INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Lausser L, Szekely R, Klimmek A, Schmid F, Kestler HA. 2020 Constraining classifiers in molecular analysis: invariance and robustness. *J. R. Soc. Interface* **17**: 20190612.
http://dx.doi.org/10.1098/rsif.2019.0612

Received: 3 January 2020
Accepted: 9 January 2020

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
bioinformatics, biomathematics

Keywords:
computational learning theory, invariances, classification, molecular profiles

Author for correspondence:
Hans A. Kestler
e-mail: hans.kestler@uni-ulm.de

†Equal contribution.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4824036.

THE ROYAL SOCIETY PUBLISHING

Constraining classifiers in molecular analysis: invariance and robustness

Ludwig Lausser[1,†], Robin Szekely[1,†], Attila Klimmek[1], Florian Schmid[1] and Hans A. Kestler[1,2]

[1]Institute of Medical Systems Biology, Ulm University, Ulm, Germany
[2]Leibniz Institute on Aging, Jena, Germany

HAK, 0000-0002-4759-5254

Analysing molecular profiles requires the selection of classification models that can cope with the high dimensionality and variability of these data. Also, improper reference point choice and scaling pose additional challenges. Often model selection is somewhat guided by *ad hoc* simulations rather than by sophisticated considerations on the properties of a categorization model. Here, we derive and report four linked linear concept classes/models with distinct invariance properties for high-dimensional molecular classification. We can further show that these concept classes also form a half-order of complexity classes in terms of Vapnik–Chervonenkis dimensions, which also implies increased generalization abilities. We implemented support vector machines with these properties. Surprisingly, we were able to attain comparable or even superior generalization abilities to the standard linear one on the 27 investigated RNA-Seq and microarray datasets. Our results indicate that *a priori* chosen invariant models can replace *ad hoc* robustness analysis by interpretable and theoretically guaranteed properties in molecular categorization.

1. Introduction

Accurate and interpretable diagnostic models are a major ingredient in modern healthcare and a key component in personalized medicine [1,2]. They facilitate the identification of optimal therapies and individual treatments. These models are derived in long-lasting and cost-intensive data-driven processes, which are based on the analysis of high-dimensional marker profiles. In general, these search spaces exceed by far the possibility of manual inspection. Computer-aided systems are required for these screening procedures.

The canonical machine learning approach for deriving diagnostic classification models is the supervised learning scheme [3–5]. Here, a predictive model, a classifier, abstracts diagnostic classes from a set of labelled training examples.

Due to the data-driven nature of this learning process, the quality of a classifier is naturally dependent on the quality and amount of available samples. It can affect the generalizability and interpretability of a model. Both characteristics are of importance for the clinical setting. An incorrect prediction can lead to an incorrect treatment decision. A non-interpretable model is not verifiable and does not provide new insights in the molecular background of a disease. Small data collections might be supplemented by existing domain knowledge on the corresponding classification task or the recording process. It can provide information about hidden relationships or dependencies, which are too complex to be extracted from the data itself [6,7]. This information can structure the training process of a classification model, increasing both its accuracy and interpretability [8,9].

In the following, we focus on incorporating invariances into classification models [10]. Other approaches focus on regression applications [11,12]. That is the classification model and its predictions should not be affected by a specific data transformation. Typically, the terms *invariance* and *tolerance* are distinguished [13]. An invariant classifier completely neglects the influence of a data transformation; a tolerant one only reduces its influences. Invariances can be gained by

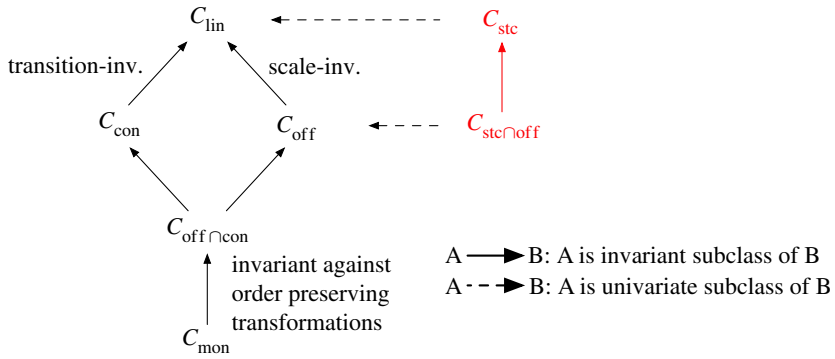

Figure 1. Invariant subclasses of linear classifiers. Linear classifiers $\mathcal{C}_{lin}$ can be organized in a hierarchy of four structural subgroups that imply different invariances. Each invariance counteracts the effects of a specific type of data transformation and preserves the predictions of the corresponding classification models. Some of these invariances can also be transferred to univariate predictors. This half-order is also reflected by a decrease in the Vapnik–Chervonenkis dimension from top to bottom, implying increased generalization ability. (Online version in colour.)

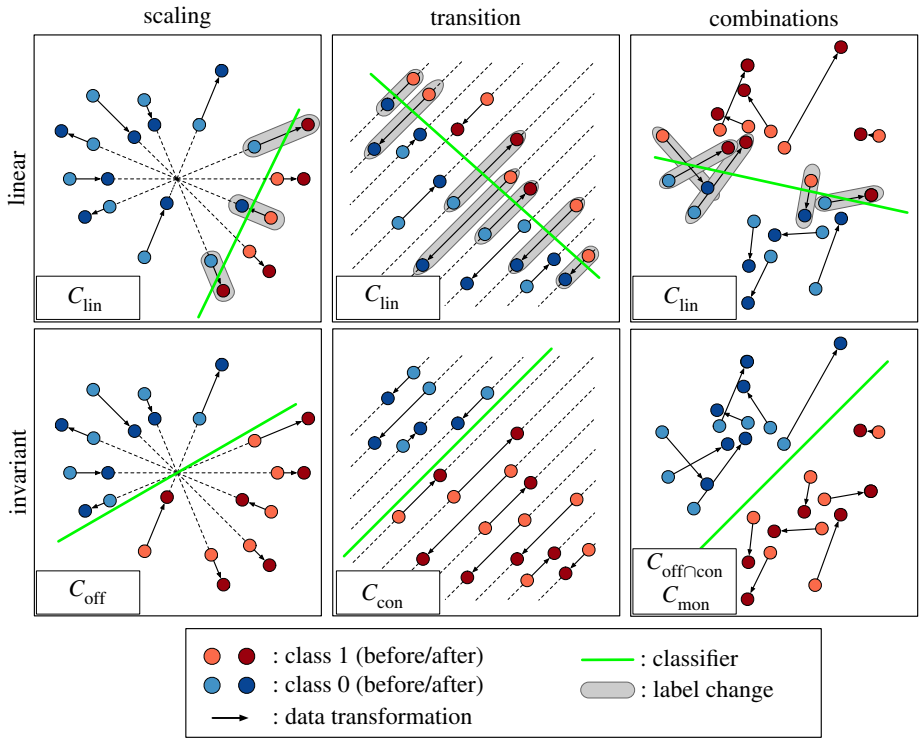

Figure 2. Structural properties of invariant linear classifiers: the first row gives examples of general linear classifiers $\mathcal{C}_{lin}$; the second row gives examples of the invariant concept classes $\mathcal{C}_{off}$, $\mathcal{C}_{con}$ and $\mathcal{C}_{off \cap con}$ ($=\mathcal{C}_{mon}$ if $\mathcal{X} = \mathbb{R}^2$). Each column provides a dataset that is affected by a specific type of data transformation. From the left to the right, the datasets are affected by *global scaling*, *global transition* and the *combination* thereof. Data points that receive a different class label due to the data transformation are marked by a grey halo. (Online version in colour.)

model restrictions [14] or by initial data transformations [15,16]. They can also be enforced during the training process of a classifier [17–20]. For example, invariances can be learned by incorporating additional artificial samples in the training process of a classification model [21,22].

Here, we impose invariance as a property of the underlying concept class of a classifier [23,24]. We generate four subclasses of linear classifiers that directly induce invariances to different data transformations (figure 1). Their structural characteristics are shown in figure 2 and listed in table 1. The theoretical properties of the concept classes and their implications on model complexity are elaborated in §2. The performance of invariant classifiers is evaluated in experiments with artificial datasets and gene expression profiles (§3). The corresponding results are shown in §4 and discussed in §5.

2. Material and methods

We use following notation throughout the article. A classifier will be seen as a function

$$c : \mathcal{X} \longrightarrow \mathcal{Y}, \tag{2.1}$$

mapping from the feature space $\mathcal{X}$ to the label space $\mathcal{Y}$. The class label of a single sample $\mathbf{x} \in \mathcal{X}$ is denoted by $y \in \mathcal{Y}$. Most of the discussion will be focused on binary classification problems (e.g. $\mathcal{Y} = \{1, 0\}$). We assume the feature space to be embedded in an n-dimensional Euclidian space $\mathcal{X} \subseteq \mathbb{R}^n$. A sample is represented as a vector $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^T$.

The optimal structure of a classifier c is typically unknown *a priori*. It has to be learned in an initial training phase consisting of two major steps. First, a concept class $\mathcal{C}$ has to be chosen.

Table 1. Overview in the discussed subclasses of linear classifiers. The concept classes are reported by their name, their structural properties, their invariances and their requirements on available measurements.

name structural properties	invariant to	required features $\|\mathbf{w}\|_0$
(standard) linear classifier: $\mathcal{C}_{\text{lin}} = \{\mathbb{1}_{[\langle \mathbf{w},\mathbf{x}\rangle \geq t]} \mid \mathbf{w} \in \mathbb{R}^n,\ t \in \mathbb{R}\}$	—	[1; n]
single threshold classifier: $\mathcal{C}_{\text{stc}} = \{\mathbb{1}_{[wx^{(i)} \geq t]} \mid w = \pm 1, i \in \{1, \ldots, n\}, t \in \mathbb{R}\}$	—	1
offset-free linear classifier: $\mathcal{C}_{\text{off}} = \{\mathbb{1}_{[\langle \mathbf{w},\mathbf{x}\rangle \geq 0]} \mid \mathbf{w} \in \mathbb{R}^n\}$	$f_a : \mathbf{x} \mapsto a \cdot \mathbf{x}$, with $a \in \mathbb{R}^+$	[1; n]
offset-free single threshold classifier: $\mathcal{C}_{\text{stc}\cap\text{off}} = \{\mathbb{1}_{[wx^{(i)} \geq 0]} \mid w = \pm 1, i \in \{1, \ldots, n\}\}$	$f_a : \mathbf{x} \mapsto a \cdot \mathbf{x}$, with $a \in \mathbb{R}^+$	1
linear contrast classifiers: $\mathcal{C}_{\text{con}} = \{\mathbb{1}_{[\langle \mathbf{w},\mathbf{x}\rangle \geq t]} \mid \langle \mathbf{w}, \mathbf{1}\rangle = 0,\ \mathbf{w} \in \mathbb{R}^n,\ t \in \mathbb{R}\}$	$f_b : \mathbf{x} \mapsto \mathbf{x} + \mathbf{b}$, with $\mathbf{b} = b \cdot \mathbf{1},\ b \in \mathbb{R}$	[2; n]
offset-free linear contrast classifier: $\mathcal{C}_{\text{off}\cap\text{con}} = \{\mathbb{1}_{[\langle \mathbf{w},\mathbf{x}\rangle \geq 0]} \mid \langle \mathbf{w}, \mathbf{1}\rangle = 0,\ \mathbf{w} \in \mathbb{R}^n\}$	$f_{a,b} : \mathbf{x} \mapsto a\mathbf{x} + \mathbf{b}$, with $\mathbf{b} = b \cdot \mathbf{1}$, $a \in \mathbb{R}^+$, $b \in \mathbb{R}$	[2; n]
pairwise comparisions: $\mathcal{C}_{\text{mon}} = \{\mathbb{1}_{[x^{(i)} - x^{(j)} \geq 0]} \mid i \neq j,\ i, j \in \{1, \ldots, n\}\}$	$f_g(\mathbf{x}) : \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{pmatrix} \mapsto \begin{pmatrix} g(x^{(1)}) \\ \vdots \\ g(x^{(n)}) \end{pmatrix}$, with $\forall x^{(i)}, x^{(j)} \in \mathbb{R} : g(x^{(i)}) < g(x^{(j)}) \iff x^{(i)} < x^{(j)}$	2

It describes the structural properties and data-independent characteristics of a classifier.

