An overview of variance-based importance measures in the linear regression context: comparative analyses and numerical tests

Laura Clouvel[1], Bertrand Iooss[2*], Vincent Chabridon[2], Marouane Il Idrissi[2], and Frédérique Robin[1]

[1] EDF R&D, PERICLES Department, Saclay, France

[2] EDF R&D, PRISME Department, Chatou, France & SINCLAIR AI Lab., Saclay, France

Abstract

One of the most fundamental issues in many socio-environmental studies is the identification of causal effects and influential variables related to phenomena of interest. In the context of regression analysis, importance measures are effective tools for feature selection and model interpretation, allowing for the ranking of the most influential regressors. In particular, variance-based importance measures (VIMs) are a prominent topic in the field of statistics, as well as in the emerging field of global sensitivity analysis. This is due to their accessible interpretation as variance shares of the explained variable. This work focuses on the linear regression model and aims to provide an updated overview of the most well-founded methods, mainly from comparative analyses and numerical tests on various toy cases. The paper also addresses some of the practical challenges that arise, including the case of dependent inputs and high input dimensionality. The practical relevance of these tools is demonstrated through empirical studies on simulated data and public datasets. The Supplementary Material C also presents the use of VIMs in a classification context, specifically via the logistic linear regression model.

Keywords

multicollinearity; proportional values; relative weight analysis; sensitivity analysis; variance decomposition

Code availability

The codes and datasets used in this paper are available at: https://gitlab.com/LauraClouvel/toydata/.

1. Introduction

Identifying causal effects and influential variables related to various phenomena is a fundamental concern in many socio-environmental studies (Razavi et al., 2020). In the context of regression analysis, *importance measures* are valuable tools for insightful feature selection and model interpretation, as they enable the ranking of explanatory variables (also known as "inputs" or "regressors") based on their influence (Kruskal, 1987; Grömping, 2015). Numerous methods exist to quantify the relative importance of inputs in models predicting a specific explained variable of interest (also known as the "output"). Among these methods, *variance-based importance measures* (VIMs) are particularly popular due to their clear interpretation as shares of the output's variance (Genizi, 1993; Budescu, 1993; Johnson & LeBreton, 2004; Bi, 2012; Iooss et al., 2022). Practically, VIMs are essential for data analysis and the post-hoc interpretation of learned models (Darlington & Hayes, 2017; Molnar et al., 2020; Lepore et al., 2022). Furthermore, their properties have spurred their adoption in the emerging field of *global sensitivity analysis* (GSA) of model outputs, where their versatility and ease of

Correspondence:
Contact B. Iooss at bertrand.iooss@edf.fr

Cite this article as:
Clouvel, L., Iooss, B., Chabridon, V., Idrissi, M.I., & Robin, F.
An overview of variance-based importance measures in the linear regression context: comparative analyses and numerical tests
Socio-Environmental Systems Modelling, vol. 7, 18681, 2025, doi:10.18174/sesmo.18681

Socio-Environmental Systems Modelling
An Open-Access Scholarly Journal
http://www.sesmo.org

estimation offer significant practical advantages (Saltelli et al., 2000; Da Veiga et al., 2021; Antoniadis et al., 2021).

In GSA, VIMs derived from linear regression analysis often form the foundational elements of any preliminary study, as highlighted in various methodological reviews (e.g., Helton et al., 2006; Iooss & Lemaître, 2015; Wei et al., 2015; Borgonovo & Plischke, 2016). However, a lack of awareness or poor understanding of GSA among practitioners can lead to significant flaws in data or model interpretation (Saltelli et al., 2020). For instance, the controversial findings of Sovacool et al. (2020), which suggested that higher nuclear energy adoption in a country does not correlate with lower carbon emissions unlike renewable energy, exemplify such issues. Their conclusions, drawn from multiple regression analyses of datasets from 123 countries (encompassing carbon emissions, renewable electricity production fraction, and nuclear electricity production fraction), have faced criticism for numerous statistical biases and errors (Wagner, 2021; Perez, 2022). Furthermore, a substantial portion of the GSA literature focused on VIMs in linear models seems to overlook crucial aspects that emerged during the historical development of these measures within the statistical research community. For example, many studies neglect to address the *desirable criteria* that an importance measure should satisfy to be considered well-defined (e.g., Johnson & LeBreton, 2004; Grömping, 2015).

This work aims to revisit some well-established VIMs for linear regression from the statistical literature. Despite being developed some time ago, some of these measures remain relatively unknown and underutilized in practice. A particular focus is put on their properties, conditions of use, and subsequent interpretation. Additionally, the discussion emphasizes the importance of a clear definition of *relative importance* (also known as "relative weight" in the literature: Johnson, 2000; Tonidandel & LeBreton, 2015; Nathans et al., 2012) from the user's perspective. Specifically, VIMs are associated with the concept of *dispersion importance*, introduced by Achen (1982), which relates to the influence of inputs on the output variance. In the context of linear regression models, the *coefficient of determination* (R^2), which quantifies the percentage of output variability explained by the model and thus, provides a validation metric of the linear regression, is also a key metric for constructing VIMs. According to Johnson & LeBreton (2004), a VIM associated with a specific regressor is defined as: "the proportionate contribution each variable makes to R^2 (the ratio of explained variance to total response variance), considering both its direct effect (i.e., its correlation with the response) and its effect when combined with other variables in the model". In line with this definition, we reintroduce the approach of *general dominance analysis*, which involves defining an R^2 decomposition by establishing a hierarchy among regressors based on certain *dominance criteria* (Budescu, 1993).

In this context, a primary challenge arises: how to meaningfully allocate shares of R^2 among statistically dependent inputs? Understanding multicollinearity is crucial in addressing this question, as it pertains to the situation where two or more inputs exhibit significant linear relationships. This concept extends from simple collinearity to encompass cases where multiple variables are highly correlated, complicating the interpretation of their individual contributions to the explained variance of the model. To illustrate this challenge, we first present an intuitive representation of multicollinearity using Venn diagrams, inspired by previous work (see, e.g., Clouvel, 2019; Il Idrissi et al., 2021), focusing initially on the case of a two-input regression model. Furthermore, after discussing classic metrics that address multicollinearity but do not directly enable R^2 decomposition, we justify the introduction of more sophisticated VIMs. These measures aim to isolate and quantify the individual effects of each variable on the output, offering a more nuanced understanding. Among the various methods for partitioning R^2, the LMG indices (Lindeman et al., 1980) and PMVD indices (Feldman, 2005) are highlighted as prominent VIMs. However, differentiating these methods and establishing their appropriate conditions of use requires defining basic desirability criteria. Moreover, in high-dimensional settings, the exponential computational complexity of LMG and PMVD indices poses a significant challenge. Addressing this issue, we emphasize the Johnson indices (Johnson, 2000; also refer to Genizi, 1993), which utilize relative weight analysis to effectively mitigate computational complexities.

While this work is not intended as an exhaustive review (for that, see Grömping, 2015), it offers several novel contributions. First, we emphasize the connections between VIMs in the statistical literature and the field of GSA. Second, we highlight recent advancements in VIMs that facilitate more meaningful and theoretically sound interpretations of linear models, particularly in the context of highly correlated inputs. Without being exaustive, we decided to focus only on a subset of sound and robust importance measures, while excluding from this work a panel of ill-defined or non-robust ones according to previous recommendations from the literature (e.g., first/last methods, Pratt, CAR scores, Wefila, as studied in Grömping, 2015; Wallard, 2015, 2019; Blanchard, 2023). Finally, we examine the implementation of these VIMs in several R packages, especially the '**sensitivity**' package (Iooss et al., 2023), and demonstrate their numerical behavior on both simulated and public datasets. From these empirical studies, we derive practical recommendations. All R scripts used in this study are also made available for reproducibility purposes (see Section 7.1).

The structure of the paper is as follows. Section 2 provides some basics about the multivariate linear regression model. Section 3 develops standard VIMs based on variance decomposition obtained with independent inputs. Section 4 introduces the effects and issues that multicollinearity can bear on variance decomposition. Then, Section 5 develops several VIMs adapted to correlated inputs, obtained from allocation rules, while Section 6 presents the Johnson indices. Section 7 applies all the studied metrics on several simulated or public datasets. Finally, Sections 8 and 9 provide a methodological synthesis and some conclusions, and draw some prospects regarding the remaining challenges. Generalizations of these VIMs for classification tasks (i.e., logistic regression) are provided in the Supplementary Material C.

Table 1 and Table 2 provide respectively a table of acronyms and a table of notations used all along the paper.

Table 1: Main acronyms.

Acronym	Definition
CC	Correlation Coefficient (or Pearson coefficient)
CC^2	Squared Correlation Coefficient
CI	Confidence Interval
GSA	Global Sensitivity Analysis
LMG	Lindeman-Merenda-Gold indices (or Shapley effects for linear models)
PCC	Partial Correlation Coefficient
PCC^2	Squared Partial Correlation Coefficient
PMVD	Proportional Marginal Variance Decomposition
SPCC	Semi-Partial Correlation Coefficient
$SPCC^2$	Squared Semi-Partial Correlation Coefficient
SRC	Standardized Regression Coefficient
SRC^2	Squared Standardized Regression Coefficient
SVD	Singular Value Decomposition
VIF	Variance Inflation Factor
VIM	Variance-based Importance Measure

Table 2: Main notations.

Notation	Definition
x_j	j-th deterministic variable
X_j	j-th random variable
$\boldsymbol{x} := (x_1, \dots, x_d)$	Vector of deterministic variables
$\boldsymbol{X} := (X_1, \dots, X_d)$	Random vector
$x_j^{(i)}$	i-th observation of the variable x_j
$\boldsymbol{x}^{(i)}$	i-th observation of the vector $\boldsymbol{x}$
$\boldsymbol{X}^n := \left(x_1^{(i)}, \dots, x_d^{(i)}\right)_{i=1,\dots,n}$	n-observation design matrix
$\hat{\beta}$	Estimator of the parameter β
$\mathrm{E}[\cdot]$	Expectation operator
$\mathrm{VAR}(\cdot)$	Variance operator
$\mathrm{COV}(\cdot,\cdot)$	Covariance operator

2. Basics of multivariate linear regression

In this section, the multivariate linear regression framework is recalled. Consider an experimental design with n observations of an explained real-valued output random variable Y and of d explanatory input random variables $\boldsymbol{X} = (X_1, \dots, X_d)$, denoted by:

$$(\boldsymbol{X}^n, \boldsymbol{y}^n) = (x_1^{(i)}, \dots, x_d^{(i)}, y^{(i)})_{i=1,\dots,n} \tag{1}$$

For simplicity and without any loss of generality, we use the following usual assumption.

Assumption 1 (Centering). *Both inputs and output are centered such that*:

$$\mathrm{E}[X_j] = 0 \textit{ for } j = 1, \dots, d, \textit{ and } \mathrm{E}[Y]=0.$$

The relationship between the random inputs $\boldsymbol{X}$ and the random output Y is modeled as being linear such that:

$$Y = \boldsymbol{X}\beta + \varepsilon, \tag{2}$$

where $\beta = (\beta_1, \dots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^d$ is an unknown vector of coefficients, and ε is a random error assumed to be Gaussian and centered, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and such that $\mathrm{E}[\varepsilon|\boldsymbol{X}] = 0$. Specifically, for each observation $\boldsymbol{x}^{(i)}$ of $\boldsymbol{X}$ and $y^{(i)}$ of Y, the previous relationship can be written as $y^{(i)} = \boldsymbol{x}^{(i)}\beta + \varepsilon^{(i)}$, where for all $i = 1, \dots, n$, the $\varepsilon^{(i)}s$ are independent and identically distributed (i.i.d.) according to the same centered Gaussian distribution with variance σ_ε^2. We thus deduce that:

$$\mathrm{E}\left[Y|\boldsymbol{X} = \left(x_1^{(i)}, \dots, x_d^{(i)}\right)\right] = \boldsymbol{x}^{(i)}\beta, \quad \textit{for } i = 1, \dots, n.$$

If the sample size is large enough (i.e., $n \gg d$), and $(\boldsymbol{X}^n)^{\mathrm{T}}\boldsymbol{X}^n$ is a positive-definite matrix, the ordinary least squares method (see, e.g., Christensen (1990)) can be used to estimate the vector of parameters β by using the unbiased maximum likelihood estimator given by:

$$\hat{\beta} = ((\boldsymbol{X}^n)^{\mathrm{T}}\boldsymbol{X}^n)^{-1}(\boldsymbol{X}^n)^{\mathrm{T}}\boldsymbol{y}^n. \tag{3}$$

Statistical techniques then allow for checking whether the use of a linear model is licit or not. An important goodness-of-fit metric is the *coefficient of determination* R^2 which quantifies the percentage of output variability captured by the linear regression model. Its theoretical value is given by:

$$R^2 = R^2_{Y(\boldsymbol{X})} := 1 - \frac{\mathrm{E}[\mathrm{VAR}(Y|\boldsymbol{X}))]}{\mathrm{VAR}(Y)} = \frac{\mathrm{VAR}(\mathrm{E}[Y|\boldsymbol{X}])}{\mathrm{VAR}(Y)}, \tag{4}$$

where the $Y(\boldsymbol{X})\ in\ R^2_{Y(\boldsymbol{X})}$ is a standard notation to explicitly mention that this quantity is obtained from the full regression of Y with respect to $\boldsymbol{X}$. Such a notation is particularly handy in the case of partial correlations and is thus introduced here for the sake of clarity. Provided a consistent estimator $\hat{\beta}$ of β, one can build a plug-in consistent estimator of R^2 based on the design matrix described by Eq. (1), leading to the following formula:

$$\widehat{R^2} = \frac{\sum_{i=1}^n (\widehat{y^{(i)}} - \bar{y})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2}, \text{where } \bar{y} = \frac{1}{n}\sum_{i=1}^n y^{(i)} \text{ and } \widehat{y^{(i)}} = x^{(i)}\hat{\beta}.$$

Remark 1. *If the sample size is close to the number of inputs, there is a risk of overfitting. In that sense, an adjusted coefficient, such as* $R^2_{adj} = 1 - |1 - R^2| \left|\frac{n-1}{n-(1+d)}\right|$ *(see Karch (2020) for an overview of the various formulations of adjusted coefficients) can be used in order to penalize this dimension drawback. Moreover, cross-validation techniques can also be used to validate the regression model so as to avoid overfitting. It mainly consists in computing a predictivity coefficient* Q^2 *based on a validation sample extracted from the learning sample using dedicated techniques (Marrel et al., 2008; Fekhari et al., 2023).*

3. Variance-based importance measures

Importance measures in regression models (Darlington & Hayes, 2017) broadly consist in quantifying the relative importance of the inputs to the output. In the field of GSA, importance measures are usually called *sensitivity indices*, and many different metrics (e.g., the variance, the entropy, a dependence measure or a dissimilarity measure between an input and the output) have been proposed to define them mathematically (Saltelli et al., 2000; Da Veiga et al., 2021). A first approach consists in quantifying the amount of input uncertainty that creates dispersion in the output, the "dispersion" being traditionally quantified by the variance. Hence, the importance of an input can be naturally understood as the amount of uncertainty (i.e., in terms of variance) it brings to the system.

