

Interpreting Machine-Learning Models In Transformed Feature Space With An Application To Remote-Sensing Classification

Dr. R. Sugumar, Assistant Professor, Department Of Data Science, SICET, Hyderabad

S.D. Anushna, Assistant Professor, Department Of Data Science, SICET, Hyderabad

N.Shekhar, Assistant Professor, Department Of Data Science, SICET, Hyderabad

Abstract

Model-agnostic tools for the post-hoc interpretation of machine-learning models struggle to summarize the joint effects of strongly dependent features in high-dimensional feature spaces, which play an important role in semantic image classification, for example in remote sensing of landcover. This contribution proposes a novel approach that interprets machine-learning models through the lens of feature-space transformations. It can be used to enhance unconditional as well as conditional post-hoc diagnostic tools including partial-dependence plots, accumulated local effects (ALE) plots, permutation feature importance, or Shapley additive explanations (SHAP). While the approach can also be applied to non-linear transformations, linear ones are particularly appealing, especially principal component analysis (PCA) and a proposed partial orthogonalization technique. Moreover, structured PCA and model diagnostics along user-defined synthetic features offer opportunities for representing domain knowledge. The new approach is implemented in the R package *wiml*, which can be combined with existing explainable machine-learning packages. A case study on remote-sensing landcover classification with 46 features is used to demonstrate the potential of the proposed approach for model interpretation by domain experts. It is most useful in situations where groups of feature are linearly dependent and PCA can provide meaningful multivariate data summaries.

Keywords Interpretable machine learning · Dataset-level post-hoc interpretation · Predictive modelling · Model visualization · Feature-space transformation · Remote sensing

1 Introduction

Interpreting complex nonlinear machine-learning models is an inherently difficult task. A common approach is the post-hoc analysis of black-box models for dataset-level interpretation (Murdoch et al., 2019) using model-agnostic techniques such as the permutation-based variable importance, and graphical displays such as partial-dependence plots that visualize main effects while integrating over the remaining dimensions (Molnar et al., 2020).

These tools are mostly limited to displaying the relationship between the response and one (or sometimes two) predictor(s), while attempting to control for the influence of the other predictors. This can be rather unsatisfactory when dealing with a large number of highly correlated predictors, which are often semantically grouped. While the literature on explainable machine learning has often focused on dealing with dependencies affecting *individual* features, for example by introducing conditional diagnostics (Strobl et al., 2008; Molnar et al., 2023), practical solutions for model interpretation in high-dimensional feature spaces with strong dependencies are a current area of research (Molnar et al., 2020, 2022; Seedorff & Brown, 2021; Au et al., 2022).

High-dimensional situations with strongly dependent features routinely occur in environmental remote sensing and other geographical and ecological analyses (Landgrebe, 2002; Zortea et al., 2007), which motivated the present proposal to enhance existing model interpretation tools by offering a new, transformed perspective. Similar issues occur in biomedical applications involving, for example, speech signal processing (Sakar et al., 2019) and Raman spectroscopy (Guo et al., 2020). With regards to remote sensing, for example, vegetation ‘greenness’ as a measure of photosynthetic activity is often used to classify landcover or land use from satellite imagery acquired at multiple time points throughout the growing season (Peña & Brenning, 2015). Spectral reflectances of equivalent spectral bands (the features) are usually strongly correlated within the same phenological stage since vegetation characteristics vary gradually. Similarly, when using texture features to characterize image structure based on a filter bank, features with similar filter settings can be strongly correlated, as in our case study (Brenning et al., 2012).

Turning our attention back to the importance of individual features, an orthogonalization technique can be used to single out the effect of individual features on model predictions, avoiding the sometimes complex structure of PCs. A similar algorithm has been proposed in previous work (Adebayo & Kagal, 2016) in an isolated form that can be accommodated within the proposed general framework. This approach can, as proposed in this contribution, be applied to paths through feature space, such as nonlinear curves defined by domain-specific perspectives, or to data-driven transitions between clusters of observations.

Considering the outlined challenges and existing partial solutions, the objective of the present work is to establish a general formal framework for the post-hoc interpretation of black-box models in transformed space. The proposed framework can be combined with commonly used plot types and diagnostics including partial dependence plots, accumulated local effects (ALE) plots, permutation-based variable importance, and Shapley additive explanations (SHAP), among other model-agnostic techniques that only have access to the trained model (Apley & Zhu, 2020; Molnar, 2022). While the focus of this contribution is on visualizing main effects and their predictive importance, analyses of conditional relationships may also benefit from this perspective (Strobl et al., 2008; Molnar et al., 2023). The framework is implemented in an extensible, open-source package in R, the `wiml` package, which can be combined with existing model interpretation toolboxes.

