Explainable Boosting Algorithms: Sparse-Group and Interaction-Aware Variable Selection in Complex Data

Fabian Lukas Obster

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von

Fabian Lukas Obster
aus Gräfelfing, Deutschland

München, 26.05.2025

Explainable Boosting Algorithms: Sparse-Group and Interaction-Aware Variable Selection in Complex Data

Fabian Lukas Obster

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von

Fabian Lukas Obster
aus Gräfelfing, Deutschland

München, 26.05.2025

Erstgutachter: Prof. Dr. Christian Heumann

Zweitgutachter: Prof. Dr. Helmut Küchenhoff

Drittgutachter: Prof. Dr. Torsten Hothorn

Tag der mündlichen Prüfung: 18.09.2025

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Nr. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 26.05.2025

_____

Fabian Obster

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Christian Heumann, for the exceptional guidance, continuous support, and valuable insights throughout all stages of this thesis. Your encouragement, patience, and critical feedback helped me grow both as a researcher and as a person. I truly appreciated the freedom you gave me to explore my ideas, while always being available to offer direction when needed.

I would also like to sincerely thank Prof. Dr. Helmut Küchenhoff for serving as the second reviewer of this thesis. Without you and the statistical consulting unit, I probably would not have pursued a career in statistics. I am also grateful to Prof. Dr. Torsten Hothorn for serving as the third reviewer—your work has greatly influenced my thinking throughout this thesis. Furthermore, I want to thank Prof. Dr. Volker Schmid and PD Dr. Fabian Scheipl for being part of the examination committee and for their time and engagement with my work.

Many thanks go to all mentors, collaborators, and colleagues I have had the pleasure of working with during my doctoral studies. In particular, I am grateful to Dr. Paul Pechan, Prof. Dr. Andreas Humpe, and Dr. Hans-Peter Hübner for their insightful comments, shared knowledge, and fruitful joint work.

I also wish to thank Prof. Dr. Rafaela Kraus and founders@unibw for providing a supportive and stimulating research environment. Thank you for your encouragement of my start-up activities and for your driving force toward innovation and autonomy.

A heartfelt thank you goes to my fellow PhD students and office mates for the camaraderie, the countless motivating conversations, and for making the long days more enjoyable. Your encouragement and humor meant a lot during this journey.

Beyond academia, I owe a great deal to my friends for their unwavering support, understanding, and for providing balance and perspective.

Finally, I would like to express my deepest gratitude to my family. Thank you for always believing in me, for your unconditional support, and for standing by me through every challenge. To Marnie, thank you for your love, patience, and motivation throughout this journey. I could not have done this without you

Abstract

High-dimensional datasets often exhibit complex group structures and interactions, posing challenges to traditional variable selection methods. This dissertation addresses these challenges through five interrelated papers, each advancing statistical boosting for complex data.

The first paper introduces methodological extensions for boosting to enable sparse-group variable selection, called sparse-group boosting. The method is inspired by the sparse-group lasso and utilizes component-wise and group-componentwise ridge regression combined through a mixing parameter. Theoretical properties of the group/variable selection process are studied.

Building on this theoretical development, the second paper operationalizes the sparse-group boosting method by introducing the R package 'sgboost', which implements sparse-group boosting and associated model interpretability tools. These include sparse group-variable importance and coefficient paths. Practical guidelines, including R code for using sparse-group boosting, are provided. In addition, a new method for reducing group selection bias for boosting is presented. The aim is to prevent the group size and structure from distorting the selection chances of specific groups.

The third paper illustrates the applicability of sparse-group boosting in economic and environmental data analysis. Here, the importance of groups and individual variables is analyzed to explain their contribution to the financial well-being of farmers in Chile and Tunisia.

The fourth paper deals with the problem of identifying interactions in high-dimensional data while preserving a stable selection of the main effects using a two-step boosting approach. The method uses componentwise boosting, only considering the main effects. After the first model is stopped, the base-learners are changed such that only interaction effects are boosted, starting with the negative gradient of the first model in the first iteration. The method is used to predict farmers' vulnerability to five different climate hazards.

The fifth paper also deals with the problem of stable selection of interaction effects via boosting through a 2-step approach. Instead of fitting a boosted additive model to the observed outcome, the same model is fitted to the predictions of a random forest. The idea is tested in a case study predicting zoo visitors.

Zusammenfassung

Hochdimensionale Datensätze weisen oft komplexe Gruppenstrukturen und Interaktionen auf, was herkömmliche Methoden zur Variablenauswahl vor Herausforderungen stellt. Durch fünf miteinander verbundene Arbeiten, befasst sich diese Dissertation sich mit den jewiligen Herausforderungen, um das statistische Boosting für komplexe Daten weiterentwickeln.

Die erste Arbeit präsentiert methodische Erweiterungen des Boostings zur sparsamen Auswahl von Gruppenvariablen, das sogenannte Sparse-Group Boosting. Die Methode ist vom Sparse-Group Lasso inspiriert und kombiniert komponentenweise sowie gruppenweise ridge regression durch einen Mischparameter. Die theoretischen Eigenschaften des Selektionsprozesses von Gruppen und Variablen werden untersucht.

Im zweiten Beitrag wird das R-Paket „sgboost" vorgestellt, welches das Sparse-Group Boosting und damit verbundene Werkzeuge zur Modellinterpretation implementiert. Dazu gehören Metriken und Visualisierungen zur Gruppenvariablen-Wichtigkeit und Koeffizientenpfade. Zusätzlich werden praktische Leitlinien einschließlich R-Code für die Verwendung von Sparse-Group Boosting bereitgestellt. Zudem wird eine neue Methode zur Reduktion von Gruppen-Selektionsbias für boosting vorgestellt. Dabei soll verhindert werden, dass die Gruppengröße und Struktur die Auswahlchance einzelner Gruppen verzerrt.

Die dritte Arbeit zeigt die Anwendbarkeit von Sparse-Group Boosting bei der Analyse ökonomischer und ökologischer Daten. Dabei wird untersucht, welchen Beitrag Gruppen- und Einzelvariablen zum finanziellen Wohlbefinden von Landwirt:innen in Chile und Tunesien leisten.

Die vierte Arbeit widmet sich dem Problem, Interaktionen in hochdimensionalen Daten zu identifizieren, ohne dabei die stabile Auswahl der Haupteffekte zu verlieren. Hierzu wird ein zweistufiger Boosting-Ansatz entwickelt: In der ersten Phase erfolgt komponentenweises Boosting der Haupteffekte. Nach dem Stopp des ersten Modells werden nur noch Interaktionen berücksichtigt, wobei das Modell mit dem negativen Gradienten aus der ersten Phase startet. Die Methode wird für die Vorhersage der Vulnerabilität von Landwirten gegenüber fünf verschiedenen Klimarisiken verwendet.

Die fünfte Arbeit befasst sich ebenfalls mit dem Problem der stabilen Auswahl von Interaktionseffekten mittels Boosting durch einen zweistufigen Ansatz. Anstatt ein geboostetes additives Modell an die beobachtete Zielgröße anzupassen, wird das gleiche Modell an die Vorhersagen eines Random Forest angepasst. Die Methode wird in einer Fallstudie zur Prognose von Zoobesuchern getestet.

Contents

Part I

Summary and Discussion

Chapter 1

Structured Variable Selection with Boosting: An Overview of Contributions

Group-structured variable selection lies at the heart of many scientific and engineering challenges, from genomics to climate modeling, economics, and medicine. In an age of data abundance, researchers increasingly face the task of identifying a small, meaningful subset of predictors from thousands of potential candidates, often embedded in complex structures such as functional groups or interactions. This problem, known as variable selection, is central not only to statistical modeling but also to ensuring interpretability, reproducibility, and scientific insight. While machine learning approaches often prioritize predictive performance, statistical methods such as regularized regression and boosting remain highly competitive in high-dimensional scenarios due to their ability to enforce sparsity and structure, leading to more stable and interpretable models [George, 2000, Bühlmann and Hothorn, 2007, Heinze et al., 2018].

This thesis contributes to this ongoing effort by enhancing boosting algorithms for structured, interpretable variable selection in high-dimensional data.
Applications of variable selection methods - such as the lasso or stepwise regression - can be found across disciplines: from the social sciences [Hindman, 2015, Haehner et al., 2024, Ofori et al., 2024], physical sciences [Gholami et al., 2023, Geng et al., 2023, Robbins et al., 2024], and life sciences [Fei et al., 2023, Wu and Zeng, 2024, Guo et al., 2024], to technology and engineering [Wang et al., 2023, Zhou et al., 2024, Yan et al., 2024], and even the humanities [Greb et al., 2018, Yaworsky et al., 2020, Anglisano et al., 2022]. Beyond regression, sparsity plays a central role in many branches of statistical learning, including sparse covariance estimation [Bien and Tibshirani, 2011], sparse principal component analysis (PCA) [Zou et al., 2006], and sparse representations in neural networks [Gripon and Berrou, 2011].

Achieving a sparse model in high-dimensional spaces is inherently complex due to the combinatorial explosion of potential predictor subsets, often complicated by multicollinearity and structured covariates [Heinze et al., 2018]. These challenges are exacerbated by the so-called "curse of dimensionality" [Heinze and Dunkler, 2017, Smith, 2018]. Model instability can undermine the reliability of clinical predictions [Efthimiou et al., 2024], while in climate research, unstable variable selection impacts predictions of environmental phenomena such as droughts or wind speeds, with potential policy consequences [Rekha Sankar and Pancha-

pakesan, 2024].

High-dimensional data frequently exhibit a grouped structure, either intrinsically - as in gene pathways or psychometric constructs - or induced by the covariance structure or through encoding categorical variables [Agarwal, 2011, Gogol et al., 2014]. For instance, in genomics, gene expressions are grouped into pathways representing biological processes [Caspi et al., 2012]. Methods like the group lasso [Yuan and Lin, 2006] and group-wise boosting [Kneib et al., 2009] have been proposed to address these structures through group-level selection.

However, modeling interactions introduces an additional layer of complexity. Considering all possible interactions vastly increases the dimensionality of the design matrix and complicates the distinction between main and interaction effects [Zhou et al., 2021]. This dissertation focuses on addressing these challenges within the flexible framework of boosting to advance variable selection in high-dimensional settings.

Boosting-based methods offer unique advantages for structured, high-dimensional data. Unlike traditional regression approaches, boosting fits models sequentially to residuals, incrementally capturing complex structures in the data [Bühlmann and Hothorn, 2007]. This iterative nature allows for adaptive and flexible modeling of grouped predictors and interactions, while maintaining a balance between interpretability and predictive performance. These properties make boosting especially attractive for applications that require sparse and explainable models. Examples include biomedical research, environmental modeling, or social sciences, where inference and interpretability are crucial.

1.1 Research themes and objectives

High-dimensional data are omnipresent across modern scientific disciplines, from genomics and climate science to economics and behavioral research. Discovering meaningful patterns in such data often depends on effective variable selection methods. This dissertation contributes to this challenge through five papers centered on boosting algorithms tailored to structure, interpretability, and robustness. The following points can summarize the common broader themes connecting all papers:

- Variable Selection in High-Dimensional Data: Papers focus on identifying relevant predictors and interactions, addressing sparsity challenges.

- Methodological Advancements: Introduction of novel frameworks like sparse-group boosting and k-step boosting.

- Application-Focused: Real-world examples in climate science, agriculture, and tourism demonstrate practical utility.

- Explainability and Interpretability: Emphasis on explainable AI techniques and interpretable boosting models.

- Interdisciplinary Approach: Combining machine learning with domain-specific challenges.

Building upon the themes, the research objectives pursued by this thesis can be summarized into two branches:

- **Sparse-group variable selection in boosting**:

  - Simultaneous sparsity within and between groups (similar to sparse-group lasso)

- Theoretical bounds for group vs. individual variable selection
- Robust applicability to real-world datasets
- Adjustment mechanisms for unequal group sizes
- Algorithms to satisfy group balancing conditions

- **Boosting Interactions**
  - Stable main effect selection under high-dimensionality
  - Enforcement of strong or weak heredity constraints
  - Reliable detection and estimation of (nonlinear) interaction effects

Table 1.1 contrasts core variable selection strategies by their ability to handle the intertconnected challenges of sparsity, hierarchy, balance, and nonlinearity. The comparison motivates the development of hybrid approaches that unite their respective strengths. Penalized Regression encompasses many methods, depending on multiple variations of penalty terms. Each penalty variation has its preferences in the selection process, and comes with unique advantages and disadvantages. The Group bridge, for example, is very flexible because of its complex loss function and hyperparameters, but is harder to fit because the loss function is non-convex. Classical statistical boosting and stepwise regression, on the other hand, encompass relatively fewer variations, exposing them to more modeling restrictions. By combining the different modeling strategies of penalized regression, stepwise regression, and Machine Learning and integrating them into the boosting framework, these persisting "modeling blind spots" imposed by the complexities of high-dimensional data shall be covered.