Besides GSA, and more generally in statistics, variance decomposition plays a central role in practical studies (e.g., in uncertainty analysis as illustrated in Kurowicka & Cooke, 2006), where it has been deemed to be an appropriate measure of information for a long time. In a nutshell, a VIM aims at quantifying the contribution of each input X_i to the variance of the output Y, denoted by VAR(Y).

3.1 The variance decomposition

In the context of a multivariate linear regression model, the VIMs are based on the variance decomposition given by the law of total variance:

$$\mathrm{VAR}(Y) = \underbrace{\mathrm{VAR}(\mathrm{E}[Y|\boldsymbol{X}])}_{\text{explained variance}} + \underbrace{\mathrm{E}[\mathrm{VAR}(Y|\boldsymbol{X})]}_{\text{residual variance}}, \tag{5}$$

which is valid for any real-valued random variable Y. The first term is usually called the *explained variance*, while the second term is usually interpreted as the *residual variance* which can be due to unaccounted inputs in the modelling, or to measurement errors. In particular, with Eq. (2), this decomposition gives:

$$\mathrm{VAR}(\mathrm{E}[Y|\mathbf{X}])=\beta^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{X}}\beta, \quad \mathrm{E}[\mathrm{VAR}(Y|\boldsymbol{X})]=\sigma_\varepsilon^2, \tag{6}$$

Where $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{X}} = (\mathrm{COV}(X_i,X_j))_{1\le i,j\le d}$ is the variance-covariance matrix of the inputs. Finally, one can notice the direct link between the explained variance and the theoretical definition of the R^2 coefficient in Eq. (4), which is nothing more than a percentage of the total variance explained by the inputs.

3.2 Criteria for R^2 decomposition

As seen above, the R^2 is directly linked to the notion of explained variance and to the variance decomposition (Eqs. (4) and (5)). Thus, historical developments of VIMs in the literature of linear regression analysis naturally focused on partitioning the R^2 among the d inputs (Johnson & LeBreton, 2004; Grömping, 2007). Many decomposition types have been proposed, leading to various R^2 partitioning strategies (and thus, to various meanings). To sum up, several authors defined some *desirability criteria* (i.e., properties) of what a "relevant decomposition" should be. For instance, according to Grömping (2007), four basic desirability criteria can be sought after for a VIM resulting from an R^2 decomposition:

- *(C_1) Proper decomposition*: the sum of all shares should be equal to the total variance (or to the R^2 itself in the case of normalized shares);
- *(C_2) Nonnegativity*: all shares should be nonnegative;
- *(C_3) Exclusion*: if $\beta_j = 0$, then the share of X_j should be zero;
- *(C_4) Inclusion*: if $\beta_j \neq 0$, then the share of X_j should be nonzero.

Criteria *(C_1)* and *(C_2)* constitute the fundamental properties that VIMs should verify as they allow for a proper interpretation as a percentage of R^2. The criterion *(C_3)* is strongly relevant if the goal is to identify variables that are not influential or spurious and which should not appear in the model. The criterion *(C_4)* seems also fundamental to highlight inputs with direct influence in the model.

For the sake of completeness, one can mention an additional criterion that is sometimes mentioned in the literature, but more related to regularization-based techniques (Zou & Hastie, 2005; Wallard, 2019):

- *(C_5) Grouping*: all shares should tend to equate for highly correlated inputs.

However, as it will be shown in Sections 5 and 7, for the VIMs that are considered in this paper, the grouping property (C_5) can be contradictory to the exclusion property (C_3). Thus, the choice of a specific VIM should depend on the case of study and on the desired criteria. If the interpretation is focused on the direct influence of the inputs on the model output, then the exclusion property (C_3) seems to be appropriate; if the correlations among data can carry necessary information for the interpretation (as sometimes in GSA), it can be useful to consider the (C_5) property instead.

3.3 Regression coefficients and Pearson correlation for independent inputs

Provided that the inputs are mutually independent, the law of total variance in Eq. (5) becomes:

$$\mathrm{VAR}(Y) = \sum_{j=1}^{d} \beta_j^2 \sigma_j^2 + \sigma_\varepsilon^2 \,,$$

and naturally allows to partition the output variance with respect to any input X_j , with j = 1, . . . , d, by means of a *standardized regression coefficient* (SRC) ${\beta_j^*}^2$ defined as:

$$\beta_j^* = \beta_j \frac{\sigma_j}{\sigma_Y},$$

where σ_Y and σ_j are the standard deviations associated with Y and the input X_j , resp. Hence, the *squared SRC* (SRC^2), denoted by ${\beta_j^*}^2$, can then be used as a VIM (Grömping, 2006; Antoniadis et al., 2021). It can be understood as the share of variance explained by each input X_j , since:

$$R^2 = \sum_{j=1}^{d} {\beta_j^*}^2.$$

One can see that the SRC^2 respects the four desirability criteria (C_1), (C_2), (C_3) and (C_4) mentioned previously. Moreover, one can notice that the SRC is strongly connected to the input-output Pearson *correlation coefficient* (CC), denoted by r_{Y,X_j}, which allows to measure the linear correlation between an input X_j and the output Y:

$$r_{Y,X_j} = \frac{\mathrm{COV}(Y, X_j)}{\sigma_Y \sigma_j}.$$

In fact, for independent inputs, both quantities are equal, i.e., $r_{Y,X_j} = \beta_j^*$ and thus, one obtains:

$$R^2 = \sum_{j=1}^{d} r_{Y,X_j}^2 \tag{7}$$

Remark 2. *As a reminder, if the input X_j admits a perfect linear relationship with the output Y, r_{Y,X_j} is equal to 1 or −1. If X_j and Y are independent, r_{Y,X_j} is equal to 0. However, a r_{Y,X_j} equals to 0 does not imply that X_j and Y are independent as the dependency between X_j and Y might be nonlinear.*

As soon as inputs are not independent anymore, the SRC^2 is no longer an admissible VIM, since it does not take the contribution due to the covariance between the inputs of Eq. (6) into account. Thus, the VIM desirability criterion (C_1) is not respected anymore. The following sections are dedicated to study alternatives which can be used when the independence is no more ensured.

4. Dealing with multicollinearity

In a regression setting, *multicollinearity* occurs whenever two or more inputs exhibit a statistically significant linear dependence. It generalizes the notion of *collinearity* (Belsley et al., 1980) to encompass a linear link between more than two variables. Two variables X_1 and X_2 are said to be perfectly collinear if and only if the CC $r_{X_1 X_2}$ is equal to 1 or -1 (see Remark 2). Similarly, there is a *perfect multicollinearity* when there are two or more inputs perfectly collinear. In practice, we speak of multicollinearity when there are several (possibly highly) correlated variables with each other.

Several drawbacks can arise due to a high degree of multicollinearity. For instance, the least-square estimates of the linear coefficients can be impacted (this consequence is sometimes known as the "aliasing effect", see, e.g., McCullagh & Nelder, 1989). Even if the matrix $(\boldsymbol{X}^n)^{\mathrm{T}}\boldsymbol{X}^n$ appearing in Eq. (3) is theoretically invertible, a computer algorithm may be unsuccessful or inaccurate enough to obtain a precise approximation of the inverse matrix due to ill-conditioning. Several methods exist to circumvent this phenomenon, such as regularization techniques (see, e.g., Deng et al., 2015).

Another issue can occur during the estimation of the impact of an input variable on the output Y. The greater the multicollinearity effect, the more difficult it is to separate the individual effects of each variable on the output variable. This section focuses on this difficulty by investigating several classic metrics proposed in the literature to deal with multicollinear inputs. However, as it will be shown, these metrics do not rely on the R^2 decomposition and are, consequently, not able to separate the individual effects of each input variable on the output variable. In the following sections, these metrics are to be distinguished from the VIMs, which are based on a variance decomposition.

4.1 An illustrative example: a two-input regression model

This subsection aims at providing a first simple example which will be used throughout the paper for illustration purposes of several metrics (and the corresponding properties).

Example: two-input regression model. *Consider the linear regression model of the Eq. (2) (with d = 2) of the output Y modeled by two inputs X_1 and X_2. For the sake of simplicity, let us introduce the following notations:*

$$b_1 := \beta_1\sigma_1, \qquad b_2 := \beta_2\sigma_2, \qquad \text{and } r := r_{X_1,X_2}\,.$$

From Eqs. (2) and (4), recalling that $\mathrm{COV}(X_1, X_2) = r\sigma_1\sigma_2$, *one has*:

$$R^2 = \frac{b_1^2 + 2b_1b_2r + b_2^2}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2}. \tag{8}$$

Similarly, we can easily determine the squared CC (CC^2):

$$r_{Y,X_1}^2 = \frac{(b_1 + rb_2)^2}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2} \quad and \quad r_{Y,X_2}^2 = \frac{(b_2 + rb_1)^2}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2}. \tag{9}$$

Both Eqs. (8) and (9) highlight the fact that, when the inputs are correlated (i.e., $r \neq 0$), the CC^2 do not satisfy the R^2 decomposition as in the case of independent inputs given by Eq. (7). Therefore, CC^2 do not satisfy the criterion (C_1). Moreover, assuming that $r = 1$, both CC^2 will be the same, even if either b_1 or b_2 are set to zero, which makes the criterion (C_3) not fulfilled.

4.2 The variance inflation factor

A standard and well-known metric of multicollinearity is the *variance inflation factor* (VIF) (Chatterjee & Price, 1977; Fox & Monette, 1992; Johnson & LeBreton, 2004) defined as:

$$\mathrm{VIF}_j = \frac{1}{1 - R^2_{X_j(\boldsymbol{X}_{-j})}} \tag{10}$$

where $\boldsymbol{X}_{-j}$ is the vector of all the inputs except X_j , and where $R^2_{X_j(\boldsymbol{X}_{-j})}$ represents the R^2 from the linear regression where X_j is considered as the output, and by taking $\boldsymbol{X}_{-j}$ as inputs. The smallest value of VIF is 1 and corresponds to the absence of collinearity. A standard rule of thumb is that a VIF value exceeding 5 or 10 indicates a substantial amount of collinearity (James et al., 2014).

Example: two-input regression model (Section 4.1, continued).

From Eq. (10), one simply has $\mathrm{VIF}_1 = \mathrm{VIF}_2 = \frac{1}{1-r^2}$ *with* $r \neq \pm 1$.

Remark 3. *The generalized variance inflation factor (GVIF) has been proposed by Fox & Monette (1992) in order to provide a similar metric of multicollinearity as the VIF in the case of categorical inputs. The GVIF also works if one desires to group polynomial terms related to the same input*.

4.3 The partial correlation coefficient

It can also be interesting to quantify the degree of association between the output Y and an input X_j by cancelling the effect of other inputs, gathered in $\boldsymbol{X}_{-j}$. It is in that spirit that the *partial correlation coefficient* (PCC) has been introduced and used in the GSA community (see, e.g., Helton, 1993; Saltelli et al., 2000; Helton et al., 2006). It is defined as:

$$r_{(Y,X_j)|\boldsymbol{X}_{-j}} = r_{\varepsilon_{Y|\boldsymbol{X}_{-j}},\varepsilon_{X_j|\boldsymbol{X}_{-j}}}, \tag{11}$$

where $\varepsilon_{Y|\boldsymbol{X}_{-j}}$ (resp. $\varepsilon_{X_j|\boldsymbol{X}_{-j}}$) represents the random error in the linear regression model of Y (resp. X_j) with respect to $\boldsymbol{X}_{-j}$. In other words, the PCC measures the residual information of X_j on Y which is not explained by the variables $\boldsymbol{X}_{-j}$.

Example: two-input regression model (Section 4.1, continued).

Eq. (11) can be written as a function of the coefficient of determination and the CC such as:

$$r^2_{(Y,X_1)|X_2} = \frac{R^2_{Y(X_1,X_2)} - r^2_{Y,X_2}}{1 - r^2_{Y,X_2}}.$$

Using Eqs. (8) and (9), we have

$$r^2_{(Y,X_1)|X_2} = \frac{b_1^2(1-r^2)}{b_1^2(1-r^2)+\sigma_\varepsilon^2} \text{ and } r^2_{(Y,X_2)|X_1} = \frac{b_2^2(1-r^2)}{b_2^2(1-r^2)+\sigma_\varepsilon^2}.$$

Note that the *squared PCC* (PCC^2) is equal to 1 if the model is perfectly linear (i.e., if $\sigma_\varepsilon^2 = 0$ with $b_j \neq 0$) and equal to zero if X_1 and X_2 are perfectly correlated. Thus, even if it respects the *exclusion criterion* (C_3), it does not respect the *inclusion* criterion (C_4). Finally, even if in the GSA literature (Saltelli et al., 2000; Helton et al., 2006; Iooss & Lemaître, 2015), the PCC^2 has been proposed as a substitute for the SRC^2 in the case of dependent inputs, it does not respect the fundamental desirability criterion (C_1), i.e., the proper R^2 decomposition, and thus should not be used as an admissible VIM.

