Valid Interpretation of Feature Relevance for Linear Data Mappings

Benoît Frénay[1*] Daniela Hofmann[2*] Alexander Schulz[2*] Michael Biehl[3]
Barbara Hammer[2]
[1] Machine Learning Group, ICTEAM Institute, Université catholique de Louvain,
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
[2] Bielefeld University - CITEC centre of excellence, Germany
[3] University of Groningen, Mathematics and Computing Science,
P.O. Box 407, 9700 AK Groningen, The Netherlands

Preprint of the publication [1], as provided by the authors. DOI=10.1109/CIDM.2014.7008661.

1 Introduction

Machine learning (ML) methods constitute core technologies in the era of big data [2]: successful applications range from everyday tasks such as spam classification up to advanced biomedical data analysis. Further, today's most significant machine learning models are supported by strong theoretical guarantees such as their universal approximation capability and generalisation ability. Still, it is a long way to enable the direct use of advanced ML technology in complex industrial applications or settings where a human has to take responsibility for the results. Most popular ML models act as black boxes and do not reveal insight into why a decision has been taken [3]. Hence the accuracy on the given data is the sole information based on which practitioners can decide to use a model. Despite strong theoretical results under idealised assumptions, this can be extremely problematic, since these assumptions are usually not met in practice. Further, black box models are restricted to a mere functional inference. Auxiliary information is not extracted, albeit often aimed for e.g. in biomedical data analysis. These facts have caused a strong interest in interpretable ML models, with first promising results in specific domains such as biomedical data analysis [4–10].

Linear (or locally linear) data transformations constitute a particularly prominent element in machine learning which seemingly combines efficient and well founded training algorithms with interpretable model components. Global linear models such as ridge regression, linear discriminant analysis, or principal component analysis

*Those authors contributed equally to this work.

constitute premier techniques in many application domains in particular if high data dimensionality is involved [11]. Besides, the very active field of metric learning usually aims for an adaptive quadratic form, which essentially corresponds to a linear transformation of the data. Many different successful approaches have recently been proposed in this context, see e.g. [12, 13]. One of the striking properties of linear models is that they seemingly allow an interpretation of the relevance of input features by inspecting their corresponding weighting; in a few cases, such techniques have led to striking semantic insights of the underlying process [14]. Thus, these models carry the promise of fast and flexible learning algorithms, which directly address a simultaneous, quantitative, and interpretable weighting of the given features, provided linear data modelling is appropriate.

Recent results, however, have shown that the interpretation of linear weights as relevance terms can be extremely misleading in particular for high-dimensional data [15]: those data likely display correlations of the features, hence relevance terms can be high due to purely statistical effects of the data. Conversely, highly correlated but very important features can be ranked low due to the fact that they share their impact. In the contribution [15] a first cure to partially avoid these effect by a L_2 regularisation has been proposed; in particular in the case of feature correlations, the approach still fails to provide efficient bounds for the minimum and maximum feature relevance, hence it offers a partial solution of the problem only. In this contribution we propose a L_1 regularisation instead, which allows an efficient formalisation of the minimum and maximum feature relevance as a linear programming problem. Since many recent datasets are characterised by their high dimensionality, this constitutes a crucial step for feature relevance interpretability in many modern domains.

Very high data dimensionality is becoming more and more prominent. For example, in omics studies, many genes are simultaneously considered [16, 17]. Even if having more information may seem beneficial at first glance, this wealth of features can also be problematic. Indeed, machine learning in high-dimensional space suffers from the curse of dimensionality [18, 19], also known as the empty space phenomenon. This is due to the fact that the size of a dataset should scale exponentially with its dimensionality, what cannot be achieved in practice. Other counterintuitive phenomena like the concentration of distances [20] occur, what causes distances to be less useful in high-dimensional spaces. Eventually, high-dimensional data are harder to analyse and to visualise for human experts. As argued above, direct feature ranking in linear maps can easily loose its interpretability in this situation.

Feature selection [21] is a common preprocessing for high-dimensional data, and we will compare our modelling to classical feature selection. Feature selection consists in selecting a few relevant features which allow reaching good prediction performances with easy-to-interpret models. For example, least angle regression (LARS) [22, 23] obtains sparse feature subsets for linear regression. Many methods have been proposed for non-linear models, based e.g. on mutual information [24–30]. Such solutions improve the performances of subsequently used machine learning algorithms. In our setting, we are not so much interested in a sparse linear representation, rather we address the question, given a linear mapping, what is the relevance of features for the given mapping, taking into account all possible invariances inherent in the data. Con-

cerning this question, classical feature selection, though very powerful, is not entirely satisfying when it comes to interpretability. Indeed, most feature selection algorithms only provide either a unique subset of features or a path of feature subsets of increasing size. This leaves out an important part of the information. For example, if two relevant features are linearly dependent, the LARS algorithm may arbitrarily include any of them in the feature subset, what may incorrectly suggest that the other feature is irrelevant. Also, most feature selection methods do not specify which features are strictly necessary, what may be interesting to understand the system under study.

These limitations of feature selection can be alleviated using the concept of strong and weak relevance [31–33]. Strongly relevant features provide new information, even if all other features are already used. Weakly relevant features may provide new information, but only if certain features (e.g. redundant ones) are not simultaneously considered. In general, the determination of weakly relevant features requires exhaustive search over all feature subsets [33]. In this paper, we restrict to linear mappings only, ignoring possible nonlinear effects. We are interested in the relevance of the features for the given mapping, aiming at both, strong and weak feature relevance. We do not strictly follow the formal definition of strong and weak feature relevance for linear settings, but we will use a different formalisation which is inspired by these terms but allows efficient modelling. Essentially, we will consider two weight vectors of a given mapping as equivalent, if they have the same (or a similar) classification behaviour and the same (or similar) length of the weight vector, thus accounting for a similar signal to noise ratio or generalisation ability, respectively. Then we propose a measurement similar to weak and strong feature relevance by the minimum and maximum weight of a feature in this equivalence class. These bounds give an interpretable interval for the feature relevance.