Method Class	Predictive Power	Sparse-Group Selection	Hierarchical Interactions	Group Balance	Model Stability	Nonlinear Effects
Penalized Regression	✓	(limited)	(limited)	✗	(limited)	(limited)
Stepwise Selection	✗	✗	✓	✓	✗	(limited)
Classical Boosting	✓	✗	✗	✗	(limited)	✓
k-step + sgboost	✓	✓	✓	✓	✓	✓

Table 1.1: Simplified comparison of major selection approaches regarding key challenges. Group Balance refers to a fair selection across differently sized groups. (limited) indicates, the method addresses the issue under specific assumptions or requires tuning of specific parameters for optimal performance. k-step + sgboost are the two main methods developed in this thesis.
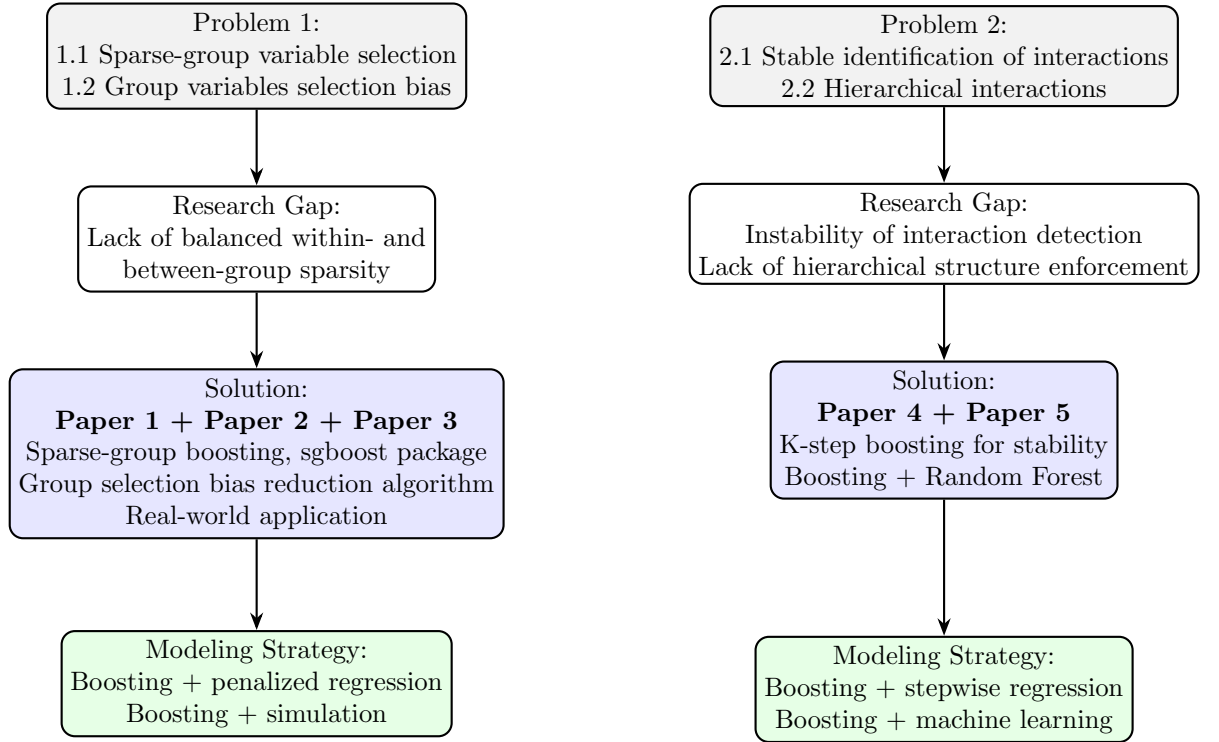
Figure 1.1: Overview of research problems, gaps, solutions, and modeling strategies addressed in this thesis.

Figure 1.1 provides an overview of the key contributions and how they relate to one another, structured according to a problem–solution–strategy framework that is elaborated in the next section.

1.2 Outline

This dissertation consists of two main parts, each addressing distinct but complementary challenges in high-dimensional statistical learning with structured data. The first part focuses on the development and application of *sparse-group boosting*, which enables simultaneous selection of relevant groups and individual predictors. It also introduces an algorithm to correct for group selection bias, thereby ensuring more interpretable and equitable modeling. The second part develops novel strategies to detect and stabilize interaction effects in structured data, especially in the presence of nonlinearity and hierarchical constraints, through the proposal of *k-step boosting* and hybrid modeling strategies. Collectively, the methods enhance interpretability, fairness, and predictive robustness across applied domains.

1.2.1 Thesis Structure and Contributions

The thesis is organized into five papers, each contributing to a distinct methodological or applied objective. Figure 1.1 provides a high-level mapping of papers to research goals. Below, the content and contributions of each paper is summarized in detail.

Paper 1: Sparse-group boosting methodology [Obster and Heumann, 2024] lays the theoretical and algorithmic foundation for sparse-group boosting. By integrating componentwise and group-componentwise base-learners within a unified boosting framework, the method achieves both within-group and between-group sparsity. The formulation includes a sparsity-controlling mixing parameter α, analogous to the sparse-group lasso, allowing fine-tuned trade-offs between individual and group-level variable selection. This paper establishes theoretical selection properties and defines bounds on when and how a group or individual variable is selected during model fitting.

Paper 2: Software implementation via the `sgboost` R package and group balancing algorithm [Obster and Heumann, 2025] presents the R package `sgboost` [Obster, 2024], which operationalizes the sparse-group boosting methodology for broader accessibility. In addition to tools for estimating sparse group-variable importance, visualizing group-aware coefficient paths, and controlling hyperparameters, the package implements the novel group balancing algorithm to mitigate (group-) variable selection bias and accounts for group size imbalances. By enhancing usability, reproducibility, and methodological fairness, this work enables the wider adoption of advanced boosting techniques in interdisciplinary data science.

Paper 3: Interdisciplinary application to climate-agriculture data [Obster et al., 2024a] applies sparse-group boosting to a dataset of 801 farmers from Chile and Tunisia [Pechan et al., 2023a, Pechan et al., 2023b], aiming to model financial well-being under climate stress. The analysis reveals complex relationships between socioeconomic, environmental, and behavioral factors. This paper demonstrates the practical relevance of sparse-group boosting in policy-driven research.

Paper 4: Boosting of pairwise interaction effects via k-step boosting This paper proposes *k-step boosting*, a two-phase procedure for identifying interaction effects while preserving stable main effect estimation. In the first phase, only main effects are selected using standard componentwise boosting. In the second phase, interaction effects are fitted starting from the negative gradient of the stopped first model. The method is applied to environmental vulnerability data and demonstrates reliable interaction discovery without overfitting.

Paper 5: Boosted GAM-RF hybrid for nonlinear interactions [Obster et al., 2023a] addresses the risk of overfitting when modeling nonlinear interactions. It introduces a two-step strategy that fits a boosted generalized additive model (GAM) not to the observed outcome, but to the predictions of a random forest. This "response shifting" enables a sparse and interpretable surrogate model. A case study on zoo visitor forecasting illustrates the potential of this method for smart systems and explainable AI applications.

1.2.2 Summary

Together, these five papers advance state-of-the-art interpretable statistical modeling through boosting. The first part develops methods for structured sparsity in grouped predictors, addresses group-related bias, and demonstrates real-world applicability through interdisciplinary collaboration. The second part introduces novel approaches to interaction detection and model stabilization, combining the strengths of statistical learning and machine learning. These contributions underscore how modern statistical methods can uncover interpretable and policy-relevant insights in complex, high-dimensional data.

1.3   Overview of contributing papers

- Obster, F., & Heumann, C. (2024). "Sparse-group boosting: Unbiased group and variable selection". *The American Statistician*, 1–22. *https://doi.org/10.1080/00031305.2024.2408007*

- Obster, F., & Heumann, C. (2024). "Sparse-Group Boosting with Balanced Selection Frequencies: A Simulation-Based Approach and R Implementation". *ArXiv e-prints arXiv:2405.21037*

- Obster, F., Bohle, H. & Pechan, P.M. (2024). "The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change". *Communications Earth & Environment*, 5, 16. *https://doi.org/10.1038/s43247-023-01128-2*

- Obster, F., Heumann, C., Bohle, H. & Pechan, P.M. (2024). "Using interpretable boosting algorithms for modeling environmental and agricultural data". *Scientific Reports*, 13, 12767. *https://doi.org/10.1038/s41598-023-39918-5*

- Obster, F., Brand, J., Ciolacu M., & Humpe, A. (2023). "Improving Boosted Generalized Additive Models with Random Forests: A Zoo Visitor Case Study for Smart Tourism". *Procedia Computer Science*, 217, 187-197. *https://doi.org/10.1016/j.procs.2022.12.214*

Chapter 2

Background

2.1 Methodological setting

This section introduces the methodological foundations by outlining a progression from linear regression to its extensions: generalized linear models, generalized additive models, and ridge regression. Each progression adds flexibility or regularization to better handle complex, high-dimensional data.

2.1.1 Linear models

Consider a *response variable, outcome variable*, or *dependent variable* $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ with its $n > 0$ realizations $y_1, ..., y_n$ and corresponding design matrix $X \in \mathbb{R}^{n \times p}$, consisting of the $p \geq 1$ *independent variables* or *predictor variables* $X_j = X_{\bullet j} = (X_{1j}, ..., X_{nj})^T \in \mathbb{R}^n$. The n rows of the design matrix are $X_{i\bullet} = (X_{i1}, ..., X_{ip})$. The i-th observation is (x_i, y_i), using $x_i = X_{i\bullet}$. Regression analysis aims to explain Y in terms of X through a functional relationship $\mu_i = \mathbb{E}[Y_i \mid x_i] = f(x_i), \quad f \colon \mathbb{R}^p \to \mathbb{R},$. Classical linear regression assumes a linear relationship $\mu_i = x_i\beta$ and a full rank of $X : \operatorname{rank}(X) = p$, so that $X^T X$ is invertible.

$$Y_i = X_{i1}\beta_1 + ... + X_{ip}\beta_p + \epsilon_i.$$

The *regression parameter* is denoted as $\beta = (\beta_1, ..., \beta_p)^T$. Furthermore, the errors $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ are modeled as independent and identically distributed (iid) Gaussian random variables $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \Rightarrow Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, with I_n being the identity matrix and σ^2 the error variance. This means that X is assumed fixed and the errors ϵ_i are independently and identically distributed, with

$$\operatorname{cov}[\epsilon_i, \epsilon_l] = \begin{cases} \sigma^2 & i = l \\ 0 & i \neq l, \end{cases}$$

so that the ϵ_i are iid $N(0, \sigma^2)$. The *ordinary least squares* (OLS) estimator is the functional

$$\widehat{\beta}_{OLS} = T(Y) = \arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n}(Y_i - x_i^T b)^2 = \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2.$$

On observing $Y = y$, the realized *estimate* is

$$\widehat{\beta}_{OLS}(y) = T(y) = (X^T X)^{-1} X^T y.$$

One can also include (diagonal) weights $W \in \mathbb{R}^{n \times n}$ leading to the weighted-least-squares (WLS) estimator and its estimate

$$\widehat{\beta}_{WLS}(y) = (X^T W X)^{-1} X^T W y,$$

as used in the IRLS Algorithm 1. The Gaussian *negative log-likelihood* $\ell(\beta, \sigma^2; Y)$ for $Y \sim N_n(X\beta, \sigma^2 I_n)$ is

$$\ell(\beta, \sigma^2; Y) = -\log p(Y \mid X, \beta, \sigma^2) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|Y - X\beta\|_2^2.$$

Because the Gaussian likelihood is minimized by the same solution that minimizes squared error, the OLS estimator coincides with the maximum likelihood estimator. Therfore, both satisfy the *normal equation*

$$X^T X \widehat{\beta}(y) = X^T Y \}$$

It is common practice to refer to estimates and estimators as $\widehat{\beta}$, which is also the case throughout this thesis. Predictions of y denoted as $\widehat{y}$ can be derived by replacing the parameters with estimates based on the data and the *residual sum of squares* (RSS) are defined by comparing the realizations of Y with the predictions as RSS $= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \|y - \widehat{y}\|_2^2$.. Using the *hat matrix* H, y can be linked with $\widehat{y}$

$$\widehat{y} = Hy = X(X^T X)^{-1} X^T y.$$

The hat matrix is a projection matrix because it is symmetric $H^T = H$ and idempotent, satisfying $H = H^2$,

$$H^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T,$$

implying that the residuals $\widehat{\epsilon}$ can be seen as a projection of Y onto the orthogonal complement of the column space of X:

$$\widehat{\epsilon} = y - X(X^T X)^{-1} X^T y = (I - H)y.$$

More information, results, and examples of linear regression can be found in textbooks e.g. [Draper and Smith, 1998, Ruppert et al., 2003, Fahrmeir et al., 2013, Wood, 2017].