In a second step, a classifier $c \in \mathcal{C}$ has to be adapted to the classification task. A training algorithm l has to be chosen that fits the classifier according to a set of labelled training examples $\mathcal{S}_{\text{tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$,

$$l(\mathcal{S}_{\text{tr}}, \mathcal{C}) \mapsto c_{\mathcal{S}_{\text{tr}}} \in \mathcal{C}. \tag{2.2}$$

We omit the subscript $\mathcal{S}_{\text{tr}}$ if the training set is known from the context.

The most important characteristic of a trained classifier is its generalization performance in predicting the class label of new unseen samples. It is typically estimated on an independent set of test samples $\mathcal{S}_{\text{te}} = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{m'}$. A possible quality measure would be the classifiers empirical accuracy

$$A_{\text{emp}}(c, \mathcal{S}_{\text{te}}) = \frac{1}{|\mathcal{S}_{\text{te}}|} \sum_{(\mathbf{x},y)\in\mathcal{S}_{\text{te}}} \mathbb{1}_{[c(\mathbf{x})=y]}. \tag{2.3}$$

Here, $\mathbb{1}_{[p]}$ denotes the indicator function, which is equal to 1 if p is true and equal to 0 otherwise.

2.1. Invariant concept classes

Besides the overall generalization performance of a classifier, the invariances of its underlying concept class can be used for model selection. The predictions of the derived invariant classifiers will be unaffected by a family of data transformations [10]. For our analysis, we will use the following definition [14]:

Definition 2.1. A classifier $c : \mathcal{X} \to \mathcal{Y}$ is called *invariant* against a parameterized class of data transformations $f_\theta : \mathcal{X} \to \mathcal{X}$ if

$$\forall\, \theta \in \Theta, \forall\, \mathbf{x} \in \mathcal{X}: \quad c(f_\theta(\mathbf{x})) = c(\mathbf{x}). \tag{2.4}$$

A concept class $\mathcal{C}$ is called invariant against f_θ if each $c \in \mathcal{C}$ is invariant against f_θ.

Definition 2.1 calls a classifier invariant if its predictions are invariant against the influence of a parameterized class of data transformations. That is the classifier must be invariant against the influence of a data transformation for an unknown value of $\theta \in \Theta$. This implies that an invariant classifier is able to handle sample wise transformations. For a given test set $\mathcal{S}_{\text{te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m'}$, an invariant classifier can neglect the effects of m' distinct data transformations

$$\forall i \in \{1, \ldots, m'\} : c(f_{\theta_i}(\mathbf{x}_i)) = c(\mathbf{x}_i). \tag{2.5}$$

A common parameter $\bar{\theta}$ that holds for all samples in $\mathcal{S}_{\text{te}}$ does not have to be estimated. A classifier invariant against f_θ is additionally invariant against sequences of data transformations

$$\forall \theta_i,\ \theta_j \in \Theta : c(f_{\theta_i}(f_{\theta_j}(\mathbf{x}))) = c(f_{\theta_j}(\mathbf{x})) = c(\mathbf{x}). \tag{2.6}$$

A concept class $\mathcal{C}$ that is invariant against f_θ summarizes all classifiers that share this invariance property. If this invariance can be traced back to a common structural characteristic of the classifiers the concept class can directly be used for training a classification model that is guaranteed to be invariant against f_θ.

Here, we present structural subclasses of linear classifiers that directly lead to different invariances (table 1). Note that classifiers which constantly predict one particular class label (e.g. $\forall \mathbf{x} : c(\mathbf{x}) = 1$ or $\forall \mathbf{x} : c(\mathbf{x}) = 0$) are invariant against all possible data transformations $f_\theta : \mathcal{X} \to \mathcal{X}$ but otherwise do not make any sense. Constant classifiers will, therefore, be excluded from the following analysis.

2.2. Linear classifiers

Linear classifiers separate the feature space via linear hyperplanes into two classes $\mathcal{Y} = \{0, 1\}$. They are given by two parameters. The norm vector $\mathbf{w}/\|\mathbf{w}\|_2$, $\mathbf{w} \in \mathbb{R}^n$ determines the direction of the hyperplane. The threshold $t \in \mathbb{R}$ can be seen as the distance from the hyperplane to the origin.

Definition 2.2. The concept class of linear classifiers $\mathcal{C}_{\text{lin}}$ is given by

$$\mathcal{C}_{\text{lin}} = \{\mathbb{1}_{[\langle \mathbf{w},\mathbf{x}\rangle \geq t]} \mid \mathbf{w} \in \mathbb{R}^n,\ t \in \mathbb{R}\}. \tag{2.7}$$

The concept class $\mathcal{C}_{\text{lin}}$ is one of the oldest ones for classification [25]. Its theoretical properties were, for example, analysed by Minsky & Papert [26] who demonstrated that Boolean functions exist that cannot be learned by linear classifiers (XOR problem). The flexibility of linear classifiers was first analysed by Cover [27]. It was proven that the probability of finding a linear classifier that perfectly separates a randomly labelled dataset increases with the dataset's dimensionality.

Linear classification models are the underlying concept class for many popular training algorithms. For example, the perceptron [28], the linear discriminant analysis [25] and the support vector machine [29] were initially designed for linear classifiers. Although these training algorithms assume $\mathcal{C}_{\text{lin}}$ to be homogeneous, there exist different ways for separating the concept class into distinct subclasses. For example, linear classifiers can be distinguished by the number of features that are involved in their decision processes $\|\mathbf{w}\|_0 = \sum_{i=1}^{n} \mathbb{1}_{[w^{(i)} \neq 0]}$. Features that receive a weight of zero do not influence the decision process and can be omitted. The exclusion of noisy or meaningless features [30], the search for highly predictive markers [31] or the reduction of the model complexity [32] are possible reasons for a feature reduction to $\|\mathbf{w}\|_0 \leq k < n$.

Linear classifiers that rely on exactly one feature ($\|\mathbf{w}\|_0 = 1$) are summarized in the concept class of single threshold classifiers $\mathcal{C}_{\text{stc}}$ [33].

Definition 2.3. The concept class of single threshold classifiers $\mathcal{C}_{\text{stc}} \subset \mathcal{C}_{\text{lin}}$ is defined as

$$\mathcal{C}_{\text{stc}} = \{\mathbb{1}_{[wx^{(i)} \geq t]} \mid w = \pm 1, i \in \{1, \ldots, n\}, t \in \mathbb{R}\}. \tag{2.8}$$

These classifiers are typically used as base learners for classifier ensembles [33–35]. In this context, they are also called decision stumps or single rays. Single threshold classifiers are the only linear classifiers suitable for analysing single features independently.

Classifiers with $\|\mathbf{w}\|_0 = 0$ are typically omitted. For technical reasons, we will treat a linear classifier with $\|\mathbf{w}\|_0 = 0$ as a constant classifier (e.g. $\forall \mathbf{x} : c(\mathbf{x}) = 1$ or $\forall \mathbf{x} : c(\mathbf{x}) = 0$) in our analysis.

2.3. Invariant subclasses of linear classifiers

The following section provides an overview on the analysed invariant subclasses of linear classifiers. For each concept class, a theoretical proof on their invariance properties is given. An illustration of these concept classes can be found in figure 2. Their properties are summarized in table 1.

2.3.1. Offset-free linear classifiers

The first invariant subclass of $\mathcal{C}_{\text{lin}}$ is the concept class of offset-free linear classifiers $\mathcal{C}_{\text{off}}$, which is characterized by fixing the threshold to $t = 0$.

Definition 2.4 ($\mathcal{C}_{\text{off}}$). The concept class of offset-free linear classifiers $\mathcal{C}_{\text{off}} \subset \mathcal{C}_{\text{lin}}$ is defined as

$$\mathcal{C}_{\text{off}} = \{\mathbb{1}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq 0]} \mid \mathbf{w} \in \mathbb{R}^n\}. \tag{2.9}$$

Fixing the threshold $t = 0$ forces the hyperplanes of offset-free linear classifiers through the origin, which leads to invariances different from those of general linear classifiers.

Theorem 2.5. *A non-constant linear classifier $c \in \mathcal{C}_{\text{lin}}$ is invariant against global scaling*

$$f_a : \mathbf{x} \mapsto a \cdot \mathbf{x}, \tag{2.10}$$

with $a \in \mathbb{R}^+$ if and only if $c \in \mathcal{C}_{\text{off}}$.

Proof of Theorem 2.5. In order to prove the invariance of a linear classifier to a certain type of data transformation f_θ, we have to prove that

$$\forall \mathbf{x} \forall \theta : \langle \mathbf{w}, f_\theta(\mathbf{x}) \rangle \geq t \iff \langle \mathbf{w}, \mathbf{x} \rangle \geq t. \tag{2.11}$$

For global scaling, we get

$$\langle \mathbf{w}, a \cdot \mathbf{x} \rangle \geq t \iff a \cdot \langle \mathbf{w}, \mathbf{x} \rangle \geq t \tag{2.12}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle \geq \frac{t}{a}. \tag{2.13}$$

For a general linear classifier $c \in \mathcal{C}_{\text{lin}}$ with $t \neq 0$, there exists at least one $a \in \mathbb{R}^+$ for which $t/a \neq t$ (e.g. $a = |t|$). For the case of $t = 0$, a linear classifier is offset-free $c \in \mathcal{C}_{\text{off}}$. ∎

Omitting an offset ($t = 0$) makes a linear classifier invariant against the global scaling of test samples, while a standard linear classifier $c \in \mathcal{C}_{\text{lin}}$ might be misguided here.

Offset-free linear classifiers can be constructed independently of the number of involved features $\|w\|_0 \geq 1$. In particular, single threshold classifiers can fulfil the structural property of $\mathcal{C}_{\text{off}}$.

Definition 2.6 ($\mathcal{C}_{\text{stc}\cap\text{off}}$). The concept class of offset-free single threshold classifiers $\mathcal{C}_{\text{stc}\cap\text{off}} \subset \mathcal{C}_{\text{lin}}$ is defined as $\mathcal{C}_{\text{stc}} \cap \mathcal{C}_{\text{off}}$,

$$\mathcal{C}_{\text{stc}\cap\text{off}} = \{\mathbb{1}_{[wx^{(i)} \geq 0]} \mid w = \pm 1, i \in \{1, \ldots, n\}\}. \tag{2.14}$$

Although single threshold classifiers $c \in \mathcal{C}_{\text{stc}\cap\text{off}}$ allow a scale-invariant classification according to single features, their applicability is limited due to the fixed threshold of $t = 0$. An alternative might be the usage of offset-free linear classifiers with $\|w\|_0 = 2$, which are, for example, used for constructing fold-change classifiers [36].