4.4 The semi-partial correlation coefficient

Instead of controlling the potential linear effects of $\boldsymbol{X}_{-j}$ with X_j , as done with the PCC, the *semi-partial correlation coefficient* (SPCC) quantifies the additional explanatory power of a variable X_j on the variance of Y (Johnson & LeBreton, 2004). The SPCC is defined as the proportion of the output variance explained by X_j after removing the "information brought" by $\boldsymbol{X}_{-j}$ (as a difference of explained variance). It is formally given by the CC (noted $r_{Y,(X_j|\boldsymbol{X}_{-j})}$) between Y and the residuals of the regression of X_j on $\boldsymbol{X}_{-j}$. The *squared SPCC* (SPCC^2) is intrinsically linked to the R^2 since it can be written as:

$$r^2_{Y,(X_j|\boldsymbol{X}_{-j})} = R^2_{Y(\boldsymbol{X})} - R^2_{Y(\boldsymbol{X}_{-j})}. \tag{12}$$

In the case of independent inputs, the SPCC is equal to the usual CC.

Example: two-input regression model (Section 4.1, continued).

Eq. (12) can be written as a function of the coefficient of determination and the CC such as:

$$r^2_{Y,(X_1|X_2)} = R^2_{Y(X_1,X_2)} - r^2_{Y,X_2} ,$$

and using Eqs. (8) and (9), one has:

$$r^2_{Y,(X_1|X_2)} = \frac{b_1^2(1-r^2)}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2} \quad \text{and} \quad r^2_{Y,(X_2|X_1)} = \frac{b_2^2(1-r^2)}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2} . \tag{13}$$

On the one hand, notice that in Genizi (1993), the SPCC^2 is called the *marginal reduction* due to X_j . This comes from the fact that it quantifies the *loss of* R^2 induced by removing X_j from the linear model. On the other hand, as illustrated in the above example, the SPCC can become artificially small in situations of highly correlated inputs. It subsequently renders any importance ranking task quite difficult. Moreover, one can see that the SPCC^2 is not an admissible VIM, since it does not respect the fundamental criterion (C_1).

4.5 Illustration with Venn diagrams

In order to provide an intuitive understanding of the multicollinearity phenomenon and how it impacts the various metrics already introduced, we propose to use Venn diagrams in the case of a standard linear regression with two inputs (Clouvel, 2019; Il Idrissi et al., 2021a).

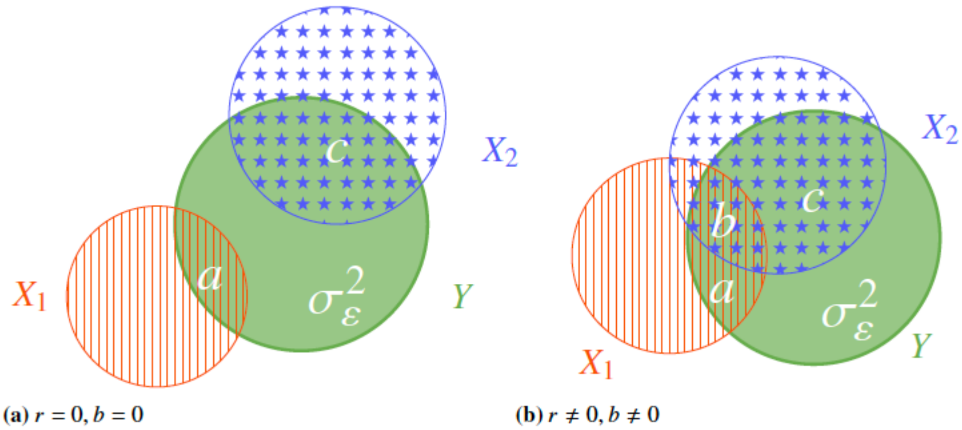

Figure 1: Illustration of the multicollinearity effects with an output Y and two inputs X_1 and X_2.

Figure 1 is to be understood as follows: in both sub-figures, the total variance of Y is represented as the green area by $a + b + c + \sigma_\varepsilon^2$ with σ_ε^2 the unexplained share of variance (i.e., the model error). The hatched orange area represents the variance of X_1, while the star-filled blue area represents the variance of X_2. The area a (resp. c) represents the additional explanatory power of the variable X_1 (resp. X_2) on the regression model $Y(\boldsymbol{X})$ (defined by Eq. (2)) given by the nominator of the SPCC^2 (Eq. (13)). We thus can write that:

$$\begin{cases} a = b_1^2(1-r^2), \\ c = b_2^2(1-r^2), \\ b = b_1^2r^2 + 2b_1b_2r + b_2^2r^2. \end{cases} \tag{14}$$

Independent case. In Figure 1a, the variables X_1 and X_2 are independent. The hatched orange and the star filled blue areas do not overlap ($r = 0, b = 0$). In this case, the CC^2 (Eq. (9)) and the SPCC^2 (Eq. (13)) are equal:

$$r^2_{Y,(X_1|X_2)} = r^2_{Y,X_1} = \frac{a}{a+c+\sigma_\varepsilon^2} \quad \text{and} \quad r^2_{Y,(X_2|X_1)} = r^2_{Y,X_2} = \frac{c}{a+c+\sigma_\varepsilon^2} \ .$$

The hatched orange area a and the star-filled blue area c finally represent the proportion of the variance in Y resp. explained by X_1 and X_2, and allow sharing the determination coefficient R^2:

$$R^2 = \frac{a+c}{a+c+\sigma_\varepsilon^2} \ .$$

Correlated case. In Figure 1b, the variables X_1 and X_2 are correlated. The hatched orange and the star-filled blue areas do overlap ($r \neq 0, b \neq 0$). The proportion of the variance in Y explained by X_1 (resp. X_2) is now equal to $a+b$ (resp. $c+b$). The CC^2 (Eq. (9)) are thus equal to:

$$r^2_{Y,X_1} = \frac{a+b}{a+b+c+\sigma_\varepsilon^2} \quad \text{and} \quad r^2_{Y,X_2} = \frac{c+b}{a+b+c+\sigma_\varepsilon^2} \ ,$$

and the SPCC^2 (Eq. (13)) are equal to:

$$r^2_{Y,(X_1|X_2)} = \frac{a}{a+b+c+\sigma_\varepsilon^2} \quad \text{and} \quad r^2_{Y,(X_2|X_1)} = \frac{c}{a+b+c+\sigma_\varepsilon^2} \ .$$

We can notice that neither the sum of the CC^2 nor the sum of the SPCC^2 is equal to the determination coefficient R^2:

$$R^2 = \frac{a+b+c}{a+b+c+\sigma_\varepsilon^2} .$$

Therefore, neither the CC^2 nor the SPCC^2 cannot be used as VIMs knowing that they do not meet the criterion (C_1) for the R^2 decomposition.

Finally, the notion of multicollinearity can be appreciated in this example as follows: the higher the absolute value of r, the larger the overlap area b. Similarly, the areas a and c are getting smaller (see Eq. (14)) and both the CC^2 and the SPCC^2 are far from meeting the criterion (C_1). This explains why the presence of multicollinearity makes the R^2 decomposition difficult, and why it is necessary to use more complex VIMs (than the previously presented metrics) to separate the individual effects of each input on the output variable.

5. Importance measures from allocation rules

As shown in the previous section, building relevant and meaningful VIMs that handle correlated predictors is a challenge. In the literature, such a topic has already been widely addressed by several authors from various communities such as, e.g., Lindeman et al. (1980) in the statistical learning community, or Budescu (1993) (with the proposition of *general dominance analysis* following proper *dominance criteria*) in quantitative psychology. However, it appears that all these techniques are very similar and can be gathered into a more generic framework of *cooperative game theory*, while reducing to the definition of specific *allocation rules* among the regressors. In this section, the links and connections between these old VIMs are exhibited and some relevant extensions are discussed in order to tackle some specific desirability criteria.

5.1 Lindeman-Merenda-Gold indices

A particular VIM tackling the problem of correlated inputs is the so-called “Lindeman-Merenda-Gold” (LMG) index, named after the initials of the authors’ names in Lindeman et al. (1980). This VIM has been studied extensively by several authors (see, e.g., Budescu, 1993; Johnson & LeBreton, 2004; Grömping, 2006) and relies on the averaging sequential sums of squares over all orderings of inputs.

Formally, let u denote a subset of indices in the set of all subsets of $\{1, \dots, d\}$ and $\boldsymbol{X}_u = (X_j : j \in u)$ represents a subset of inputs. From Budescu (1993), one can interpret this index from the concept of “dominance analysis”. It is based on the measure of the elementary contribution of any given variable $X_j (j \in \{1, \dots, d\})$ to a given subset model $Y(\boldsymbol{X}_u)$ by the increase in R^2 that results from adding that predictive variable to the regression model. Formally, the LMG$_j$ index associated to X_j can be defined as follows:

$$\mathrm{LMG}_j = \frac{1}{d!} \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\dots,d\}}} r^2_{Y,(X_j|\boldsymbol{X}_\pi)} \tag{15}$$

where the SPCC² $r^2_{Y,(X_j|\boldsymbol{X}_\pi)} = R^2_{Y(\boldsymbol{X}_{v\cup\{j\}})} - R^2_{Y(\boldsymbol{X}_v)}$ is to be understood with v being the indices preceding j in the order π. An equivalent formula is given by:

$$\mathrm{LMG}_j = \frac{1}{d} \sum_{i=0}^{d-1} \sum_{\substack{u \subseteq -\{j\} \\ |u|=i}} \binom{d-1}{i}^{-1} r^2_{Y,(X_j|\boldsymbol{X}_u)} = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} r^2_{Y,(X_j|\boldsymbol{X}_u)} \tag{16}$$

with $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ and $r^2_{Y,(X_j|\boldsymbol{X}_u)} = R^2_{Y(\boldsymbol{X}_{u\cup\{j\}})} - R^2_{Y(\boldsymbol{X}_u)}$.

In Eq. (16) (resp. Eq. (15)), this averaging process over all combinations (resp. permutations) is carried out in the absence of order between the inputs. This VIM has been extensively studied in the literature (see, e.g., Kruskal, 1987; Genizi, 1993). The main drawback in regards to its broad utilization in practice is its exponential complexity (i.e., one needs to perform $2^d - 1$ different linear regressions to compute the summands in Eq. (16)), which can be challenging even for moderate size d.

Example: two-input regression model (Section 4.1, continued).

From Eq. (15), one gets:

$$\mathrm{LMG}_1 = \frac{1}{2}\left(R^2_{Y(X_1,X_2)} - R^2_{Y(X_2)} + R^2_{Y(X_1)}\right), \quad \mathrm{LMG}_2 = \frac{1}{2}\left(R^2_{Y(X_1,X_2)} - R^2_{Y(X_1)} + R^2_{Y(X_2)}\right),$$

and using Eqs. (8) and (9):

$$\mathrm{LMG}_1 = \frac{b_1^2 + b_1 b_2 r + \frac{r^2}{2}(b_2^2 - b_1^2)}{b_1^2 + 2b_1 b_2 r + b_2^2 + \sigma_\varepsilon^2}, \quad \mathrm{LMG}_2 = \frac{b_2^2 + b_1 b_2 r + \frac{r^2}{2}(b_1^2 - b_2^2)}{b_1^2 + 2b_1 b_2 r + b_2^2 + \sigma_\varepsilon^2} \tag{17}$$

This result is already given in Grömping (2007).

Going back to the Venn diagram illustration (see Fig. 1 and in Il Idrissi et al. (2021a)), one obtains:

$$\mathrm{LMG}_1 = \frac{a + \frac{b}{2}}{a + b + c + \sigma_\varepsilon^2},$$
$$\mathrm{LMG}_2 = \frac{c + \frac{b}{2}}{a + b + c + \sigma_\varepsilon^2}.$$

Focusing on the numerators, one can notice that the LMG redistributes b equally between X_1 and X_2 (each variable gets half of the variance due to their correlation).

Note also what happens in the two following particular cases:

- If $|r|$ tends to 1 (i.e., the inputs are collinear), LMG$_1$ and LMG$_2$ tends to be both equal to 0.5. The grouping property (C_5) in Section 3.2 is respected;
- If one input is not in the model, e.g., X_2 (then $\beta_2 = 0$ and $b_2 = c = 0$), then LMG$_2$ cannot be zero as long as X_2 is correlated with X_1. The exclusion property (C_3) in Section 3.2 is not respected in this case.

In conclusion, the LMG index respects most of the fundamental VIM desirability criteria such as R^2 decomposition (C_1), nonnegativity (C_2), but also, as stated in Feldman (2005) and Grömping (2007), inclusion (C_4) and grouping (C_5). However, it does not respect the exclusion criterion (C_3).

5.2 The proportional marginal variance decomposition

By analogy with the LMG indices, Feldman (2005) proposed the so-called *proportional marginal variance decomposition* (PMVD). This index also makes use of sequential sum of squares, but differ from the LMG index on the averaging process over the different orderings of inputs. These indices have been extensively studied in Grömping (2007, 2015) and used in a logistic regression context in Il Idrissi et al. (2021b). The PMVD indices are defined as follows:

$$\mathrm{PMVD}_j = \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\dots,d\}}} \frac{L(\pi)}{\sum_{\pi'} L(\pi')} r^2_{Y,(X_j|\boldsymbol{X}_\pi)}, \tag{18}$$

where the two sums are performed over all possible permutations of $\{1, \dots, d\}$ and

$$L(\pi) = \prod_{i=1}^{d-1} \left[r^2_{Y,(\boldsymbol{X}_{\pi_{i+1},\dots,\pi_d}|\boldsymbol{X}_{\pi_1,\dots,\pi_i})} \right]^{-1} .$$

Example: two-input regression model (Section 4.1, continued).