2 Proposed method

Let's consider a regression or classification model

$$\hat{f}: \mathbf{x} \mapsto \hat{f}(\mathbf{x}) \in \mathbb{R}$$

that was fitted to a training sample L in the (original, untransformed) p -dimensional feature space $X \subset \mathbb{R}^p$. I will assume $\hat{f}(\mathbf{x}) \in \mathbb{R}$; in the case of classification problems, $\hat{f}(\mathbf{x})$ shall therefore represent predictions of some real-valued quantity such as the probability or logit of a selected target class. One of the features, referred to as x_s , is selected as the feature of interest, and the remaining features are denoted by $\mathbf{x}_c$.

2.1 Example: partial-dependence plots

In this situation, the partial-dependence plot of $\hat{f}$ with respect to x_s can formally be defined as

$$\begin{aligned} \hat{f}_{x_s, PDP}(\mathbf{x}_c) &= E_{\mathbf{x}_c} \left[\hat{f}(\mathbf{x}_c, x_s) \right] \\ &= \int_{\mathbf{x}_c} \hat{f}(\mathbf{x}_c, x_s) dP(\mathbf{x}_c) \end{aligned}$$

(Molnar, 2022). This plot, which can be generalized to more than one x_s dimension, was introduced by Friedman (2001) to visualize main effects of predictors in machine-learning models.

The approach outlined in this section can be applied to ALE plots and related model-agnostic tools, including permutation-based variable importance and their conditional modifications, or Shapley additive explanations (see reviews by Molnar et al, 2020; Molnar, 2022).

Partial-dependence plots have some disadvantages such as the extrapolation of $\hat{f}$ beyond the region in X for which training data is available (Apley & Zhu, 2020; Molnar et al., 2022). This is especially the case when predictors are strongly correlated, as in our case study. Nevertheless, without loss of generality, this simple plot type helps to illustrate the proposed approach.

2.2 Transformed feature space

When several predictors are strongly correlated and/or express the same domain-specific concept such as ‘early-season vegetation vigour’ in vegetation remote-sensing, we may be more interested in exploring the overall effect of these predictors. Principal component analysis (PCA) and related data transformation techniques such as factor analysis are tools that are often used by practitioners to synthesize and interpret multivariate data (for example, Basille et al, 2008; Rousson & Gasser, 2004; Cunningham & Ghahramani, 2015).

More generally speaking, we could think of an invertible transformation function

$$\mathbf{T} : X \rightarrow W \subset \mathbb{R}^p, \quad \mathbf{w} = \mathbf{T}(\mathbf{x})$$

that can be used to re-express the features in our data set. We will assume that $\mathbf{T}$ is continuous and differentiable. PCA is one such example, which has been considered recently by Seedorff and Brown (2021) with a focus on a practical algorithm and its implementation.

Through the composition of the back transformation $\mathbf{T}^{-1}$ and the model function $\hat{f}$, we can now formally define a model $\hat{g}$ on W ,

$$\hat{g} := \hat{f} \circ \mathbf{T}^{-1},$$

which predicts the real-valued response based on ‘data’ in W although it was trained using a learning sample $L \subset X$ in the untransformed space.

We can use this to formally re-express the partial-dependence plot as a function of w_s :

$$\begin{aligned} \hat{f}_{w_s, PDP}(w_s) &= E_{w_C}(\hat{f} \circ \mathbf{T}^{-1})(w_s, w_C) \\ &= \int_{w_C} (\hat{f} \circ \mathbf{T}^{-1})(w_s, w_C) dP_{w_C} \end{aligned}$$

Note that $\mathbf{T}^{-1}$, when used only on data in $\mathbf{T}(X)$, does not create values outside the data-supported region X , and it therefore avoids extrapolation of $\hat{f}$.

Also, when choosing PCA for $\mathbf{T}$ as a data-driven approach, the variables in $\mathbf{T}(L)$ are linearly independent, and statistically independent if L arises from a multivariate normal distribution. Thus, the PCA approach overcomes one of the limitations of partial-dependence plots and broadens their applicability, at least in the case of linear dependencies.

2.3 Partial orthogonalization

In some instances, PCs (and other multivariate transformations) of large and complex feature sets can be difficult to interpret, and analysts would therefore like to focus on individual features that are perhaps ‘representative’ of a larger group of features — for example, vegetation greenness in mid-June may be a good proxy for vegetation greenness a few weeks earlier and later, as expressed by other features in the feature set (Peña & Brenning, 2015).

This can be addressed by proposing a transformation of X in which $w_s := x_s$ is retained, while making the remaining base vectors linearly independent of x_s . This can be achieved through *partial orthogonalization*,

$$w_i := x_i - a_i - b_i x_s, \quad (1)$$

where a_i and b_i are the intercept and regression slope of a simple linear regression of x_s on x_i as the response. For simplicity of notation, it is assumed that the data is centered and standardized beforehand, in which case it simplifies to $a_i = 0$, and b_i equals the Pearson correlation coefficient.