2.3.2. Linear contrast classifiers

The second invariant subclass is the concept class of linear contrast classifiers $\mathcal{C}_{\text{con}}$ [14].

Definition 2.7 ($\mathcal{C}_{\text{con}}$). The concept class of linear contrast classifiers $\mathcal{C}_{\text{con}} \subset \mathcal{C}_{\text{lin}}$ is defined as

$$\mathcal{C}_{\text{con}} = \left\{ \mathbb{1}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq t]} \mid \sum_{i=1}^{n} w^{(i)} = 0,\ \mathbf{w} \in \mathbb{R}^n,\ t \in \mathbb{R} \right\}. \tag{2.15}$$

The norm vector of a linear contrast classifier is additionally constrained by $\sum_{i=1}^{n} w^{(i)} = 0$. In the context of variation analysis, such linear mappings $\mathbf{w}$ are called contrasts [37,38]. The structural properties of a linear contrast classifier induce the invariance of $\mathcal{C}_{\text{con}}$.

Theorem 2.8. *A non-constant linear classifier $c \in \mathcal{C}_{\text{lin}}$ is invariant against global transition*

$$f_b : \mathbf{x} \mapsto \mathbf{x} + \mathbf{b} \tag{2.16}$$

with $b \in \mathbb{R}$, $\mathbf{b} = b \cdot \mathbf{1}$ if and only if $c \in \mathcal{C}_{\text{con}}$.

Proof of Theorem 2.8. A global transition affects the decision of a linear classifier in the following way:

$$\langle \mathbf{w}, \mathbf{x} + \mathbf{b} \rangle \geq t \iff \langle \mathbf{w}, \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{b} \rangle \geq t \tag{2.17}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^{n} w^{(i)} b \geq t \tag{2.18}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle + b \sum_{i=1}^{n} w^{(i)} \geq t \tag{2.19}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle \geq t - b \sum_{i=1}^{n} w^{(i).} \tag{2.20}$$

For a linear contrast classifier $c \in \mathcal{C}_{\text{con}}$ ($\sum_{i=1}^{n} w^{(i)} = 0$), the second term on the right-hand side is equal to zero. The scalar product is equivalent to $\langle \mathbf{w}, \mathbf{x} \rangle$ and the classification of the transformed sample is equivalent to the classification of the original sample.

For a general linear classifier $c \in \mathcal{C}_{\text{lin}}$ ($\sum_{i=1}^{n} w^{(i)} \neq 0$), there exists a $b \in \mathbb{R}$ (e.g. $b = 1$) for which $d = b \sum_{i=1}^{n} w^{(i)} \neq 0$. This corresponds to a replacement of the original threshold t by $t - d \neq t$. ∎

The predictions of the linear contrast classifier $c \in \mathcal{C}_{\text{con}}$ are not affected by the individual transitions of the single samples while predictions of a general linear classifier $c \in \mathcal{C}_{\text{lin}}$ can be switched in both directions.

It is worth noting that there are no single threshold classifiers that can fulfil the additional constraint of $\mathcal{C}_{\text{con}}$. As a consequence, at least $\| w \|_0 \geq 2$ features are needed for constructing a linear classifier that is invariant against global scaling. In the two-dimensional case $\| \mathbf{w} \|_0 = 2$, the concept class is restricted to classifiers of type $c(\mathbf{x}) = \mathbb{1}_{[w^{(i)}x^{(i)} + w^{(j)}x^{(j)} \geq t]}$, $w^{(i)} = -w^{(j)}$, $i \neq j$, i, $j \in \{1, \ldots, n\}$, $t \in \mathbb{R}$.

2.3.3. Offset-free contrast classifiers

The third invariant concept class consists of those linear classifiers that fulfil the constraints of both $\mathcal{C}_{\text{off}}$ and $\mathcal{C}_{\text{con}}$. It can be seen as the intersection of both concept classes.

Definition 2.9 ($\mathcal{C}_{\text{off} \cap \text{con}}$). The concept class of offset-free contrast classifiers $\mathcal{C}_{\text{off} \cap \text{con}} \subset \mathcal{C}_{\text{lin}}$ is defined as $\mathcal{C}_{\text{con}} \cap \mathcal{C}_{\text{off}}$,

$$\mathcal{C}_{\text{off} \cap \text{con}} = \left\{ \mathbb{1}_{[\langle \mathbf{w}, \mathbf{x} \rangle \geq 0]} \mid \sum_{i=1}^{n} w^{(i)} = 0,\ \mathbf{w} \in \mathbb{R}^n \right\}. \tag{2.21}$$

As a classifier $c \in \mathcal{C}_{\text{off} \cap \text{con}}$ fulfils the structural properties of $\mathcal{C}_{\text{con}}$ and $\mathcal{C}_{\text{off}}$, it is invariant to both global scaling and global transition. In addition, it is invariant against combined effects.

Theorem 2.10. *A non-constant linear classifier $c \in \mathcal{C}_{\text{lin}}$ is invariant against linear transformation that combine a global scaling and a global transition*

$$f_{a,b} : \mathbf{x} \mapsto a\mathbf{x} + \mathbf{b} \tag{2.22}$$

with $a \in \mathbb{R}^+$, $b \in \mathbb{R}$, $\mathbf{b} = b \cdot \mathbf{1}$ if and only if $c \in \mathcal{C}_{\text{off} \cap \text{con}}$.

Proof of Theorem 2.10. In case of linear transformations as described in equation (2.22), the decision of a linear classifier is influenced in the following way:

$$\langle \mathbf{w}, a\mathbf{x} + \mathbf{b} \rangle \geq t \iff \langle \mathbf{w}, a\mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{b} \rangle \geq t \tag{2.23}$$

$$\iff a \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^{n} w^{(i)} b \geq t \tag{2.24}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle + \frac{b}{a} \sum_{i=1}^{n} w^{(i)} \geq \frac{t}{a} \tag{2.25}$$

$$\iff \langle \mathbf{w}, \mathbf{x} \rangle \geq \frac{t}{a} - \frac{b}{a} \sum_{i=1}^{n} w^{(i)}. \tag{2.26}$$

For $a = 1$, the proof is now equivalent to the proof of theorem 2.8 for the invariance of $\mathcal{C}_{\text{con}}$. For all other $a \in \mathbb{R}^+ \setminus \{1\}$, the classifier is invariant if

$$t = \underbrace{-\frac{b}{a-1}}_{:=d} \sum_{i=1}^{n} w^{(i)}, \tag{2.27}$$

where $d \in \mathbb{R}$ can be either positive or negative for different data transformations. The only unique threshold can be generated by forcing $\sum_{i=1}^{n} w^{(i)} = 0$, which results in $t = 0$. The general linear classifier is, therefore, only invariant against $f_{a,b}$, if $c \in \mathcal{C}_{\text{off} \cap \text{con}}$. ∎

As $\mathcal{C}_{\text{off} \cap \text{con}} \subset \mathcal{C}_{\text{con}}$, the concept class again requires a minimal number of $\| \mathbf{w} \|_0 \geq 2$ features for constructing a non-constant classifier. For the two-dimensional case $\| \mathbf{w} \|_0 = 2$, the concept class is restricted to classifiers of type $c(\mathbf{x}) = \mathbb{1}_{[w^{(i)}x^{(i)} + w^{(j)}x^{(j)} \geq 0]}$, $w^{(i)} = -w^{(j)}$, $i \neq j$, i, $j \in \{1, \ldots, n\}$.

2.3.4. The concept class of pairwise comparisons

We change the line of argumentation for introducing the fourth invariant concept class, which we call $\mathcal{C}_{\text{mon}}$. We first specify $\mathcal{C}_{\text{mon}}$ by its invariances and show afterwards that this subclass of linear classifiers can be defined by its structural properties.

Definition 2.11 ($\mathcal{C}_{\text{mon}}$). The concept class $\mathcal{C}_{\text{mon}} \subset \mathcal{C}_{\text{lin}}$ is defined as the subset of non-constant linear classifiers that is invariant against feature-wise strictly monotone increasing functions f_g, where

$$f_g(\mathbf{x}) : \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{pmatrix} \mapsto \begin{pmatrix} g(x^{(1)}) \\ \vdots \\ g(x^{(n)}) \end{pmatrix}, \tag{2.28}$$

and $g : \mathbb{R} \to \mathbb{R}$ fulfills

$$\forall x^{(i)}, x^{(j)} \in \mathbb{R} : g(x^{(i)}) < g(x^{(j)}) \iff x^{(i)} < x^{(j)}. \tag{2.29}$$

The concept class $\mathcal{C}_{\text{mon}}$ consists of linear classifiers that are invariant against all feature-wise strictly monotone increasing effects. This set of data transformations especially includes feature-wise nonlinear effects as, for example, strictly monotone polynomial or exponential transformations. The concept class $\mathcal{C}_{\text{mon}}$ is, therefore, at least as restrictive as $\mathcal{C}_{\text{off} \cap \text{con}}$ and shares its invariance property with rank-based classifiers [15]. Theorem 2.12 states that $\mathcal{C}_{\text{mon}}$ is a real subset of $\mathcal{C}_{\text{off} \cap \text{con}}$.

Theorem 2.12. *The concept class $\mathcal{C}_{\text{mon}}$ is given by*

$$\mathcal{C}_{\text{mon}} = \left\{ \mathbb{1}_{[w^{(i)}x^{(i)} + w^{(j)}x^{(j)} \geq 0]} \mid w^{(i)} = -w^{(j)},\ i \neq j, i, j \in \{1, \ldots, n\} \right\}. \tag{2.30}$$

Proof of Theorem 2.12. The proof of Theorem 2.12 is split into three parts. First, we show that no non-constant linear classifier $c \in \mathcal{C}_{\text{mon}}$ with $\| \mathbf{w} \|_0 = 1$ exists. In a second step, we prove that the structural properties of a classifier $c \in \mathcal{C}_{\text{mon}}$ with $\| \mathbf{w} \|_0 = 2$ match exactly the description given in equation (2.30). Finally, we prove that there is no non-constant classifier $c \in \mathcal{C}_{\text{mon}}$ with $\| \mathbf{w} \|_0 \geq 3$.

Case $\| \mathbf{w} \|_0 = 1$: a linear classifier $c \in \mathcal{C}_{\text{mon}}$ has to be invariant to all feature-wise strictly monotone increasing functions f_g. In particular, it has to be invariant to global scaling and global transition $\mathcal{C}_{\text{mon}} \subseteq \mathcal{C}_{\text{off} \cap \text{con}}$. As there is no non-constant linear classifier $c \in \mathcal{C}_{\text{off} \cap \text{con}}$ with $\| \mathbf{w} \|_0 = 1$, there cannot be a non-constant linear classifier $c \in \mathcal{C}_{\text{mon}}$ with $\| \mathbf{w} \|_0 = 1$.