Eq. (18) becomes (Grömping, 2007):

$$\mathrm{PMVD}_1 = \frac{b_1^2 + \left[\frac{b_1^2}{b_1^2 + b_2^2}\right] 2b_1b_2r}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2}\,, \quad \mathrm{PMVD}_2 = \frac{b_2^2 + \left[\frac{b_2^2}{b_1^2 + b_2^2}\right] 2b_1b_2r}{b_1^2 + 2b_1b_2r + b_2^2 + \sigma_\varepsilon^2}.$$

Moreover, one can notice that:

- *If we let $|r| \to 1$, we obtain, for the quantities derived above, that:*

$$\mathrm{PMVD}_1 \to b_1^2 \frac{(b_1+b_2)^2}{(b_1^2 + b_2^2)(b_1^2 + 2b_1b_2 + b_2^2 + \sigma_\varepsilon^2)}\,, \; \mathrm{PMVD}_2 \to b_2^2 \frac{(b_1+b_2)^2}{(b_1^2 + b_2^2)(b_1^2 + 2b_1b_2 + b_2^2 + \sigma_\varepsilon^2)}.$$

  These two values can be strongly different in cases of large differences between b_1 and b_2. This shows that the grouping criterion (C_5) of Section 3.2 is not respected. Note that the pathological case $|r| = 1$ corresponds to the perfect collinearity case. From a statistical perspective, the regression coefficients β_1 and β_2 in b_1 and b_2 are aliased, making them difficult to identify (i.e., their value is nonunique). As a result, these two indices become difficult to interpret in practice;

- *If one input is not in the model, for example, X_2, then $\beta_2 = 0$ and $b_2 = c = 0$, and subsequently,* $\mathrm{PMVD}_2 = 0$. *Therefore, the exclusion property (C_3) is respected;*

- *If $\sigma_\varepsilon^2 = 0$, the equations simplify to:*

$$\mathrm{PMVD}_1 = \frac{b_1^2}{b_1^2 + b_2^2}\,, \quad \mathrm{PMVD}_2 = \frac{b_2^2}{b_1^2 + b_2^2}\,.$$

Going back to the Venn diagram analogy (see Fig. 1), one has:

$$\mathrm{PMVD}_1 = a \frac{\left[1 + \frac{b}{a + c}\right]}{a + b + c + \sigma_\varepsilon^2}\,,$$
$$\mathrm{PMVD}_2 = c \frac{\left[1 + \frac{b}{a + c}\right]}{a + b + c + \sigma_\varepsilon^2}\,.$$

In this case, the share b due to the correlation between inputs is not equally shared (as already illustrated, e.g., in Hérin et al. (2022)), as for the LMG indices, but rather "proportionally" shared with respect to the magnitude of the shares a and c. In the particular case where $\sigma_\varepsilon^2 = 0$, the above equations simplify to:

$$\mathrm{PMVD}_1 = \frac{\mathrm{a}}{\mathrm{a} + \mathrm{c}}\ ,\ \ \mathrm{PMVD}_2 = \frac{\mathrm{c}}{\mathrm{a} + \mathrm{c}}\ .$$

and one can notice that the PMVD does not depend on b anymore. While this behavior is known when dealing with two inputs, Grömping (2007) shows that it does not generalize to situations with more inputs.

In conclusion, the PMVD respects almost all the fundamental VIM desirability criteria: R^2 decomposition (C_1) and nonnegativity (C_2), but also, as stated by Feldman (2005) and Grömping (2007), exclusion (C_3) and inclusion (C_4).

5.3 Synthesis and discussion

5.3.1. Synthesis about the two-input regression model

Table 3 synthesizes the analytical expressions of the discussed VIMs based on the illustration of Figure 1. The equations for CC^2, PCC^2 and SPCC^2 are displayed as functions of a, b, c (using Eqs. (14)). As seen in this case, CC^2, PCC^2 and SPCC^2 are not admissible VIMs because they do not sum to R^2. Contrarily, LMG and PMVD are admissible. Moreover, LMG does not respect the exclusion property but respects the inclusion property, while the PMVD respects both. The behavior of these indices is illustrated and studied on more general examples and on real datasets in Section 7.

Table 3: Metrics and VIMs associated with the decomposition of $R^2 = (a + b + c)/(a + b + c + \sigma_\varepsilon^2)$.

Input	CC^2	PCC^2	SPCC^2	LMG	PMVD
X_1	$\frac{a+b}{a+b+c+\sigma_\varepsilon^2}$	$\frac{a}{a+\sigma_\varepsilon^2}$	$\frac{a}{a+b+c+\sigma_\varepsilon^2}$	$\frac{a+\frac{1}{2}b}{a+b+c+\sigma_\varepsilon^2}$	$\frac{a+\frac{a}{a+c}b}{a+b+c+\sigma_\varepsilon^2}$
X_2	$\frac{c+b}{a+b+c+\sigma_\varepsilon^2}$	$\frac{c}{c+\sigma_\varepsilon^2}$	$\frac{c}{a+b+c+\sigma_\varepsilon^2}$	$\frac{c+\frac{1}{2}b}{a+b+c+\sigma_\varepsilon^2}$	$\frac{c+\frac{c}{a+c}b}{a+b+c+\sigma_\varepsilon^2}$

Remark 4. *In Il Idrissi et al. (2021b), estimation schemes for LMG and PMVD indices have been proposed for both linear and logistic models and applied to regression and classification tasks, resp. In the Supplementary Material C, some results for logistic regression are provided as a complementary content to the regression part presented in the core paper.*

5.3.2. Discussion about the links between the studied VIMs, GSA and game theory

The two above-presented VIMs (LMG and PMVD) are inherently linked with cooperative game theory. The sequential approach (i.e., the formulations using permutations) is related to the notion of *random order allocation* introduced by Weber (1988) and Feldman (2005). These allocations (also called "solution concepts" in the game theory literature) allow decomposing a quantity (in this case, the R^2) by means of quantifying the "value" of each player using a value function (here, the SPCC^2). Through this lens, the LMG indices are none other than the so-called *Shapley values* of the cooperative game (Shapley, 1953). This value is "egalitarian" in its redistribution, i.e., the behavior of splitting b in half in the Venn diagram analogy actually holds in higher dimensions. On the other side, the PMVD is analogous to the *proportional values* (Feldman, 2000), allowing for a proportional redistribution.

In GSA, when dealing with a linear numerical model, the only difference with the present study is the fact that σ_ε^2 is equal to 0 (Saltelli et al., 2000; Helton et al., 2006). GSA actually encompasses the definition of VIMs of more general models (i.e., not necessarily linear). For instance, whenever the inputs are assumed to be independent, the SRC^2 is actually equal to the *first-order Sobol'* index, which is defined outside of the realm of linear models (Sobol', 1993). Additionally, provided that the error is null, the R^2 can be directly comparable to the *closed Sobol' indices* (see, e.g., Il Idrissi et al., 2021a; Da Veiga et al., 2021), which need not be restricted to linear models to be defined.

As for the use of Shapley values to define sensitivity indices, this idea has been introduced by Owen (2014) in the context of variance-based GSA. In this work, the value function is the closed Sobol' index of a subset of players, leading to the so-called *Shapley effects*. Several authors (Song et al., 2016; Benoumechiara & Elie-Dit-Cosaque, 2019; Iooss & Prieur, 2019; Plischke et al., 2021) proposed and studied numerous dedicated estimation algorithms for nonlinear models. Analytical formulas have also been exhibited for linear models with Gaussian inputs (Owen & Prieur, 2017), and can be efficiently computed by finely tuned algorithms (Broto et al., 2019). However, these techniques require the knowledge and the ability to draw samples from the joint density of the inputs. Especially, one needs to know how to model the dependence structure (i.e., the copula) between the inputs. Typically, such a condition is not met in common statistical learning (or machine learning) practice when only scarce data is available. Therefore, the estimation of such VIMs often appears to be difficult (either because of the inherent cost, especially with respect to the input dimension, or since only a few data is available). Recently, *given-data* strategies have been proposed to leverage this issue (Broto et al., 2020; Bénard et al., 2022).

It has also been noticed that, theoretically, the Shapley effects can grant *exogenous inputs* (i.e., which are not explicitly included in the structural equations of the model) some importance, especially when these inputs are correlated to *endogenous inputs* (i.e., effectively present in the model). Inspired by the PMVD index, Hérin et al. (2022, 2024) proposed to use the Proportional values instead of the Shapley values as a baseline. As a result, a set of novel GSA indices, called the *proportional marginal effects* (PME) have been defined Hérin et al. (2022, 2024). These indices allow the detection of exogenous inputs, despite the correlation, in a nonlinear setting (which is analogous to the exclusion criterion (C_3)).

6. Dealing with high-dimensional inputs via the relative weight analysis

When d becomes "large" (e.g., several tens), the previous VIMs (i.e., LMG and PMVD) may suffer from their estimation cost (due to the cardinality of the permutations). In order to circumvent these issues, Johnson (2000) proposed the so-called *relative weight measures* (later called *Johnson's relative weights* or simply *Johnson indices* in the rest of the paper). To put it briefly, the basic idea is to transform the correlated inputs into uncorrelated variables using a singular value decomposition (SVD) method, and then to use an appropriate reweighting process in order to get the indices. Note that this approach has been proposed earlier by several authors (see associated references, e.g., in Nimon & Oswald (2013) and Grömping (2015)). As an example, one can mention the work of Genizi (1993) which led to the so-called *Genizi's approach*. All in all, these VIMs based on a preliminary transformation of inputs are known to be adapted to large input dimension as well as providing similar results to those obtained via LMG indices (Johnson & LeBreton, 2004; Clouvel, 2019; Clouvel et al., 2019), at a highly reduced computational cost. Thus, the goal of this section is to explain how to build and estimate the Johnson indices.

6.1 Johnson indices

The Johnson indices (Johnson, 1966, 2000)[1] are part of a wider set of methods called *Relative Weights Analysis*. In the case of the Johnson indices (Johnson, 2000), the matrix $\boldsymbol{X}^n \in \mathbb{R}^{n\times d}$ of the design of experiments is transformed into an orthogonal matrix $\boldsymbol{Z}^n \in \mathbb{R}^{n\times d}$ in the least squares sense. Figure 2 summarizes the global approach of the Johnson indices. For Johnson (1966), it consists in finding $\boldsymbol{Z}^n$ and $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ such that:

$$\begin{cases} \boldsymbol{X}^n = \boldsymbol{Z}^n \boldsymbol{W} \\ (\boldsymbol{Z}^n)^{\mathrm{T}} \boldsymbol{Z}^n = \boldsymbol{I} \\ \boldsymbol{Z}^n = \arg\min\limits_{\boldsymbol{\Pi}^n} Tr\{(\boldsymbol{X}^n - \boldsymbol{\Pi}^n)^{\mathrm{T}} (\boldsymbol{X}^n - \boldsymbol{\Pi}^n)\} \end{cases} \tag{19}$$

where $\boldsymbol{I} \in \mathbb{R}^{d\times d}$ is the identity matrix and $Tr\{\cdot\}$ the trace operator. Johnson shows that the solution matrices of Eq. (19) are given by:

[1] Note, that there are two different authors with the same name. Johnson (2000) suggested determining the matrix W as the weights of the regression of X^n on Z^n contrary to Johnson (1966) which regressed Z^n on X^n.

$$\boldsymbol{Z}^n = \boldsymbol{P}^n \boldsymbol{Q}^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{Q} \Delta \boldsymbol{Q}^{\mathrm{T}} \tag{20}$$

where $\boldsymbol{P}^n \in \mathbb{R}^{n \times d}$ and $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ are two matrices defined by the following SVD:

$$\boldsymbol{X}^n = \boldsymbol{P}^n \Delta \boldsymbol{Q}^{\mathrm{T}} \tag{21}$$

Thus, $\boldsymbol{P}^n$ and $\boldsymbol{Q}$ contain, resp., the eigenvectors of $\boldsymbol{X}^n \boldsymbol{X}^{n\,\mathrm{T}}$ and $\boldsymbol{X}^{n\,\mathrm{T}} \boldsymbol{X}^n$. As for $\Delta \in \mathbb{R}^{d \times d}$, it is a diagonal matrix which, itself, contains the singular values $\delta_1 \geq \ldots \geq \delta_d > 0$ of $\boldsymbol{X}^n$. In that sense, the new set of uncorrelated variables $z_1, \ldots, z_d$ is maximally correlated with the original set of correlated variables $x_1, \ldots, x_d$ (i.e., the columns of $\boldsymbol{X}^n$).

Remark 5. *Note also that Eq. (20) gives:*

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{X}} = \boldsymbol{W}^2 . \tag{22}$$

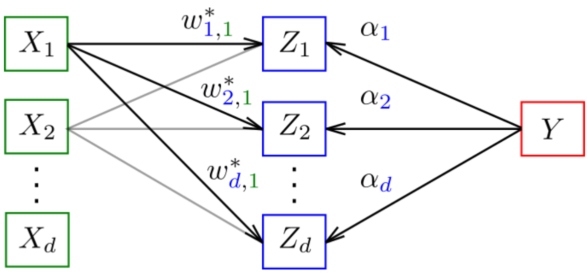

Figure 2: Representation of the Johnson (Johnson, 2000) relative weight calculation associated with the input X_1.