This paper is organised as follows. First, Section 2 discusses the problem of weak and strong relevance for linear relationships. The concept of bounds for feature relevance is introduced, as well as a simple, generic reference algorithm. Section 3 proposes a new algorithm to find strongly and weakly relevant features for linear models (and the corresponding feature relevance bounds). Experiments are performed in Section 4 and Section 5 concludes this paper.

2 Definition and Measure of Feature Relevance

This section defines the concept of feature relevance and discusses a simple algorithm to quantify it, aiming at approximations of the formal concept of weak and strong feature relevance. For linear mappings, a similar mathematical definition is proposed in Section 3 which resembles the underlying ideas but directly gives rise to an efficient solution.

2.1 Feature Relevance

The question what means feature relevance has been extensively discussed, see e.g. the survey [34] and the approaches [35, 36]. The notion of strong and weak feature relevance has been defined in [31–33]. Assume the task is to predict a target Y based on d features $X_1 \dots X_d$, which can be either continuous (regression) or discrete (classification). A variable Y is *conditionally independent* of a variable X_j given a set

of variables S, if $P(Y|X_j, S) = P(Y|S)$. This is denoted as $Y \perp\!\!\!\perp X_j|S$. A feature X_j is *strongly relevant* to predict Y iff

$$Y \not\perp\!\!\!\perp X_j|X_{(j)} \tag{1}$$

where $X_{(j)}$ is the set of all features except X_j. Strongly relevant features are strictly necessary to achieve good prediction, since they contain some information which is not provided by any other feature. Finding theses features is particularly interesting to understand the studied process, since these features are likely to play a key role.

A feature X_j is defined as *weakly relevant* to predict Y iff it is not strongly relevant and

$$Y \not\perp\!\!\!\perp X_j|S \tag{2}$$

for some feature subset $S \subset X_{(j)}$. A weakly relevant feature is not necessarily useful, since it provides information which is also contained in other features. Indeed, $Y \perp\!\!\!\perp X_j|X_{(j)}$ holds if the feature X_j is not strongly relevant (first part of the definition). This can occur if X_j is redundant with other features, for example. Nonetheless, experts are often still interested in such features: some weakly relevant features are often necessary for a good model accuracy, albeit the choice is not necessarily unique. Further, weakly relevant features are often crucial to understand the complex relationships between the features and the target. One example is explained in [33]: in gene expression analysis, experts '*are primarily interested in identifying all features (genes) that are somehow related to the target variable, which may be a biological state such as "healthy" vs. "diseased"*' [37, 38].

2.2 Searching for Relevant Features

Under reasonable assumptions, generic (but potentially time consuming) algorithms are proposed in [33] to find strongly and weakly relevant features. We recall this procedures for convenience. Strongly relevant features can be found by selecting all features whose removal lowers the prediction performance. Assume there is given a classifier with prediction error $c(S)$ based on the feature set S. Then these features corresponds to the subset $\left\{X_j | c\left(X_{(j)}\right) > c(X) + \epsilon\right\}$ where the parameter $\epsilon > 0$ controls the trade-off between prediction and recall [33]. This backward procedure is efficient, since this criterion must only be estimated d times.

Weakly relevant feature are much harder to find. When directly testing the definition, one has to consider the $\mathcal{O}\left(2^d\right)$ possible feature subsets $S \subset X_{(j)}$ for the conditional dependence $Y \not\perp\!\!\!\perp X_j|S$. In practice, such an exhaustive search is not affordable and one has to rely on heuristics to find weakly relevant features. For example, the recursive independence test (RIT) algorithm [33] first finds the features X_j satisfying $Y \not\perp\!\!\!\perp X_j$. Then, it recursively adds all the other features $X_{j'}$ which are pairwise dependent with respect to those features, i.e. $X_j \not\perp\!\!\!\perp X_{j'}$. For each step, a (specific) statistical independency test is required.

2.3 Bounds for Feature Relevance

The algorithms described in Section 2.2 find sets of relevant features, whereby weakly relevant features can only approximately be determined efficiently. We are interested in a yet different setting: on the one hand, we do not necessarily consider a clear objective such as the classification error, rather our goal is to interpret the

relevance of features for a given linear mapping and data set. In addition, we are not only interested in qualitative results, indicating a feature as relevant or irrelevant, respectively. Rather, we would like to identify an interval for every feature which quantifies the minimum and maximum relevance the feature might have for the given mapping. Thus, such bounds should not only indicate whether features are strongly or weakly relevant, but also *how much* they are relevant. A non-zero lower bound indicates that a feature is strongly relevant, whereas a large upper bound points out that the feature is at least weakly relevant.

In the following, we will focus on linear relationships, which are common in biomedicine or social sciences, and particularly interesting for the case of high data dimensionality, i.e. a potentially large number of correlated features. In this section, inspired by the formal notion of strong and weak feature relevance, we propose a generic approach which is suitable for low dimensionalities and which can serve as a basic comparison. Afterwards, in Section 3, we propose another efficient method to compute feature relevance bounds. This is then tested in Section 4.

2.4 Generic Approach to Compute Feature Relevance Bounds

Using the same idea as the algorithm in [33] which finds strongly relevant features (see Section 2.2), the following algorithm computes lower bounds for the feature relevance.