Case $\| \mathbf{w} \|_0 = 2$: the structural properties of $\mathcal{C}_{\text{off} \cap \text{con}} \supseteq \mathcal{C}_{\text{mon}}$ for $\| \mathbf{w} \|_0 = 2$ lead to the description of $\mathcal{C}_{\text{mon}}$ given in equation (2.30). The decision criterion can be rewritten as $c(\mathbf{x}) = \mathbb{1}_{[x^{(i)} \geq x^{(j)}]}$. As g is strictly monotone increasing

$$c(f_g(\mathbf{x})) = \begin{cases} 1 & \text{if } g(x^{(i)}) > g(x^{(j)}) \iff x^{(i)} > x^{(j)} \\ 1 & \text{if } g(x^{(i)}) = g(x^{(j)}) \iff x^{(i)} = x^{(j)} \\ 0 & \text{if } g(x^{(i)}) < g(x^{(j)}) \iff x^{(i)} < x^{(j)} \end{cases}, \tag{2.31}$$

which corresponds to $c(f_g(\mathbf{x})) = c(\mathbf{x})$.

Case $\| \mathbf{w} \|_0 \geq 3$: for simplicity, we will omit feature dimensions that do not have any influence on the decision rule ($w^{(i)} =$

0). We will prove that for each linear classifier $c \in \mathcal{C}_{\text{off}} \supset \mathcal{C}_{\text{mon}}$ with $\|\mathbf{w}\|_0 = n \geq 3$ a sample $\mathbf{x} \in \mathbb{R}^n$ and a strictly monotone function g exist for which $c(\mathbf{x}) \neq c(f_g(\mathbf{x}))$. Without loss of generality, we will show that

$$\exists \mathbf{x} \exists g : \sum_{i=1}^{n} w^{(i)} x^{(i)} \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} w^{(i)} g(x^{(i)}) < 0. \tag{2.32}$$

As $\|\mathbf{w}\|_0 = n \geq 3$, there are at least two weights which share the same sign. By permuting the ordering of the features, we can ensure that $\text{sign}(w^{(1)}) = \text{sign}(w^{(n)})$. We construct a sample $\mathbf{x} \in \mathbb{R}^n$ with

$$x^{(1)} < x^{(2)} = \cdots = x^{(n-1)} = 0 < x^{(n)}. \tag{2.33}$$

We furthermore construct a strictly monotone function g with $g(0) = 0$. This implies $g(x^{(1)}) < 0$ and $g(x^{(n)}) > 0$. The decision criterion in equation (2.32) can now be reduced to

$$\underbrace{-\frac{w^{(1)}}{w^{(n)}} x^{(1)}}_{>0} \leq x^{(n)} \quad \text{and} \quad \underbrace{-\frac{w^{(1)}}{w^{(n)}} g(x^{(1)})}_{>0} > g(x^{(n)}). \tag{2.34}$$

As $x^{(n)}$ and $g(x^{(n)})$ can be randomly chosen from $\mathbb{R}^+$, we can find a pair of numbers that fulfil these equations. Similar proofs can be given for samples of class 0. ∎

In contrast to the other invariant concept classes $\mathcal{C}_{\text{mon}}$ is directly coupled to a fixed number of features $\|\mathbf{w}\|_0 = 2$. It is restricted to the unweighted pairwise comparison of two measurements $x^{(i)}$ and $x^{(j)}$. As a consequence, the training for a classifier $c \in \mathcal{C}_{\text{mon}}$ is directly coupled to a feature selection process for higher dimensional settings ($n > 2$). For a two-dimensional subspace, exactly two classification models exist ($w^{(i)} = -w^{(j)}$, $w^{(i)} \lneqq 0$). They both share the same decision boundary.

2.3.5. Vapnik–Chervonenkis dimension

Motivated by the need for invariance, we can further show that the identified subclasses also form a half-order of complexity classes which in turn can lead to an increased generalization ability. In general, the complexity of the invariant concept classes decreases with imposing additional invariances (figure 1). This, in turn, leads to a decrease in their susceptibility to overfitting [39].

The invariant concept classes can be seen as real subclasses of $\mathcal{C}_{\text{lin}}$. Here, we provide their Vapnik–Chervonenkis dimension (*VCdim*) as a combinatorial complexity measure [29] and show that they are lower than the VCdim of $\mathcal{C}_{\text{lin}}$. The VCdim is closely related to the probably approximately correct (PAC) learning framework [40], where it can be used to provide upper bounds on the generalization performance of a classifier. In the case of two classifiers with equal empirical performance, the classifier with the lower VCdim should be preferred [41].

A $\text{VCdim}(\mathcal{C}) = m$ gives the maximal number of arbitrarily chosen but fixed data points m that can be given all 2^m possible labellings when classified by members $c \in \mathcal{C}$.

Our proofs are mainly based on the following theorem [29], where $X = \mathbb{R}^n$:

Theorem 2.13. *Let X be a finite-dimensional real vector space and let U be a finite-dimensional vector space of functions from X to $\mathbb{R}$.*

Let further

$$V = \{v : X \to \{-1, 1\} : v(\mathbf{x}) = \text{sign}\,(u(\mathbf{x})),\ u \in U,\ \mathbf{x} \in X\}.$$

Then $\text{VCdim}(V) = \dim(U)$.

Proof. We follow the original proof here [29]: we first prove $\dim(U) \leq \text{VCdim}(V)$ by showing that for all $d \leq \dim(U)$, there are points $\mathbf{x}_1, \ldots, \mathbf{x}_d$ such that for arbitrary labellings $y_i \in \{-1, 1\}$, $i = 1, \ldots, d$ of these points, there is a function $u \in U$ with $u(\mathbf{x}_i) = y_i$.

Pick d linearly independent functions $u_1, \ldots, u_d \in U$. Then, as these functions are linearly independent, there are points $x_1, \ldots, x_d \in X$ such that the vectors

$$\begin{pmatrix} u_1(\mathbf{x}_1) \\ \vdots \\ u_d(\mathbf{x}_1) \end{pmatrix}, \ldots, \begin{pmatrix} u_1(\mathbf{x}_d) \\ \vdots \\ u_d(\mathbf{x}_d) \end{pmatrix} \in \mathbb{R}^d$$

are linearly independent in $\mathbb{R}^d$. Therefore, their span is the whole $\mathbb{R}^d$ and there are coefficients $a_i \in \mathbb{R}$ with

$$y_i = \sum_{j=1}^{d} a_j u_j(\mathbf{x}_i), \quad i = 1, \ldots, d.$$

Setting $u(\mathbf{x}) = \sum_{j=1}^{d} a_j u_j(\mathbf{x}) \in U$ proves the claim.

We now prove $\text{VCdim}(V) \leq \dim(U)$. Set $k = \dim(U) + 1$ and assume the contrary, namely $\text{VCdim}(V) \geq k$.

Thus, for any set of labels $y_i \in \{-1, 1\}$, there is a function $v \in V$, $v(\mathbf{x}) = \text{sign}(u(\mathbf{x}))$, $u \in U$ and points $\mathbf{x}_i \in X$ such that

$$\text{sign}(u(\mathbf{x}_i)) = y_i, \quad i = 1, \ldots, k. \tag{2.35}$$

For these points $\mathbf{x}_1, \ldots, \mathbf{x}_k$, define the vector space

$$\tilde{U} = \left\langle \left\{ \begin{pmatrix} u(\mathbf{x}_1) \\ \vdots \\ u(\mathbf{x}_k) \end{pmatrix} \in \mathbb{R}^k : u \in U \right\} \right\rangle \subset \mathbb{R}^k, \tag{2.36}$$

where $\langle \cdot \rangle$ denotes the linear span. By assumption, $\dim(\tilde{U}) \leq \dim(U) < k$. Hence, there is a non-zero vector $\mathbf{a} \in \tilde{U}^{\perp}$ in the orthogonal complement of $\tilde{U}$, i.e.

$$0 = \sum_{i=1}^{k} a^{(i)} u(\mathbf{x}_i), \quad \text{for all } u \in U. \tag{2.37}$$

Then, by equation (2.35), there is a function u with $\text{sign}\, u(\mathbf{x}_i) = \text{sign}\,(a^{(i)})$, $i = 1, \ldots, k$. Thus,

$$0 = \sum_{i=1}^{k} a^{(i)} \text{sign}\,(a^{(i)}). \tag{2.38}$$

As $\mathbf{a} \neq \mathbf{0}$, we have a contradiction. ∎

Using theorem 2.13, we are now able to provide the VCdim of the invariant concept classes of linear classifiers:

Theorem 2.14. *Let n be the dimensionality of the input space $\mathcal{X} \subseteq \mathbb{R}^n$. The VC dimensions of the major concept classes given above* (table 1) *are*

(a) $\text{VCdim}\,(\mathcal{C}_{\text{lin}}) = n + 1$.
(b) $\text{VCdim}\,(\mathcal{C}_{\text{off}}) = n$.
(c) $\text{VCdim}\,(\mathcal{C}_{\text{con}}) = n$.
(d) $\text{VCdim}\,(\mathcal{C}_{\text{off} \cap \text{con}}) = n - 1$.
(e) $\text{VCdim}\,(\mathcal{C}_{\text{mon}}) \leq \max\{m | 2^m \leq n(n-1)\}$.

Proof of Theorem 2.14. In the proof, we make use of theorem 2.13, using a different vector space of functions U in every case.

(a) This result for general linear classifiers is well known in the literature [39].
(b) For the concept class $\mathcal{C}_{\text{off}}$, we chose the space of linear mappings $u : \mathbb{R}^n \to \mathbb{R}$ for U. It is well known that this space has dimension n [42]. Then theorem 2.13 implies the assertion.
(c) Consider the vector space $X = \langle (1, \ldots, 1) \rangle^{\perp}$ which is the orthogonal complement of the space spanned by

$(1, \ldots, 1) \in \mathbb{R}^n$. Note that $dim(X) = n - 1$ and there holds

$$\sum_{i=1}^{n} w^{(i)} = 0, \quad \mathbf{w} \in X. \tag{2.39}$$

In theorem 2.13, we take for U the space of affine mappings from X to $\mathbb{R}$ [42], which has dimension $(n-1)+1 = n$.

(d) We argue exactly as in step (c), except that we take for U the space of linear mappings from X to $\mathbb{R}$ [42], which has dimension $n-1$.

(e) For a fixed set of m samples $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^m$ and fixed pair of feature dimensions $i \neq j$, with $\forall k : x_k^{(i)} \neq x_k^{(j)}$ the classifiers in $\mathcal{C}_{\text{mon}}$ can result in at most two labellings

$$(a) \quad y_k = \mathbb{1}_{[x_k^{(i)} \geq x_k^{(j)}]}$$
$$(b) \quad \bar{y}_k = \mathbb{1}_{[x_k^{(j)} \geq x_k^{(i)}]} = \mathbb{1}_{[x_k^{(i)} < x_k^{(j)}]} = \neg \mathbb{1}_{[x_k^{(i)} \geq x_k^{(j)}]},$$

which can be seen as a random labelling and its negation. In this way, $\mathcal{C}_{\text{mon}}$ can generate at most $n(n-1)$ distinct labellings in $\mathbb{R}^n$. The set $\mathcal{X}$ can, therefore, receive all 2^m distinct labellings if $2^m \leq n(n-1)$. The maximal set size $\max\{m \mid 2^m \leq n(n-1)\}$ is therefore an upper limit to $\text{VCdim}(\mathcal{C}_{\text{mon}})$. ∎

2.4. Support vector machines

In the following, we consider (linear) support vector machines (SVMs) [29] as training algorithms for the invariant concept classes. SVMs are standard training algorithms for linear classifiers. In its original form, it is designed for maximizing the margin between the training samples and the hyperplane of a linear classifier. Several modifications of the original training algorithm exist [43]. For our experiments, we have chosen two $L1$ soft-margin SVMs.