2.1.2 Additive models

Additive models extend linear regression by replacing each linear term with a smooth function. They overcome the limitation of a linear relationship between covariates and the response, as assumed in the linear regression setting. As in the previous section, let $Y = (Y_1, \ldots, Y_n)^T$ be the random response with realization $y = (y_1, \ldots, y_n)^T$. We model

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

The p covariates $X_j \in \mathbb{R}^n$ are modeled through smooth functions f_j for $j \leq p$, approximated and represented through basis functions allowing for estimation similar to linear models. To ensure identifiability, each f_j is constrained to have zero mean $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ for each j, and a global intercept β_0 is included. In practice, one represents $f_j(X_j) \approx B_j \alpha_j$ via a spline basis $B_j \in \mathbb{R}^{n \times K_j}, \alpha_j \in \mathbb{R}^{K_j}$, such that $(B_j \alpha_j)_i \approx f_j(x_{ij})$. All α_j are estimated by

penalized least squares:

$$\min_{\beta_0 \in \mathbb{R}, \, \alpha_j \in \mathbb{R}^{K_j}} \left\| y - \beta_0 \mathbf{1} - \sum_{j=1}^{p} B_j \alpha_j \right\|_2^2 + \sum_{j=1}^{p} \lambda_j \, \alpha_j^T D_j \, \alpha_j,$$

$$\text{s.t. } \sum_{i=1}^{n} (B_j \alpha_j)_i = 0.$$

Here, $D_j \in \mathbb{R}^{K_j \times K_j}$ is the positive semi-definite penalty matrix with null-space corresponding to low-order polynomials to penalize roughness, and λ_j controls smoothness. Penalization is also covered in Section 2.1.4. A backfitting algorithm cycles through the p terms until convergence [Wood, 2017]. More information on basis representation and the estimation of additive models can also be found in [Wood, 2017].

2.1.3 Generalized linear models

Another limiting assumption in classical linear regression is normality, which can be extended through *generalized linear models* (GLM). GLMs consist of multiple components:

- Linear predictor: Let $x_i \in \mathbb{R}^p$ be the ith row of X. Then the linear predictor is

$$\eta_i = x_i^T \beta, \quad \eta = X\beta \in \mathbb{R}^n.$$

- Random component: For $Y_i \mid x_i \sim \text{EF}(\theta_i, \phi)$, the density is

$$p(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} \, + \, c(y_i, \phi) \right).$$

Here $b(\cdot)$ is the cumulant function, $a(\phi)$ the dispersion parameter, and $c(\cdot, \cdot)$ the base measure. This parametrization is also refered to as canonical form [McCullagh and Nelder, 1993], but also other parametrizations exist, e.g. $f_{\exp}(y_i, \theta_i) = \exp\left[y_i b(\theta_i + c(\theta_i) + d(y_i) \right]$ [Dobson and Barnett, 2008]. By properties of the exponential family, $\mu_i = \mathbb{E}[Y_i \mid x_i] = b'(\theta_i)$ and $\text{Var}(Y_i \mid x_i) = \phi \, b''(\theta_i) = \phi \, v(\mu_i)$.

- Link function: $g(\mu_i) = \eta_i$, invertible to $\mu_i = g^{-1}(\eta_i)$.

- Variance: $\text{Var}(Y_i \mid x_i) = \phi \, v(\mu_i)$, where $v(\cdot)$ is the variance function derived from $b''(\theta)$.

Generalized linear models can be estimated via Fisher scoring [Jennrich and Sampson, 1976], which is a second-order optimization method, or iteratively reweighted least squares (IRLS), which is the practical implementation of Fisher scoring when using the canonical link. IRLS is presented below to illustrate the similarities to statistical boosting in Section 2.2. The goal is to update the linear predictor η iteratively

---

Algorithm 1 iteratively reweighted least squares

1: Initialize $m \leftarrow 0$ and $\eta_m \equiv 0$ or some other starting point

2:
$$\mu^{[m]} = h(\eta^{[m]})$$

3: **while** $m \leq M$ or until convergence **do**

4:     Compute the working response

$$z_i^{[m]} = \eta_i^{[m]} + \frac{y_i - \mu_i^{[m]}}{d\mu/d\eta(\eta_i^{[m]})} = \eta_i^{[m]} + (y_i - \mu_i^{[m]}) \, g'(\mu_i^{[m]}).$$

   The derivative is evaluated at $\widehat{\eta}^{[m]}$.

5:     Regress the covariates on $z^{[m]}$ using the weights

$$W_{ii}^{[m]} = \frac{1}{\mathrm{Var}(Y_i \mid x_i) \, [d\eta/d\mu(\mu_i^{[m]})]^2} = \frac{[g'(\mu_i^{[m]})]^2}{\phi \, v(\mu_i^{[m]})}.$$

6:     Retrieve the estimates $\widehat{\beta}_m$

7:

8:     $m \leftarrow m + 1$

9:     Update

$$\beta^{[m+1]} = \arg\min_b \|W^{[m]^{1/2}} (z^{[m]} - Xb)\|_2^2,$$
$$\eta^{[m+1]} = X \, \beta^{[m+1]},$$
$$\mu^{[m+1]} = g^{-1}(\eta^{[m+1]}).$$

10: **end while**

The algorithm can be stopped either after a fixed number of iterations M, or until the parameter changes are smaller than some tolerance.

2.1.4 Ridge regression

More information on ridge regression can be found in [Wieringen, 2023]. Ridge regression was originally proposed to address multicollinearity, meaning the covariates are strongly linearly dependent [Hoerl and Kennard, 1970a, Hoerl and Kennard, 1970b]. In such cases, $X^T X$ is ill-conditioned or singular [Lesaffre and Marx, 1993]. This problem is especially present in high-dimensional data and has affected Machine Learning in general [Chan et al., 2022].

Definition 2.1.1. For a given $\lambda \geq 0$, design matrix X and outcome y the *ridge regression estimator* $\widehat{\beta}(\lambda)$ is given by
$$\widehat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T y.$$

The set $\{\widehat{\beta}(\lambda) : \lambda \in [0, \infty)\}$ is called the *regularization path* or *solution path* of coefficients and can be plotted as a function of λ.

The ridge estimator can also be derived by minimizing the regularized minimization problem
$$\widehat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Typically, one uses centered X and y or even standardization. If λ is strictly positive, the ridge regression estimator is well-defined also for high-dimensional and multicollinear design matrices. Practically, larger values of λ lead to stronger regularization, meaning smaller coefficients in absolute value.

Similarly to linear regression, the ridge hat matrix $H_\lambda = X(X^T X + \lambda I_p)^{-1} X^T$ can be defined.

$$\widehat{y} = X(X^T X + \lambda I_p)^{-1} X^T y = H_\lambda y$$

An important observation is that H_λ is not a projection matrix as the hat matrix in classical linear regression being H_0. This can be seen using the full singular value decomposition of $X = UDV^T$, where $U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}$ are (column) unitary matrices and $D = \mathrm{diag}(d_1, ..., d_r, 0, ..., 0)$ is a diagonal rectangular matrix containing the singular values of the design matrix $X \in \mathbb{R}^{n \times p}$ of rank $r \leq p$. Then, using $V^T V = VV^T = I_p$ and $U^T U = I_n$:

$$\begin{aligned}
H_\lambda &= X(X^T X + \lambda I_p)^{-1} X^T \\
&= UDV^T (VDU^T UDV^T + \lambda I_p)^{-1} (UDV^T)^T \\
&= UDV^T (VD^2 V^T + \lambda I_p)^{-1} VDU^T \\
&= UDV^T (V(D^2 + \lambda I_p)V^T)^{-1} VDU^T \\
&= UD(D^2 + \lambda I_p)^{-1} DU^T \\
&= U\tilde{D}U^T,
\end{aligned}$$

with $\tilde{D} = \mathrm{diag}[\tilde{d}_1, ..., \tilde{d}_r, 0...0]$ and $\tilde{d}_j = \frac{d_j^2}{(d_j^2 + \lambda)}, j \leq r$. Hence,

$$\begin{aligned}
H_\lambda^2 &= U\tilde{D}U^T U\tilde{D}U^T \\
&= U\tilde{D}^2 U^T.
\end{aligned}$$

Positive λ in the diagonal elements prevent H_λ from being idempotent, meaning $H_\lambda^2 \neq H_\lambda$. Even though the ridge hat matrix is symmetric, it is not idempotent and therefore not a projection matrix. The same holds for $I - H_\lambda$. However, it also makes the ridge estimator well-defined even if the rank of X is not full. λ shrinks the eigenvalues of the hat matrix and makes $(X^T X + \lambda I)$ invertible and improves the condition number through the choice of λ. Using the same decomposition as for the hat matrix, the ridge estimator $\widehat{\beta}_\lambda$ can be rewritten as

$$\begin{aligned}
\widehat{\beta}_\lambda &= (X^T X + \lambda I_p)^{-1} X^T y \\
&= (VD^2 V^T + \lambda I_p)^{-1} VDU^T \\
&= V(D^2 + \lambda I_p)^{-1} DU^T y,
\end{aligned}$$

also shrinking the coefficients through $\frac{d_j^2}{(d_j^2 + \lambda)}, j \leq r$. This formulation also shows the limits depending on λ:

$$\lim_{\lambda \downarrow 0} \frac{d_j^2}{(d_j^2 + \lambda)} = \begin{cases} 0 & d_j = 0 \\ 1 & d_j \neq 0, \end{cases}$$

meaning that the estimator converges to the Moore–Penrose minimum-norm solution, which coincides with the MLE in the classical setting when all $d_j > 0$: $\lim_{\lambda \downarrow 0} \widehat{\beta}_\lambda = \widehat{\beta}_{\mathrm{ML}} = X^+ y$. X^+ is the Moore–Penrose pseudoinverse. Increasing λ shrinks the estimator to zero in the limit: $\lim_{\lambda \to \infty} \widehat{\beta}_\lambda = (0, ..., 0)^T$, as for each $j \leq p$: $\lim_{\lambda \to \infty} \frac{d_j^2}{(d_j^2 + \lambda)} = 0$. While $\frac{d_j^2}{(d_j^2 + \lambda)}$ is strictly decreasing in λ, the behaviour of each component of $\widehat{\beta}_\lambda$ may not necessarily be so.

The *effective degrees of freedom* can be defined as the trace of the hat matrix ((5.16) and (3.50) in elements of stat learning [Hastie et al., 2009]) and can be expressed in terms of the singular values of the design matrix.

$$\mathrm{df}[\lambda] = \mathrm{tr}(H_\lambda) = \sum_{j=1}^{p} \tilde{d}_j.$$

Other definitions for the *degrees of freedom* exist, such as $\mathrm{tr}(2H_\lambda - H_\lambda^2)$ [Hofner et al., 2011] or the Satterthwaite-Welch approximation [Satterthwaite, 1946] leading to

$$\frac{\mathrm{tr}(H^T H)^2}{\mathrm{tr}(H^T H H^T H)},$$

as described in [Adluru et al., 2012]. In linear regression, the hat matrix is a projection, so all definitions coincide.

More information on the bias of ridge regression and two ways to de-bias it can be found in [Bühlmann, 2013, Zhang and Politis, 2022, Zhang and Politis, 2023].