Algorithm 1 Compute lower bounds for feature relevance

Input: criterion c and dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1\ldots n}$
Output: lower bound l_j for each feature X_j

compute $c(\mathcal{D})$
for $j = 1 \ldots d$ **do**
  $l_j \leftarrow c\left(\mathcal{D}_{X_{(j)}}\right) - c(\mathcal{D})$
end for

Here, $\mathcal{D}_{X_{(j)}}$ is the dataset restricted to the features $X_{(j)}$ and c measures the relevance of a feature subset to predict Y. Hence, the difference $c\left(\mathcal{D}_{X_{(j)}}\right) - c(\mathcal{D})$ can be interpreted as the minimum contribution of X_j to the total relevance. This quantity is used as a lower bound l_j to the relevance of feature X_j. It is non-zero if X_j is strongly relevant.

For upper bounds, an exhaustive search would be necessary, but intractable in practice. Instead, a greedy forward-backward search is used in the following algorithm.
Here, $\mathcal{C}$ and $\mathcal{S}$ are the subsets of candidate and selected features, respectively. If c is the mean square error, the quantity $c(\mathcal{D}_\emptyset)$ is defined as the target variance. Also, NB_FB_STEPS is the number of backward and forward steps which are performed. Using greedy algorithms like the above forward-backward search is a standard approach in feature selection. Even if it is not optimal, it often gives good results. The particularity of the above greedy search is that the search criterion is the upper bound itself. In other words, the algorithm searches for the feature subset which allows a given feature to be as useful as possible. The number of steps is deliberately limited because (i) weakly relevant features are unlikely to be highly relevant when a lot of other features are simultaneously considered and (ii) the estimation of c is often less reliable when the dimensionality increases. Also, computing the upper bounds with Alg. 2 requires to evalute $\mathcal{O}(d^2 \times \text{NB\_FB\_STEPS})$ times

Algorithm 2 Compute upper bounds for feature relevance

Input: criterion c, dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1\dots n}$
lower bounds l_j for every feature X_j
Output: upper bound u_j for each feature X_j

compute $c(\mathcal{D}_\emptyset)$
for $j = 1 \dots d$ **do**
  // initialise upper bound
  $u_j \leftarrow \max\left(l_j, c(\mathcal{D}_\emptyset) - c\left(\mathcal{D}_{X_j}\right)\right)$
  $\mathcal{C} \leftarrow \{1 \dots d\} \setminus \{j\}$
  $\mathcal{S} \leftarrow \emptyset$

  // forward search steps
  for $s = 2 \dots$ NB_FB_STEPS **do**
    // find next feature to add to $\mathcal{S}$
    for $k \in \mathcal{C}$ **do**
      $\Delta c_k = c\left(\mathcal{D}_{X_{\mathcal{S}\cup\{k\}}}\right) - c\left(\mathcal{D}_{X_{\mathcal{S}\cup\{j,k\}}}\right)$
    end for
    $k^* \leftarrow \arg\max_{k\in\mathcal{C}} \Delta c_k$
    $u_j \leftarrow \max(u_j, \Delta c_{k^*})$

    $\mathcal{C} = \mathcal{C} \setminus \{k^*\}$
    $\mathcal{S} = \mathcal{S} \cup \{k^*\}$
  end for

  // backward search steps
  for $s =$ NB_FB_STEPS $\dots 2$ **do**
    // find next feature to remove from $\mathcal{S}$
    for $k \in \mathcal{S}$ **do**
      $\Delta c_k = c\left(\mathcal{D}_{X_{\mathcal{S}\setminus\{k\}}}\right) - c\left(\mathcal{D}_{X_{\mathcal{S}\setminus\{k\}\cup\{j\}}}\right)$
    end for
    $k^* \leftarrow \arg\max_{k\in\mathcal{C}} \Delta c_k$
    $u_j \leftarrow \max(u_j, \Delta c_{k^*})$

    $\mathcal{S} = \mathcal{S} \setminus \{k^*\}$
  end for
end for

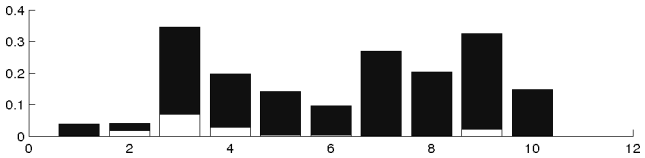

Figure 1: Lower and upper bounds of feature relevance given by Alg. 1 and Alg. 2 for the diabetes dataset. c is the mean square error of a linear regression.

the criterion c. It is therefore necessary to use a small value for NB_FB_STEPS. Here, we use NB_FB_STEPS $= 6$ as a compromise between accuracy and efficiency.

Fig. 1 shows the lower and upper bounds obtained for the diabetes dataset used in the original LARS paper [22]. The 10 features for the 442 patients are the age, the sex, the body mass index (BMI), the blood pressure (BP) and 6 blood serum measurements $X_5 \dots X_{10}$. The goal is to predict a measure Y of diabetes progression one year after feature acquisition. Fig. 1 shows that the BMI X_3, the BP X_4 and the serum measurement X_9 are particularly informative; this is confirmed by the results of LARS obtained by Efron et al. [22].