2.4.1. R2-support vector machines

The original SVM algorithm maximizes the margin by a regularization of the Euclidean norm $\|\mathbf{w}\|_2$. It will be denoted as *R2-SVM* in the following. The training algorithm can be summarized by the following constrained optimization criterion:

$$\min_{\mathbf{w},t,\xi} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i \tag{2.40}$$

$$\text{s.t.} \quad \forall i : y_i(\mathbf{w}^T\mathbf{x}_i - t) \geq 1 - \xi_i \tag{2.41}$$

$$\forall i : \xi_i \geq 0. \tag{2.42}$$

In this context, we assume class labels $\mathcal{Y} = \{+1, -1\}$. The parameter ξ_i denotes the slack variables that enable the use of SVMs in the non-separable case by measuring deviation from the ideal condition. C is the cost parameter which induces a trade-off between margin maximization and minimization of the classification error.

2.4.2. R1-support vector machines

A feature selecting version of the SVM replaces the regularization of the Euclidian norm by the regularization of the Manhattan norm $\|\mathbf{w}\|_1$. We will use the term *R1-SVM* throughout the manuscript. The corresponding objective replaces equation (2.40) by

$$\min_{\mathbf{w},t,\xi} \ \|\mathbf{w}\|_1 + C\sum_{i=1}^{n} \xi_i. \tag{2.43}$$

The Manhattan norm is more sensitive to small weights near zero. The corresponding features will be removed from the linear decision boundary ($w^{(i)} = 0$).

2.4.3. Training invariant support vector machines

The SVM training algorithm for linear classifiers can be restricted to invariant subclasses by additional constraints. These constraints reflect the structural properties of the subclasses.

$$\text{s.t.} \quad t = 0 \qquad \text{if } c \in \mathcal{C}_{\text{off}} \tag{2.44}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w^{(i)} = 0 \qquad \text{if } c \in \mathcal{C}_{\text{con}} \tag{2.45}$$

$$\text{s.t.} \quad \|\mathbf{w}\|_0 = 2 \qquad \text{if } c \in \mathcal{C}_{\text{mon}} \tag{2.46}$$

The trained SVMs will be denoted as SVM_{off}, SVM_{con}, $\text{SVM}_{\text{off}\cap\text{con}}$ and SVM_{mon}. Note that a constraint has to be added for an invariant subclass and subclasses thereof. For example, if the SVM training algorithm should be applied to a classifier $c \in \mathcal{C}_{\text{off}\cap\text{con}}$ both constraints for $\mathcal{C}_{\text{off}}$ and $\mathcal{C}_{\text{con}}$ have to be added.

3. Experiments

We have conducted experiments on artificial and real datasets in order to characterize how the choice of an invariant concept class influences the training of a linear SVM. All experiments were performed with help of the TunePareto software [44].

3.1. Experiments on artificial datasets

The performance of the invariant concept classes was examined in a sequence of controlled experiments on artificial datasets. A summary on all parameters is given in table 2. For these experiments, two normal distributions $\mathcal{N}(\mathbf{c}_y, \mathbf{I})$, $y \in \mathcal{Y}$ were chosen as class wise distributions. Here, the class wise centroids are given by $\mathbf{c}_y \in \mathbb{R}^n$. The covariance of the classes is given by the identity matrix $\mathbf{I} \in \mathbb{R}^{n\times n}$. The centroid of the positive class $\mathbf{c}_1 = (c_1^{(1)}, \ldots, c_1^{(n)})^T$ is randomly selected according to a feature wise uniform distribution with $c_1^{(i)} \sim \mathcal{U}(0, 10)$, $i = 1, \ldots, n$. With that, it is ensured that the components of the centroid of the positive class are always positive. The centroid of the negative class is chosen in dependency of c_1. Is is given by $\mathbf{c}_0 = \mathbf{c}_1 + d\mathbf{w}/\|\mathbf{w}\|_2$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. In this way, the Euclidean distance between both centroids is ensured to be $\|\mathbf{c}_1 - \mathbf{c}_0\|_2 = d$.

A single experiment is parameterized by the dimensionality of the feature vectors $n \in \{2, 10, 100\}$ and the distance between the class centroids d. A set of 2×50 (two classes with 50 samples each) training samples was used for adapting the SVM classifiers and a set of 2×50 test samples was used for evaluating their accuracy. For each dimensionality n and distance d, the experiment was repeated for 10 different pairs of class centroids $r \in \{1, \ldots, 10\}$.

3.1.1. Experiments without noise

In this experiment, the training and test sets were analysed in their original form. The distance between the class centroids was varied $d \in \{1, 1.1, \ldots, 5\}$. The performance of an invariant SVM is compared to its standard version. That means, an invariant R2-SVM is compared to the standard version of the R2-SVM and an invariant version of the R1-SVM is compared to the standard version of the R1-SVM.

3.1.2. Experiments with noise

The artificial datasets were also used for experiments with different types of noise (table 2). For this purpose, the samples of a dataset were partially replaced by noisy copies. The influence of a noise type was regulated by a common noise

Table 2. Summary of the analysed experiments on artificial datasets.

experiments without noise			
tested classifiers:			
concept classes:	$\mathcal{C} \in \{\mathcal{C}_{\text{lin}}, \mathcal{C}_{\text{off}}, \mathcal{C}_{\text{con}}, \mathcal{C}_{\text{off}\cap\text{con}}, \mathcal{C}_{\text{mon}}\}$		
training algorithms:	R2-SVM, R1-SVM		
dataset parameters (varied):		**dataset parameters (constant):**	
dimensionality:	$n \in \{2, 10, 100\}$	samples:	$m = 2 \times 50$
distance of centroids:	$d \in \{1, 1.1, \ldots, 5\}$		
repetitions:	$r \in \{1, \ldots, 10\}$	**summary:**	
		number of experiments:	1 23 000
experiments with noise			
tested classifiers:			
concept classes:	$\mathcal{C} \in \{\mathcal{C}_{\text{lin}}, \mathcal{C}_{\text{off}}, \mathcal{C}_{\text{con}}, \mathcal{C}_{\text{off}\cap\text{con}}, \mathcal{C}_{\text{mon}}\}$		
training algorithms:	R2-SVM, R1-SVM		
dataset parameters (varied):		**random parameters (per sample):**	
experiment:	$ex \in \{cl., sam.\}$		
noise types:	$id \in \{1, \ldots, 5\}$	$a \sim \mathcal{U}(10^{-5}, p)$	
noise parameter:	$p \in \{0, \ldots, 5\}$	$a \sim \mathcal{U}(10^{-5}, p)$	
dimensionality:	$n \in \{2, 10, 100\}$	$b \sim \mathcal{U}(-p, p)$	
repetitions:	$r \in \{1, \ldots, 10\}$	$c \sim \mathcal{U}(10^{-5}, p)$	
dataset parameters (constant):		**summary:**	
samples:	$m = 2 \times 50$	number of experiments:	18 000
distance of centroids:	$d = 4$		
noise types (*id*)			
1. none:	$f : \mathbf{x} \mapsto \mathbf{x}$		
2. scaling:	$f_a : \mathbf{x} \mapsto a \cdot \mathbf{x}$		
3. transition:	$f_b : \mathbf{x} \mapsto \mathbf{x} + \mathbf{b}$, with $\mathbf{b} = b \cdot \mathbf{1}$, $b \in \mathbb{R}$		
4. scaling and transition:	$f_{a,b} : \mathbf{x} \mapsto a \cdot \mathbf{x} + \mathbf{b}$, with $\mathbf{b} = b \cdot \mathbf{1}$, $b \in \mathbb{R}$		
5. exponential:	$f_c : \mathbf{x} \mapsto e^{0.2c \cdot \mathbf{x}}$		

parameter p. Experiments for six different noise levels were conducted ranging from $p = 0$ (no noise) to $p = 5$ (maximal noise). The distance between the class centroids was fixed to $d = 4$. Experiments were conducted for two different settings:

Sample wise noise: In this experiment, the individual samples of a test set $\mathcal{S}_{te}$ were affected by individual noise effects $\theta_i \in \Theta$ resulting in

$$\mathcal{S}'_{te} = \{(f_{\theta_i}(\mathbf{x}'_i), y'_i)\}_{i=1}^{m'}. \tag{3.1}$$

Class wise noise: Here, the samples of a pair of training and test sets $\mathcal{S}_{tr}$, $\mathcal{S}_{te}$ were affected by class wise noise effects. These effects were chosen individually for training and test samples $\theta_y, \psi_y \in \Theta$ resulting in

$$\mathcal{S}'_{tr} = \{(f_{\theta_{y_i}}(\mathbf{x}_i), y_i)\}_{i=1}^{m}, \quad \text{and} \quad \mathcal{S}'_{te} = \{(f_{\psi_{y'_i}}(\mathbf{x}'_i), y'_i)\}_{i=1}^{m'}. \tag{3.2}$$

3.2. Experiments on transcriptome datasets

We have conducted experiments on 27 gene expression datasets, consisting of 22 microarray and five RNA-Seq datasets. A summary of the datasets is given in table 3. We used standard and established preprocessing methodologies for the transcriptome data [67]: RMA is used for gene expression measurements based on microarrays (luminescence measurements) and includes an internal log-transformation [68], for the count data from RNA-Seq experiments, we used RSEM which does not include an internal log-transformation [69,70].

As reference k-nearest neighbours classifiers [71] (kNN) with $k \in \{1, 3, 5\}$, random forests [72] (RF) with $nt \in \{100, 200, 300\}$ trees and stacked auto-encoders [73] (SAE) with three layers of u, $\lceil u/4 \rceil$, $\lceil u/16 \rceil$ units and $u \in \{100, 500, 1000\}$ were chosen.

All classifiers were evaluated in 10×10 cross-validations [3]. For this experiment, a dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ is split into 10 folds of approximately equal size. Nine of them are combined to a training set $\mathcal{S}_{tr}$ while the remaining one is used as a test set $\mathcal{S}_{te}$ for evaluation. The procedure is repeated for 10 permutations of $\mathcal{S}$.

