\mathbf{X}, \mathbf{z}) \in \mathbb{R}^{n \times (p+1)}$. The OLS estimated regression coefficients are given by

$$\begin{pmatrix} \hat{\beta}_x \\ \hat{\beta}_z \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{z} \\ \mathbf{z}^T\mathbf{X} & \mathbf{z}^T\mathbf{z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{z}^T \end{pmatrix} \mathbf{y}$$

Then, using the formula for the inverse of a block matrix, we have that

$$\begin{pmatrix} \hat{\beta}_x \\ \hat{\beta}_z \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1} + \frac{1}{k}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}\mathbf{z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} & -\frac{1}{k}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z} \\ -\frac{1}{k}\mathbf{z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} & \frac{1}{k} \end{pmatrix} \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{z}^T\mathbf{y} \end{pmatrix}$$

with

$$k = \mathbf{z}^T\mathbf{z} - \mathbf{z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z} = \mathbf{z}^T\mathbf{z} - \hat{\mathbf{z}}_x^T\hat{\mathbf{z}}_x = (\mathbf{z} - \hat{\mathbf{z}}_x)^T(\mathbf{z} - \hat{\mathbf{z}}_x),$$

where $\hat{\mathbf{z}}_x = \mathbf{H}_x\mathbf{z}$ and $\mathbf{H}_x = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the *hat matrix* in any linear regression over $\mathbf{X}$. Then, calling $\hat{\beta}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, $\hat{\alpha} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$, we have

$$\begin{pmatrix} \hat{\beta}_x \\ \hat{\beta}_z \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 + \frac{1}{k}\hat{\alpha}\hat{\mathbf{z}}_x^T\mathbf{y} - \frac{1}{k}\hat{\alpha}\mathbf{z}^T\mathbf{y} \\ \frac{1}{k}(\mathbf{z} - \hat{\mathbf{z}}_\mathbf{x})^T\mathbf{y} \end{pmatrix}$$

and finally

$$\hat{\beta}_x = \hat{\beta}_0 - \hat{\alpha}\hat{\beta}_z.$$

Therefore, the following updating formula is derived:

$$\hat{\mathbf{y}}_{x,z} = \mathbf{X}\hat{\beta}_x + \mathbf{z}\hat{\beta}_z = \mathbf{X}\hat{\beta}_0 - (\mathbf{X}\hat{\alpha})\hat{\beta}_z + \mathbf{z}\hat{\beta}_z = \hat{\mathbf{y}}_x + (\mathbf{z} - \hat{\mathbf{z}}_x)\hat{\beta}_z.$$

A.2 Proof of Theorem 1

By definition, the *relevance by a ghost variable* of the variable Z is

$$\mathrm{RV}_{\mathrm{gh}}(Z) = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2} (\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X.\hat{z}})^T (\hat{\mathbf{y}}_{2.X.z} - \hat{\mathbf{y}}_{2.X.\hat{z}}) =$$

$$\frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2} \hat{\beta}_z (\mathbf{z}_2 - \hat{\mathbf{z}}_{2.2})^T (\mathbf{z}_2 - \hat{\mathbf{z}}_{2.2}) \hat{\beta}_z = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{1}{n_2} \hat{\beta}_z^2 k_{\mathrm{gh}} = \frac{1}{\hat{\sigma}^2_{n_1,n_2}} \frac{\hat{\sigma}^2_{n_1} F_z}{n_1} \frac{k_{\mathrm{gh}}/n_2}{k/n_1},$$

where $k_{\text{gh}} = (\mathbf{z}_2 - \hat{\mathbf{z}}_{2.2})^T(\mathbf{z}_2 - \hat{\mathbf{z}}_{2.2})$ and k has been defined in Appendix A.1. Observe that both, $\hat{\sigma}^2_{n_1}$ and $\hat{\sigma}^2_{n_1,n_2}$, are consistent estimators of the residual variance in the linear regression of Y over (X, Z). The proof concludes by observing that $\hat{\sigma}^2_{z.x,n_2} = k_{\text{gh}}/n_2$ is an estimator of the residual variance in the linear regression model $Z = X^T\alpha + \varepsilon_z$, as they also are $k/(n_1 - p)$ and $\hat{\sigma}^2_{z.x,n_1} = k/n_1 = ((n_1 - p)/n_1)(k/(n_1 - p))$. The expression involving the O_p notation is derived by standard arguments for the limit of a quotient.