2.5 Notes on the Error Criterion and the Proposed Algorithms

In this paper, c is the mean square error, since we focus on linear regression. However, the above discussion and the two proposed algorithms remain valid for non-linear regression using e.g. a kNN like in [33]. Also, other criteria can be used, like the (estimated) conditional entropy $c(\mathcal{D}) = \hat{H}(Y|X)$. The difference $c\left(\mathcal{D}_{X_{(j)}}\right) - c(\mathcal{D})$ becomes the (estimated) conditional mutual information $\hat{I}\left(X_j; Y|X_{(j)}\right) =$

$\hat{I}\left(X_{(j)} \cup \{X_j\}; Y\right) - \hat{I}\left(X_{(j)}; Y\right)$, i.e. the additional information in X_j about Y. Entropies can be estimated with the Kozachenko-Leonenko estimator [26, 27, 39, 40]. Similar approaches exist in feature selection [41, 42], but they do not derive bounds.

The above algorithms have several drawback. First, the criterion c has to be computed for each feature subsets. Second, when the number of feature d increases, the lower bounds tend to zero because of overfitting. Third, the used algorithm for the upper bounds is a heuristic, since forward-backward search is not exhaustive. Eventually, the overall computational cost is quadratic w.r.t. the dimensionality d. However, these two algorithms can still provide excellent points of comparison in Section 4 due to their strong resemblance of the weak and strong relevance of features.

3 Linear Bounds

We are interested in the interpretation of a given linear mapping $f(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x} \in \mathbb{R}$ with $\boldsymbol{\omega} \in \mathbb{R}^d$, which we assume to map to a one-dimensional space, for simplicity. Generalisations to higher dimensions such as present in metric transformation, for example, are immediate (i.e. treat each one-dimensional mapping independently and aggregate the results). We assume that this mapping either comes from a regression or classification task such as ridge regression, LARS, LASSO, or it arises from a quadratic metric adaptation method which corresponds to a linear transformation of the data space. For a given linear mapping, the value $|\omega_j|$ is often taken as a direct indicator of the relevance of feature X_j provided the input features have the same scaling, i.e. the values delivered by a linear mapping are directly interpreted. As pointed out in [15], this is highly problematic: for high-dimensional data and hence high feature correlation, the absolute value ω_j can be very misleading. The approach [15] bases this observation on the formalisation of mapping invariances for the given data.

First, we define the central notion of invariance, which will substitute the role of a criterion c. Given a mapping $f(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x}$ and data X consisting of a matrix with data vectors $\mathbf{x}_i$ we define that $\boldsymbol{\omega}$ is *equivalent* to $\boldsymbol{\omega}'$ iff

$$\boldsymbol{\omega}^\top X = (\boldsymbol{\omega}')^\top X \tag{3}$$

i.e. the mapping of the data is not changed when substituting $\boldsymbol{\omega}$ by $\boldsymbol{\omega}'$. Unlike a pre specified criterion c such as the accuracy, this notion directly relates to the behaviour of the mapping on the given data only. The approach [15] exactly characterises under which condition $\boldsymbol{\omega}$ is equivalent to $\boldsymbol{\omega}'$: two vectors $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ are equivalent iff the difference vector $\boldsymbol{\omega} - \boldsymbol{\omega}'$ is contained in the null space of the data covariance matrix $XX^\top$. The covariance matrix has eigenvectors $\mathbf{v}_i$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_I > \lambda_{I+1} = \ldots = \lambda_d = 0$ sorted according to their size, whereby I denotes the number of non zero eigenvalues.

In [15] it is proposed to choose one canonic representation $\boldsymbol{\omega}'$ of the equivalence class induced by a given $\boldsymbol{\omega}$ before interpreting the values: one considers the vector $\boldsymbol{\omega}'$ which results by dividing the null space; $\boldsymbol{\omega}$ becomes $\boldsymbol{\omega}' = \Psi\boldsymbol{\omega}$ where

$$\Psi = \mathrm{Id} - \sum_{i=I+1}^{d} \mathbf{v}_i \mathbf{v}_i^\top$$

denotes the matrix which corresponds to the projection of $\boldsymbol{\omega}$ to the eigenvectors with non zero eigenvalues only induced by the eigenvectors $\mathbf{v}_i$ of the matrix $XX^\top$. Hence the eigenvectors with eigenvalue zero are divided out. It has been

shown in the approach [15] that this choice of a representative corresponds to the vector in the equivalence class with smallest L_2 norm.

This has the result, that it is no longer possible to assign a high value ω_j to an irrelevant feature based on random effects of the data, i.e. strongly relevant features are identified. While providing a unique representative of every equivalence class, this choice is problematic as concerns the direct interpretability of the values: Weakly relevant features share the total relevance of the features uniformly. Hence a feature which is highly correlated to a large number of others is always weighted low, independent of the fact that the information provided by this feature (or any equivalent one) might be of high relevance for the linear mapping prescription. In the following, we propose an alternative to choose representatives which are equivalent to $\boldsymbol{\omega}$ but which allow a direct interpretation of the weight vector. Essentially, we will not consider the representative with smallest L_2 norm, but use the L_1 norm instead. Unlike the former, the latter induces a set of equivalent weights which have minimal L_1 norm. We can infer the minimum and maximum relevance of a feature by looking at the minimum and maximum weighting of the feature within this set. Now we formalise this intuition.

3.1 Formalising the Objective

Given a parameter vector $\boldsymbol{\omega}$ of a linear mapping, we are interested in equivalent vectors, i.e. vectors of the form

$$\boldsymbol{\omega}' = \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \tag{4}$$

for real valued parameters α_i which add the null space of the mapping to the vector $\boldsymbol{\omega}$. We want to avoid random scaling effects of the null space, therefore we choose minimum vectors only, similar to the approach [15] . Unlike the L_2 norm, however, we use the L_1 norm:

$$\mu \leftarrow \min_{\boldsymbol{\alpha}} \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 . \tag{5}$$

The value of the minimum μ is unique per definition. This is not the case for the corresponding vector $\boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i$. A very simple case illustrates this fact: assume identical features $X_i = X_j$ and a weighting ω_i and ω_j. Then any weighting $\omega_i' = t \cdot \omega_i + (1-t)\omega_j$ and $\omega_j' = (1-t)\omega_i + t\omega_j$ yields an equivalent vector with the same L_1 norm.