4. Results

4.1. Results on artificial datasets

The results for the noise-free experiments on artificial datasets are shown in figure 3. The accuracy differences between SVM_{lin} and the invariant SVMs are given. A positive value denotes a higher accuracy of the SVM_{lin}. In general, R2-

Table 3. Summary of the used transcriptome microarray and RNA-Seq datasets. The classes, class wise sample sizes and number of features are shown.

id	tissue	class labels (y_0, y_1)	samples (m_0, m_1)	features (n)
d_1:	bone marrow [45]	acute myeloid leukaemia (AML), mutated AML	21, 57	22 215
d_2:	breast [46]	non-inflammatory, inflammatory	69, 26	22 215
d_3:	bladder [47]	Ta, T1∪T2+	20, 20	7129
d_4:	tongue [48]	normal mucosa, oral tongue squamous cell carcinoma	26, 31	12 558
d_5:	soft tissue [49]	dedifferentiated liposarcoma, well-differentiated liposarcoma	40, 52	22 215
d_6:	lymph node [50]	intermediate, monoclonal B-cell lymphocytosis	48, 44	22 215
d_7:	brain [51]	healthy, schizophrenia	15, 13	12 558
d_8:	kidney [52]	non-tumour kidney tissue, renal cell carcinoma (RCC)	23, 69	22 215
d_9:	brain [53]	inbred alcohol-preferring, inbred alcohol-non-preferring	29, 30	8740
d_{10}:	head and neck [54]	normal mucosa, head and neck squamous cell carcinoma	22, 22	12 558
d_{11}:	lung [55]	normal tissue, adenocarcinoma	49, 58	22 215
d_{12}:	lung [56]	adenocarcinoma, squamous cell carcinoma	14, 18	12 558
d_{13}:	blood [57]	healthy, severe asthma	18, 17	32 321
d_{14}:	blood [58]	diffuse large B-cell lymphoma, follicular lymphoma	19, 58	7129
d_{15}:	prostate [59]	non-tumour prostate tissue, prostate tumour	50, 52	12 558
d_{16}:	intestinal mucosa [60]	non-cystic fibrosis, cystic fibrosis	13, 16	22 215
d_{17}:	fibroblasts [61]	healthy, macular degeneration	18, 18	12 558
d_{18}:	prostate [62]	non-recurrent cancer, recurrent cancer	40,39	22 215
d_{19}:	colon [63]	microsatellite instable tumour, microsatellite stable tumour	13, 38	7071
d_{20}:	stomach [64]	non-cardia tumour tissue, cardia tumour tissue	72, 62	22 215
d_{21}:	stomach [64]	normal gastric glands, tumour tissue	134, 134	22 215
d_{22}:	skin [65]	melanoma, metastasis	25, 24	22 215
TCGA RNA-Seq [66]				
d_{23}:	kidney	chrom. RCC (ChRCC), clear cell RCC (CCRCC)	91, 606	20 655
d_{24}:	kidney	ChRCC, papillary RCC (PRCC)	91, 323	20 632
d_{25}:	kidney	CCRCC, PRCC	606, 323	20 684
d_{26}:	bile duct, pancreas	cholangiocarcinoma, pancreatic cancer	45, 183	20 439
d_{27}:	liver, pancreas	HCC, pancreatic cancer	424, 183	20 657

SVMs and R1-SVMs react comparably on the test scenarios. It can be observed that the accuracy differences decrease with higher numbers of dimensions. Higher differences occur for larger distances of the class centroids. Over all R2-SVMs and R1-SVMs, both bias and variance decrease for increasing dimensionality. For $n = 2$, SVM_{off}, SVM_{con}, $SVM_{con\cap off}$ achieve mean differences of 9.9% (IQR: [17.0%, 1.0%]), 10.1% (IQR: [17.0%, 1.0%]), 29.4%, (IQR: [42.0%, 18.0%]). For $n = 100$, they decrease to 0.2% (IQR: [1.0%, −1.0%]), 0.2% (IQR: [1.0%, −1.0%]), 0.2%, (IQR: [2.0%, −1.0%]).

The behaviour of the SVM_{mon} can be seen as an exception to these observations. Restricted to exactly two input dimensions, the SVM_{mon} cannot take advantage of the high-dimensional setting. Here, the bias and variance do not decline for higher dimensionality. For $n = 2$, a mean difference of 29.4% (IQR: [42.0%,18.0%]) can be observed. For $n = 100$, it achieves 14.9% (IQR: [21.0%,8.0%]).

The results of the noise experiments on artificial data are shown in figure 4. Figure 4*a* provides the results for the sample wise noise. In general, these experiments confirm the theoretical invariances against data transformations. It can be seen that for global scaling, SVM_{off}, $SVM_{off\cap con}$ and SVM_{mon} achieved equal accuracies for all noise levels. The performance of the SVM_{lin} variants of R2-SVM and R1-SVM drop rapidly. For the lowest noise level $p = 1$, mean accuracy losses of 34.6% (IQR: [40.5%, 33.8%]) are observed for the low-dimensional setting ($n = 2$) and 30.2% (IQR: [36.5%, 28.5%]) for the high-dimensional setting ($n = 100$). For global transition, the same invariant behaviour can be observed for the classifiers SVM_{con}, $SVM_{off\cap con}$ and SVM_{mon}. Here, the lowest noise level $p = 1$ results in mean losses in accuracy of 2.4% (IQR: [4.0%, 0.0%]) for the SVM_{lin} variants in the low-dimensional setting ($n = 2$) and 4.6% (IQR: [6.0%, 0.8%]) for the high-dimensional setting ($n = 100$). The combination of global scaling and global transition resulted in equal accuracies for $SVM_{off\cap con}$ and SVM_{mon} for every dimension and noise level. The SVM_{lin} variants showed mean accuracy differences of 34.7% (IQR: [42.3%, 31.0%]) in the low-dimensional setting ($n = 2$) and 29.6% (IQR: [35.8%, 30.0%]) in the high-dimensional setting ($n = 100$). After performing an exponential

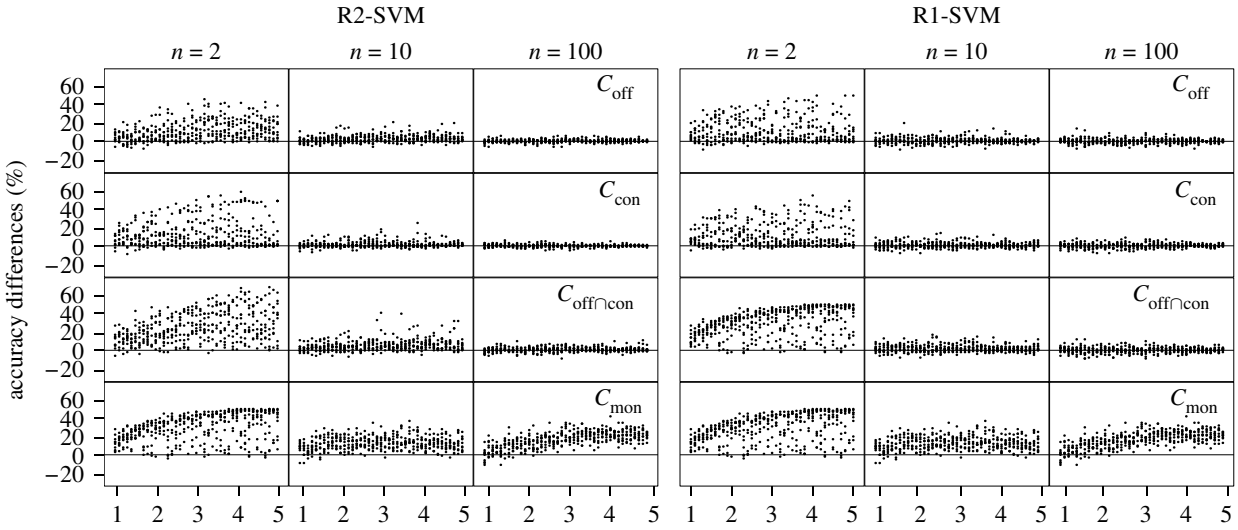

Figure 3. Evaluation of experiments on artificial datasets: the accuracy differences between SVM_{lin} and the invariant SVMs in noise-free experiments are shown. The rows show the different invariant classifiers. The columns provide the dimensionality of the underlying datasets $n = \{2, 10, 100\}$. The experiments are organized ascending according to the distances of the class centroids d (x-axis). The y-axis provides the accuracy difference. A positive value denotes a higher accuracy of the SVM_{lin}. For each value of d, 10 experiments with different class centroids are shown.

transformation on the test data, only SVM_{mon} led to equal accuracies for every dimension. The performance of the SVM_{lin} variants decreased by 44.8% (IQR: [48.0%, 45.8%]) in the low-dimensional setting ($n = 2$) and 38.6% (IQR: [43.3%, 38.0%]) in the high-dimensional setting ($n = 100$).

Figure 4*b* shows the results for the class wise noise. For global scaling, the SVM_{off} variants outperformed the SVM_{lin} variants in mean by 19.8% (IQR: [40.3%, 0.0%]) accuracy over all noise levels and all repetitions in the low-dimensional setting ($n = 2$). For the high-dimensional setting, a mean improvement of 35.3% (IQR: [48.3%, 3.0%]) accuracy was observed. For the global transition, the SVM_{con} gained in mean 8.1% (IQR: [27.3%, −12.5%]) accuracy for $n = 2$ and 33.5% (IQR: [47.3%, 0.0%]) for $n = 100$. In case of global scaling and transition, the $SVM_{off\cap con}$ variants achieved in mean −2.8% (IQR: [11.0%, −24.8%]) less accuracy in the low-dimensional settings and 40.7% (IQR: [77.3%, 15.8%]) more accuracy in the high-dimensional setting. For the exponential transformation, the SVM_{mon} variants showed a performance decreased in mean by −14.5% (IQR: [0.0%, −40.3%]) for $n = 2$. It was in mean increased by 11.1% (IQR: [23.8%, −1.3%]) for $n = 100$.

4.2. Results on transcriptome datasets

The accuracies achieved on the microarray and RNA-Seq datasets are shown in figure 5 and tabularized in the electronic supplementary material.

The R2-SVM_{lin} outperformed the kNN ($k \in \{1, 3, 5\}$) on {25, 25, 26} datasets. It was inferior in {2, 2, 1} cases. The R1-SVM_{lin} was better than the kNN in {19, 21, 20} cases. In {7, 6, 7} settings the kNN was superior. In comparison to the RFs with $nt \in \{100, 200, 300, 1000\}$ trees the R2-SVM_{lin} achieved better accuracies on {19, 20, 19, 18} datasets. Its accuracy was inferior on {7, 6, 6, 7} cases. The R1-SVM_{lin} outperformed the RFs on {12, 12, 12, 11} datasets. The RFs had higher accuracies on {15, 15, 15, 16} datasets. Th R2-SVM_{lin} showed better performance than SAE with $u \in \{100, 500, 1000\}$ in {25, 25, 25} cases. They were outperformed on {2, 2, 2} datasets. For the R1-SVM, better performances were observed in {26, 26, 26} cases. Lower performance was gained on {1, 1, 1} datasets.

Overall, the respective invariant SVMs achieved better or equal results compared to the linear one in 41 of 54 cases. At the level of individual invariant linear SVMs, it can be observed that for 20 out of 27 datasets, an invariant R2-SVM was able to achieve the same or a higher mean accuracy than R2-SVM_{lin} (R1-SVMs: 21 datasets). R2-SVM_{off} outperformed R2-SVM_{lin} in four cases (R1-SVMs: 14 cases), achieved the same accuracy in 14 cases (R1-SVMs: two cases) and achieved a lower accuracy in nine cases (R1-SVMs: 11 cases). R2-SVM_{con} was able to achieve higher accuracies than R2-SVM_{lin} for 0 datasets (R1-SVMs: 18 datasets), equal accuracies on 17 datasets (R1-SVMs: 0 datasets) and lower accuracies for 10 datasets (R1-SVMs: nine datasets). R2-$SVM_{off\cap con}$ was capable of achieving a higher accuracy than R2-SVM_{lin} in six cases (R1-SVMs: 14 cases), an equal accuracy in 12 out of 27 cases (R1-SVMs: 0 cases) and a lower accuracy in nine cases (R1-SVMs: 13 cases). The internally feature selecting R2-SVM_{mon} was never able to achieve a higher accuracy than R2-SVM_{lin}, but the R1-SVM_{mon} outperformed its linear variant in four cases. For two (R1-SVM: 0) out of 27 datasets, R2-SVM_{mon} achieved the same accuracy as R2-SVM_{lin} and for 25 datasets (R1-SVM: 23 datasets) it led to a lower accuracy.