A first least squares regression of $\boldsymbol{y}^n$ (n-size sample of the variable $Y \in \mathbb{R}$) on $\boldsymbol{Z}^n$ allows determining the vector $\alpha \in \mathbb{R}^d$ for which a consistent estimator $\hat{\alpha} = (\hat{\alpha}_j)_{1 \leq j \leq d}$ is given by:

$$\hat{\alpha} = ((\boldsymbol{Z}^n)^{\mathrm{T}} \boldsymbol{Z}^n)^{-1} (\boldsymbol{Z}^n)^{\mathrm{T}} \boldsymbol{y}^n = (\boldsymbol{Z}^n)^T \boldsymbol{y}^n .$$

Since the new transformed predictors z_j are uncorrelated to one another, the predictable variance of Y can be decomposed such as:

$$VAR(E[\boldsymbol{Y}|\boldsymbol{Z}]) = \sum_{j=1}^{d} \alpha_j^2 . \tag{23}$$

The α_j^2's are considered to be close approximations to the relative weights of the original set of correlated variables $x_1, \ldots, x_d$, but they do not give close representations, particularly if some original variables are highly correlated. To take into account the correlation effects, Johnson (2000) thus suggests to compute the regression coefficients of $\boldsymbol{X}^n$ on $\boldsymbol{Z}^n$.

Remark 6. *Using Eqs. (3), (19) and (22), one can prove that:*

$$\hat{\alpha} = \boldsymbol{W} \hat{\beta}. \tag{24}$$

The d linear combinations between $\boldsymbol{X}^n$ and $\boldsymbol{Z}^n$ therefore allows determining the matrix of the weights $\boldsymbol{W}$ for which a consistent estimator $\widehat{\boldsymbol{W}} = (\widehat{w}_{ij})_{1 \leq i,j \leq d}$ is given by:

$$\widehat{\boldsymbol{W}} = (\boldsymbol{Z}^n)^T \boldsymbol{X}^n .$$

Using Eq. (20), it can be shown that the standardized matrix $\boldsymbol{W}^*$ is composed of the CCs r_{Z_i, X_j} such that:

$$w^*_{i,j} = \frac{w_{ij}}{\sqrt{\sum_k w_{kj}^2}} = r_{Z_i X_j} , \tag{25}$$

and thus, for all j , one has:

$$\sum_{i=1}^{d}(w_{i,j}^{*})^2 = 1 \ . \tag{26}$$

Thus, $w_{i,j}^{*}$ represents the proportion of variance in Z_i accounted by X_j .

Finally, the proportionate contribution of X_j to Y can then be estimated by multiplying the proportion $\hat{\alpha}_i^2$ of variance in Y accounted for by Z_i by the proportion $(\widehat{w}_{i,j}^{*})^2$ of each Z_i accounted for by X_j . The Johnson index associated with the input X_j can thus be expressed as:

$$J_j = \sigma_Y^{-2}\sum_{i=1}^{d}\alpha_i^2\, {w_{ij}^{*}}^2 \ , \tag{27}$$

for which a natural plug-in estimator can be easily obtained via

$$\hat{J}_J = \hat{\sigma}_Y^{-2}\sum_{i=1}^{d}\hat{\alpha}_i^2 {\widehat{w}_{ij}^{*}}^2 \quad \text{with} \quad \widehat{w}_{i,j}^{*} = \frac{\widehat{w}_{ij}}{\sqrt{\sum_k \widehat{w}_{kj}^2}} \ .$$

6.2 Standardized Johnson indices for the variance decomposition

As discussed in the previous Section 3.1, the R^2 decomposition (C1), which is linked to the variance decomposition, is a fundamental property that VIM should fulfill. Moreover, as presented in Section 2, the R^2 can be decomposed and estimated thanks to the covariance matrices $\widehat{\boldsymbol{\Sigma}}_{Y,\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{X}}$ (Grömping, 2006):

$$\hat{R}^2 = \hat{\sigma}_Y^{-2}\widehat{\boldsymbol{\Sigma}}_{Y,\boldsymbol{X}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{Y}} \ ,$$

and using $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{Y}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{X}}\hat{\beta}$, the latter equation gives:

$$\hat{R}^2 = \hat{\sigma}_Y^{-2}\hat{\beta}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X},\boldsymbol{X}}\hat{\beta} \tag{28}$$

Using Eqs.(22) and (24), Eq. (28) thus gives the decomposition[2]:

$$\hat{R}^2 = \hat{\sigma}_Y^{-2}\hat{\alpha}^T\hat{\alpha} \tag{29}$$

In the paper of Johnson (2000), it is quickly said that the input samples are “expressed in standard score form”. Using Eqs. (25) and (22), the standardization of the predictors imply that $w_{ij}^{*} = w_{ij}$ and, by the symmetry of $\boldsymbol{W}$, that $\sum_{i=1}^{d} {w_{ij}^{*}}^2 = 1$. The sum of the d relative weights $\sum_{i=1}^{d}\alpha_i^2 {w_{ij}^{*}}^2$ thus forms the variance decomposition of Eq. (23) and finally:

$$\sum_{j=1}^{d} J_j = \sigma_Y^{-2}\alpha^T\alpha$$

With Eq. (29), the standardization of the inputs thus gives:

$$\hat{R}^2 = \sum_{j=1}^{d}\hat{J}_j \tag{30}$$

[2] Note that, by construction, the vector α and β are associated with the same quadratic minimization problem of the function $S(\beta) = \|\boldsymbol{y}^n - \boldsymbol{X}^n\beta\|^2 = \|\boldsymbol{y}^n - \boldsymbol{P}^n\boldsymbol{Q}^T\boldsymbol{Q}\Delta\boldsymbol{Q}^T\beta\|^2 = \|\boldsymbol{y}^n - \boldsymbol{Z}^n\boldsymbol{W}\beta\|^2 = \|\boldsymbol{y}^n - \boldsymbol{Z}^n\alpha\|^2 = S(\alpha)$

Finally, it is important to note that the standardization of the inputs is equivalent to directly calculate the matrix $\boldsymbol{W}^*$ and α^* thanks to the multivariate correlation matrices $\boldsymbol{R_{X,X}}$ and $\boldsymbol{R_{X,Y}}$, as in the initial paper of Johnson (1966). The eigendecomposition of the correlation matrix $\boldsymbol{R_{X,X}}$ gives:

$$\boldsymbol{R_{X,X}} = \boldsymbol{Q}^* \Delta^{*2} \boldsymbol{Q}^{*T}.$$

The matrix $\boldsymbol{W}^*$ is then given (similarly to the Eq. (20)) by[3]:

$$\boldsymbol{W}^* = \boldsymbol{Q}^* \Delta^* \boldsymbol{Q}^{*T},$$

and the vector α^* is determined thanks to the relation:

$$\alpha^* = \boldsymbol{W}^{*-1} \boldsymbol{R_{X,Y}} .$$

Remark 7. *As previously with Eq. (24), we can also write:*

$$\alpha^* = \boldsymbol{W}^* \beta^* , \tag{31}$$

with β^* the vector of standardized coefficients presented in Section 3.3.

Finally, the *standardized Johnson index* associated with the input X_j is directly given by:

$$J_j^* = \sum_{i=1}^{d} \alpha_i^{*2} w_{ij}^{*\,2} . \tag{32}$$

Similarly to the previous case, a natural plug-in estimator can be easily derived for this standardized index.

The index in Eq. (32) respects the fundamental VIM desirability criteria such as positivity (C_2) and R^2 decomposition (C_1). Moreover, as for the LMG index, it respects both the inclusion (C_4) and the grouping (C_5) criteria, but not the exclusion (C_3) one. Indeed, Eq. (27) intuitively shows that the correlation structure of the inputs carried by $\boldsymbol{W}^*$ is distributed over the Johnson indices. The similar behavior between the LMG and Johnson indices has been confirmed in Thomas et al. (2014) who show their strict equality in the two-dimensional case (also refer to the proof in Supplementary Material A).

7. Applications on toy functions and public datasets

In this section, various metrics (VIF, PCC^2, SPCC^2) and VIMs (SRC^2, LMG, PMVD, Johnson) are computed and compared across multiple datasets. It is important to recall that, if the inputs are not independent, only LMG, PMVD, and Johnson indices are relevant since they adhere to the R^2 decomposition property. All estimations include associated confidence intervals (CI) to capture uncertainties in the estimates arising from finite sample sizes. The standard bootstrap technique is employed to derive these CIs at a 95% level, typically utilizing 100 replicates.

Table 4 summarizes the datasets used in this section, including their characteristics: dataset name and corresponding subsection, input dimension d, number of observations n, indication of quantitative vs. qualitative inputs (qt/ql), and dataset source. The first three rows pertain to simulated toy cases, while the remaining rows correspond to public datasets. Note that the "+1" mentioned in the input dimension column indicates the inclusion of a dummy correlated variable (though not explicitly part of the model).

The first three application cases are based on scenarios defined directly from linear models. The *independent case* includes, as its name suggests, independent inputs, illustrating a scenario without any collinearity. In the collinear case, a strong correlation is introduced between two variables. The third model involves a *dummy correlated input*, i.e., an input not included in the model but correlated with another model input.

[3] As a reminder, in Johnson (1966), one has $\boldsymbol{W}^* = \boldsymbol{Q}^* {\Delta^*}^{-1} {\boldsymbol{Q}^*}^T$ because $\boldsymbol{Z}^n$ is regressed on $\boldsymbol{X}^n$.

Table 4: Summary of the toy and public use cases.

Name	*§*	*d*	*n*	*qt/ql*	*Source*
Independent	7.2.1	3	100	qt	-
Collinear	7.2.2	4	100	qt	-
Dummy correlated	7.2.3	1+1	100	qt	-
Air quality	7.3	5	111	qt/ql	`airquality` dataframe (built-in `R`)
Car prices	7.4	15	804	qt/ql	`cars` dataframe (`caret` package)
Ames housing	7.5	79	2930	qt/ql	`AmesHousing` dataframe (`AmesHousing` package)

The three other applications illustrate the relevance of each VIM using real datasets. The analysis of the *car prices* dataset (here, in the context of regression, but classification is also studied in the Supplementary Material C) highlights the ability of the various VIMs to discern variable influence in the context of significant dimensionality and high multicollinearity between predictors. In this case, the approximation by a linear model is relatively validated compared to the *air quality* dataset. Lastly, the *Ames housing* dataset illustrates cases in high dimensions where it is impossible to directly determine LMG and PMVD indices, requiring the use of approximate methods such as Johnson indices.

In addition to the results presented below, Supplementary Material B provides data matrix plots, also known as *pairs plots*. These plots display the CC for each pair of variables, kernel density estimates (or histograms) for marginal distributions, and scatter plots with fitted local polynomial regressions, for each pair of variables. Additionally, the Supplementary Material includes tables containing the numerical values of the metrics and VIMs computed for each use case, which are visually represented as graphs in this section.

7.1 Computational details for reproducibility

The results in this paper (as well as in the Supplementary Material) are obtained using `R`. Both codes and datasets are available at: https://gitlab.com/LauraClouvel/toydata/. Several `R` packages are used and are briefly described below.

The sensitivity package (Iooss et al., 2023). This package[4] contains a collection of functions for GSA, from factor screening, ranking to robustness analysis. Most of the functions have to be applied on a model with scalar output, but several functions support multidimensional outputs. Single-analysis metrics (see Section 4) and multiple-analysis ones (see Section 5) are provided by this package, via the functions: `src()` (for SRC^2), `pcc()` (for PCC^2 and SPCC^2), `lmg()` (for LMG), `pmvd()` (for PMVD) and `johnson()` (for Johnson indices). The correlation ratio (see the Supplementary Material C) is computed using the `correlRatio()` function.

The car package. This package provides the VIF and GVIF metrics (`vif()` function) for multicollinearity detection (see Section 4.2).

Other standard R packages. The package boot is used for computing bootstrap confidence intervals for several metrics while the package `ggplot2` is used for visualization and displaying graphics.

7.2 Simulation data from linear models

7.2.1. Independent inputs'case (without noise)

We simulate a 100-size sample of $X = (X_1, X_2, X_3)$ with $X_1 \sim \mathcal{U}([0.5,1.5])$, $X_2 \sim \mathcal{U}([1.5,4.5])$ and $X_3 \sim \mathcal{U}([4.5,13.5])$. We study the model:

$$Y = X_1^2 + X_2 + X_3$$

The linear regression between the output and the inputs gives $R^2 = 0.999$ and $Q^2 = 0.999$.

[4] The `sensitivity` package (information: `https://cran.r-project.org/web/packages/sensitivity`, sources: `https://github.com/cran/sensitivity`) is maintained by EDF R&D (with B. Iooss as the maintainer) under a GPL-2 license.

In this initial example, given the additive model structure, X_3 is expected to exert significant influence on output variability. As depicted in Figure S1 in Supplementary Material B, both scatter plots and CC illustrate this straightforward linear impact. From Table 5, it is evident that all VIF values are unity, indicating the absence of collinearity, as anticipated.

Table 5: VIFs for the independent inputs' case.

Input	X_1	X_2	X_3
VIF	1.02	1.06	1.07

Figure 3 (also refer to Table S1 in Supplementary Material B) presents both mean estimates and bootstrap-based CIs for the metrics and VIMs. SRC^2, SPCC^2, LMG, Johnson, and PMVD successfully capture the substantial influence of X_3, while PCC^2 primarily assesses the linearity of inputs with respect to the output. In this scenario with negligible multicollinearity and a linear relationship between inputs and output ($R^2 = 0.999$), the outcomes from LMG and Johnson indices are identical.