This then defines a linear transformation $\mathbf{T} : X \rightarrow W$, which can be represented by its coefficient matrix. Note that $\mathbf{T}$ can be inverted using

$$x_i = w_i + b_i w_s, \quad (2)$$

since $x_s = w_s$, and assuming that all $b_i < 1$, which is the case when there are no duplicated features. A related iterative orthogonalization approach has previously been proposed in the context of feature ranking (Adebayo & Kagal, 2016).

2.4 Partial orthogonalization for dependence plots along synthetic features

Domain scientists may more generally want to visualize the effect of a real-valued function of multiple features. As an example, knowing that several features are strongly correlated, how does the response vary with their average, or, more generally, a linear or nonlinear function of the features? This information is sometimes hidden in an ocean of individual main-effects plots or variable-importance measures.

In other situations, there may be simple process-based models that have the potential to provide deeper insights into black-box models based on domain knowledge. These models may be candidates for an enhancement of feature space, or they might express specific theories or hypotheses.

Any of these transformations can be thought of as a real-valued function of the other features in the data set, $h(\mathbf{x})$, which is added to the feature set as a new feature $x_{p+1} := h(\mathbf{x})$ to augment the feature space by one dimension. While this feature is not actually used by $\hat{f}$, the partial orthogonalization technique offers an entry point to examine how $h(\mathbf{x})$, through its (linear) contribution to $x_1, \dots, x_p$, impacts the predictions produced by $\hat{f}$.

In different use cases there may be different ways of constructing synthetic features of interest to the domain scientist:

- A group of strongly positively correlated features could be averaged to obtain an overall signal (examples: daily mean temperature when the actual features are hourly temperatures).
- Contrasts between groups of features could be calculated (example: average of daytime temperature features minus average of nighttime temperature features as a measure of diurnal temperature amplitude).
- A linear path can be drawn from one cluster centre to another, where cluster centres $\mathbf{c}_1, \dots, \mathbf{c}_k \in X$ are obtained by unsupervised clustering in feature space (for example, k -means). The path between clusters 1 and 2 is simply defined as $t\mathbf{c}_1 + (1 - t)\mathbf{c}_2$, etc. Here, an instance's distance to a cluster centre could serve as a synthetic feature.

- A linear path between user-defined points in feature space; in remote sensing, for example, so-called endmembers representing spectral characteristics of ‘pure’ surface types such as asphalt or water (Somers et al., 2016).

Evidently, these synthetic features could also be added to the feature set in a feature engineering step and used for model training. Nevertheless, the proposed approach provides an opportunity for a post-hoc assessment and visualization of the influence of such features on the model’s output.

Technically, partial orthogonalization along synthetic features is achieved by (linearly) partialing the effect of x_{p+1} out of the features $x_1, \dots, x_p$ (Eq. (1)) — either out of all features, or out of the subset of features effectively involved in the calculation of $h(\mathbf{x})$. In applying interpretation tools such as ALE plots or permutation methods, these features then need to be reconstructed from new values of x_{p+1} , which is achieved by inverting the partial orthogonalization (Eq. (2)).

2.5 Two-dimensional model-agnostic plots

The proposed approaches are not limited to one-dimensional model interpretation along one selected feature $x_s \in \mathbb{R}$ — the methods equally apply to bivariate relationships ($\mathbf{x}_s \in \mathbb{R}^2$), which can be used to display pairwise interactions. Clearly, in a high-dimensional situation, the need to reduce dimensionality in post-hoc model interpretation is even more pressing when interpreting up to $p(p-1)/2$ pairwise interactions, and the proposed approach offers a practical tool to address this in situations where dimension reduction is viable.

3 Implementation

The proposed methods have been implemented in the R package `wiml` (code available at <https://github.com/alexanderbrenning/wiml>). It implements transformation functions called ‘warpers’ based on PCA (of all features or a subset of features), structured PCA (for multiple groups of features), and partial feature orthogonalization, all of which are based on rotation matrices and therefore share a common core. Due to the modular and object-oriented structure, users can implement their own transformations without requiring changes to the package.

These warpers can be used to implement the composition $\hat{f} \circ \mathbf{T}^{-1}$ by ‘warping’ a fitted machine-learning model. The resulting object behaves like a regular fitted machine-learning model in R, offering an adapted `predict` method. From a user’s perspective, the resulting model feels like it had been fitted to the transformed data $\mathbf{T}^{-1}(L)$, except that it hasn’t. This ‘warped’ fitted model can, in principle, be used with any model-agnostic tool that doesn’t require retraining. An implementation of the composition $\hat{f} \circ \mathbf{T}^{-1}$ involving the untrained model $\hat{f}$ is also available; this can be used for drop and relearn or permute-and-relearn techniques (Hooker et al., 2021).