Besides the two-dimensional SVM_{mon} classifiers the R1-SVMs yields at the reduction of features that influence the final decision boundary. An overview on the mean percentage of used features is shown in the electronic supplementary material. In all experiments, no classifier selects more than 1% of the available features. The unconstrained SVM_{lin} constructed decision boundaries based on 0.06% to 0.51% of all features. The absolute mean size of these signatures lies in between 7.36 and 104.65 features. The invariant SVMs select comparable percentages of features. They lie in the ranges of 0.07% and 0.50% (SVM_{off}), 0.07% and 0.86% (SVM_{con}) and 0.07% and 0.51% ($SVM_{off\cap con}$). This translates to a mean signature size of 9.93 and 102.76 (SVM_{off}), 9.57 and 105.08 (SVM_{con}) and 11.04 and 103.37 ($SVM_{off\cap con}$).

Figure 4. Accuracies achieved under the influence of data transformations: the figure provides the results of noise experiments with invariant R2-SVMs and R1-SVMs on artificial datasets. (*a*) The effects of sample wise data transformations on the test samples. (*b*) The influence of distinct class wise data transformations for training and test samples. The results are organized in blocks (from left to right), which correspond to the types of applied data transformations. Each column provides the results of a subclass of invariant classifiers. The rows give the dimensionality of the data $n = \{2, 10, 100\}$. Each box contains the result of 10 repetitions $r \in \{1, \ldots, 10\}$ and six increasing noise parameter $p \in \{0, \ldots, 5\}$. (Online version in colour.)

5. Discussion

In this work, we derived four invariant types of linear classifiers. The structural properties of these models allow guaranteeing invariances in the presence of small collections of molecular profiles, where malicious variation might not even be detected.

From bench to bioinformatics, the extraction of molecular profiles requires multiple preprocessing steps which have to fulfil strict protocols and often need the collaboration of different experts or institutes. Deviations or differences of these protocols can lead to noise and bias, which might lead to imprecise estimates and wrong conclusions [38]. Invariances applied can be preventive in this context. A particular type of information, which is assumed to be affected, will be neglected in subsequent modelling processes. This work is related somehow to work by the group of Rainer Spang on zero-sum regression [11,12]; in fact, our classifier $\mathcal{C}_{\text{con}}$ corresponds to this concept class. Here, we extend and generalize this approach and also embed it into the PAC learning framework.

However, ignoring a specific type of information might result in diminished classification accuracies. Our experiments with invariant support vector machines indicate that incorporating invariances against global scaling and transition leads to approximately equal performance in high-dimensional biomarker settings. In this case, the differences in the complexity of the concept classes decrease. Decreased accuracies were only observed in experiments with low dimensionality. By contrast, restriction to exactly two input variables, which is required for the strictest invariant subclass, can affect a classifier's performance.

Also, sparsity and invariance principles can be combined harmonically. The general findings described above can be observed for the feature selecting, invariant manhattan norm support vector machine. These results show that invariances can be incorporated into feature selection processes and might be used for constructing invariant marker signatures. In this case, the invariance on the full feature space is transferred to the reduced representation. The signatures of the invariant manhattan norm support vector machines have approximately the same length as their non-invariant counterpart. In our

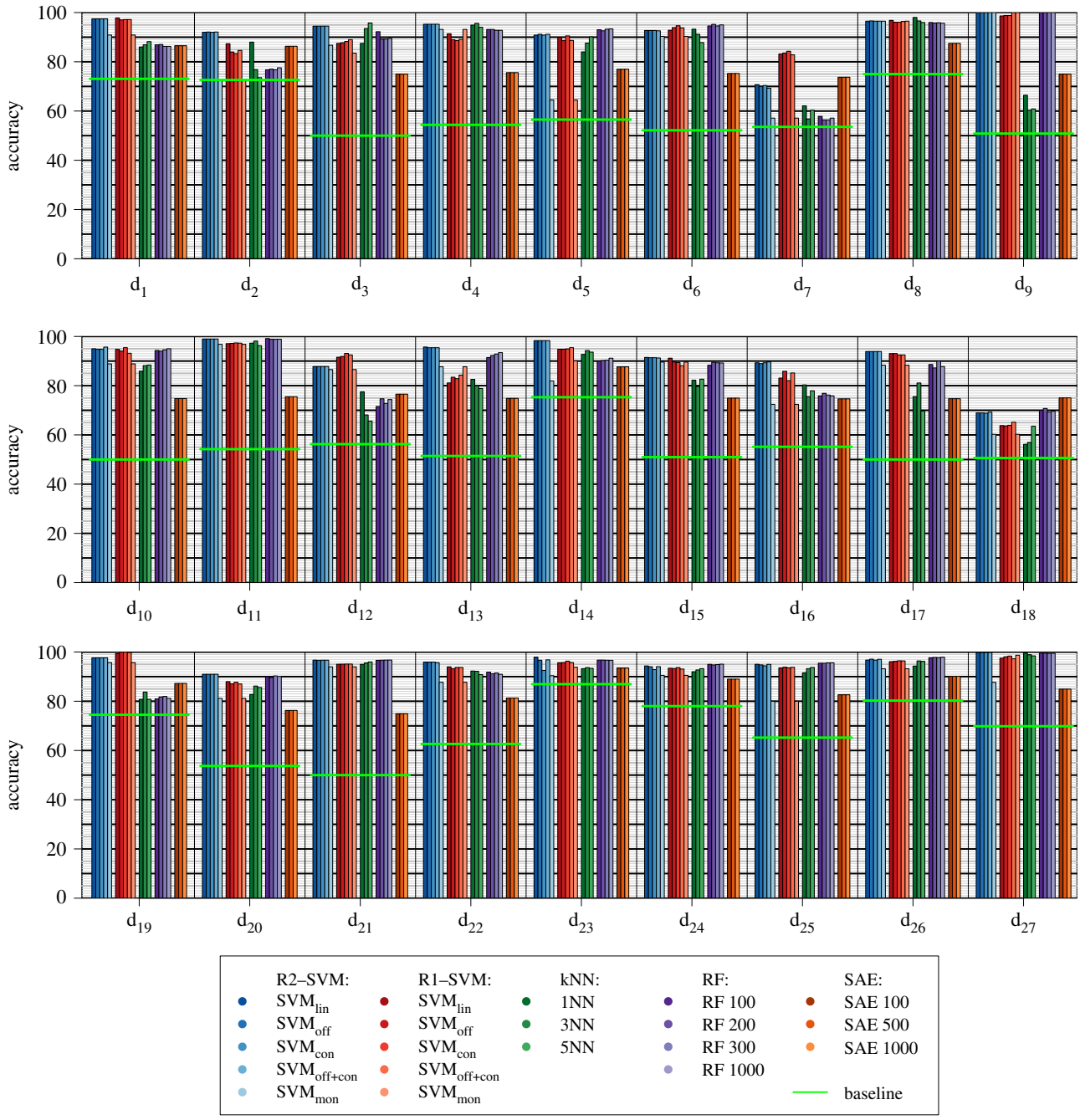

Figure 5. Results of 10 × 10 cross-validation experiments for transcriptome data: the mean accuracy is shown for the five concept classes of linear support vector machines (R2 and R1), for *k*NN with $k \in \{1, 3, 5\}$, for random forests with $nt \in \{100, 200, 300\}$ trees and the stacked auto-encoders SAE with $u \in \{100, 500, 1000\}$ units. Baseline denotes the performance of the classifier that always choses the larger class. (Online version in colour.)

experiments, invariances against global scaling or transition result in signatures comprising in mean 0.07% to 0.51% of all available biomarkers, i.e. we obtain invariant signatures of mean length of 15.77 to 105.08 markers.

Our theoretical analysis, i.e. estimating the VC dimension of the four invariant concept classes, also reveals construction principles for other invariant concepts or more complex invariant classification models. The analysed hierarchy of concept classes does reflect not only an accumulation of invariances but also a reduction of the VC dimension. These analyses indicate that a restriction to invariant classification models also reduces the complexity of the corresponding concept classes and the risk of overfitting. Suitable models might be chosen according to the PAC learning framework.

Invariances can lead to constraints on the dimensionality of the input space of a linear classifier. While invariance against global scaling require multivariate profiles, the invariance against order-preserving functions is only guaranteed for the use of two covariates. Univariate linear classifiers do not match both criteria. These invariances do not, therefore, hold for architectures that are based on single-threshold classifiers. Among these architectures are standard implementations of hierarchical systems such as classification or regression trees or ensemble classifiers such as boosting ensembles. However, these systems can gain the desired invariances by completely replacing all univariate linear classifiers by higher dimensional invariant ones. Identifying suitable combinations of fusion architectures and invariant concept classes can be seen as a natural extension of this work.

Data accessibility. Additional data can be found in the electronic supplementary material and under https://sysbio.uni-ulm.de/?Software:InvariantSVM.

Authors' contributions. L.L. and H.K. conceived the idea, L.L. and R.S. conceived the experiments, R.S. performed data acquisition, L.L.

and A.K. performed theoretical analysis, L.L. and R.S. analysed the results, R.S. and F.S. implemented the algorithms, L.L. and F.S. drafted the manuscript, H.A.K. supervised and guided the study. L.L., R.S. and H.A.K. wrote the manuscript. All authors reviewed the manuscript.

Competing interests. We declare we have no competing interests.

Funding. The research leading to these results has received funding from the German Research Foundation (D.F.G., SFB 1074 project Z1), the Federal Ministry of Education and Research (BMBF, e:Med, CONFIRM and DIFUTURE, Medical Informatics Initiative), the Ministry of Research and Art of Baden-Württemberg, Germany (Project ZIV) all to H.A.K.