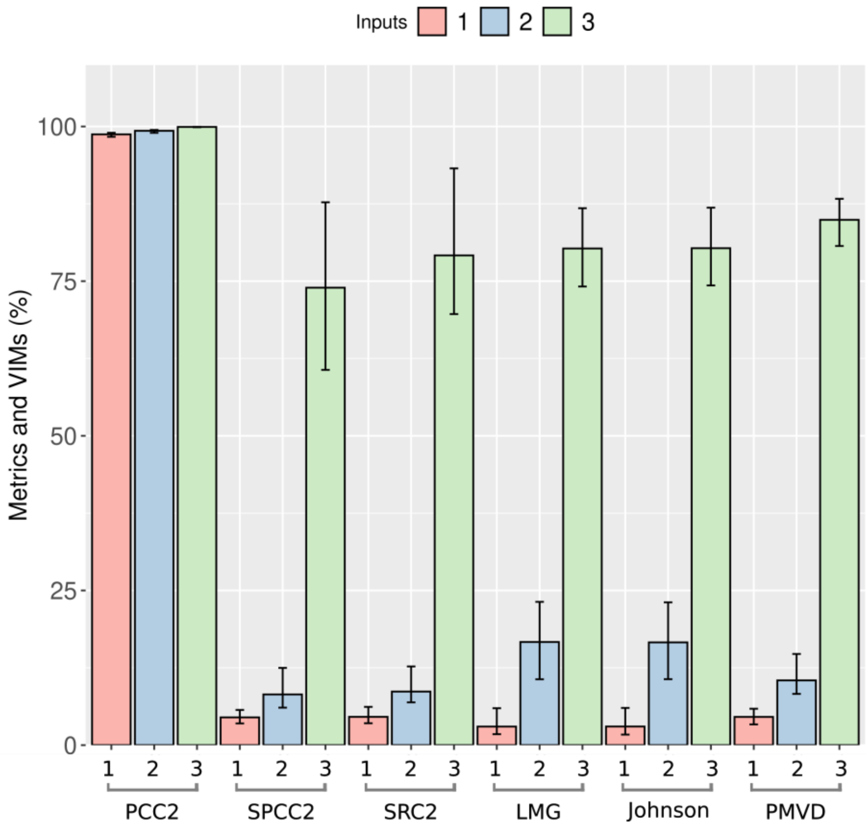

Figure 3: Estimates (with bootstrap) of the metrics (PCC^2, SPCC^2) and the VIMs (SRC^2, LMG, PMVD, Johnson) in the independent linear regression case.

7.2.2. Collinear case (without noise)

We simulate a 100-size sample of $X = (X_1, X_2, X_3, X_4)$ with $X_1 \sim \mathcal{U}([0.5,1.5])$, $X_2 \sim \mathcal{U}([1.5,4.5])$, $X_3 \sim \mathcal{U}([4.5,13.5])$, $X_4 = X_3 + \eta, \eta \sim \mathcal{N}(0,1)$. We study the model:

$$Y = X_1^2 + X_2 + X_3 + X_4$$

The linear regression between the output and the inputs gives $R^2 = 1.000$ and $Q^2 = 1.000$.

In this example, collinearity is introduced within the model between X_3 and X_4. From Table 6, one can notice that VIF values associated with X_3 and X_4 are above 10, which clearly indicates the collinearity between these two regressors. Figure S2 in Supplementary Material B presents the data matrix plot, where the CC clearly indicates a strong correlation between X_3 and X_4. Additionally, scatter plots reveal the linear influence of these two inputs on the model output.

Table 6: VIFs for the collinear case.

Input	X_1	X_2	X_3	X_4
VIF	1.02	1.06	1.07	9.46

Figure 4 (also refer to Table S2 in Supplementary Material B) displays both mean estimates and bootstrap-based CIs for the metrics and VIMs. SRC^2 confirms the earlier findings and help identify the collinearity (as their sums are far from R^2). PCC^2 primarily highlights the linear relationships between inputs and output, while SPCC^2 does not effectively capture either collinearity or relative importance. Here, SRC^2, LMG, Johnson and PMVD are able to capture that X_3 and X_4 have a similar influence, as anticipated by the strong collinearity between these two inputs. One can also observe the consistency between LMG and Johnson indices, which produce identical rankings. In this case, PMVD, LMG, and Johnson indices yield comparable results.

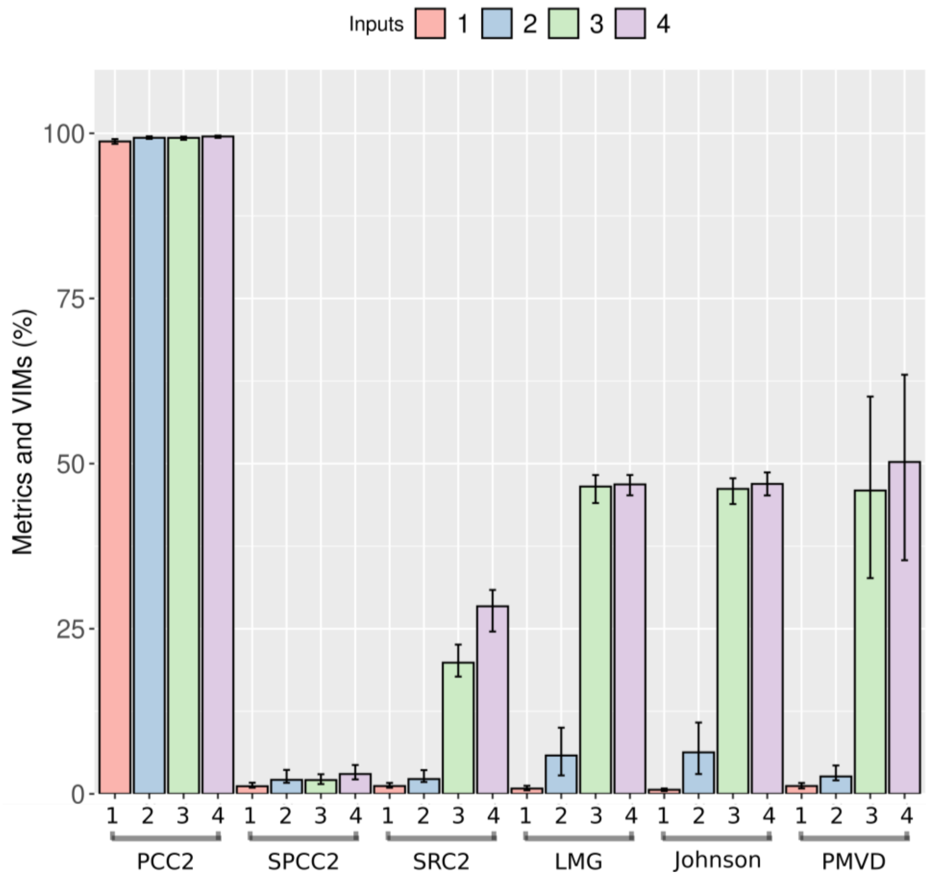

Figure 4: Estimates (with bootstrap) of the metrics (PCC^2, SPCC^2) and the VIMs (SRC^2, LMG, PMVD, Johnson) in the collinear case.

7.2.3. Model with a dummy (not included in the model) correlated input

We simulate a 100-size sample of $X = (X_1, X_2)$ with $X \sim \mathcal{N}_2\left(\binom{0}{0}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$. We study the model:

$$Y = X_1 + \eta \text{ , with } \eta \sim \mathcal{N}(0, 0.01) \text{ .}$$

The linear regression between the output and the inputs gives $R^2 = 0.992$ and $Q^2 = 0.992$.

This case introduces collinearity by means of the variable X_2 which is not directly included in the regression model, while being strongly correlated to X_1. Table 7 provides the VIFs associated with each input.

Table 7: VIFs for the collinear case.

Input	X_1	X_2
VIF	6.05	6.05

Figure 5 (also refer to Table S3 in Supplementary Material B) gives both mean estimates together with bootstrap estimates of the CIs for the metrics and the VIMs. One can see that VIF manages to catch a strong collinearity between the two inputs, while SRC^2, PCC^2, SPCC^2 and PMVD only measure the effect of X_1 (seen as a pure linear relationship with Y). Finally, the VIMs LMG, Johnson and PMVD emphasize two different interesting behaviors. LMG and Johnson allocate parts of contribution to both X_1 and $X2$, while PMVD assigns the full contribution to X_1 only. This highlights a fundamental difference between PMVD and LMG (recalled in Section 5): the PMVD formulation forces to get a null index for a non-included correlated input.

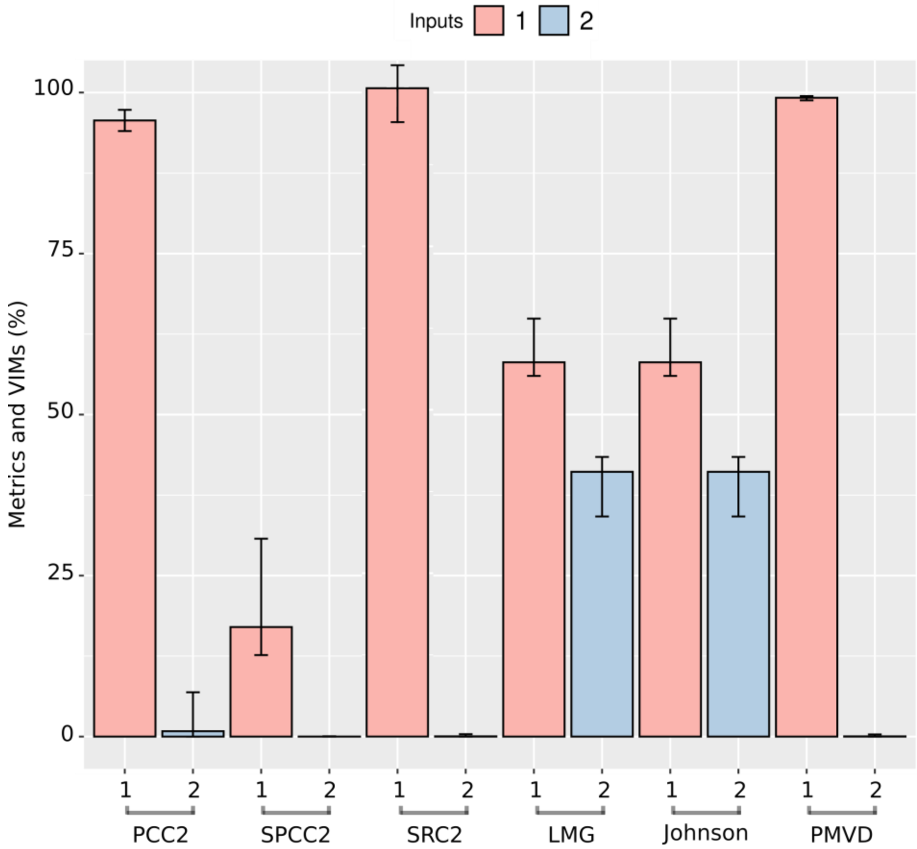

Figure 5: Estimates (with bootstrap) of the metrics and the VIMs in the dummy-correlated-variable regression case.

This test case mostly illustrates that LMG, as already pointed out for the Shapley effects (Iooss & Prieur, 2019; Hérin et al., 2022, 2024), attributes a weight to a dummy variable as soon as it is somehow correlated to another input. This behavior is also found using the Johnson indices. Such a fact is at odds with the exclusion criterion (C_3) recalled in Section 3.2.

7.3 Public dataset on air quality

We use the R dataframe "airquality", which contains some measures of the air quality of New-York in 1973. There are 153 observations but only n = 111 without missing data. In our analysis, we have only considered lines with non-missing data. The output is Ozone and the d = 5 inputs are Solar.R, Wind, Temp, Month and Day. Table 8 provides the VIFs associated with each input and no strong collinearity is detected.

Table 8: VIFs for the air quality data.

Input	Solar.R	Wind	Temp	Month	Day
VIF	1.15	1.33	1.72	1.26	1.01

The linear regression between the output and the inputs gives R^2 = 0.625 and Q^2 = 0.582 (see Fig. S5 in Supplementary Material B.3). Figure 6 provides the VIMs and metric results (also refer to Table S4 in Supplementary Material B.3). The matrix plot, given in Figure S4 in Supplementary Material B.3, clearly indicates that two inputs, Wind and Temp, are highly linearly correlated to the output (Temp has a positive influence and Wind a negative one). However, analyzing the relative influence and inferring collinearity with this matrix plot become more difficult as the dimension increases (d = 5) and the patterns of the scatter plots become rather complex.

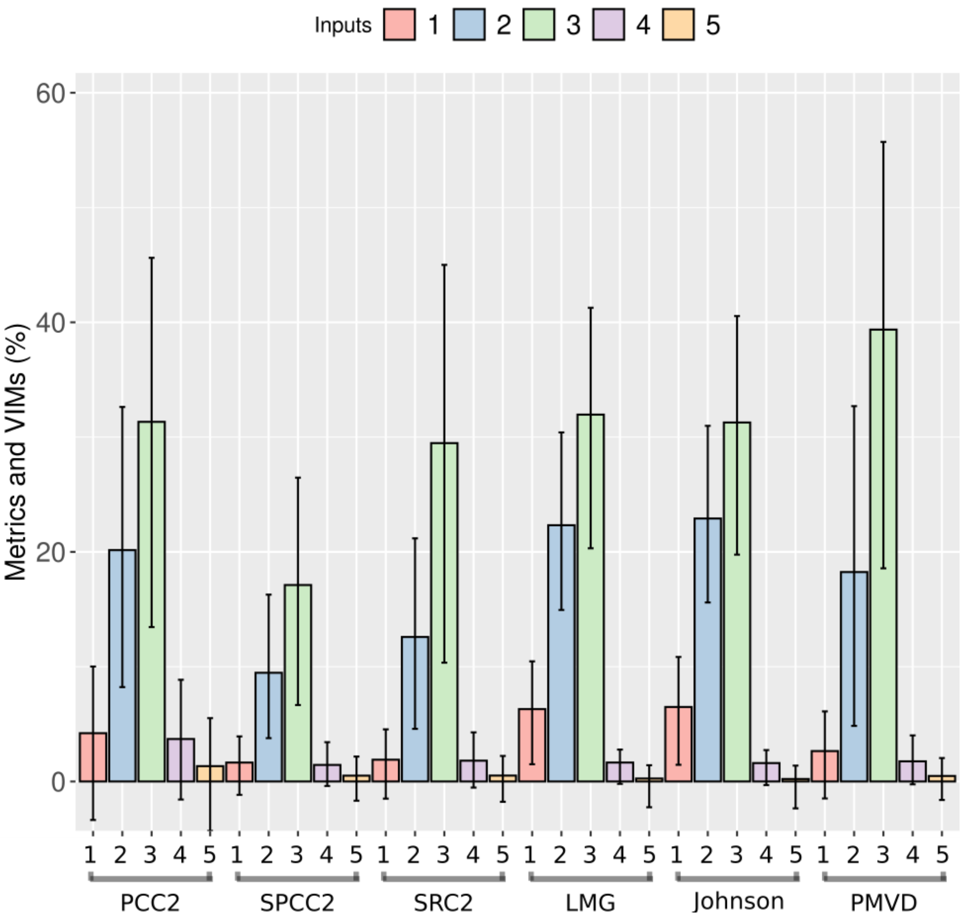

Figure 6: Estimates (with bootstrap) of the metrics (PCC^2, SPCC^2) and the VIMs (SRC^2, LMG, PMVD, Johnson) for the air quality dataset. Inputs are numbered as follows: Solar.R (1), Wind (2), Temp (3), Month (4), Day (5).