The package has been tested and works well with the `iml` package for interpretable machine learning (Molnar et al., 2018), but it can also be combined with other frameworks since it only builds thin wrappers around standard R model functions. Initial tests with the DALEX framework for explainable machine-learning (Biecek, 2018) and its interactive environment `modelStudio` (Baniecki & Biecek, 2019) have been successful, as have

been tests with the `pdp` package (Greenwell, 2017). The `wiml` package does not re-implement existing model interpretation routines.

4 Case study

The potential of the proposed methods is demonstrated in a case study from land cover classification, which is a common machine-learning task in environmental remote sensing (for example, Mountrakis et al, 2011; Peña & Brenning, 2015). One particularly challenging task is the detection of rock glaciers, which, unlike ‘true’ glaciers, do not present visible ice on their surface; they are rather the visible expression of creeping ice-rich mountain permafrost. In the present case study, we look at a subset of a well-documented data set consisting of a sample of 1000 labelled point locations (500 presence and 500 absence locations of flow structures on rock glaciers) in the Andes of central Chile (Brenning et al., 2012).

There are 46 features in total, which are divided into two unequal subsets: Six features are terrain attributes (local slope angle, potential incoming solar radiation, mean slope angle of the catchment area, and logarithm of catchment height and catchment area), which are proxies for processes related to rock glacier formation. The other 40 features are Gabor texture features (Clausi & Jernigan, 2000), which are designed to detect the furrow-and-ridge structure in high-resolution (1 m × 1 m) satellite imagery, in this case panchromatic IKONOS imagery (see Brenning et al, 2012, for details). The 40 Gabor features correspond to different filter bandwidths (5, 10, 20, 30 and 50 m), anisotropy factors (1 or 2), and types of aggregation over different filter orientations (minimum, median, maximum, and range). Sample maps of features and IKONOS imagery are shown in Fig. 2 of Brenning et al. (2012).

Texture features with similar filter settings are often strongly correlated with each other. This is especially true for minimum and median aggregation with otherwise equal settings, and for maximum and range aggregation. Overall, the median of each feature’s strongest Pearson correlation is 0.92 (minimum: 0.80). Correlations among terrain attributes are much smaller (median strongest correlation: 0.60). Terrain attributes and texture features are weakly correlated (maximum correlation: 0.30). Correlation statistics are very similar for Spearman’s rank-based correlation.

To explore the feature sets, PCAs is performed for the entire set of 46 feature and for the subset of 40 Gabor features (Fig. 1). In the entire feature set, 63.6% of the variance is concentrated in the first two PCs (first six PCs: 83.7%). In the more strongly correlated Gabor feature set, in contrast, the first two PCs make up 72.2% of the variance (first six PCs: 89.5%). The main PCs turned out to be interpretable by domain experts. PC #1 of the Gabor feature set (‘Gabor1’, in the figures) is basically an overall average of all texture features, meaning that it expresses the overall presence of striped patterns of any characteristics. Gabor PC #2 represents the contrast between minimum and median aggregated anisotropic Gabor features and the rest; large values are interpreted as incoherent patterns with no distinct, repeated stripe pattern. Gabor PC #3 expresses differences between large-wavelength range or maximum-aggregated features versus the short-wavelength features, which represents the heterogeneity in the width of stripes, and thus the size of linear surface structures. Large values correspond to distinct patterns of large amplitude.

To test the partial orthogonalization along synthetic features, the role of a widely used terrain attribute that is not included in the original feature set is examined. The

Table 1 Overall accuracies of random-forest classification of rock glaciers using the entire feature set, and omitting either the terrain attributes or the texture features

	All features	Only terrain attributes	Only texture features
Accuracy	0.808	0.757	0.697
Decrease in accuracy	0.000	0.051	0.111