References

1. Jameson JL, Longo DL. 2015 Precision medicine — personalized, problematic, and promising. *N. Engl. J. Med.* **372**, 2229–2234. (doi:10.1056/NEJMsb1503104)
2. Kraus JM, Lausser L, Kuhn P, Jobst F, Bock M, Halanke C, Hummel M, Heuschmann P, Kestler HA. 2018 Big data and precision medicine: challenges and strategies with healthcare data. *Int. J. Data Sci. Anal.* **6**, 241–249. (doi:10.1007/s41060-018-0095-0)
3. Bishop C. 2006 *Pattern recognition and machine learning (Information Science and Statistics)*. Heidelberg, Germany: Springer.
4. Webb A, Copsey K. 2011 *Statistical pattern recognition*. Chichester, UK: Wiley.
5. Hastie T, Tibshirani R, Friedman J. 2001 *The elements of statistical learning: data mining, inference, and prediction*. Heidelberg, Germany: Springer.
6. Lattke R, Lausser L, Müssel C, Kestler HA. 2015 Detecting ordinal class structures. In *Multiple Classifier Systems, 12th Int. Workshop, MCS 2015, Günzburg, Germany, 29 June–1 July* (eds F Schwenker, F Roli, J Kittler), vol. 9132, pp. 100–111. Springer.
7. Lausser L, Schäfer LM, Schirra LR, Szekely R, Schmid F, Kestler HA. 2019 Assessing phenotype order in molecular data. *Sci. Rep.* **9**, 11746. (doi:10.1038/s41598-019-48150-z)
8. Lausser L, Schmid F, Platzer M, Sillanpää MJ, Kestler HA. 2016 Semantic multi-classifier systems for the analysis of gene expression profiles. *Arch. Data Sci., Ser. A* **1**, 157–176.
9. Taudien S *et al.* 2016 Genetic factors of the disease course after sepsis: rare deleterious variants are predictive. *EBioMedicine* **12**, 227–238. (doi:10.1016/j.ebiom.2016.08.037)
10. Haasdonk B, Burkhardt H. 2007 Invariant kernel functions for pattern analysis and machine learning. *Mach. Learn.* **68**, 35–61. (doi:10.1007/s10994-007-5009-7)
11. Altenbuchinger M *et al.* 2017 Molecular signatures that can be transferred across different omics platforms. *Bioinformatics* **33**, i333–i340. with erratum Sept. 2017 (doi:10.1093/bioinformatics/btx241)
12. Zacharias HU, Rehberg T, Mehrl S, Richtmann D, Wettig T, Oefner PJ, Spang R, Gronwald W, Altenbuchinger M. 2017 Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints. *J. Proteome Res.* **16**, 3596–3605. (doi:10.1021/acs.jproteome.7b00325)
13. Wood J. 1996 Invariant pattern recognition: a review. *Pattern Recognit.* **29**, 1–17. (doi:10.1016/0031-3203(95)00069-0)
14. Schmid F, Lausser L, Kestler H. 2014 Linear contrast classifiers in high-dimensional spaces. In *Artificial neural networks in pattern recognition* (eds NE Gayar, F Schwenker, C Suen), vol. LNAI 8774, pp. 141–152.
15. Lausser L, Schmid F, Schirra LR, Wilhelm AFX, Kestler HA. 2018 Rank-based classifiers for extremely high-dimensional gene expression data. *Adv. Data Anal. Classif.* **12**, 823–825. (doi:10.1007/s11634-016-0277-3)
16. Burkovski A, Schirra LR, Schmid F, Lausser L, Kestler HA. 2017 Ordinal prototype-based classifiers. *Arch. Data Sci., Ser. A* **2**, 3–21.
17. Schölkopf B, Smola A, Müller KR. 1998 Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319. (doi:10.1162/089976698300017467)
18. Chapelle O, Schölkopf B. 2001 Incorporating invariances in non-linear support vector machines. In *NIPS* (eds T Dietterich, S Becker, Z Ghahramani), pp. 609–616. Cambridge, MA: MIT Press.
19. Tsuda K. 1999 Support vector classifier with asymmetric kernel functions. In *Proc. of ESANN'99 – European Symp. on Artificial Neural Networks* (ed. M Verleysen) pp. 183–188. D Facto.
20. Simard P, LeCun Y, Denker JS, Victorri B. 1998 Transformation invariance in pattern recognition-tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*, pp. 239–27. Berlin, Germany: Springer.
21. Schölkopf B, Burges C, Vapnik V. 1996 Incorporating invariances in support vector learning machines. In *Artificial Neural Networks — ICANN'96* (eds C von der Malsburg, W von Seelen, J Vorbrüggen, S Sendhoff), pp. 47–52. Springer Lecture Notes in Computer Science, vol. 1112.
22. Niyogi P, Poggio T, Girosi F. 1998 Incorporating prior information in machine learning by creating virtual examples. *IEEE Proc. Intell. Signal Process.* **86**, 2196–2209. (doi:10.1109/5.726787)
23. Anthony A, Biggs N. 1992 *Computational learning theory*. Cambridge, UK: Cambridge University Press.
24. Chase H, Freitag J. 2018 Modell theory and machine learning. See http://arxiv.org/abs/1801.06566.
25. Fisher RA. 1936 The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188. (doi:10.1111/j.1469-1809.1936.tb02137.x)
26. Minsky M, Papert S. 1988 *Perceptrons: an introduction to computational geometry*. Cambridge, MA: MIT Press.
27. Cover TM. 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **14**, 326–334. (doi:10.1109/PGEC.1965.264137)
28. Rosenblatt F. 1958 The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–407. (doi:10.1037/h0042519)
29. Vapnik V. 1998 *Statistical learning theory*. Chichester, UK: Wiley.
30. Guyon I, Elisseeff A. 2003 An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182.
31. Bhattacharyya C, Grate LR, Rizki A, Radisky D, Molina FJ, Jordan MI, Bissell MJ, Mian IS. 2003 Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Process.* **83**, 729–743. (doi:10.1016/S0165-1684(02)00474-7)
32. Kearns J, Vazirani U. 1994 *An introduction to computational learning theory*. Cambridge, MA: MIT Press.
33. Kestler H, Lausser L, Lindner W, Palm G. 2011 On the fusion of threshold classifiers for categorization and dimensionality reduction. *Comput. Stat.* **26**, 321–340. (doi:10.1007/s00180-011-0243-7)
34. Breiman L, Friedman J, Olshen R, Stone C. 1984 *Classification and regression trees*. Monterey, CA: Wadsworth Publishing Company.
35. Freund Y, Schapire R. 1995 A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory* (ed. P Vitányi) vol. 904, *Lecture notes in artificial intelligence*, pp. 23–37, Berlin, Germany: Springer.
36. Lausser L, Kestler H. 2014 Fold change classifiers for the analysis for the analysis of gene expression profiles. In *Proc. volume of the German/Japanese Workshops in 2010 (Karlsruhe) and 2012 (Kyoto), Studies in Classification, Data Analysis, and Knowledge Organization* (eds W Gaul, A Geyer-Schulz, Y Baba, A Okada), pp. 193–202.
37. Casella G, Berger R. 2002 *Statistical inference*. Pacific Grove, CA: Duxbury.
38. Lin W, Shi P, Feng R, Li H. 2014 Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–797. (doi:10.1093/biomet/asu031)
39. Kearns M, Vazirani U. 1994 *An introduction to computational learning theory*. Cambridge, MA: MIT Press.
40. Valiant LG. 1984 A theory of the learnable. *Commun. ACM* **27**, 1134–1142. (doi:10.1145/1968.1972)

41. Burges CJC. 1998 A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167. (doi:10.1023/A:1009715923555)
42. Hogben L ed. 2006 *Handbook of linear algebra*. Boca Raton, FL: CRC Press.
43. Abe S. 2010 *Support vector machines for pattern classification*. Heidelberg, Germany: Springer.
44. Müssel C, Lausser L, Maucher M, Kestler H. 2012 Multi-objective parameter selection for classifiers. *J. Stat. Softw.* **46**, 1–27. (doi:10.18637/jss.v046.i05)
45. Alcalay M *et al.* 2005 Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance. *Blood* **106**, 899–902. (doi:10.1182/blood-2005-02-0560)
46. Boersma B, Reimers M, Yi M, Ludwig J, Luke B, Stephens R, Yfantis H, Lee D, Weinstein J, Ambs S. 2008 A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer* **122**, 1324–1332. (doi:10.1002/ijc.23237)
47. Dyrskjøt L, Thykjaer T, Kruhøffer M, Jensen J, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft T. 2003 Identifying distinct classes of bladder carcinoma using microarrays. *Nat. Genet.* **33**, 90–96. (doi:10.1038/ng1061)
48. Estilo C *et al.* 2009 Oral tongue cancer gene expression profiling: identification of novel potential prognosticators by oligonucleotide microarray analysis. *BMC Cancer* **9**, 11. (doi:10.1186/1471-2407-9-11)
49. Gobble RM, Qin LX, Brill ER, Angeles CV, Ugras S, O'Connor RB, Moraco NH, DeCarolis PL, Antonescu C, Singer S. 2011 Expression profiling of liposarcoma yields a multigene predictor of patient outcome and identifies genes that contribute to liposarcomagenesis. *Cancer Res.* **71**, 2697–2705. (doi:10.1158/0008-5472.CAN-10-3588)
50. Hummel M *et al.* 2006 A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* **354**, 2419–2430. (doi:10.1056/NEJMoa055351)
51. Iwamoto K, Kakiuchi C, Bundo M, Ikeda K, Kato T. 2004 Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol. Psychiatry* **9**, 406–416. (doi:10.1038/sj.mp.4001437)
52. Jones J *et al.* 2005 Gene signatures of progression and metastasis in renal cell cancer. *Clin. Cancer Res.* **11**, 5730–5739. (doi:10.1158/1078-0432.CCR-04-2225)
53. Kimpel MW, Strother WN, McClintick JN, Carr LG, Liang T, Edenberg HJ, McBride WJ. 2007 Functional gene expression differences between inbred alcohol-preferring and non-preferring rats in five brain regions. *Alcohol* **41**, 95–132. (doi:10.1016/j.alcohol.2007.03.003)
54. Kuriakose M. 2004 Selection and validation of differentially expressed genes in head and neck cancer. *Cell. Mol. Life Sci.* **61**, 1372–1383. (doi:10.1007/s00018-004-4069-0)
55. Landi MT *et al.* 2008 Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE* **3**, e1651. (doi:10.1371/journal.pone.0001651)
56. Lu Y *et al.* 2006 A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med.* **3**, e467. (doi:10.1371/journal.pmed.0030467)
57. Orsmark-Pietras C *et al.* 2013 Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. *Eur. Respir. J.* **42**, 65–78. (doi:10.1183/09031936.00077712)
58. Shipp M *et al.* 2002 Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74. (doi:10.1038/nm0102-68)
59. Singh D *et al.* 2007 Gene expression correlates of clinical prostate cancer behavior. *J. Neurosci.* **1**, 203–209.
60. Stanke F, van Barneveld A, Hedtfeld S, Wölfl S, Becker T, Tümmler B. 2014 The CF-modifying gene EHF promotes p.Phe508del-CFTR residual function by altering protein glycosylation and trafficking in epithelial cells. *Eur. J. Hum. Genet.* **22**, 660. (doi:10.1038/ejhg.2013.209)
61. Strunnikova N, Hilmer S, Flippin J, Robinson M, Hoffman E, Csaky K. 2005 Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. *Free Radical Biol. Med.* **39**, 781–796. (doi:10.1016/j.freeradbiomed.2005.04.029)
62. Sun Y, Goodison S. 2009 Optimizing molecular signatures for predicting prostate cancer recurrence. *Prostate* **69**, 1119–1127. (doi:10.1002/pros.20961)
63. Vilar E *et al.* 2009 Gene expression patterns in mismatch repair-deficient colorectal cancers highlight the potential therapeutic role of inhibitors of the phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin pathway. *Clin. Cancer Res.* **15**, 2829–2839. (doi:10.1158/1078-0432.ccr-08-24320)
64. Wang G *et al.* 2013 Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PLoS ONE* **8**, e63826. (doi:10.1371/journal.pone.0063826)
65. Xu L *et al.* 2008 Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. *Mol. Cancer Res.* **6**, 760–769. (doi:10.1158/1541-7786.MCR-07-0344)
66. The Cancer Genome Atlas (TCGA) Research Network. 2008 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068. (doi:10.1038/nature07385)
67. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. 2013 Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS ONE* **8**, 1–10. (doi:10.1371/journal.pone.0071462)
68. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003 Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264. (doi:10.1093/biostatistics/4.2.249)
69. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2009 RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500. (doi:10.1093/bioinformatics/btp692)
70. O'Hara RB, Kotze DJ. 2010 Do not log-transform count data. *Methods Ecol. Evol.* **1**, 118–122. (doi:10.1111/j.2041-210X.2010.00021.x)
71. Fix E, Hodges JL. 1951 Discriminatory analysis: nonparametric discrimination: consistency properties. Technical report project 21-49-004, report number 4 USAF School of Aviation Medicine, Randolf Field, Texas.
72. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
73. Bengio Y, Lamblin P, Popovici D, Larochelle H. 2007 Greedy layer-wise training of deep networks. In *Advances in neural information processing systems 19* (eds B Schölkopf, JC Platt, T Hoffman), pp. 153–160. Cambridge, MA: MIT Press.