A.3 Proof of Theorem 2

Lemma 1 Let $\mathbf{a}$ and $\mathbf{b}$ be two non-null vectors of $\mathbb{R}^d$. Let $\mathbb{P}_{\mathbf{b}}(\mathbf{a})$ be the projection vector of $\mathbf{a}$ over $\mathbf{b}$, and let $\alpha(\mathbf{a}, \mathbf{b})$ be the angle between $\mathbf{a}$ and $\mathbf{b}$. Then

$$\cos\left(\alpha\left(\mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a}), \mathbf{b} - \mathbb{P}_{\mathbf{a}}(\mathbf{b})\right)\right) = -\cos(\alpha(\mathbf{a}, \mathbf{b})).$$

Proof Given that $\cos(\alpha(\mathbf{a}, \mathbf{b})) = \mathbf{a}^T\mathbf{b}/(\|\mathbf{a}\|\,\|\mathbf{b}\|)$ and $\mathbb{P}_{\mathbf{b}}(\mathbf{a}) = \cos(\alpha(\mathbf{a}, \mathbf{b}))\|\mathbf{a}\|(\mathbf{b}/\|\mathbf{b}\|) = (\mathbf{a}^T\mathbf{b})\mathbf{b}/\|\mathbf{b}\|^2$, it follows that

$$\mathbf{a}^T\mathbb{P}_{\mathbf{a}}(\mathbf{b}) = \mathbb{P}_{\mathbf{b}}(\mathbf{a})^T\mathbf{b} = \mathbf{a}^T\mathbf{b},\ \ \mathbb{P}_{\mathbf{b}}(\mathbf{a})^T\mathbb{P}_{\mathbf{a}}(\mathbf{b}) = \cos^2(\alpha(\mathbf{a}, \mathbf{b}))\,\mathbf{a}^T\mathbf{b}$$

and

$$\|\mathbb{P}_{\mathbf{b}}(\mathbf{a})\|^2 = \mathbb{P}_{\mathbf{b}}(\mathbf{a})^T\mathbb{P}_{\mathbf{b}}(\mathbf{a}) = \cos^2(\alpha(\mathbf{a}, \mathbf{b}))\|\mathbf{a}\|^2.$$

By the Pythagoras Theorem,

$$\|\mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a})\|^2 = \|\mathbf{a}\|^2 - \|\mathbb{P}_{\mathbf{b}}(\mathbf{a})\|^2 = \sin^2(\alpha(\mathbf{a}, \mathbf{b}))\|\mathbf{a}\|^2.$$

Finally,

$$\cos\left(\alpha\left(\mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a}), \mathbf{b} - \mathbb{P}_{\mathbf{a}}(\mathbf{b})\right)\right) = \frac{(\mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a}))^T(\mathbf{b} - \mathbb{P}_{\mathbf{a}}(\mathbf{b}))}{\|\mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a})\|\|\mathbf{b} - \mathbb{P}_{\mathbf{a}}(\mathbf{b})\|} =$$

$$\frac{\mathbf{a}^T\mathbf{b} - \mathbb{P}_{\mathbf{b}}(\mathbf{a})^T\mathbf{b} - \mathbf{a}^T\mathbb{P}_{\mathbf{a}}(\mathbf{b}) + \mathbb{P}_{\mathbf{b}}(\mathbf{a})^T\mathbb{P}_{\mathbf{a}}(\mathbf{b})}{\sin^2(\alpha(\mathbf{a}, \mathbf{b}))\|\mathbf{a}\|\|\mathbf{b}\|} =$$

$$\frac{-(1 - \cos^2(\alpha(\mathbf{a}, \mathbf{b})))\mathbf{a}^T\mathbf{b}}{\sin^2(\alpha(\mathbf{a}, \mathbf{b}))\|\mathbf{a}\|\|\mathbf{b}\|} = -\cos(\alpha(\mathbf{a}, \mathbf{b})).$$

□

Proof of Theorem 2.

We start proving that the matrix

$$\mathbf{G} = \frac{1}{n_2}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)^{\text{T}}(\mathbf{X}_2 - \hat{\mathbf{X}}_2)$$

has generic non-diagonal element $g_{jk} = \hat{\rho}_{jk.R}\hat{\sigma}_{[j]}\hat{\sigma}_{[k]}$ for $j \neq k$, where $\hat{\rho}_{jk.R}$ is the partial correlation coefficient between variables j and k when controlling by the rest of variables, and $\hat{\sigma}^2_{[j]}$ is the j-th element in the diagonal of G:

$$g_{jj} = \hat{\sigma}^2_{[j]} = \frac{1}{n_2}(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})^{\text{T}}(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j}).$$

It is equivalent to proof that

$$\hat{\rho}_{jk.R} = -\frac{g_{jk}}{\sqrt{g_{jj}g_{kk}}} = -\frac{(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})^{\mathrm{T}}(\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k})}{\sqrt{(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})^{\mathrm{T}}(\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j})}\sqrt{(\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k})^{\mathrm{T}}(\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k})}},$$

that is, we have to prove that the cosinus of the angle between the vector of residuals $\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j}$ and $\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k}$ is equal to minus the cosinus of the angle between the vector of residuals $\mathbf{a} = \mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j.R}$ and $\mathbf{b} = \mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k.R}$, obtained when regressing $\mathbf{x}_{2.j}$ and $\mathbf{x}_{2.k}$ respectively over $\mathbf{R} = \mathbf{X}_{2.[jk]}$, the matrix with columns $\mathbf{x}_{2.1}, \ldots, \mathbf{x}_{2.p}$, except $\mathbf{x}_{2.j}$ and $\mathbf{x}_{2.k}$.