This observation enables us to formalise a notion of minimum and maximum feature relevance for a given linear mapping: the *minimum feature relevance* of feature X_j is the smallest value of a weight $|\omega_j'|$ such that $\boldsymbol{\omega}'$ is equivalent to $\boldsymbol{\omega}$ and $|\boldsymbol{\omega}'|_1 = \mu$. The *maximum feature relevance* of feature X_j is the largest value of a weight $|\omega_j'|$ such that $\boldsymbol{\omega}'$ is equivalent to $\boldsymbol{\omega}$ and $|\boldsymbol{\omega}'|_1 = \mu$. In mathematical terms, this corresponds to the following optimisation problems:

$$\begin{aligned} \underline{\omega}_j \leftarrow \min_{\boldsymbol{\alpha}} & \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right| \\ \text{s.t.} \quad & \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 = \mu \end{aligned} \tag{6}$$

and

$$\begin{aligned} \overline{\omega}_j \leftarrow \max_{\boldsymbol{\alpha}} & \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right| \\ \text{s.t.} \quad & \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 = \mu. \end{aligned} \tag{7}$$

where $(\mathbf{v}_i)_j$ refers to component j of $\mathbf{v}_i$. This framework yields a pair $(\underline{\omega}_j, \overline{\omega}_j)$ for each feature X_j indicating the minimum and maximum weight of this feature for all equivalent mappings with the same L_1 norm. This strongly resembles the notion of strong and weak feature relevance in the special case of linear mappings and the mapping invariance as objective.

Note that this framework does not realise the notion of strong and weak feature relevance in a strict sense due to the following reason: we aim for scaling terms as observed in the linear mapping, which are subject to L_1 regularisation. This has the consequence that two features which have the same information content but which are scaled differently are not treated as identical by this formalisation. Rather, the feature with the better signal to noise ratio which corresponds to a smaller scaling of the corresponding weight is preferred. Qualitative feature selection would treat such variables identically.

There exist natural relaxations of this problem as follows: In Eq. (4), we can incorporate eigenvectors which correspond to small eigenvalues, thus enabling an only approximate preservation of mapping equivalence. Further, we can relax the equality in Eq. (5) to allow values which do not exceed $\mu+\epsilon$ instead of μ for some small $\epsilon > 0$. Such relaxations with small values ϵ are strongly advisable for practical applications to take into account noise in the data. We will use these straight-forward approximations in experiments.

3.2 Reformalisation as Linear Programming Problem

For an algorithmic solution, we rephrase these problems as linear optimisation problems (LP). We reformulate problem (6) as the following equivalent LP where we introduce a new variable $\tilde{\omega}_k$ for every k which takes the role of $|\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k|$:

$$\begin{aligned} \underline{\omega}_j \leftarrow \min_{\tilde{\boldsymbol{\omega}}, \boldsymbol{\alpha}} & \tilde{\omega}_j, \\ \text{s.t.} \quad & \sum_{i=1}^{d} \tilde{\omega}_i \leq \mu \\ & \tilde{\omega}_k \geq \omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k, \forall k \\ & \tilde{\omega}_k \geq -\left(\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k \right), \forall k, \end{aligned} \tag{8}$$

where μ is computed in (5) and the variables $\tilde{\omega}_i$ must be non negative due to the constraints. For the optimum solution, we can assume that equality holds for one of the two constraints for every k; otherwise, the solution could be improved due to the weaker constraints and the minimisation of the objective. For problem (7), we use the equivalent formulation

$$\begin{aligned} \max_{\tilde{\boldsymbol{\omega}}, \boldsymbol{\alpha}} & \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right|, \\ \text{s.t.} \quad & \sum_{i=1}^{d} \tilde{\omega}_i \leq \mu \\ & \tilde{\omega}_k \geq \omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k, \forall k \\ & \tilde{\omega}_k \geq -\left(\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k \right), \forall k, \end{aligned} \tag{9}$$

where, again, new variables $\tilde{\omega}_k$ are introduced. Again, these take the role of the absolute value $|\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k|$: any solution for which equality does not hold for one of the constraints can be improved due to the weaker constraints and maximisation as the objective. This is not

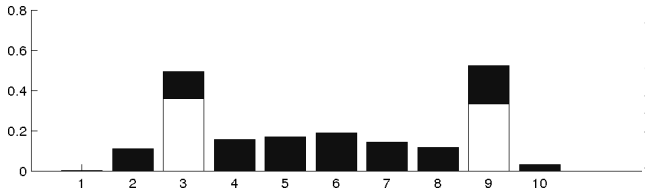

Figure 2: Lower and upper bounds of feature relevance given by the linear programming method for the diabetes dataset.

yet a LP since an absolute value is optimised. For its solution, we can simply solve two LPs where we consider the positive and negative value of the objective:

$$\overline{\omega}_j^{\pm} \leftarrow \max_{\tilde{\boldsymbol{\omega}},\boldsymbol{\alpha}} \pm \left(\omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right),$$

and we add the corresponding non negativity constraint

$$\pm \left(\omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right) \geq 0$$

At least one of these LPs has a feasible solution, and the final upper bound can be derived thereof as the maximum value

$$\overline{\omega}_j = \max\{\overline{\omega}_j^+, \overline{\omega}_j^-\}$$

This approach requires to solve LP problems containing $2d$ constraints and $I + 1$ variables. Standard solver can be applied.