2.2 Statistical boosting

Building on the previous models, the core modeling framework of boosting is introduced in this section. The goal of boosting in general is to find a real-valued function that minimizes a typically differentiable and convex loss function $l(\cdot, \cdot)$. Throughout the thesis, we will consider the negative log-likelihood as a loss function to estimate f^* as

$$f^*(\cdot) = \arg \min_{f:\mathbb{R}^p \to \mathbb{R}} \mathbb{E}_{Y|X}[l(Y, f(X))].$$

As in previous sections, X is assumed fixed, aligning with classical statistical modeling. However, in predictive machine learning contexts, X is often considered random, and one minimizes the expected loss over the joint distribution of (X, Y), leading to:

$$f^*(\cdot) = \arg \min_{f:\mathbb{R}^p \to \mathbb{R}} \mathbb{E}_{(X,Y)}[l(Y, f(X))]$$

Contrary to having one "strong" learner, which has high predictive performance on its own, boosting achieves predictive performance by aggregating "weak" learners, having a low predictive performance by themselves [Freund, 1995]. However, through the aggregation, predictive performance is increased iteratively, as outlined in Algorithm 2.

Algorithm 2 General Functional Gradient Descent Algorithm [Friedman, 2001]

1: Define a space of candidate functions $\mathcal{F}$, which is spanned by base learners h of the form $h : \mathbb{R}^p \rightarrow \mathbb{R}$.
2: Initialize $m \leftarrow 0$ and $\widehat{f}^{[0]} \equiv 0$ or $\widehat{f}^{[0]} \equiv c, \quad c \in \mathbb{R}$ as a constant function.
3: **while** $m \leq M$ **do**
4:   $m \leftarrow m + 1$
5:   Compute the negative gradient $\frac{\partial}{\partial f} l(y, f)$ and evaluate it at $\widehat{f}^{[m-1]}$, yielding pseudo-residuals $u_1, \ldots, u_n$:

$$u_i^{[m]} = \left. \frac{\partial}{\partial f} l(y_i, f) \right|_{f = \widehat{f}^{[m-1]}(x_i)} \qquad \text{for } i = 1, \ldots, n$$

6:   Fit the base-learner h with response $(u_1^{[m]}, \ldots, u_n^{[m]})^T$ to the data, yielding $\widehat{h}^{[m]}$, an approximation of the negative gradient.
7:   Update:

$$\widehat{f}^{[m]} \leftarrow \widehat{f}^{[m-1]} + \nu \cdot \widehat{h}^{[m]}$$

8:   where ν is the learning rate, $\nu \in (0, 1)$.
9: **end while**

The functional derivative in step 5 is computed with respect to f and evaluated point-wise using the data. The algorithm is sequential and flexible. It starts with a simple model and iteratively fits functions to the data, gradually increasing the model's complexity. The base learners, or weak learners, are usually simple functions like small trees or regression models [Schapire, 1990]. The model $\widehat{f}^{[m]}$ continues to improve until it reaches the maximum iteration M. Early stopping helps balance predictive performance and model complexity by halting training before overfitting occurs [Adam J. Grove, 1998, Jiang, 2004, Zhang and Yu, 2005]. This aligns with the principle of sparsity, akin to Occam's razor, by trimming non-predictive components. The optimal stopping point is typically determined by comparing out-of-sample predictive performance across iterations, stopping when no further improvement is observed. Boosting has primarily been used in the context of machine learning as a predictive black-box model using regression/classification trees.

2.2.1   Adaptive Boosting

The first prominent example was Adaptive Boosting (AdaBoost) [Freund and Schapire, 1996], which was first utilized for binary classification and minimized the exponential loss function [Ridgeway, 1999]. The exponential loss function heavily penalizes misclassified observations. This motivates the reweighting mechanism and ensures focus on difficult/misclassified examples.

Algorithm 3 Adaptive Boosting (AdaBoost)

1: Initialize weights for each observation: $w_i^{[1]} = \frac{1}{n}$ for $i = 1, \ldots, n$
2: **for** $m = 1$ to M **do**
3:     Fit a base-learner $h^{[m]} : \mathbb{R}^p \to \{-1, 1\}$ to the data using the weights $\{w_i^{[m]}\}_{i=1}^n$
4:     Compute the weighted classification error:

$$\epsilon^{[m]} = \frac{\sum_{i=1}^n w_i^{[m]} \mathbb{1}_{(y_i \neq h^{[m]}(x_i))}}{\sum_{i=1}^n w_i^{[m]}}$$

5:     Compute the model weight:

$$\alpha^{[m]} = \frac{1}{2} \ln \left(\frac{1 - \epsilon^{[m]}}{\epsilon^{[m]}} \right)$$

6:     Update observation weights:

$$w_i^{[m+1]} = w_i^{[m]} \cdot \exp \left(-\alpha^{[m]} y_i h^{[m]}(x_i) \right), \quad i = 1, \ldots, n$$

7:     Normalize weights so that $\sum_{i=1}^n w_i^{[m+1]} = 1$
8: **end for**
9: Final model:

$$\widehat{f}(x) = \text{sign} \left(\sum_{m=1}^M \alpha^{[m]} h^{[m]}(x) \right)$$

First, the weights are initialized, while the observations have equal weight. At each iteration, a weak learner $h^{[m]}$ is trained using the weighted data, and the misclassification error $\epsilon^{[m]}$ is computed. Misclassification is used as the error. The weight $\alpha^{[m]}$ reflects the base learner's contribution to the ensemble, down-weighting poor classifiers. Misclassified observations are assigned higher weights for the next iteration to focus on previously misclassified observations. The final prediction of the model output aggregates the weak learners via a weighted majority vote, which corresponds to the empirical risk minimization using the exponential loss.

Fast implementations like 'XGboost' [Chen and Guestrin, 2016] of gradient boosting combined with their high predictive performance compared to other Machine Learning algorithms contribute to the popularity of boosting algorithms. Later, the concept of the algorithm was adapted to the field of statistical modeling [Ridgeway, 2000].

2.2.2 Boosting ridge regression

One important boosting algorithm is L^2 Boosting applying boosting algorithms to linear models optimized for squared error loss in high-dimensional settings [Bühlmann and Yu, 2003]. Using component-wise linear least squares base learners, it updates only one variable per iteration, offering computational efficiency. Asymptotic consistency in high dimensions has been shown [Bühlmann, 2006], establishing it as a reliable tool for variable selection and prediction in statistical modeling.

The success of boosting algorithms in the statistical sciences can be attributed to three factors [Mayr et al., 2014]:

- Automated variable selection and model choice through the fitting process [Li and Luan, 2005]

- flexibility regarding the predictor variables e.g random effects, nonparametric effects [Binder et al., 2013]

- Stability for the analysis of high-dimensional data [Mayr and Schmid, 2014]

One important boosting algorithm for this dissertation is boosting ridge regression, which combines boosting and regularization and is therefore described separately.

Let $h : \mathbb{R} \mapsto \mathbb{R}$ be the strictly increasing and invertible response function (inverse link function) of a generalized linear model, where $\mathbb{E}[y|X] = \mu = h(\eta)$. Note that in this case h is the response function and not the whole base-learner $h : \mathbb{R}^p \to \mathbb{R}$ as in the functional gradient descent algorithm. Here, $y \mid x$ follows the simple exponential family in its canonical form, with the linear predictor $\eta = X\beta$. Let $p(y|x, \eta)$ be the conditional density of the exponential family. For L base-learners, define the l-th candidate sets consisting of p_l columns as $V_l = \{(v_l)_1, ..., (v_l)_{p_l}\} \subseteq \{1, ..., p\}$. Contrary to [Tutz and Binder, 2007], the candidate sets do not have to be disjoint. Therefore, the groups can overlap, which will be utilized for the sparse-group boosting. More information on group variables and other methods dealing with groups can be found in Sections 2.3 and 2.4

Algorithm 4 Boosting ridge regression

1: Initialize $m = 0, \widehat{\beta}^{[0]} = \mathbf{0}_p, \widehat{\eta}^{[0]} = X\widehat{\beta}^{[0]}$, and $\widehat{\mu}^{[0]} = h(\widehat{\eta}^{[0]})$
2: **while** $m < M$ **do**
3:     Set $m = m + 1$

4:     **for** For each candidate set V_l, $l \leq L$ **do**
5:         Fit the model:
$$\mu = h(\widehat{\eta}^{[m-1]} + X_{V_l}\beta_{V_l}),$$

by minimizing the penalized negative log-likelihood:

$$\mathcal{L}_l^{[m]}(\beta_{V_l}) = -\sum_{i=1}^{n} \log p\left(y_i \mid x_i, \widehat{\eta}_i^{[m-1]} + (X_{V_l}\beta_{V_l})_i\right) + \lambda\|\beta_{V_l}\|_2^2.$$

with offset $\widehat{\eta}^{[m-1]} = X\widehat{\beta}^{[m-1]}$ derived from the previous iteration. p is the conditional density. This can be done by Fisher scoring or iterative weighted least squares. For the l-th base-learner denote the estimate of β_{V_l} as $\widehat{\widehat{\beta}}_{V_l}$ and the estimate of the negative log-likelihood as $\widehat{\mathcal{L}}_l^{[m]}$.

6:     **end for**
7:     Select the candidate set which evaluates the lowest negative log-likelihood $l^* = \arg\min_{l \leq L} \widehat{\mathcal{L}}_l^{[m]}$.
8:     Update for all $l \leq L$

$$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}_{V_l}^{[m-1]} + \nu\widehat{\widehat{\beta}}_{V_l} & l = l^*, \\ \widehat{\beta}_{V_l}^{[m-1]} & l \neq l^* \end{cases}$$

and:

$$\widehat{\eta}^{[m]} = X\widehat{\beta}^{[m]},$$
$$\widehat{\mu}^{[m]} = h(X\widehat{\beta}^{[m]}).$$

    Here ν can be seen as learning rate with $\nu \in]0, 1[$.
9: **end while**
10: **Output:** Retrieve $\widehat{\beta}^{[M]}$ as the global estimate.

Instead of solving the ridge regression optimization problem directly, the algorithm employs functional gradient descent, as outlined in Algorithms 2 to build the model iteratively. Large coefficients are penalized, preventing overfitting additionally to the learning rate. This is especially useful in high-dimensional settings, where the design matrix of one candidate set has a large number of predictors.

Therefore, boosting ridge regression combines the strengths of regularization through the penalty term $\lambda\beta^T\beta$ and iterative model building, handling multicollinearity and preventing overfitting, especially in high-dimensional settings. Its iterative nature provides flexibility and interpretability, with extensions possible for generalized linear models, as outlined in Algorithm 4. However, careful hyperparameter tuning is required, which can be computationally intensive, especially for large datasets. It may also lack the sparsity of methods like lasso regression, as the coefficients within the candidate set are not shrunken to exactly zero. Contrary to classical ridge regression, boosting ridge regression utilizes candidate sets allowing grouped covariates, which are discussed in the next section.

2.2.3 Boosting and interpretability

While boosting can be used as a black-box model, the intrinsically interpretable variants can balance interpretability and predictive performance well, compared to other Machine Learning algorithms [Obster et al., 2024b].

One can interpret the model coefficients in L^2 boosting as in linear regression. This is also the case using non-linear effects or boosting ridge regression. However, the ability to perform inference after variable selection is limited and needs adjustments [Rügamer and Greven, 2020, Kueck et al., 2023, Rasines and Young, 2023].

Because of the sequential nature of boosting, one can also look at the evolution of the coefficients by looking at the coefficients at each boosting iteration, which is called the coefficient path. This path can be visualized by plotting the iteration versus the values of the regression parameters on the other axis. Connecting the coefficients of each base-learner with a line yields the path which also exists for Lasso regression, but not depending on the iteration but on the regularization parameter λ [Rosset et al., 2004].