Even if no strong collinearity has been detected with VIF, a rough analysis of the matrix plot led one to believe that the correlation of −0.5 between Wind and Temp, together with the correlation of 0.4 between Temp and Month, are potential sources of collinearity. In this sense, slight differences in values can be observed, but a similar hierarchy is maintained between the PMVD and LMG/Johnson indices: the PMVD indices highlight the influence of the temperature, decreasing those of the wind and the solar irradiation. This illustrates the more discriminatory power of PMVD compared to other VIM. It is worth noting the equivalence between LMG and Johnson.

7.4 Public dataset on cars prices data

We use the cars dataset of the `R` package `caret` which comes from Kelly Blue Book resale data (2005 model year). It contains suggested retail price (column Price) and various characteristics of each car. There are n = 804 observations, one output (Price in $) and 18 inputs. For our analysis, we keep d = 15 inputs (numerical problems in linear regression with the others). One input (Mileage) is quantitative and the others are qualitative: one (Cylinder) has three modes and the 13 other inputs are binary (two modes). Large multicollinearity issues are present in these data (large VIF values observed in Table 9).

Table 9: VIFs for cars data.

Input	Mileage	Cylinder	Doors	Cruise	Sound	Leather	Buick	
VIF	1.01	2.35	4.61	1.55	1.14	1.19	2.60	
Input	Cadillac	Chevy	Pontiac	Saab	Convertible	Hatchback	Sedan	
VIF	3.33	4.41	3.42	3.56	1.63	2.45	4.51	

The linear regression between the output and the inputs gives R^2 = 0.915 and Q^2 = 0.911 (see Fig. S6 in Supplementary Material B.4). Metrics and VIMs are given in Figure 7 (also refer to Table S5 in Supplementary Material B.4 where the information on the sign of the CC between the price and each input are provided in order to know the sense of variation). Table S5 provides results for all the inputs while the Figure 7 is restricted to the eight most influential (in the SRC^2 sense) for readability purpose. As in the previous examples, LMG and Johnson are very close to each other, and PMVD allows for a better inputs' influence discrimination.

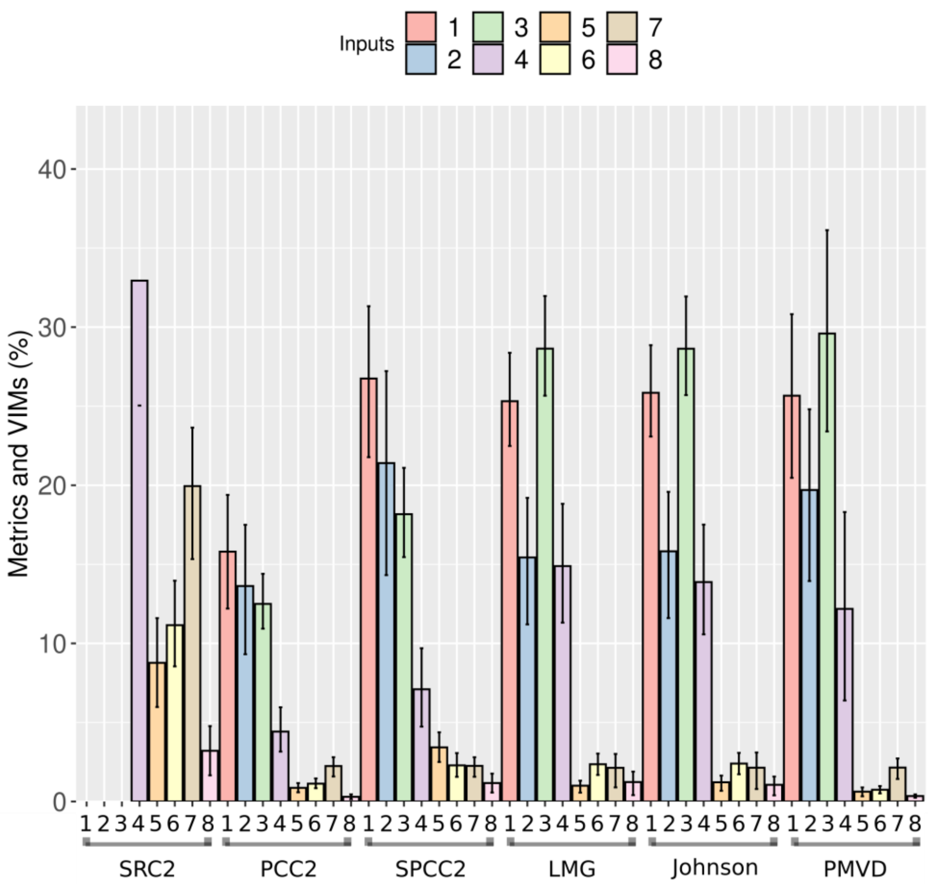

Figure 7: Estimates (with bootstrap) of the metrics and the VIMs for cars prices dataset in the regression context. Inputs are numbered as follows: Mileage (1), Cylinder (2), Cruise (3), Cadillac (4), Chevy (5), Pontiac (6), Saab (7), Convertible (8).

7.5 Public dataset: the Ames housing

We use the Ames housing dataset of the R package AmesHousing. It is a well-known dataset in the field of machine learning and data analysis. It contains d = 79 distinct features or variables that describe various aspects of residential homes in Ames, Iowa, USA. It consists of n = 2930 observations. The output variable in the dataset is "SalePrice" representing the sale price of the houses. Large multicollinearity issues are present in these data (significant VIF values are observed in Table 10).

Table 10: VIFs for the Ames housing data.

Input	SeconfFlrSF	FirstFlrF	TotalBsmtSF	YearBuilt	YearRemodAdd
VIF	3.04	4.32	3.11	2.06	1.76

Input	BedroomAbvGr	KitchenAbvGr	MasVnArea	TotRmsAbvGrd	GarageCars
VIF	2.05	1.21	1.35	4.14	1.82

The computational cost of the LMG and PMVD indices is exponential with the number of input variables. It appears impossible to calculate them for the entire set of input variables. This example shows that there are cases where it is impossible to determine the LMG and PMVD indices, and where it is necessary to use approximate methods, such as Johnson indices, to conduct sensitivity analyses. In this case, we calculated the PCC^2, SPCC^2, SRC^2, and Johnson indices for the set of 34 quantitative variables. We then determined the 10 most influential variables with respect to the SRC^2 indices. Finally, we determined all the VIMs and metrics for these 10 variables.

Remark 8. *Note that these two choices (of 34 and 10 variables) are arbitrary. Their purpose is simply to demonstrate the utility of Johnson indices on a large number of variables compared to LMG indices and to illustrate the interpretative differences brought by PMVD in relation to LMG/Johnson indices on such a dataset. Thus, analyzing such a dataset would actually require more careful consideration of the relevant variables to select in order to best interpret the results.*

The linear regression between the output and the 34 inputs gives R^2 = 0.799 and Q^2 = 0.770 (see Fig. S8 in Supplementary Material B.5). The linear regression between the output and the 10 inputs gives R^2 = 0.777 and Q^2 = 0.769. Metrics and VIMs are given in Figure 8 for the 10 variables (also refer to Tables S6 and S7 resp. for the 10 and 34 variables in Supplementary Material B.5). The matrix plot is given for the 10 first variables in Figure S7 of Supplementary Material B.5. It shows that strong dependencies exist between inputs and that quite a complex relation links the output with the inputs. Table 10 also show the multicollinearity present in these data with larger VIF for some inputs. Therefore, large differences between SRC^2 and LMG/Johnson appear. Moreover, the interest of PMVD compared to LMG/Johnson is exemplified: the PMVD indices provide a hierarchy that is generally similar to those of the LMG/Johnson indices, but with higher values for the variables SecondFlrSF and FirstFlrSF (which are the variables with a higher SRC^2 index) and lower values for the variables TotRmsAbvGrd and MasVnrArea (which are variables with a low SRC^2 index). Finally, the proximity between LMG and Johnson values is again highlighted, even with a moderate quality of the linear regression model.

8. Discussion

In this work, the relative importance has been considered as the contribution each input makes to the coefficient of determination R^2, considering both its direct effect (i.e., correlation with the output) and its indirect effect (i.e., correlation with other inputs). A distinction was made between the VIM (SRC^2) and metrics (CC^2, PCC^2 and SPCC^2) based on a single regression analysis and the VIMs requiring multiple regression analyses: the LMG corresponding to Shapley effects, the PMVD and the Johnson indices corresponding to relative weight analysis.

Figure 9 allows to distinguish VIMs regarding their respective positioning, conditions of use, intrinsic capabilities, and interpretation. Before computing VIMs, the first step consists of performing the linear regression and looking at the validation diagnostics (that has been recalled in Section 2). It allows to understand what portion of the output variance will be explained by VIMs. Computing the VIFs will also help to choose the right VIM and to provide a right explanation in case of strong multicollinearity.

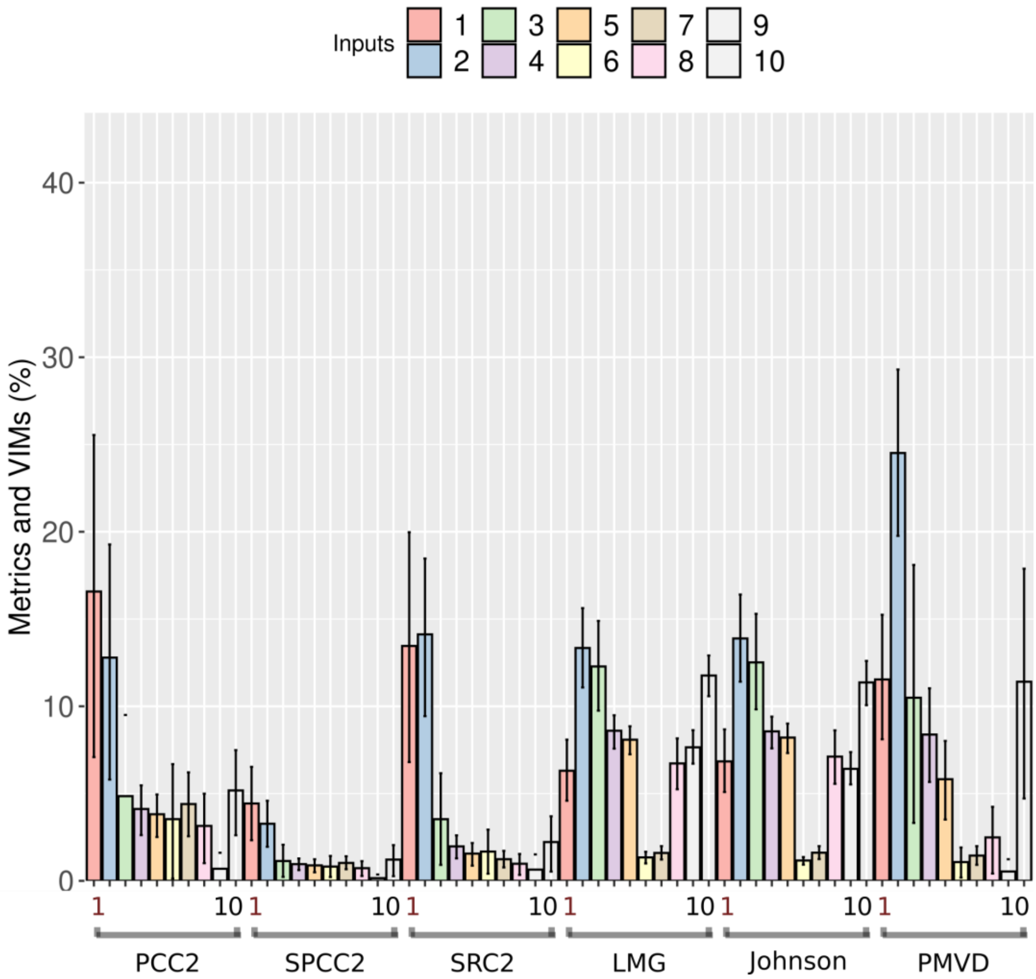

Figure 8: Estimates (with bootstrap) of the metrics (PCC², SPCC²) and the VIMs (SRC², LMG, PMVD, Johnson) for the Ames housing data. Inputs are numbered as follows: SecondFlrSF (1). FirstFlrSF (2). TotalBsmtSF (3). YearBuilt (4). YearRemodAdd (5). BedroomAbvGr (6). KitchenAbvGr (7). MasVnrArea (8) TotRmsAbvGrd (9) GarageCars (10).