4 Experiments

In this section, results accomplished by the linear bounds method and the generic approach are compared. For both methods, data are normalised beforehand to have zero expectation and unit variance. Further, we consider a relaxed LP, allowing a bound of $1.1 \cdot \mu$ instead of μ, and incorporating eigenvectors also with eigenvalues close to zero. We report the used number of eigenvectors for every data set.

Note that the methods investigated in this experiment do not reveal the strong and weak relevance, but they rely on the quantitative scaling instead. Still, upper and lower bounds allow us to distinguish three settings:

1. A feature is irrelevant: this corresponds to a small upper bound.

2. A feature is relevant for the mapping but can be substituted by others: this corresponds to a small lower bound and large upper bound.

3. A feature is relevant and cannot be substituted: this corresponds to a large lower bound.

Albeit cases 2) and 3) are not equivalent to weak and strong feature relevance in the strict sense, we will refer to these setting by these terms in the following.

As a first illustration, we display the feature relevances of the LP approach generated on the diabetes dataset as discussed in Section 2.4 in Fig. 2. Here, we utilize the smallest 3 eigenvalues. The features X_3 and X_9 are indicated as strongly relevant. Otherwise, features display similar upper bounds as predicted before, with small differences: the strongly relevant features X_2 and X_4, as detected by the baseline, are not highlighted by the LP technique. This is due to the fact that the resulting map can slightly be changed since noise due to small eigenvectors is accepted. Under these conditions, the features are no longer mandatory to explain the mapping.

Further, X_1 vanishes for the LP method, which can be attributed to the fact that the same effect to the mapping can be achieved with another feature which has a better signal to noise ratio, i.e. L_1 norm would increase when incorporating X_1.

4.1 Difference between methods

To show a major advantage of the LP method, a toy dataset was generated: unlike iterative feature selection, the LP technique simultaneously judges the relevance of all features. Hence it can better handle settings where a large number of noisy features masks weakly relevant information. In this example, the first twelve dimensions are noisy and only slightly correlated with the target, features X_{13} and X_{14} are useful but redundant, and the last two dimensions are necessary and independent. The objective for the task is to predict the sum of the last three dimensions. We choose the dimensionality 1 for the approximated null space.

Results for both methods are displayed in Fig. 3. The generic method finds the two necessary and independent dimensions. It does not single out the weak relevance of the previous two features. Better results can be obtained with the linear programming approach which disregards the first dimensions completely, shows a full lower bound for the last two features, and correctly indicates the potential relevance of the other two dimensions.

4.2 Benchmarks

We utilize several benchmark data sets from [43, 44].

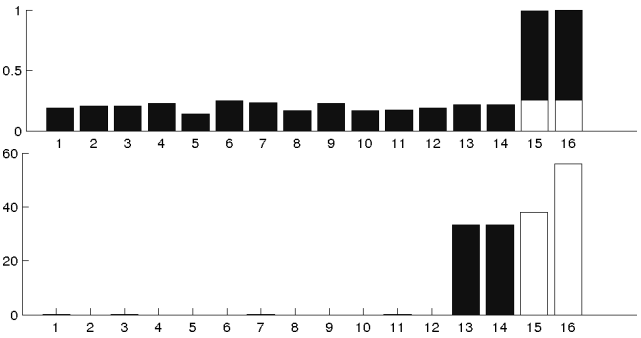

Figure 3: Lower and upper bounds of feature relevance for a toy dataset. The top figure shows the results of the generic approach, the lower one for the LP method.

Boston Housing The Boston Housing dataset [45] concerns housing values in suburbs of Boston with the median value of owner-occupied homes as target. The dimensionality of the null space is picked as 3. Like displayed in Fig. 4, features X_6 and X_{13} which correspond to the average number of rooms per dwelling and the percentage of lower status of the population are identified as most relevant. The same holds for X_4, X_{11} and X_{12} but to a lesser degree. Interestingly, the relevance of features like X_9 (index of accessibility to radial highways) can play an important role, but this information can also be gathered from other features.

Poland Electricity Consumption This dataset [46, 47] is a time series monitoring the electricity consumption in Poland based on time windows of size 30. We choose the zero space dimensionality as 3 corresponding to the extremely high correlation observed in this time series data. Fig. 5 shows that the last feature is identified by LP as the most relevant one. This is expected due to the smoothness of the time series. For the LP technique, the

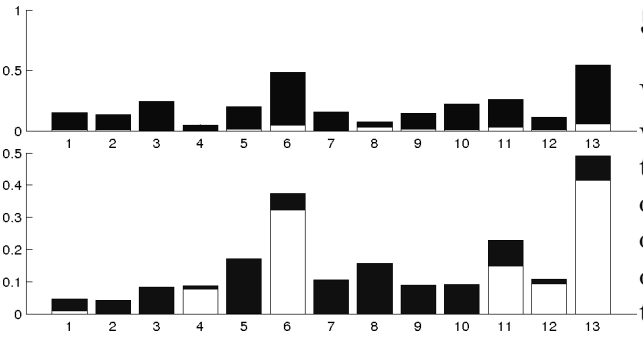

Figure 4: Lower and upper bounds of feature relevance for a Boston Housing dataset. The top figure shows the results of the generic approach, the lower one for the LP method.

feature is marked as strongly relevant since its substitution would require a too large weighting. Further, for both methods, the cyclicity of the time series is clearly observable, whereby the basic method does not identify any feature as strongly relevant but the last one. Interestingly, the LP technique identifies two consecutive features as relevant for every cycle, since two values allow the estimation of the first-order derivative for better time series prognosis [48].