The variable importance is a metric to quickly understand what are the main contributors to the model and can be visualized using bar plots [Obster et al., 2024a]. It summarizes how much each variable contributes to the model's overall fit improvement across all boosting iterations. Unlike in traditional variable importance metrics, such as coefficients or p-values, both the frequency and the impact of a variable being selected contribute to the variable's importance. In cases such as sparse-group boosting, which depends on grouped variable selection, the importance can be aggregated across groups or individual variables, depending on the model structure. Let $\Delta \hat{\mathcal{L}}_{l_m}^{[m]} = \hat{\mathcal{L}}^{[m-1]} - \hat{\mathcal{L}}_{l_m}^{[m]}$ be the reduction of log-likelihood in boosting iteration m and predictor $j \leq p$ be the base-learner which was selected in this step. Then the reduction can be attributed to this predictor. Hence, we can compute the relative contribution of this individual variable, call it j to the global model

$$\frac{\sum_{m=1}^{M} \Delta \hat{\mathcal{L}}_{\{l_m : l_m = j\}}^{[m]}}{\sum_{m=1}^{M} \Delta \hat{\mathcal{L}}_{l_m}^{[m]}}.$$

While variable importance indicates influential variables, it should also be noted that the metric is less stable if predictor variables are correlated. In such cases, importance scores can be spread across correlated variables, potentially underestimating the influence of individual predictors. However, in grouped settings, this effect is less pronounced, as variables within the same group typically exhibit stronger correlations with each other than with variables from different groups (i.e., within-group correlation exceeds between-group correlation).

2.3 Grouped variables

Most regression problems, or supervised learning algorithms in general, are of the form $\mathbb{E}[Y|X] = h(X, \beta)$. Typically there is a known design matrix $X \in R^{n \times p}$, an assumed conditional distribution of y, and some function $h : R^p \to \mathbb{R}$, linking the observed input X with the observed output y. The goal is the estimation of the parameter vector $\beta \in \mathbb{R}^p$, such that the function h describes the relationship between X and y in a "good" way. Depending on the distributional assumptions for $Y \mid X$, the choice of response function $h(\cdot)$, and the method used to estimate the parameters β, different statistical models - and even entire subfields of statistics and machine learning - can arise. However, one aspect is shared across almost all disciplines of statistics, deep learning, and even unsupervised learning like image representation: The design matrix X. Design matrices are not created in the same way, nor do they always represent the same type of instances. A lot of research has been attributed to the observations of the design matrix, especially in the form of

detecting and correcting sample selection bias [Berk, 1983], [Cortes et al., 2008], [Yang et al., 2023]. Sample selection bias occurs when the observed data are not representative of the population of interest, potentially resulting in biased parameter estimates and reduced generalizability of the model. The other dimension - the space of independent variables - may also contain structural information reflecting underlying phenomena, such as natural groupings of variables. This secondary information can take many forms, and one way of storing the similarities and differences of variables is a group structure, indicating which variables belong together.

2.3.1 Variations of grouped variables

An intuitive and probably the easiest case of a group is a categorical variable. They are often referred to as one variable, yet in the design matrix, in most cases, multiple columns are used to store the information. Example 2.3.1 shows how one categorical variable and one numerical variable are represented as a group design matrix.

Example 2.3.1. Transformation of categorical and numeric data into a grouped design matrix using reference coding with intercept (gray), with color indicating the group. One level per categorical variable (reference category) is omitted to avoid multicollinearity with the intercept. The reference category in this case is 'a'.

$$
\begin{pmatrix}
a & 1 \\
a & 2 \\
b & 3 \\
b & 4 \\
c & 5
\end{pmatrix}
\rightsquigarrow
\begin{pmatrix}
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 2 \\
1 & 1 & 0 & 3 \\
1 & 1 & 0 & 4 \\
1 & 0 & 1 & 5
\end{pmatrix}.
$$

Similarly, an interaction can be represented as one group in a grouped design matrix as illustrated in Example 2.3.2. While categorical variables are represented in column-orthogonal groups in balanced schemes, such as in effect coding, meaning the inner product of different dummy-coded variables within a group equals zero, interaction terms usually introduce non-orthogonal groups.

Here, a categorical variable with three levels and a numeric variable is transformed into a grouped design matrix using reference (baseline) coding with intercept. The main effects of the categorical variable are encoded using $k-1$ dummies, and interaction terms are defined only for the non-reference levels. This ensures consistency between the number of columns used for main and interaction effects.

Example 2.3.2. Transformation of categorical and numeric data into a grouped design matrix representing main effects and an interaction effect using reference (baseline) coding with intercept (gray). Color indicates group membership. The reference category is 'a'. Violet group: interaction terms between the non-reference levels of the categorical variable and the numeric variable.

$$
\begin{pmatrix}
a & 1 \\
a & 2 \\
b & 3 \\
b & 4 \\
c & 5
\end{pmatrix}
\rightsquigarrow
\begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 2 & 0 & 0 \\
1 & 1 & 0 & 3 & 3 & 0 \\
1 & 1 & 0 & 4 & 4 & 0 \\
1 & 0 & 1 & 5 & 0 & 5
\end{pmatrix}
$$

Notation 2.3.1. A design matrix $X \in \mathbb{R}^{n \times p}$ with G groups, each of the size $p_g, g \in \{1, ..., G\}$ with a collection of index sets $V = (V_g)_{g \leq G}$, as $V_g = \{(v_g)_1, ..., (v_g)_{p_g}\} \subseteq \{1, ..., p\}$ representing the columns belonging to each group is called a grouped design matrix. The

submatrix of group g is denoted as X_{V_g}. In grouped regression settings, the corresponding coefficient parameter vector corresponding to group g will be referred to as $\beta^{(g)}$ such that $X_{V_g}\beta^{(g)}$ is well defined.

Different versions of grouped variables exist. There are *non-overlapping groups*, represented by $(V_g)_{g \leq G}$ being a partition. In the definition, such a constraint is not imposed, yet some methods assume non-overlapping groups, as the optimization may be complicated through the imposed regularization pattern. Also, the concept of partial grouping exists, where only a subset of variables is grouped and the other subset is not grouped. In this case, one could still view this as a grouped dataset, where some groups have a group size of one. While categorical variables lead to orthogonal design matrix groups, meaning $X_{V_g}^T X_{V_g} = I_{p_g}$, groups of numerical variables are often designed or defined in a way such that there is within-group collinearity. This could be the case of constructs in a psychological survey, where the items are designed to have high correlations with each other, indicating high internal consistency [Tavakol and Dennick, 2011]. There are also many other examples of datasets that have natural group structures like gene expression data representing gene pathways [Li et al., 2018] or structural breaks in time series data [Chan et al., 2014]. Also, many nonlinear effects can be represented through a group design matrix [Varah, 1982].

2.4   Methods for grouped variable selection

The most straightforward way of selecting groups of variables is by manually selecting which groups to include in the modeling by the analyst or data collector. This selection happens implicitly in all data analysis, is subjective, and is not of interest to this thesis. Statistical methods performing group selection can be classified into two categories. One deals with knowledge-driven group structures and the other is data-driven, based on the dependence structure observed in the data without known group labels [Huang et al., 2012]. The assumption behind the latter is that "similar" variables are likely to represent the same or similar information and therefore belong to the same group [Zeng, 2009]. These data-driven models can be further divided into two approaches. One approach is to do the group classification and selection in one algorithm, and the second approach consists of two separate algorithms, where the first defines the groups and the second then performs group variable selection in the same way as the knowledge-based group selection.

2.4.1   The sparse-group lasso

The sparse-group lasso is a statistical method designed for high-dimensional data settings, such as when the number of predictors exceeds the number of observations It performs variable selection at both the individual and group levels by combining two penalties: the group lasso penalty ([Yuan and Lin, 2006]), which encourages sparsity at the group level, and the lasso penalty ([Tibshirani, 1996]), which encourages sparsity within groups.

The optimization objective for the sparse-group lasso is:

$$\arg\min_{\beta} \frac{1}{2n} \left\| y - \sum_{g=1}^{G} X_{V_g}\beta^{(g)} \right\|_2^2 + (1-\alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| \beta^{(g)} \right\|_2 + \alpha\lambda \|\beta\|_1 .$$

where $\alpha \in [0,1]$ is the mixing parameter, and $\lambda \geq 0$ controls the overall penalty strength. The idea was first proposed in [Wu and Lange, 2008] and then refined [Simon et al., 2013]. This formulation balances group-wise sparsity (the number of active groups) and within-group sparsity (the number of active variables within a group) by tuning α. There are two

special cases, including the lasso $\alpha = 1$:

$$\arg\min_{\beta} \frac{1}{2n} \left\| y - \sum_{g=1}^{G} X_{V_g} \beta^{(g)} \right\|_2^2 + \lambda \left\| \beta \right\|_1,$$

and group lasso for $\alpha = 0$:

$$\arg\min_{\beta} \frac{1}{2n} \left\| y - \sum_{g=1}^{G} X_{V_g} \beta^{(g)} \right\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| \beta^{(g)} \right\|_2.$$

In the group lasso penalty, there is a group size adjustment $\sqrt{p_g}$. For practical implementation, hyperparameter tuning such as cross-validation or bootstrapping can be used to select optimal values for the hyperparameters α and λ. Tuning both parameters simultaneously through grid search can introduce computational overhead. Therefore, fast implementations such as [Ida et al., 2019, Liang et al., 2023] are useful compared to the original optimization method [Simon et al., 2019].

The sparse-group lasso can also be used in the setting of generalized linear models by replacing the least-squares loss with the empirical average negative log-likelihoods

$$\ell(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \log L(y_i \mid x_i; \beta),$$

where $L(y_i \mid x_i; \beta)$ denotes the likelihood of the response given the predictors and model parameters of observation i. The generalized sparse-group lasso estimator then solves:

$$\arg\min_{\beta} \ell(\beta) + (1 - \alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| \beta^{(g)} \right\|_2 + \alpha\lambda \left\| \beta \right\|_1.$$

The dual-level sparsity makes the sparse-group lasso a flexible tool for high-dimensional grouped data analysis.

This method is foundational for developing the sparse-group boosting framework explored in this dissertation. By balancing sparsity levels, the sparse-group lasso provides a critical theoretical and practical basis for addressing challenges in complex, high-dimensional datasets.

2.4.2 Other (sparse-)group variable selection methods

This dissertation explores how boosting frameworks can be extended to deal with grouped variables, allowing simultaneous selection of relevant groups and important variables within groups. By integrating these concepts directly into the boosting process, the methods developed here offer a flexible and interpretable approach, especially suited for high-dimensional structured complex data. To complement the concepts of sparse group variable selection in boosting, alternative concepts of (sparse-)group variable selection methods are explained in this section. *Group bridge (G-bridge)* [Huang et al., 2009] was one of the first methods enabling dual-level sparsity, extending the bridge estimator to groups [Frank and Friedman, 1993]. Instead of using a convex combination of group and individual variable penalties, the mixing parameter is found in modifying the L_1 norm $\left\| \cdot \right\|_1^{\gamma}$ defined as $\left\| x \right\|_1^{\gamma} = \sum_{i=1}^{p} |x_i|^{\gamma}$ for $x \in \mathbb{R}^n$:

$$\arg\min_{\beta} \ell(\beta) + \lambda \sum_{g=1}^{G} c_g \left\| \beta^{(g)} \right\|_1^{\gamma}.$$

c_g, the penalty weights, allow for scaling group penalties, incorporating prior knowledge about group importance or size. There are also special cases, where $\gamma = 0$ yields ordinary least squares, $\gamma = 1$ the group lasso, and $\gamma = 2$ group ridge regression. Typically, one chooses $\gamma \in {]0, 1[}$, as values greater than one generally do not yield sparse solutions. 0.5 is a common choice, yielding the square root [Zhou and Zhu, 2010, Huang et al., 2009]. Compared to other methods like the SGL, one advantage of the G-bridge is that it has the oracle property for group selection, meaning it can select important groups with probability converging to one with increasing sample size [Huang et al., 2009]. However, the loss function is not convex for $\gamma < 1$. Therefore, the optimization is more challenging, leading to greater computation time and the necessity of a careful choice of γ and initialization in the optimization. Another disadvantage of the G-bridge is that the threshold for group vs individual variable selection cannot be set directly as with α in the sparse-group lasso or sparse-group boosting [Obster and Heumann, 2024].

The *group exponential lasso (GEL)* [Breheny, 2015] also allows this using the exponential lasso penalty with a support of $[0, \infty[$:

$$\arg \min_\beta \ell(\beta) + \sum_{g=1}^{G} \frac{\lambda^2}{\tau} \left(1 - \exp\left[-\frac{\tau \left\| \beta^{(g)} \right\|_1}{\lambda} \right] \right).$$

The hyperparameter $\tau \in [0, 1]$ describes the coupling, meaning parameters within one group are updated together rather than individually. For $\tau \to 0$ the penalty converges to the lasso [Belhechmi et al., 2020]. Small values indicate relatively more individual variable selection, and greater values indicate more group variable selection. More information on the penalty can be found in [Breheny, 2015].