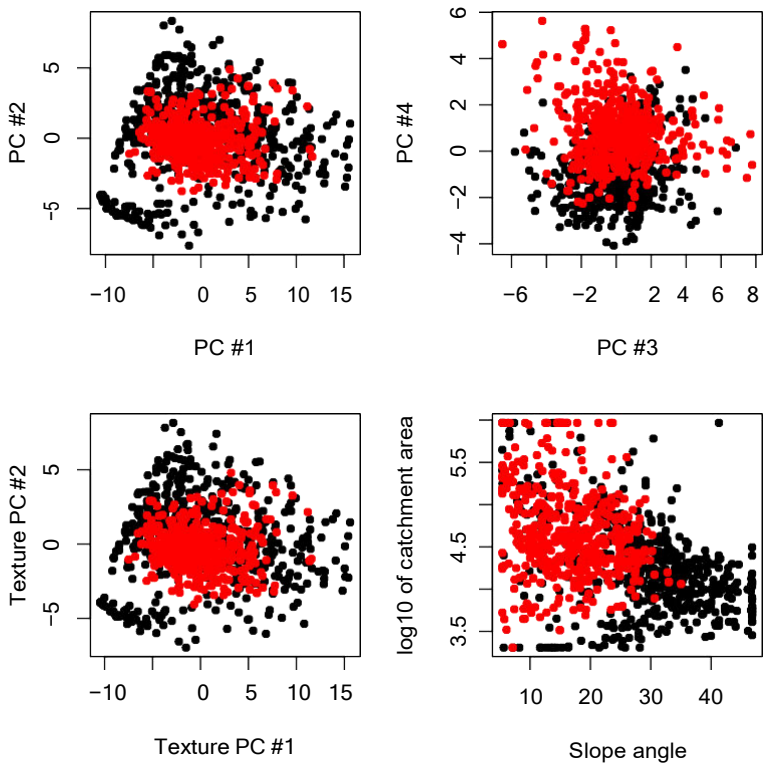


Fig. 1 Feature (sub)space diagrams. Top row: First PCs of the entire feature set. Bottom row: First PCs of the texture feature (sub-)set and the top-ranked terrain attributes

topographic wetness index (TWI) is defined as the logarithm of the ratio of catchment area and the tangent of slope angle, which means that it is linear in one feature (logarithm of catchment area; correlation 0.94), and slightly curved in another one (slope angle; correlation -0.68). Partial orthogonalization is applied with respect to slope and log. catchment area while leaving the other features unchanged. For comparison with the synthetic-feature approach, a model with TWI as an additional feature is also trained.

A random-forest classifier is used for the classification of rock glaciers based on the features introduced above. Its overall accuracy, estimated by spatial cross-validation between the two sub-regions (Brenning, 2012), is 80.8%. Omitting terrain attributes from the feature set has a greater impact on performance than omitting the texture features (Table 1).

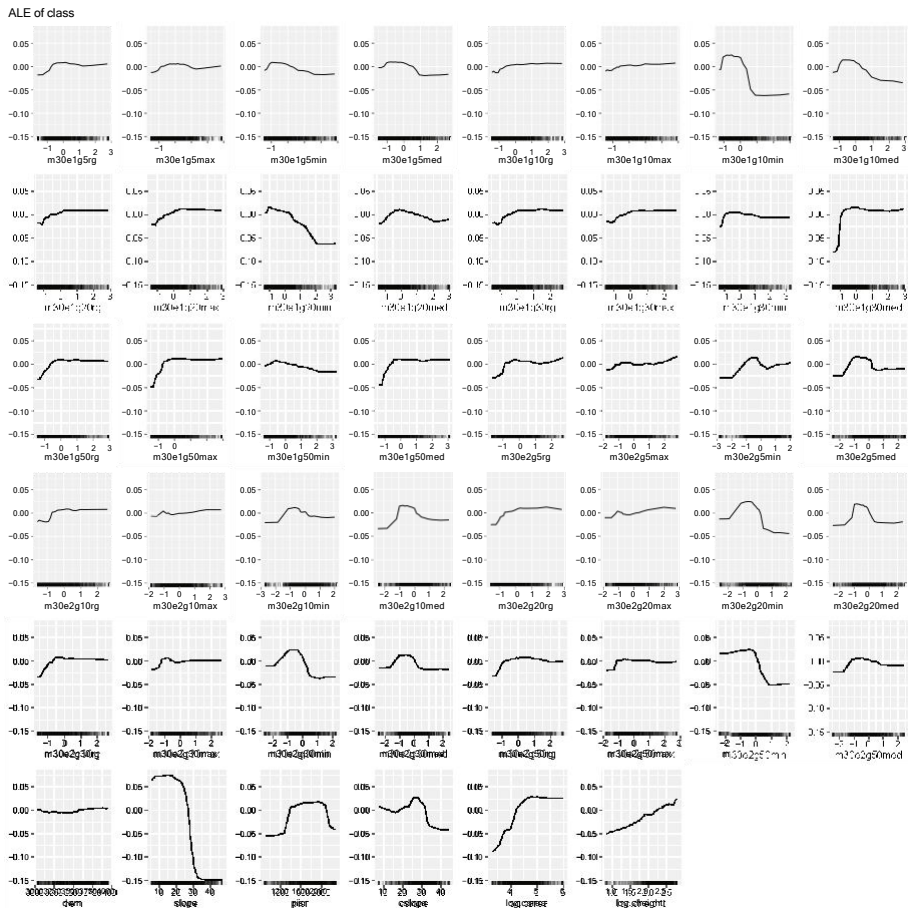


Fig. 2 Ordinary ALE plots for all 46 features

5 Results

5.1 Standard approach

With 46 features that are grouped into two semantic feature types (terrain attributes, texture features), it can be challenging to interpret the patterns represented by marginal effects plots (Fig. 2). Although there appears to be some consistency in direction among many of the texture features, it is difficult to identify an overall pattern that can be summarized verbally, and it would be unreasonable to present such detailed visual information to a conference audience that is expecting a concise and coherent narrative.