Santa Fe laser This dataset [49, 50] is a time series monitoring the physical process related to a laser with time windows of size 12; the dimensionality of the null space is chosen as 2. Interestingly, a result which is very similar to the previous one can be obtained. The features X_6 and X_{12} as well as their immediate predecessors are picked by the LP technique as strongly relevant. As can be seen in Fig. 6 both methods identify the last two features as relevant, but the LP method shows a clearer profile as concerns the past values, which coincides with findings from [48].

5 Conclusion

We have addressed the question in how far weights which arise from a linear transformation such as a linear classification, regression, or metric scaling, allow a direct interpretation of the weighting terms as relevances. We have discussed that this is usually not the case in particular for high-dimensional data, a setting with particular importance e.g. for the biomedical domain. Inspired by previous work which addresses the null space of the observed data, and the notion of weak and strong feature relevance, we have developed a framework which yields to an efficient quantitative evaluation of the minimum and maximum feature relevance for a given linear mapping. This framework is based on the hypothesis that the objective is the output of the given mapping for the given data, and only weights which are minimum in L_1 norm are of interest. Then, linear programming enables a polynomial technique to estimate these relevance intervals.

We have compared the techniques to a corresponding baseline which is directly based on forward-backward feature selection. It becomes apparent that the techniques closely resembles the notion of weak and strong feature relevance; unlike iterative methods, it does not face problems when dealing with high-dimensional data and many irrelevant features, still being capable of distinguishing this information from mere noise.

So far, we have demonstrated the techniques for various benchmarks with very promising results. It will be the subject of future work to test the suitability of this technique for biomedical applications where relevance intervals will be checked by medical experts. In addition, we are in the process of testing and improving the tech-

nique for higher dimensionality in the range of several hundred or thousand features. For these settings, efficient optimisation techniques will be needed for a feasible LP solution.

Acknowledgment

Funding by DFG under grant number HA 2719/7-1 and by the CITEC centre of excellence are gratefully acknowledged.

References

[1] B. Frénay, D. Hofmann, A. Schulz, M. Biehl, and B. Hammer, "Valid interpretation of feature relevance for linear data mappings," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 149–156.

[2] Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council, *Frontiers in Massive Data Analysis*. The National Academies Press, 2013. [Online]. Available: http://www.nap.edu/openbook.php?record_id=18374

[3] C. Rudin and K. L. Wagstaff, "Machine learning for science and society," *Machine Learning*, vol. 95, no. 1, pp. 1–9, 2014.

[4] V. V. Belle and P. Lisboa, "White box radial basis function classifiers with component selection for clinical prediction models," *Artificial Intelligence in Medicine*, vol. 60, no. 1, pp. 53–64, 2014.

[5] S. Briesemeister, "Interpretable machine learning approaches in computational biology," Ph.D. dissertation, University of Tübingen, 2011.

[6] S. Briesemeister, J. Rahnenführer, and O. Kohlbacher, "Going from where to why - interpretable prediction of protein subcellular localization," *Bioinformatics*, vol. 26, no. 9, pp. 1232–1238, 2010.

[7] P. J. G. Lisboa, "Interpretability in machine learning - principles and practice," in *WILF*, ser. Lecture Notes in Computer Science, F. Masulli, G. Pasi, and R. R. Yager, Eds., vol. 8256. Springer, 2013, pp. 15–21.

[8] S. Rüping, "Learning interpretable models," Ph.D. dissertation, University of Dortmund, 2006.

[9] J. Tikka, *Input Variable Selection Methods for Construction of Interpretable Regression Models*, ser. TKK Dissertations in information and computer science. Helsinki University of Technology, 2008. [Online]. Available: http://books.google.de/books?id=dHwpQwAACAAJ

[10] A. Vellido, J. Martin-Guerrero, and P. Lisboa, "Making machine learning models interpretable," in *ESANN*, 2012.

[11] G. K. Smyth, *Limma: linear models for microarray data.* Springer, New York, 2005, pp. 397–420.

[12] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," *ArXiv e-prints*, Jun. 2013.

[13] M. Biehl, B. Hammer, P. Schneider, and T. Villmann, "Metric learning for prototype based classification," in *Innovations in Neural Information – Paradigms and Applications*, ser. Studies in Computational Intelligence 247, M. Bianchini, M. Maggini, and F. Scarselli, Eds. Springer, 2009, pp. 183–199.

[14] W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libe, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, and P. M. Stewart, "Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors," *J Clinical Endocrinology and Metabolism*, vol. 96, pp. 3775–3784, 2011.

[15] M. Strickert, B. Hammer, T. Villmann, and M. Biehl, "Regularization and improved interpretation of linear data mappings and adaptive distance measures," in *IEEE SSCI CIDM 2013*. IEEE Computational Intelligence Society, 2013, pp. 10–17.

[16] T. D. Bie, L.-C. Tranchevent, L. M. M. van Oeffelen, and Y. Moreau, "Kernel-based data fusion for gene prioritization." in *ISMB/ECCB (Supplement of Bioinformatics)*, 2007, pp. 125–132.

[17] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nat Rev Genet*, vol. 13, no. 8, pp. 523–536, 2012.

[18] R. E. Bellman, *Adaptive control processes - A guided tour*, Princeton, New Jersey, U.S.A., 1961.

[19] M. Verleysen, "Learning high-dimensional data," *Limitations and Future Trends in Neural Computation*, vol. 186, pp. 141–162, 2003.