The *composite minimax concave penalty (cMCP)* [Breheny and Huang, 2009] addresses the issue of lasso penalties to not shrink coefficients relative to the size and promises less bias [Zhang, 2007]. This can lead to over-penalization of large coefficients. Therefore, one can adaptively weigh the penalties [Belhechmi et al., 2020], which the minimax concave penalty (MCP) does with a support of $[0, \infty[$:

$$f_{\lambda,\gamma}(\theta) = \begin{cases} \lambda|\theta| - \frac{\theta^2}{2\gamma} & \theta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \theta > \gamma\lambda. \end{cases}$$

The cMCP [Breheny and Huang, 2009] minimizes

$$\arg \min_\beta \ell(\beta) + \sum_{g=1}^{G} f_{\lambda,\gamma_1}\left(\sum_{k=1}^{p_g} f_{\lambda,\gamma_2}(|\beta_{gk}|) \right).$$

Here β_{gk} is the $k-$th parameter in group g. The MCP is applied as outer and inner penalization, working on both the group level and individual variable level. The first penalty term γ_1 is typically set as a function of the group size. The second parameter γ_2 has to be tuned, where a small γ_2 increases the region of constant penalization, while larger values yield penalization close to the LASSO [Buch et al., 2023]. Recommended parameters are $\gamma_2 = 3$, if the covariates are standardized [Zhang, 2007]. The MPC also has the oracle property [Fan and Li, 2001, Fan and Peng, 2004]. Unlike the other methods, the cMCP does not have a bounded parameter such as $\alpha \in [0, 1]$ in the sparse-group lasso, or $\tau \in [0, 1]$ in the GEL, making the model less intuitive to tune.

Bi-level stagewise estimation equation (BiSEE) and *hierarchical stagewise estimation equation (HiSEE)* use stagewise regression [Hocking, 1976] while considering group structures in

a generalized estimation equation framework [Tibshirani, 2015, Vaughan et al., 2017]. Similar to (group)-componentwise boosting, one starts with a zero model, where no covariates are included, and then iteratively adds variables based on some selection criterion, such as statistical test, AIC [Akaike, 1974] or BIC [Schwarz, 1978]. BiSEE uses the sparse-group lasso penalty within each step for the variable selection, and HiSEE uses a hierarchical approach, by first selecting the relevant group with the group lasso penalty and then selecting the relevant individual variable with the lasso penalty [Buch et al., 2023].

2.5 Interaction-aware and nonlinear variable selection

While variable selection in high-dimensional settings has been well-studied, as the previous sections show, interactions and also non-linear effects are less understood [Radchenko and James, 2010]. Approaches to consider the specialty of interaction effects include enforcing ideas like the heredity constraint [Hamada and Wu, 1992, Chipman, 1996], which is often assumed [Chipman et al., 1997], meaning that interaction effects are only allowed if all main effects are selected, called *strong heredity interaction model (SHIM)* [Choi et al., 2010]. The heredity constraint ensures that if an interaction term is selected, both main effects of that interaction must be selected as well. This concept is also referred to as *strong hierarchy* [Nelder, 1977], or *marginality* [McCullagh, 2002, Chen et al., 2020] whereas in *weak hierarchy* or *weak heredity* only one variable of an interaction effect has to be included as main effect [Nelder, 1998]. However, there can also be "non-hierarchical" interactions where only the interaction terms are associated with the outcome [Hallgrímsdóttir and Yuster, 2008] [Obster et al., 2024a]. Two types of interaction selection strategies exist [Hao and Zhang, 2017], one-step and two-step approaches. One-step approaches select main and interaction effects simultaneously while imposing the hierarchical constraint. Two-step methods, such as the two-stage *least angle regression (LARS)* [Efron et al., 2004], first, select main effects and then only consider interactions of the selected main effects, which can also have strong heredity [Yuan et al., 2007].

2.5.1 Strong heredity interaction model

This approach uses the parametrization of the standard two-way interaction model:

$$g(x) = \beta_0 + x_1\beta_1 + ... + x_p\beta_p + \gamma_{1,2}(x_1 x_2)+$$
$$\gamma_{1,3}(x_1 x_3) + ... + \gamma_{p-1,p}(x_{p-1} x_p)$$

together with the lasso-type penalty

$$\arg\min_{\beta,\gamma} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda_\beta(|\beta_1| + |\beta_2| + ... + |\beta_p|) + \lambda_\gamma(|\gamma_{1,2}| + |\gamma_{1,3}| + ... + |\gamma_{p-1,p}|).$$

Note that no quadratic terms of main effects are used here, such as in [Hao and Zhang, 2014]. $x_1, ..., x_p$ are the columns of X and not the observations $x_i, i \leq n$, as sometimes used in previous sections. Through λ_β the main effects are regularized and through λ_γ the interaction effects. If a main effect is zero, its sparsity is inherited by the interaction effect including this main effect. Additional predefined weights for each parameter can be added as in [Breiman, 1995, Zou, 2006, Zou and Zhang, 2009]. As with many other regularized linear regression models, SHIM can also be used for other regularized likelihoods and is based on non-convex optimization, which strongly limits the number of variables (p) to be used [Radchenko and James, 2010]. However, more recent implementations and extensions such as [Chen et al., 2020] improve upon the original method.

2.5.2   Group lasso for interactions

Other methods include the *Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions VANISH* and the *group lasso for interactions (glinter)* which uses a hierarchical group lasso penalty [Lim and Hastie, 2015]. In glinter the parameters of the main effects are regularized with the group lasso penalty, and the interaction effects are regularized through the parameters of the group-design matrix combined with the individual variable design matrix. This leads to overlapping groups in which each main effect coefficient appears both in its own group and in the interaction group. Each variable can have different associated coefficients in main and interaction terms [Lim and Hastie, 2015]. The glinternet penalty can be viewed as a hybrid of the group lasso for main effects and a group ridge penalty over interactions, with overlapping group structure enforcing strong hierarchy. Consider two categorical variables with L_1 and L_2 categories, represented through the group design matrices X_1 and X_2 and the interaction group design matrix denoted as $X_{1:2}$. Then, glinter is given by

$$\arg\min_{\beta,\tilde{\beta}} \left\| y_i - X_1\beta_1 - X_2\beta_2 - \begin{bmatrix} X_1 & X_2 & X_{1:2} \end{bmatrix} \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \beta_{1:2} \end{bmatrix} \right\|_2^2$$

$$+\lambda \left(\|\beta_1\|_2 + \|\beta_2\|_2 + \sqrt{L_2 \left\|\tilde{\beta}_1\right\|_2^2 + L_1 \left\|\tilde{\beta}_2\right\|_2^2 + \|\beta_{1:2}\|_2^2} \right),$$

subject to

$$\sum_{i=1}^{L_1} \beta_1^i = 0, \quad \sum_{i=1}^{L_2} \beta_2^i = 0, \quad \sum_{i=1}^{L_1} \tilde{\beta}_1^i = 0, \quad \sum_{i=1}^{L_2} \tilde{\beta}_2^i = 0,$$

and

$$\sum_{i=1}^{L_1} \beta_{1:2}^{ij} = 0 \text{ for fixed } j, \quad \sum_{j=1}^{L_2} \beta_{1:2}^{ij} = 0 \text{ for fixed } i.$$

The sum constraints, such as on the first part of the main effect $\beta_1 = (\beta_1^1, ..., \beta_1^{L_1})^T$ as denoted in [Lim and Hastie, 2015] are imposed to avoid over-parametrization and no intercept is included. Strong hierarchy is endorsed because of the part $\sqrt{L_2 \left\|\tilde{\beta}_1\right\|_2^2 + L_1 \left\|\tilde{\beta}_2\right\|_2^2 + \|\beta_{1:2}\|_2^2}$ in the penalty term. That is because either all interactions are zero $\tilde{\beta}_1 = \tilde{\beta}_2 = \beta_{1:2} = 0$, or all interactions are nonzero, meaning interactions are always selected together with both main effects. Actually, the name "group lasso for interactions" can be slightly misleading because the method is more of a hybrid between the group lasso and group ridge regression, as the group lasso is used for main effects and group ridge for the interaction terms using the same hyperparameter λ. This means there is shrinkage of the interaction effects without forcing some of them to zero, leading to shrinkage of the entire interaction block rather than selection of individual interactions. This means all interactions between a pair of variables are either jointly included, when the main effects are included, or excluded. The overlap group lasso can be solved using a simple group lasso [Lim and Hastie, 2015] and can be fitted using the R package "glinternet" [Lim and Hastie, 2021], which explains why the name glinter persists beyond historical reasons. Other penalized regression models for interactions amongst others include *lasso for hierarchical interactions (hierNet)* [Bien et al., 2013], *framework for modeling interactions with a convex penalty (FAMILY)* [Haris et al., 2016] and *hierarchical integrative group least absolute shrinkage* [Boss et al., 2021]
Another strategy for finding interactions is based on stepwise regression

2.5.3 Stepwise interaction models

The heredity constraint can also be satisfied using stepwise regression [Hamada and Wu, 1992], using significance testing to assess which variables are selected, which is also criticized [Smith, 2018]. The heredity constraint is enforced by performing stepwise selection twice. First, using stepwise regression, considering only main effects, and then selecting pairwise interaction effects only of previously selected main effects. One can also continue by then looking at interactions of a higher order of the previously selected lower-order interaction effects. This approach can also be applied to "ultrahigh-dimensional" data, using the *iFORT* algorithm [Hao and Zhang, 2014], which is scalable to larger data sets because of its efficiency. Compared to models with complex penalties, leading to computationally expensive optimization algorithms, the stepwise approach is feasible for high-dimensional settings [Wu et al., 2009, Wu et al., 2010]. However, there are shortcomings, especially in high-dimensional settings where small changes in predictor variables can lead to strong changes in the selected variables [James and McCulloch, 1990, Derksen and Keselman, 1992, Austin and Tu, 2004]. This lack of robustness and multiple testing problem limits the practicality of stepwise regression.

Boosting can also be used to fit interactions in a high-dimensional setting, eg. using componentwise boosting. The strong or weak heredity constraint is typically not enforced in statistical boosting but can be enforced through k-step boosting [Obster et al., 2023b].
One way to enforce strong or weak heredity for any given variable selection process is by refitting a nonpenalized regression model based on the predictors selected by the variable selection model and removing all interactions violating the heredity constraint. Such a strategy is employed in [Wolf et al., 2020].

2.6 Advancements introduced by this work beyond boosting

As fairness and interpretability become central concerns in statistical modeling, attention must also be paid to structural biases that arise from modeling assumptions, particularly in grouped variable selection. A central observation underlying this thesis is a persistent bias in many group selection methods - namely, the implicit and often unaddressed influence of group size on selection probability. In applications where group structures are assumed to reflect meaningful units (e.g., gene pathways, categorical variables, interaction terms), it is often desirable that, under a null model with no true signal (all coefficients equal to zero), all groups have equal probability of being selected. This fairness assumption is violated in many popular regularized regression approaches, including group boosting and the sparse-group lasso [Obster, 2024].
Methods such as the sparse-group lasso incorporate group-size adjustments, e.g., through the use of a $\sqrt{p_g}$ scaling in the outer penalty [Simon et al., 2013], but these are typically heuristic. Empirical studies demonstrating the effectiveness of such corrections in ensuring unbiased group selection are rare, and no general theory specifies what "balance" truly entails in finite samples. This motivates the following formalization of a condition that a model should ideally satisfy when aiming for unbiased group selection.

Definition 2.6.1 (Group balancing condition)**.** Let f_θ be a parametrization of a statistical model $(\mathcal{S}, \mathcal{P})$ performing group selection given a group design matrix $X \in \mathbb{R}^{n \times p}$ with the group structure $V = (V_g)_{g \leq G}$, as $V_g = \{(v_g)_1, ..., (v_g)_{p_g}\} \subseteq \{1, ..., p\}$. Denote the indicator set for active groups, indicating which groups are active, meaning a group has at least one nonzero coefficient, as $A = \{0, 1\}^G$. Then the *group-balancing condition* for X is satisfied

under the global null hypothesis $\beta = 0$, if

$$\forall_{j,k \leq G} : P(A_j = 1) = P(A_k = 1).$$

We refer to models satisfying this condition as *group-balanced*. This property is relevant both in variable selection theory and in practice, particularly for model interpretability. While some methods partially address imbalance through penalty scaling, these adjustments are typically not derived from first principles, lack tuning guidelines, and do not account for other sources of selection bias, such as collinearity or differences in group-level scaling and variability.