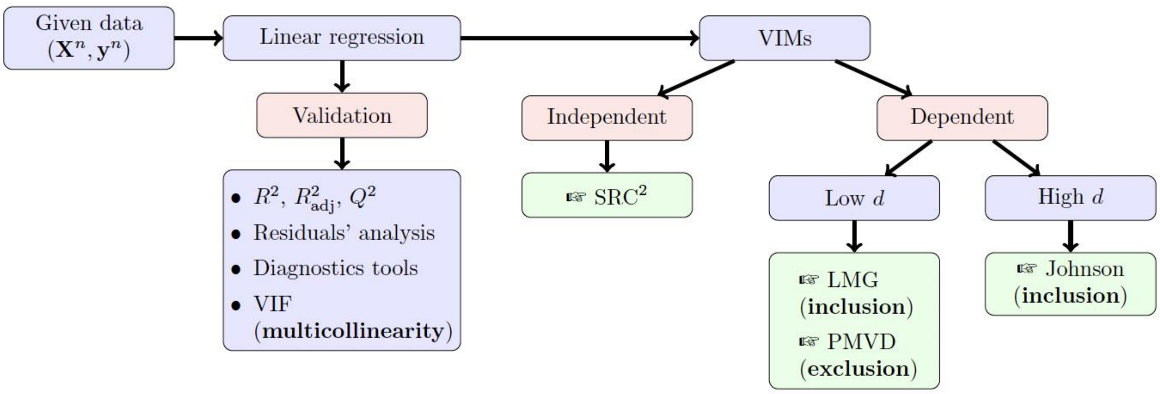

Figure 9: Summary of useful metrics and VIMs.

For independent inputs, SRC² can be used. In cases of multicollinearity (where inputs exhibit significant dependencies), LMG, PMVD and Johnson indices are recommended VIMs due to their ability to partition R^2 among the inputs (a key criterion for R^2 decomposition). The LMG and Johnson indices quantify the additional contribution of each input to R^2, classifying inputs based on their contributions while considering the weights of inter-input correlations. This means an input can hold relative importance even without direct influence on the model output. Therefore, using LMG and Johnson indices necessitates a thorough analysis of correlation information (direct, indirect, incidental) among inputs. In contrast, PMVD indices ensure that an input with a zero estimated regression coefficient contributes zero to this measure of relative importance by design.

Table 11 (inspired from Grömping, 2015) synthesizes the desirability criteria (described in detail in Section 3.1) that these three VIMs satisfy.

Table 11: Adequation between the VIMs and their desirability criteria.

VIM	*(C_1)*	*(C_2)*	*(C_3)*	*(C_4)*	*(C_5)*
LMG	x	x		x	x
PMVD	x	x	x	x	
Johnson	x	x		x	x

9. Conclusion

This work introduces various methods to assess the relative importance of predictors/inputs in linear regression models (the Supplementary Material C of this paper extends these methodologies to the classification context using logistic linear regression). Interpretations and conditions of use for these importance measures are developed based on output variance decomposition, with specific considerations for global sensitivity analysis (GSA). Ultimately, this work aims to provide a practical user guide for practitioners (see e.g., Iooss et al., 2022), highlighting the utility of such guides within the GSA community (Iooss & Lemaître, 2015).

Several datasets are employed to simulate and analyze the effects measured by these VIMs. The results confirm theoretical properties and intuitions, such as the close relationship between LMG and Johnson indices. The preference to use LMG/Johnson or PMVD thus depends on whether the user wants to consider causality effects or not (Grömping, 2015; Zhao & Hastie, 2021).

A significant practical limitation of the LMG and the PMVD methods is the complexity of their calculation which is proportional to $2d$, the number of possible subsets in a set of d inputs. It has been shown that the Johnson indices can give an excellent alternative to measure the multicollinearity effects when deriving importance measures in a regression model containing a large number (several dozens) of inputs. In this case, the LMG and the PMVD computation is practically impossible. The Johnson indices are in fact a good approximation of the LMG indices in linear regression context. Our ongoing research endeavors to develop similar approximations for PMVD.

Concerning the linear model restriction of all the metrics developed in this paper, other works have developed metrics valid in the general case of nonlinear models. For instance, Hérin et al. (2024) extended PMVD to nonlinear models through novel sensitivity indices termed PME, inspired by cooperative game theory and based on proportional value allocation rules. Extending Johnson indices to nonlinear models remains a significant challenge, with initial attempts noted in Iooss & Clouvel (2023). We consider that such results are achievable thanks to the strong connections between the statistical literature on regression analysis and the field of GSA that we have emphasized in this paper.

Supplementary Material

The Supplementary Material can be found online at: https://sesmo.org/article/view/18681/18318.

References

Achen, C. H. (1982). Interpreting and Using Regression. SAGE Publications.

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety, 206(107312).

Belsley, D., Kuh, E., & Welsch, R. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons, Inc.

Bénard, C., Biau, G., Veiga, S. D., & Scornet, E. (2022). SHAFF: Fast and consistent SHApley eFfect estimates via random Forests. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, volume 151 of PMLR, Valencia, Spain.

Benoumechiara, N. & Elie-Dit-Cosaque, K. (2019). Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms. ESAIM: Proceedings and Surveys, 65:266–293.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. Journal of Sensory Studies, 27:87–101.

Blanchard, J.-B. (2023). Sensitivity analysis with correlated inputs: comparison of indices for the linear case. International Journal for Uncertainty Quantification, 13:25–56.

Borgonovo, E. & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. European Journal of Operational Research, 248:869–887.

Broto, B., Bachoc, F., & Depecker, M. (2020). Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution. SIAM/ASA Journal on Uncertainty Quantification, 8:693–716.

Broto, B., Bachoc, F., Depecker, M., & Martinez, J.-M. (2019). Sensitivity indices for independent groups of variables. Mathematics and Computers in Simulation, 163:19–31.

Budescu, D. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. Psychological Bulletin, 114:542–551.

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example. John Wiley & Sons, New York. Christensen, R. (1990). Linear models for multivariate, time series and spatial data. Springer-Verlag.

Clouvel, L. (2019). Uncertainty quantification of the fast flux calculation for a PWR vessel. Thèse de l'Université Paris- Saclay.

Clouvel, L., Mosca, P., Martinez, J., & Delipei, G. (2019). Shapley and Johnson values for sensitivity analysis of PWR power distribution in fast flux calculation. In M&C 2019, Portland, USA.

Da Veiga, S., Gamboa, F., Iooss, B., & Prieur, C. (2021). Basics and Trends in Sensitivity Analysis. Theory and Practice in R. SIAM.

Darlington, R. & Hayes, A. (2017). Regression analysis and linear models. The Guilford Press.

Deng, X., Yin, L., Peng, S., & Ding, M. (2015). An iterative algorithm for solving ill-conditioned linear least squares problems. Geodesy and Geodynamics, 6(6):453–459.

Fekhari, E., Iooss, B., Muré, J., Pronzato, L., & Rendas, J. (2023). Model predictivity assessment: incremental test-set selection and accuracy evaluation. In Salvati, N., Perna, C., Marchetti, S., & Chambers, R., editors, Studies in Theoretical and Applied Statistics, SIS 2021, Pisa, Italy, June 21-25, pages 315–347. Springer.

Feldman, B. (2000). The proportional value of a cooperative game. In Econometric Society World Congress 2000 Con- tributed papers, number 1140. Econometric Society.

Feldman, B. (2005). Relative importance and value. SSRN Electronic Journal.

Fox, J. & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association, 87(417):178–183.

Genizi, A. (1993). Decomposition of $R2$ in multiple regression with correlated regressors. Statistica Sinica, pages 407–420.

Grömping, U. (2006). Relative importance for linear regression in R: the Package relaimpo. Journal of Statistical Software, 17:1–27.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. The American Statistician, 61(2).

Grömping, U. (2015). Variable importance in regression models. WIREs Comput Stat, 7(137-152).

Helton, J. (1993). Uncertainty and sensitivity analysis techniques for use in performance assesment for radioactive waste disposal. Reliability Engineering & System Safety, 42:327–367.

Helton, J., Johnson, J., Salaberry, C., & Storlie, C. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering & System Safety, 91:1175–1209.

Hérin, M., Il Idrissi, M., Chabridon, V., & Iooss, B. (2022). Proportional marginal effects for sensitivity analysis with correlated inputs. In Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output (SAMO 2022), Tallahassee, Florida, USA.

Hérin, M., Il Idrissi, M., Chabridon, V., & Iooss, B. (2024). Proportional marginal effects for global sensitivity analysis. SIAM/ASA Journal of Uncertainty Quantification, 12:667–692.

Il Idrissi, M., Chabridon, V., & Iooss, B. (2021a). Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. Environmental Modelling & Software, 143(105115).

Il Idrissi, M., Iooss, B., & Chabridon, V. (2021b). Mesures d'importance relative par décomposition de la performance de modèles de régression. In Actes des 52èmes Journées de Statistique de la Société Française de Statistique (SFdS), pages 497–502, Nice, France.

Iooss, B., Chabridon, V., & Thouvenot, V. (2022). Variance-based importance measures for machine learning model interpretability. In Actes du 23ème Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement ($\lambda\mu$23), Saclay, France. Institut pour la Maîtrise des Risques, URL: https://hal.science/IMDR/hal-03878431v1.

Iooss, B. & Clouvel, L. (2023). Une méthode d'approximation des effets de Shapley en grande dimension. In Actes des 54èmes Journées de Statistique de la Société Française de Statistique (SFdS), Brussels, Belgium, URL: https://hal.science/hal-04088622/document.

Iooss, B., Da Veiga, S., Janon, A., & Pujol, G. (2023). sensitivity: Global Sensitivity Analysis of Model Outputs. R package version 1.29.0.

Iooss, B. & Lemaître, P. (2015). A review on global sensitivity analysis methods. In Meloni, C. & Dellino, G., editors, Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications, pages 101–122. Springer.

Iooss, B. & Prieur, C. (2019). Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications. International Journal for Uncertainty Quantification, 9:493–514,.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). An introduction to statistical learning: With applications in R, 7th edition. Springer.

Johnson, J. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behavioral Research, 35:1–19.

Johnson, J. & LeBreton, J. (2004). History and use of relative importance indices in organizational research. Organizational Research Methods, 7:238–257.

Johnson, R. (1966). The minimal transformation to orthonormality. Psychometrika, 31:61–66. Karch, J. (2020). Improving on Adjusted R-Squared. Collabra: Psychology, 6(1):1–11.

Kruskal, W. (1987). Relative importance by averaging over orderings. The American Statistician, 41:6–10. Kurowicka, D. & Cooke, R. (2006). Uncertainty analysis with high dimensional dependence modelling. Wiley.

Lepore, A., Palumbo, B., & Poggi, J.-M., editors (2022). Interpretability for Industry 4.0: Statistical and Machine Learning Approaches. Springer.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to bivariate and multivariate analysis. Scott Foresman and Company, Glenview, IL.

Marrel, A., Iooss, B., Van Dorpe, F., & Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis, 52:4731–4744.

McCullagh, P. & Nelder, J. (1989). Generalized linear models. Chapman & Hall.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning - A brief history, state-of-the-art and challenges. In PKDD/ECML Workshops 2020, pages 417–431.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. Practical assessment, research & evaluation, 17(9):n9.

Nimon, K. & Oswald, F. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. Organizational Research Methods, 16:650–674.

Owen, A. (2014). Sobol' indices and Shapley value. SIAM/ASA Journal on Uncertainty Quantification, 2:245–251.

Owen, A. & Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. SIAM/ASA Journal on Uncertainty Quantification, 5:986–1002.

Perez, D. (2022). An attempt of reproduction of sovacool et al.'s differences in carbon emissions reduction. EPJ Nuclear Science & Technology, 8:24.

Plischke, E., Rabitti, G., & Borgonovo, E. (2021). Computing Shapley effects for sensitivity analysis. SIAM/ASA Journal on Uncertainty Quantification, 9:1411–1437.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J., Jakeman, J., Gupta, H., Melillo, N., Rabiti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Kucherenko, S., & Maier, H. (2020). The future of sensitivity analysis: An essential discipline for systems modelling and policy making. Environmental Modelling and Software, 137(104954).

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Espeland, W., Kay, J., Lo Piano, S., Mayo, D., Jr, R., Portaluri, T., Porter, T., Puy, A., Rafols, I., Ravetz, J., Reinert, E., Sarewitz, D., Stark, P., & Vineis, P. (2020). Five ways to ensure that models serve society: a manifesto. Nature, 582:482–484.

Saltelli, A., Chan, K., & Scott, E., editors (2000). Sensitivity analysis. Wiley Series in Probability and Statistics. Wiley. Shapley, L. (1953). A value for n-persons game. In Kuhn, H. & Tucker, A., editors, Contributions to the theory of games II, Annals of mathematic studies. Princeton University Press, Princeton, NJ.

Sobol', I. (1993). Sensitivity estimates for non linear mathematical models. Mathematical Modelling and Computational Experiments, 1:407–414.

Song, E., Nelson, B., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. SIAM/ASA Journal on Uncertainty Quantification, 4:1060–1083.

Sovacool, B., Schmid, P., Stirling, A., Walter, G., & MacKerron, G. (2020). Differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power. Nature Energy, 5:928.

Thomas, D., Zumbo, B., Kwan, E., & Schweitzer, L. (2014). On johnson's (2000) relative weigths method for assessing variable importance: A reanalysis. Multivariate Behavioral Research, 49:329–338.

Tonidandel, S. & LeBreton, J. M. (2015). Rwa web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. Journal of Business and Psychology, 30:207–216.

Wagner, F. (2021). CO2 emissions of nuclear power and renewable energies: a statistical analysis of european and global data. The European Physical Journal Plus, 136:562.

Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In Proceedings of the 16th Conference of the Applied Stochastic Models and Data Analysis, Le Pirée, Greece.

Wallard, H. (2019). Grouping property and decomposition of explained variance in linear regression. In Skiadas, C. & Bozeman, J., editors, Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining, pages 73–89. Wiley.

Weber, R. J. (1988). Probabilistic values for games. In The Shapley Value, pages 101–120. Cambridge University Press, 1st edition.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: a comprehensive review. Reliability Engineering & System Safety, 142:399–432.

Zhao, K. & Hastie, T. (2021). Causal Interpretations of Black-Box Models. Journal of Business & Economic Statistics, 39(1):272–281.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320.