[20] D. Francois, F. Rossi, V. Wertz, and M. Verleysen, "Resampling methods for parameter-free and robust feature selection with mutual information," *Neurocomputing*, vol. 70, no. 7-9, pp. 1276–1288, 2007.

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *JMLR*, vol. 3, pp. 1157–1182, 2003.

[22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[23] T. C. Hesterberg, N. H. Choi, L. Meier, and C. Fraley, "Least angle and l1 penalized regression: A review," *Statistics Surveys*, 2008.

[24] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, 1994.

[25] M. Verleysen, F. Rossi, and D. François, "Advances in feature selection with mutual information," in *Similarity-Based Clustering*, 2009, vol. 5400, pp. 52–69.

[26] E. Schaffernicht, R. Kaltenhaeuser, S. Verma, and H.-M. Gross, "On estimating mutual information for feature selection," in *Artificial Neural Networks – ICANN 2010*. Springer Berlin Heidelberg, 2010, vol. 6352, pp. 362–367.

[27] G. Doquire and M. Verleysen, "A comparison of multivariate mutual information estimators for feature selection," in *ICPRAM (1)*, 2012, pp. 176–185.

[28] B. Frénay, G. Doquire, and M. Verleysen, "Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification," *Neurocomputing*, vol. 112, pp. 64–78, 2013.

[29] ——, "Is mutual information adequate for feature selection in regression ?" *Neural Networks*, vol. 48, pp. 1–7, 2013.

[30] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, no. 1, pp. 175–186, 2014.

[31] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *ICML'94*, 1994, pp. 121–129.

[32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[33] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time." *Journal of Machine Learning Research*, vol. 8, pp. 589–612, 2007.

[34] D. Bell and H. Wang, "A formalism for relevance and its application in feature subset selection," *Machine Learning*, vol. 41, no. 2, pp. 175–195, 2000.

[35] I. Tsamardinos and C. F. Aliferis, "Towards Principled Feature Selection: Relevancy, Filters and Wrappers," in *in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[36] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[37] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[38] D. K. Slonim, "From pattern to pathways: gene expression data analysis comes of age," *Nature Genetics Supplement*, vol. 32, pp. 502–508, 2002.

[39] L. F. Kozachenko and N. Leonenko, "Sample estimate of the entropy of a random vector," *Problems Inform. Transmission*, vol. 23, pp. 95–101, 1987.

[40] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, 2004.

[41] P. Pudil, J. Novovicová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recogn. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.

[42] J. Novovicová, P. Somol, M. Haindl, and P. Pudil, "Conditional mutual information based feature selection for classification task," in *CIARP'07*, 2007, pp. 417–426.

[43] "Environmental and industrial machine learning group," http://research.ics.aalto.fi/eiml/datasets.shtml.

[44] D. N. A. Asuncion, "UCI machine learning repository."

[45] D. H. Jr. and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81 – 102, 1978.

[46] A. Lendasse, J. A. Lee, V. Wertz, and M. Verleysen, "Time series forecasting using CCA and kohonen maps - application to electricity consumption," in *ESANN 2000, Bruges (Belgique)*, M. Verleysen, Ed., April 2000, pp. 329–334.

[47] ——, "Forecasting electricity consumption using nonlinear projection and self-organizing maps," *Neurocomputing*, vol. 48, no. 1-4, pp. 299–311, 2002.

[48] B. Frénay, M. van Heeswijk, Y. Miche, M. Verleysen, and A. Lendasse, "Feature selection for nonlinear models with extreme learning machines," *Neurocomputing*, vol. 102, pp. 111–124, 2013.

[49] U. Hübner, N. B. Abraham, and C. O. Weiss, "Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared nh3 laser," *Phys. Rev. A*, vol. 40, pp. 6354–6365, 1989.

[50] A. Weigend and N. Gershenfeld, "Results of the time series prediction competition at the santa fe institute," in *Neural Networks, 1993., IEEE International Conference on*, 1993, pp. 1786–1793 vol.3.

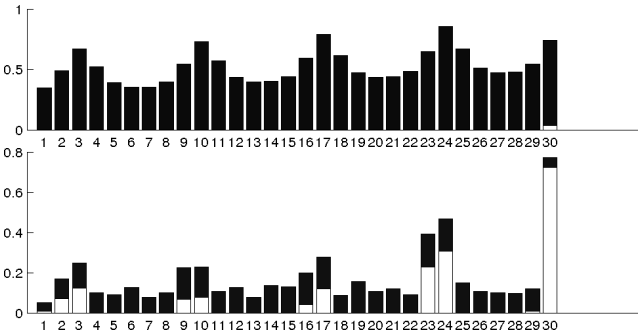

Figure 5: Lower and upper bounds of feature relevance for a Poland Electricity Consumption dataset. The top figure shows the results of the generic approach, the lower one for the LP method.

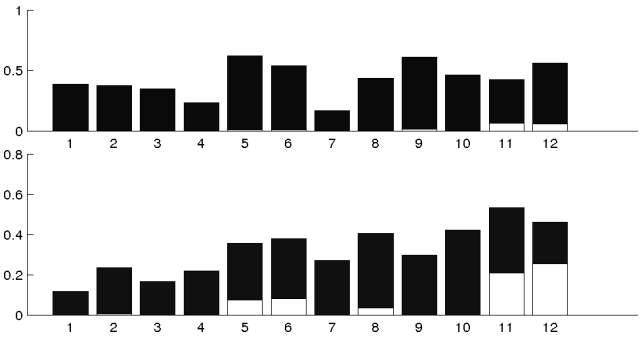

Figure 6: Lower and upper bounds of feature relevance for a Santa Fe Laser dataset. The top figure shows the results of the generic approach, the lower one for the LP method.