Models that select all or no groups trivially satisfy the balancing condition, as do certain stepwise approaches using F-statistics. Also, (sparse-) group boosting models, when corrected with the group balancing algorithm, satisfy the group balancing condition [Obster, 2024]. However, in regularized models, especially those with unknown or intractable selection distributions, satisfying this condition is non-trivial. For example, the degrees of freedom in group boosting, used for shrinkage control, can implicitly affect the selection bias and hence the group-balancing behavior [Hofner et al., 2011].

This motivates viewing group balance as a finite-sample analogue to variable selection consistency. Under the global null hypothesis, a consistent model (e.g., one satisfying the oracle property) will asymptotically select no groups, thus satisfying the group balancing condition in the limit. In this sense, variable selection consistency implies group balance asymptotically. A model can be group-balanced in *finite samples* without being variable selection consistent and vice versa. Recognizing this opens new directions for evaluating and improving variable selection methods beyond asymptotic theory.

Although the group balancing condition is defined under the null assumption, it could theoretically be extended to account for non-zero signals by comparing group-wise selection probabilities conditional on similar signal strength. However, such an extension would require redefining the notion of fairness to account for informativeness, rather than strict neutrality, and would raise new theoretical and practical questions.

The methodological developments presented in this thesis - particularly sparse-group boosting, k-step selection strategies, and the group balancing condition - offer clear practical advantages. They also establish conceptual links to broader statistical and machine learning literature.. The following discussion outlines these connections and highlights open avenues for theoretical and empirical research that emerge from this work.

Chapter 3

Discussion and open research

Many ideas introduced in this thesis build on and extend core concepts in regularized regression. For instance, k-step boosting bridges the conceptual gap between boosting and classical stepwise regression by combining the flexibility of iterative fitting with the interpretability of staged inclusion. Similarly, sparse-group boosting establishes a principled connection to the sparse-group lasso, replacing the mixed-norm penalty structure with componentwise base-learner updates, where the degrees of freedom serve as an interpretable analogue to the mixing parameter.

A particularly fruitful direction lies in the treatment of grouped variables and interactions. Prior work such as glinternet [Lim and Hastie, 2021], uses group lasso regularization to enforce strong heredity constraints when selecting interaction terms. The boosting framework developed in this thesis provides an alternative, but similar path: by defining interactions as groups - following, for example, a similar structure as in [Lim and Hastie, 2015] - it becomes possible to implement group-aware interaction modeling using sparse-group boosting. Furthermore, combining k-step and sparse-group boosting enables a hierarchical fitting procedure, in which heredity is imposed across steps while retaining control over group structure and model complexity.

The group balancing condition, introduced in this thesis as a finite-sample fairness criterion for group selection, has a range of potential applications. It is particularly relevant in high-dimensional biological data (e.g., gene expression), categorical variable modeling (e.g., factor encoding in ANOVA), or functional regression settings, where group size and structure can strongly affect selection bias. In the context of orthogonal designs, recent work suggests that the mixing parameter in sparse-group boosting governs the relative selection probabilities of groups versus individual variables [Obster and Heumann, 2024]. This relationship hints at a deeper connection between penalty design and fairness in variable selection.

Importantly, the group balancing condition is not universally desirable. In structured modeling contexts such as interaction heredity, group imbalance may be intentional, for instance, by enforcing a selection probability of one for main effects when their corresponding interactions are selected. Here, k-step boosting offers a promising approach to enforce such hierarchical constraints, while still allowing within-step group balancing to avoid selection artifacts due to group size or correlation structures.

The proposed group balancing algorithm is not limited to boosting but can be generalized to other regularized models. One extension would be to use it for tuning the outer penalty in sparse-group lasso or cMCP. Alternatively, one could define group-wise variance scaling factors, which modify the effective influence of each group prior to model fitting. By iteratively tuning these scaling factors, the balancing algorithm could be applied to ensure approximately equal selection probabilities under the null across a broad range of model

classes.

Bibliography

[Adam J. Grove, 1998] Adam J. Grove, D. S. (1998). Boosting in the Limit: Maximizing the Margin of Learned Ensembles.

[Adluru et al., 2012] Adluru, N., Ennis, C. M., Davidson, R. J., and Alexander, A. L. (2012). Max margin general linear modeling for neuroimage analyses. In *2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 105–110.

[Agarwal, 2011] Agarwal, N. K. (2011). Verifying survey items for construct validity: A two-stage sorting procedure for questionnaire design in information behavior research. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–8.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. Conference Name: IEEE Transactions on Automatic Control.

[Anglisano et al., 2022] Anglisano, A., Casas, L., Queralt, I., and Di Febo, R. (2022). Supervised Machine Learning Algorithms to Predict Provenance of Archaeological Pottery Fragments. *Sustainability*, 14(18):11214. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[Austin and Tu, 2004] Austin, P. C. and Tu, J. V. (2004). Bootstrap Methods for Developing Predictive Models. *The American Statistician*, 58(2):131–137. Publisher: ASA Website _eprint: https://doi.org/10.1198/0003130043277.

[Belhechmi et al., 2020] Belhechmi, S., Bin, R. D., Rotolo, F., and Michiels, S. (2020). Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC bioinformatics*, 21(1):277.

[Berk, 1983] Berk, R. A. (1983). An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review*, 48(3):386–398. Publisher: [American Sociological Association, Sage Publications, Inc.].

[Bien et al., 2013] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141. Publisher: Institute of Mathematical Statistics.

[Bien and Tibshirani, 2011] Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820. Publisher: [Oxford University Press, Biometrika Trust].

[Binder et al., 2013] Binder, H., Benner, A., Bullinger, L., and Schumacher, M. (2013). Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. *Statistics in Medicine*, 32(10):1778–1791.

[Boss et al., 2021] Boss, J., Rix, A., Chen, Y.-H., Narisetty, N. N., Wu, Z., Ferguson, K. K., McElrath, T. F., Meeker, J. D., and Mukherjee, B. (2021). A hierarchical integrative group least absolute shrinkage and selection operator for analyzing environmental mixtures. *Environmetrics*, 32(8):e2698. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2698.

[Breheny, 2015] Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.

[Breheny and Huang, 2009] Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369–380.

[Breiman, 1995] Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1995.10484371.

[Buch et al., 2023] Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., and Wild, P. S. (2023). A systematic review and evaluation of statistical methods for group variable selection. *Statistics in Medicine*, 42(3):331–352. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9620.

[Bühlmann, 2006] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583. Publisher: Institute of Mathematical Statistics.

[Bühlmann, 2013] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

[Bühlmann and Hothorn, 2007] Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4):477–505. Publisher: Institute of Mathematical Statistics.

[Bühlmann and Yu, 2003] Bühlmann, P. and Yu, B. (2003). Boosting With the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, 98(462):324–339. Publisher: ASA Website _eprint: https://doi.org/10.1198/016214503000125.

[Caspi et al., 2012] Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P., and Karp, P. D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753.

[Chan et al., 2022] Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., and Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, 10(8):1283. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[Chan et al., 2014] Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group LASSO for Structural Break Time Series. *Journal of the American Statistical Association*, 109(506):590–599. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2013.866566.

[Chen et al., 2020] Chen, K., Li, W., and Wang, S. (2020). An Easy-to-Implement Hierarchical Standardization for Variable Selection under Strong Heredity Constraint. *Journal of statistical theory and practice*, 14(3):38.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

[Chipman, 1996] Chipman, H. (1996). Bayesian Variable Selection with Related Predictors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(1):17–36. Publisher: [Statistical Society of Canada, Wiley].

[Chipman et al., 1997] Chipman, H., Hamada, M., and Wu, C. F. J. (1997). A Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing. *Technometrics*, 39(4):372–381. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[Choi et al., 2010] Choi, N. H., Li, W., and Zhu, J. (2010). Variable Selection With the Strong Heredity Constraint and Its Oracle Property. *Journal of the American Statistical Association*, 105(489):354–364. Publisher: ASA Website _eprint: https://doi.org/10.1198/jasa.2010.tm08281.

[Cortes et al., 2008] Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample Selection Bias Correction Theory. In Freund, Y., Györfi, L., Turán, G., and Zeugmann, T., editors, *Algorithmic Learning Theory*, pages 38–53, Berlin, Heidelberg. Springer.

[Derksen and Keselman, 1992] Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.1992.tb00992.x.

[Dobson and Barnett, 2008] Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, New York, 3 edition.

[Draper and Smith, 1998] Draper, N. R. and Smith, H. (1998). The General Regression Situation. In *Applied Regression Analysis*, pages 135–148. John Wiley & Sons, Ltd. Section: 5 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118625590.ch5.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–451. Publisher: Institute of Mathematical Statistics.

[Efthimiou et al., 2024] Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., and Salanti, G. (2024). Developing clinical prediction models: a step-by-step guide. *BMJ*, 386:e078276. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting.

[Fahrmeir et al., 2013] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). The Classical Linear Model. In Fahrmeir, L., Kneib, T., Lang, S., and Marx, B., editors, *Regression: Models, Methods and Applications*, pages 73–175. Springer, Berlin, Heidelberg.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[Fan and Peng, 2004] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3). arXiv:math/0406466.

[Fei et al., 2023] Fei, T., Funnell, T., Waters, N. R., Raj, S. S., Sadeghi, K., Dai, A., Miltiadous, O., Shouval, R., Lv, M., Peled, J. U., Ponce, D. M., Perales, M.-A., Gönen, M., and van den Brink, M. R. M. (2023). Enhanced Feature Selection for Microbiome Data using FLORAL: Scalable Log-ratio Lasso Regression. *bioRxiv*, page 2023.05.02.538599.

[Frank and Friedman, 1993] Frank, l. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1993.10485033.

[Freund, 1995] Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, 121(2):256–285.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, pages 148–156, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

[Geng et al., 2023] Geng, J., Yang, C., Li, Y., Lan, L., Zhang, F., Han, J., and Zhou, C. (2023). A bidirectional dictionary LASSO regression method for online water quality detection in wastewater treatment plants. *Chemometrics and Intelligent Laboratory Systems*, 237:104817.

[George, 2000] George, E. I. (2000). The Variable Selection Problem. *Journal of the American Statistical Association*, 95(452):1304–1308. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474336.

[Gholami et al., 2023] Gholami, H., Mohammadifar, A., Fitzsimmons, K. E., Li, Y., and Kaskaoutis, D. G. (2023). Modeling land susceptibility to wind erosion hazards using LASSO regression and graph convolutional networks. *Frontiers in Environmental Science*, 11. Publisher: Frontiers.

[Gogol et al., 2014] Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., and Preckel, F. (2014). "My Questionnaire is Too Long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, 39(3):188–205.

[Greb et al., 2018] Greb, F., Steffens, J., and Schlotz, W. (2018). Understanding music-selection behavior via statistical learning: Using the percentile-Lasso to identify the most important factors. *Music & Science*, 1:2059204318755950. Publisher: SAGE Publications Ltd.

[Gripon and Berrou, 2011] Gripon, V. and Berrou, C. (2011). Sparse Neural Networks With Large Learning Diversity. *IEEE Transactions on Neural Networks*, 22(7):1087–1096. Conference Name: IEEE Transactions on Neural Networks.

[Guo et al., 2024] Guo, Y., Li, L., Zheng, K., Du, J., Nie, J., Wang, Z., and Hao, Z. (2024). Development and validation of a survival prediction model for patients with advanced non-small cell lung cancer based on LASSO regression. *Frontiers in Immunology*, 15. Publisher: Frontiers.

[Haehner et al., 2024] Haehner, P., Bleidorn, W., and Hopwood, C. J. (2024). Examining individual differences in personality trait changes after negative life events. *European Journal of Personality*, 38(2):209–224. Publisher: SAGE Publications Ltd.

[Hallgrímsdóttir and Yuster, 2008] Hallgrímsdóttir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, 9(1):17.

[Hamada and Wu, 1992] Hamada, M. and Wu, C. F. J. (1992). Analysis of Designed Experiments with Complex Aliasing. *Journal of Quality Technology*, 24(3):130–137. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00224065.1992.11979383.

[Hao and Zhang, 2014] Hao, N. and Zhang, H. H. (2014). Interaction Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association*, 109(507):1285–1301. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2014.881741.