5.2 Interpretation in transformed feature space

The ALE plots along principal axes distill 71.6 percent of the feature variance into only three plots (Fig. 3). Nevertheless, considering the semantic differences and weak correlations between texture features and terrain attributes, it seems unnecessary to combine all features in a joint PCA, which results in PCs with an at least slightly mixed meaning in this purely data-driven transformation.

The structured PCA approach, in contrast, allows us to explicitly separate the model’s representation of effects of texture features and terrain attributes, which is desirable from a domain expert’s perspective (knowledge-driven transformation) and statistically justifiable based on the weak correlations between these feature groups. Larger overall texture signals (Gabor PC #1) are associated with higher predicted rock glacier probability, although extremely large PC #1 values are less discriminative since they may also occur along linear features such as eroded channels (Fig. 4). However, a large contrast between minimum/median anisotropic texture features and the remaining texture features, as expressed by a high Gabor PC #2 value, is more often associated with an absence of rock glaciers. In other words, the absence of coherently oriented, repeated stripes is not typical of rock glaciers — these may be more typical of non-repeated stripes (for example, erosion gullies, jagged rock slopes).

The permutation-based and SHAP-based assessments of the importance of texture PCs and terrain attributes both show that subsequent PCs contribute much less to the predictive performance, and that slope angle is the most salient feature overall (Fig. 5). Clearly, the combined importance of Gabor features as summarized by Gabor PCs #1 and #2 provides a more comprehensible summary than an incoherent litany of individual feature importances of strongly correlated features, which should not be permuted independently of each other (Fig. 6).

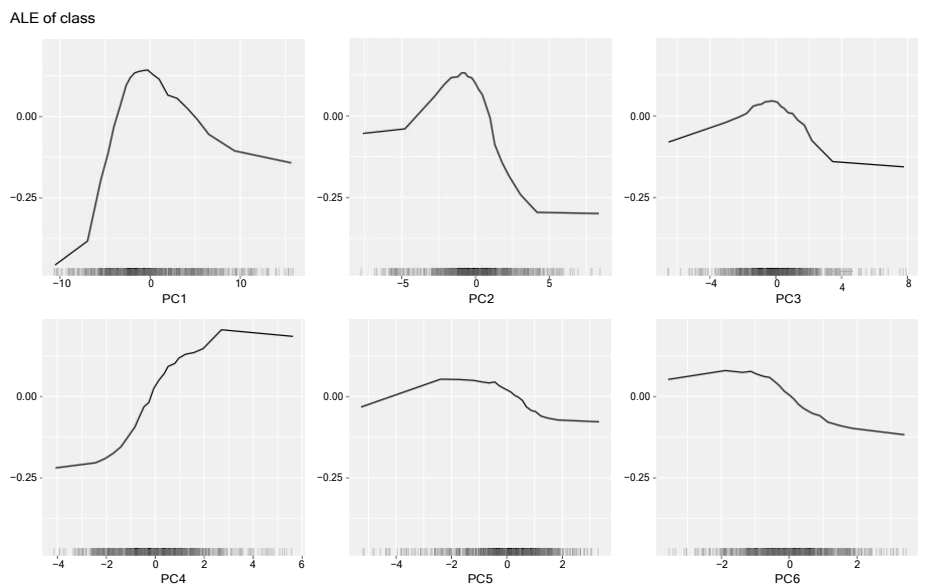


Fig. 3 ALE plots along the first six principal axes, applying PCA to the entire feature set

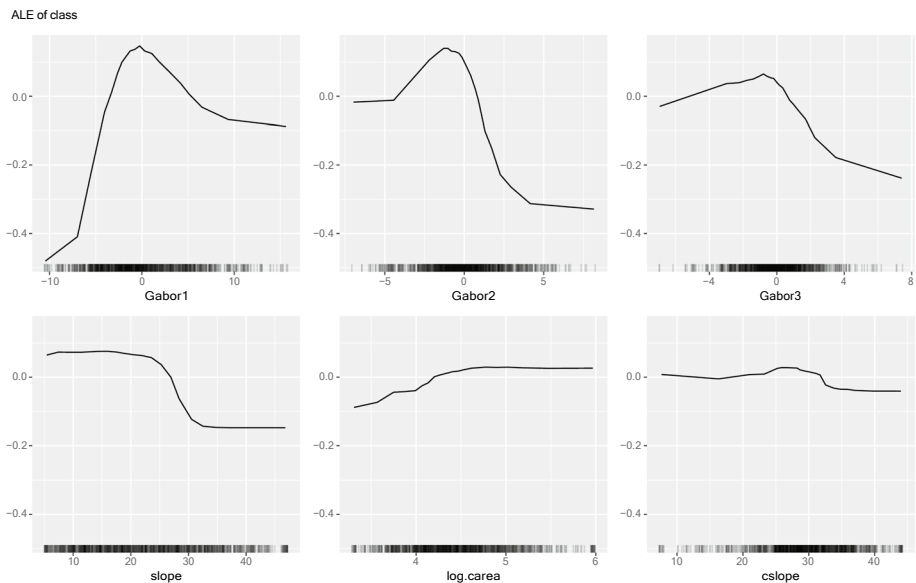


Fig. 4 ALE plots along the first principal axes of texture features, and for the most important terrain attributes