We use now the notation $\mathbb{P}_{\mathbf{U}}(\mathbf{x})$ to denote the projection of the vector $\mathbf{x}$ over the linear space $\mathbf{U}$, and $\{\mathbf{R}, \mathbf{x}\}$ for the subspace generated by the columns of the matrix $\mathbf{R}$ and the vector $\mathbf{x}$. Observe that

$$\hat{\mathbf{x}}_{2.j} = \mathbb{P}_{\{\mathbf{R},\mathbf{x}_{2.k}\}}(\mathbf{x}_{2.j}) = \mathbb{P}_{\{\mathbf{R},\mathbf{b}\}}(\mathbf{x}_{2.j}) = \mathbb{P}_{\{\mathbf{R},\mathbf{b}\}}(\hat{\mathbf{x}}_{2.j.R} + \mathbf{a}) =$$

$$\mathbb{P}_{\{\mathbf{R},\mathbf{b}\}}(\hat{\mathbf{x}}_{2.j.R}) + \mathbb{P}_{\{\mathbf{R},\mathbf{b}\}}(\mathbf{a}) = \hat{\mathbf{x}}_{2.j.R} + \mathbb{P}_{\mathbf{b}}(\mathbf{a}).$$

Therefore

$$\mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j} = \mathbf{x}_{2.j} - \hat{\mathbf{x}}_{2.j.R} - \mathbb{P}_{\mathbf{b}}(\mathbf{a}) = \mathbf{a} - \mathbb{P}_{\mathbf{b}}(\mathbf{a}).$$

Analogously, $\mathbf{x}_{2.k} - \hat{\mathbf{x}}_{2.k} = \mathbf{b} - \mathbb{P}_{\mathbf{a}}(\mathbf{b})$. A direct application of Lemma 1 finishes the proof that $\hat{\rho}_{jk.R} = -g_{jk}/\sqrt{g_{jj}g_{kk}}$. The other statements in the Theorem follow directly from the relation $\mathbf{V} = \mathrm{diag}(\hat{\beta})\mathbf{G}\mathrm{diag}(\hat{\beta})/\hat{\sigma}^2_{n_1,n_2}$.

References

Barber, R.F. and E.J. Candès. 2015. Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5): 2055–2085 .

Barredo-Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115 .

Biecek, P. and T. Burzykowski. 2021. *Explanatory model analysis: Explore, explain and examine predictive models.* Chapman and Hall/CRC.

Bishop, C.M. 1994. Mixture density networks. Aston University.

Breiman, L. 2001. Statistical modeling: The two cultures. *Statistical Science* 16: 199–231 .

Candès, E., Y. Fan, L. Janson, and J. Lv. 2018. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3): 551–577 .

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1): 1–22 .

Gregorutti, B., B. Michel, and P. Saint-Pierre. 2015. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics and Data Analysis* 90: 15–35 .

Gregorutti, B., B. Michel, and P. Saint-Pierre. 2017. Correlation and variable importance in random forests. *Statistics and Computing 27*(3): 659–678 .

Hooker, G., L. Mentch, and S. Zhou. 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing 31*(82). https://doi.org/10.1007/s11222-021-10057-z .

Johnson, R.A., D.W. Wichern, et al. 2002. *Applied multivariate statistical analysis* (5th ed.). Prentice Hall.

Kuhn, M. 2018. *caret: Classification and Regression Training*. R package version 6.0-81. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan and T. Hunt.

Lei, J., M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*(523): 1094–1111 .

Liaw, A. and M. Wiener. 2002. Classification and regression by randomforest. *R News 2*(3): 18–22 .

Masís, S. 2021. *Interpretable Machine Learning with Python*. Packt Publishing Ltd.

Molnar, C. 2019. *Interpretable Machine Learning*. Lulu. com.

Patterson, E. and M. Sesia 2022. *knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.5.

Peña, D. and V.J. Yohai. 1995. The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 145–156 .

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011.

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830 .

Ribeiro, M.T., S. Singh, and C. Guestrin. 2016a. Model-agnostic interpretability of machine learning. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA.

Ribeiro, M.T., S. Singh, and C. Guestrin 2016b. 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*, pp. 1135–1144. ACM.

Seber, G.A. and A.J. Lee. 2003. *Linear regression analysis* (2nd ed.). John Wiley & Sons.

Tansey, W., V. Veitch, H. Zhang, R. Rabadan, and D.M. Blei. 2022. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics 31*(1): 151–162. https://doi.org/10.1080/10618600.2021.1923520 .

Venables, W.N. and B.D. Ripley. 2002. *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Wood, S.N. 2017. *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC Press.

Zhu, R., D. Zeng, and M.R. Kosorok. 2015. Reinforcement learning trees. *Journal of the American Statistical Association 110*(512): 1770–1784 .

SUPPLEMENTAL MATERIALS

Additional results and data analysis: Results on relevance by *loco*, and on relevance matrix by random permutations are stated and proved. Moreover, in the real data example, the relevance analysis by ghost variables is included for the linear and the additive models.
R-scripts and datasets: The code to reproduce the computations and graphics can be found at https://github.com/pedrodelicado/GhostVariables.