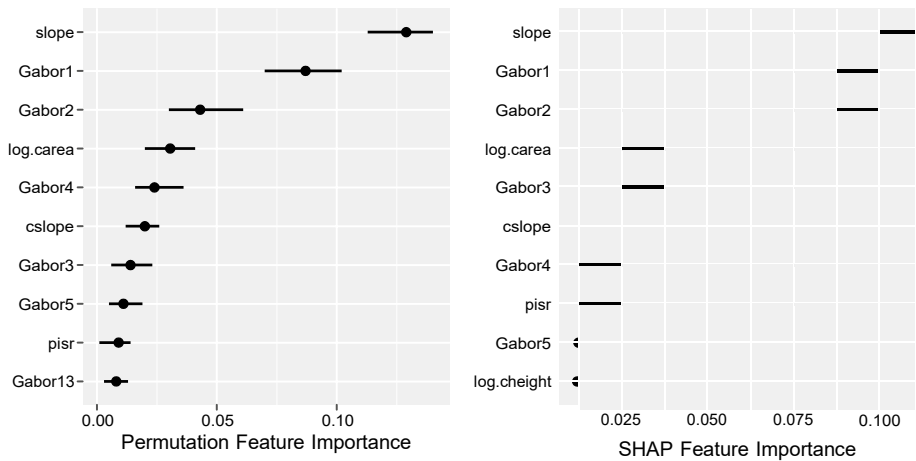


Fig. 5 Permutation and SHAP feature importances of the 10 top-ranked texture principal components and terrain attributes. Bars indicate permutation variability and approximate confidence intervals, respectively, both at the 90% level

6 Discussion

Overall, interpretation plots along the principal axes are capable of distilling complex high-dimensional relationships into low-dimensional summaries in a data-driven manner, thus providing a tidier, better structured and more focused approach to model interpretation than traditional tools that focus on individual predictors in an ocean of highly corre-

lated features. This behaviour is highly desirable from a domain expert's perspective, and applying it in a structured, knowledge-driven manner allows the analyst to honour domain knowledge and feature semantics.

One concern in model-agnostic model interpretation is, or should be, the extrapolation from a training sample to a set of data points at which the model $\hat{f}$ is evaluated. This is especially an issue for PD plots and permutation methods, which require stronger extrapolation than the locally operating ALE method, as well as for non-smooth models such as random forests. In conjunction with the proposed method, this extrapolation takes place in the transformed space W , and as a consequence, it may be exacerbated or reduced, depending on the local properties of the transformation T . This is an issue that users should be aware especially when working with sparse data, such as multimodal data distributions.

Of course fitting the classifier to PCA-transformed data as input features could have provided direct access to ALE plots along principal axes. However, we would want our feature engineering decisions to be directed towards improving predictive performance, and we would therefore prefer not to risk compromising an optimal performance to satisfy our desire to interpret our model. While this is not an issue in the present case study (overall accuracy 0.789 with PC features versus 0.808 with the original predictors), our experience shows that PCA-transformed predictors can worsen the predictive performance. Also, model-agnostic post-hoc analysis tools are precisely meant to be applicable to black-box models that are provided 'as is', without the possibility of altering their input features, in which case the proposed 'hands-off' access to transformed perspectives is particularly valuable.

The proposed use of PCA and related linear transformation technique appears to be in contradiction to the use of complex nonlinear machine-learning models. Nevertheless, it could be argued that linear cross-sections of feature space along the original feature axes are no less arbitrary and limiting, considering the often strong correlations with other features that would have to be interpreted simultaneously. From that perspective, principal axes provide a 'tidier' perspective and smarter peek into feature space than traditional ALE or partial-dependence plots. Linear transformations similar to PCA may further enhance interpretability by offering a more structured or target-oriented perspective based on simple components (Rousson & Gasser, 2004) or discriminant functions (Cunningham & Ghahramani, 2015).

Beyond linear transformations, the proposed approach provides a general framework even for nonlinear perspectives on feature space and model functions. In particular, paths proposed in Sect. 2.4 may be nonlinear, as, for example, defined by a physical model that could be used by domain experts to check model plausibility in a knowledge-driven way. Nevertheless, especially nonlinear transformations should only be used in conjunction with interpretation tools such as ALE plots and SHAP values that aim to preserve correlations among features, and non-monotonic mappings should be avoided.

7 Conclusions

Despite the inherent limitations of post-hoc machine-learning model interpretation, feature-space transformations, and structured PCA transformations in particular, are a powerful tool that allows us to distill complex nonlinear relationships into an even smaller number of univariate plots than previously possible, representing perspectives that are informed by domain knowledge. These transformations provide an intuitive access to feature space, which can be easily wrapped around existing model implementations. Model interpretation through the lens of feature transformation and dimension reduction allows us to peek into the feature space at an oblique angle — a strategy that many of us have successfully applied when checking if our kids are asleep, and a much more successful strategy than staring along the walls, that is, the original feature axes, especially when these are nearly parallel.

Acknowledgements I would like to thank Philipp Lucas (DLR Institute of Data Science) for encouraging discussions on an early version of the manuscript. Constructive feedback from two reviewers is gratefully acknowledged.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received for conducting this study.

Data availability The case study's data is included in the R package `wiml` (<https://github.com/alexanderbrenning/wiml/>) as `data(gabor)` and should be cited as Brenning et al. (2012). The package includes a vignette that walks the reader through the analysis steps (https://github.com/alexanderbrenning/wiml/blob/main/vignettes/gabor_rf.md).

References

- Adebayo, J., & Kagal, L. (2016). Iterative orthogonal feature projection for diagnosing bias in black-box models. <http://arxiv.org/abs/1611.04967>
- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>
- Au, Q., Herbinger, J., Stachl, C., et al. (2022). Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, 36, 1401–1450. <https://doi.org/10.1007/s10618-022-00840-5>
- Baniecki, H., & Biecek, P. (2019). modelStudio: Interactive studio with explanations for ML predictive models. *Journal of Open Source Software*, 4(43), 1798. <https://doi.org/10.21105/joss.01798>.
- Basille, M., Calenge, C., Marboutin, E., et al. (2008). Assessing habitat selection using multivariate statistics: Some refinements of the ecological-niche factor analysis. *Ecological Modelling*, 211(1), 233–240. <https://doi.org/10.1016/j.ecolmodel.2007.09.006>
- Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. *Journal of Machine Learning Research*, 19(84), 1–5.
- Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorst. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, pp 5372–5375, <https://doi.org/10.1109/IGARSS.2012.6352393>.
- Brenning, A. (2023). Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. *International Journal of Geographical Information Science*, 37, 584–606. <https://doi.org/10.1080/13658816.2022.2131789>.
- Brenning, A., Long, S., & Fieguth, P. (2012). Detecting rock glacier flow structures using Gabor filters and IKONOS imagery. *Remote Sensing of Environment*, 125, 227–237. <https://doi.org/10.1016/j.rse.2012.07.005>
- Clausi, D. A., & Jernigan, M. E. (2000). Designing Gabor filters for optimal texture separability. *Pattern Recognition*, 33(11), 1835–1849. [https://doi.org/10.1016/S0031-3203\(99\)00181-8](https://doi.org/10.1016/S0031-3203(99)00181-8)
- Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(89), 2859–2900.
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, 9(1), 421–436.
- Guo, S., Rösch, P., Popp, J., et al. (2020). Modified PCA and PLS: Towards a better classification in Raman spectroscopy-based biological applications. *Journal of Chemometrics*, 34, e3202. <https://doi.org/10.1002/cem.3202>
- Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31, 82. <https://doi.org/10.1007/s11222-021-10057-z>.
- Landgrebe, D. (2002). Hyperspectral image data analysis. *IEEE Signal Processing Magazine*, 19(1), 17–28. <https://doi.org/10.1109/79.974718>
- Molnar, C. (2022). Interpretable machine learning — a guide for making black box models explainable. <https://christophm.github.io/interpretable-ml-book/>.
- Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, 3(26), 786. <https://doi.org/10.21105/joss.00786>.
- Molnar, C., Casalicchio, G., & Bischl, B., et al. (2020). Interpretable machine learning – a brief history, state-of-the-art and challenges. In I. Koprińska, M. Kamp, & A. Appice (Eds.), *ECML PKDD 2020 Workshops* (pp. 417–431). Cham: Springer International Publishing.
- Molnar, C., König, G., Herbinger, J., et al. (2022). *General pitfalls of model-agnostic interpretation methods for machine learning models*. Cham: Springer International Publishing.
- Molnar, C., König, G., Bischl, B., et al. (2023). Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery*. <https://doi.org/10.1007/s10618-022-00901-9>.
- Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001>
- Murdoch, W. J., Singh, C., Kumbier, K., et al. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>.
- Peña, M. A., & Brenning, A. (2015). Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo valley, Chile. *Remote Sensing of Environment*, 171, 234–244. <https://doi.org/10.1016/j.rse.2015.10.0>