[Hao and Zhang, 2017] Hao, N. and Zhang, H. H. (2017). A Note on High-Dimensional Linear Regression With Interactions. *The American Statistician*, 71(4):291–297. Publisher: ASA Website _eprint: https://doi.org/10.1080/00031305.2016.1264311.

[Haris et al., 2016] Haris, A., Witten, D., and Simon, N. (2016). Convex Modeling of Interactions with Strong Heredity. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 25(4):981–1004.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

[Heinze and Dunkler, 2017] Heinze, G. and Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1):6–10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tri.12895.

[Heinze et al., 2018] Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201700067.

[Hindman, 2015] Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):48–62. Publisher: SAGE Publications Inc.

[Hocking, 1976] Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49. Publisher: International Biometric Society.

[Hoerl and Kennard, 1970a] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[Hoerl and Kennard, 1970b] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634.

[Hofner et al., 2011] Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011). A Framework for Unbiased Model Selection Based on Boosting. *Journal of Computational and Graphical Statistics*, 20(4):956–971.

[Huang et al., 2012] Huang, J., Breheny, P., and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, 27(4):481–499. Publisher: Institute of Mathematical Statistics.

[Huang et al., 2009] Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

[Ida et al., 2019] Ida, Y., Fujiwara, Y., and Kashima, H. (2019). Fast Sparse Group Lasso. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[James and McCulloch, 1990] James, F. C. and McCulloch, C. E. (1990). Multivariate Analysis in Ecology and Systematics: Panacea or Pandora's Box? *Annual Review of Ecology, Evolution, and Systematics*, 21(Volume 21, 1990):129–166. Publisher: Annual Reviews.

[Jennrich and Sampson, 1976] Jennrich, R. I. and Sampson, P. F. (1976). Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation. *Technometrics*, 18(1):11–17. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[Jiang, 2004] Jiang, W. (2004). Process consistency for AdaBoost. *The Annals of Statistics*, 32(1):13–29. Publisher: Institute of Mathematical Statistics.

[Kneib et al., 2009] Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable Selection and Model Choice in Geoadditive Regression Models. *Biometrics*, 65(2):626–634. Publisher: John Wiley & Sons, Ltd.

[Kueck et al., 2023] Kueck, J., Luo, Y., Spindler, M., and Wang, Z. (2023). Estimation and inference of treatment effects with L2-boosting in high-dimensional settings. *Journal of Econometrics*, 234(2):714–731.

[Lesaffre and Marx, 1993] Lesaffre, E. and Marx, B. D. (1993). Collinearity in generalized linear regression. *Communications in Statistics - Theory and Methods*, 22(7):1933–1952. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610929308831126.

[Li and Luan, 2005] Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409.

[Li et al., 2018] Li, J., Dong, W., and Meng, D. (2018). Grouped Gene Selection of Cancer via Adaptive Sparse Group Lasso Based on Conditional Mutual Information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):2028–2038. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[Liang et al., 2023] Liang, X., Cohen, A., Heinsfeld, A. S., Pestilli, F., and McDonald, D. J. (2023). sparsegl: An R Package for Estimating Sparse Group Lasso. arXiv:2208.02942 [stat].

[Lim and Hastie, 2015] Lim, M. and Hastie, T. (2015). Learning Interactions via Hierarchical Group-Lasso Regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].

[Lim and Hastie, 2021] Lim, M. and Hastie, T. (2021). glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization.

[Mayr et al., 2014] Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms. From machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6):419–427.

[Mayr and Schmid, 2014] Mayr, A. and Schmid, M. (2014). Boosting the Concordance Index for Survival Data – A Unified Framework To Derive and Evaluate Biomarker Combinations. *PLoS ONE*, 9(1):e84483.

[McCullagh, 2002] McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, 30(5):1225–1310. Publisher: Institute of Mathematical Statistics.

[McCullagh and Nelder, 1993] McCullagh, P. and Nelder, J. A. (1993). Generalized Linear Models (2nd ed.). *Journal of the American Statistical Association*, 88(422):698.

[Nelder, 1977] Nelder, J. A. (1977). A Reformulation of Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 140(1):48–77. Publisher: [Royal Statistical Society, Oxford University Press].

[Nelder, 1998] Nelder, J. A. (1998). The Selection of Terms in Response-Surface Models-How Strong is the Weak-Heredity Principle? *The American Statistician*, 52(4):315–318. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[Obster, 2024] Obster, F. (2024). sgboost: Sparse-Group Boosting.

[Obster et al., 2024a] Obster, F., Bohle, H., and Pechan, P. M. (2024a). The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change. *Communications Earth & Environment*, 5(1):1–12. Number: 1 Publisher: Nature Publishing Group.

[Obster et al., 2023a] Obster, F., Brand, J., Ciolacu, M., and Humpe, A. (2023a). Improving Boosted Generalized Additive Models with Random Forests: A Zoo Visitor Case Study for Smart Tourism. *Procedia Computer Science*, 217:187–197.

[Obster et al., 2024b] Obster, F., Ciolacu, M. I., and Humpe, A. (2024b). Balancing Predictive Performance and Interpretability in Machine Learning: A Scoring System and an Empirical Study in Traffic Prediction. *IEEE Access*, 12:195613–195628. Conference Name: IEEE Access.

[Obster and Heumann, 2024] Obster, F. and Heumann, C. (2024). Sparse-Group Boosting: Unbiased Group and Variable Selection. *The American Statistician*, 0(0):1–14. Publisher: ASA Website _eprint: https://doi.org/10.1080/00031305.2024.2408007.

[Obster and Heumann, 2025] Obster, F. and Heumann, C. (2025). Sparse-Group Boosting with Balanced Selection Frequencies: A Simulation-Based Approach and R Implementation. arXiv:2405.21037 [stat].

[Obster et al., 2023b] Obster, F., Heumann, C., Bohle, H., and Pechan, P. (2023b). Using interpretable boosting algorithms for modeling environmental and agricultural data. *Scientific Reports*, 13(1):12767. Number: 1 Publisher: Nature Publishing Group.

[Ofori et al., 2024] Ofori, I. K., Obeng, C. K., and Asongu, S. A. (2024). What Really Drives Economic Growth in Sub-Saharan Africa? Evidence from the Lasso Regularization and Inferential Techniques. *Journal of the Knowledge Economy*, 15(1):144–179.

[Pechan et al., 2023a] Pechan, P. M., Bohle, H., and Obster, F. (2023a). Reducing vulnerability of fruit orchards to climate change. *Agricultural Systems*, 210:103713.

[Pechan et al., 2023b] Pechan, P. M., Obster, F., Marchioro, L., and Bohle, H. (2023b). Climate change impact on fruit farm operations in Chile and Tunisia. *agriRxiv*, 2023:20230025166. Publisher: CABI.

[Radchenko and James, 2010] Radchenko, P. and James, G. M. (2010). Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553. Publisher: ASA Website _eprint: https://doi.org/10.1198/jasa.2010.tm10130.

[Rasines and Young, 2023] Rasines, D. G. and Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614.

[Rekha Sankar and Panchapakesan, 2024] Rekha Sankar, S. and Panchapakesan, M. (2024). Hybrid feature selection model for accurate wind speed forecasting from numerical weather prediction dataset. *Expert Syst. Appl.*, 248(C).

[Ridgeway, 1999] Ridgeway, G. (1999). The State of Boosting. In *Computing Science and Statistics*, pages 172–181.

[Ridgeway, 2000] Ridgeway, G. (2000). Additive Logistic Regression: A Statistical View of Boosting: Discussion. *The Annals of Statistics*, 28(2):393–400. Publisher: Institute of Mathematical Statistics.

[Robbins et al., 2024] Robbins, C. J., Sadler, J. M., Trolle, D., Nielsen, A., Wagner, N. D., and Scott, J. T. (2024). Does polymixis complicate prediction of high-frequency dissolved oxygen in lakes and reservoirs? *Limnology and Oceanography*. Publisher: John Wiley & Sons, Ltd.

[Rosset et al., 2004] Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a Regularized Path to a Maximum Margin Classifier. *J. Mach. Learn. Res.*

[Ruppert et al., 2003] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). Parametric Regression. In *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, pages 15–56. Cambridge University Press, Cambridge.

[Rügamer and Greven, 2020] Rügamer, D. and Greven, S. (2020). Inference for $$L_2$$-Boosting. *Statistics and Computing*, 30(2):279–289.

[Satterthwaite, 1946] Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2(6):110–114. Publisher: [International Biometric Society, Wiley].

[Schapire, 1990] Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464. Publisher: Institute of Mathematical Statistics.

[Simon et al., 2019] Simon, N., Friedman, J., Hastie, T., and Tibshirani, a. R. (2019). SGL: Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization.

[Simon et al., 2013] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

[Smith, 2018] Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(1):32.

[Tavakol and Dennick, 2011] Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2:53. Publisher: IJME.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Tibshirani, 2015] Tibshirani, R. J. (2015). A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.*, 16(1):2543–2588.

[Tutz and Binder, 2007] Tutz, G. and Binder, H. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, 51(12):6044–6059.

[Varah, 1982] Varah, J. M. (1982). A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46. Publisher: Society for Industrial and Applied Mathematics.

[Vaughan et al., 2017] Vaughan, G., Aseltine, R., Chen, K., and Yan, J. (2017). Stagewise generalized estimating equations with grouped variables. *Biometrics*, 73(4):1332–1342. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12669.

[Wang et al., 2023] Wang, M., Zhou, H., Li, X., Chen, S., Gao, D., and Zhang, Y. (2023). Motor imagery classification method based on relative wavelet packet entropy brain network and improved lasso. *Frontiers in Neuroscience*, 17. Publisher: Frontiers.

[Wieringen, 2023] Wieringen, W. N. v. (2023). Lecture notes on ridge regression. arXiv:1509.09169 [stat].

[Wolf et al., 2020] Wolf, B. J., Jiang, Y., Wilson, S. H., and Oates, J. C. (2020). Variable selection methods for identifying predictor interactions in data with repeatedly measured binary outcomes. *Journal of Clinical and Translational Science*, 5(1):e59.

[Wood, 2017] Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC, New York, 2 edition.

[Wu et al., 2010] Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and Clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–285.

[Wu and Zeng, 2024] Wu, M. and Zeng, S. (2024). Exploring factors influencing farmers' health self-assessment in China based on the LASSO method. *BMC Public Health*, 24(1):333.

[Wu et al., 2009] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

[Wu and Lange, 2008] Wu, T. T. and Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalized Regression. *The Annals of Applied Statistics*, 2(1):224–244. Publisher: Institute of Mathematical Statistics.

[Yan et al., 2024] Yan, H., Ma, J., Chen, W., Yue, H., Li, L., and Liu, W. (2024). Enhanced Error Compensation Method for Robotic Machining System via Two-Step Kinematic Parameters Calibration. In *2024 9th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 159–165.

[Yang et al., 2023] Yang, S., Guo, X., Yu, K., Huang, X., Jiang, T., He, J., and Gu, L. (2023). Causal Feature Selection in the Presence of Sample Selection Bias. *ACM Trans. Intell. Syst. Technol.*, 14(5):78:1–78:18.

[Yaworsky et al., 2020] Yaworsky, P. M., Vernon, K. B., Spangler, J. D., Brewer, S. C., and Codding, B. F. (2020). Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument. *PLOS ONE*, 15(10):e0239424. Publisher: Public Library of Science.

[Yuan et al., 2007] Yuan, M., Joseph, V. R., and Lin, Y. (2007). An Efficient Variable Selection Approach for Analyzing Designed Experiments. *Technometrics*, 49(4):430–439. Publisher: ASA Website _eprint: https://doi.org/10.1198/004017007000000173.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

[Zeng, 2009] Zeng, L. (2009). Group variable selection methods and their applications in analyses of genomic data. *Theses and Dissertations Available from ProQuest*, pages 1–93.

[Zhang, 2007] Zhang, C.-H. (2007). Penalized linear unbiased selection. *Technical Report 2007-003*.

[Zhang and Yu, 2005] Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4). arXiv:math/0508276.

[Zhang and Politis, 2022] Zhang, Y. and Politis, D. N. (2022). Ridge regression revisited: Debiasing, thresholding and bootstrap. *The Annals of Statistics*, 50(3):1401–1422. Publisher: Institute of Mathematical Statistics.

[Zhang and Politis, 2023] Zhang, Y. and Politis, D. N. (2023). Debiased and thresholded ridge regression for linear models with heteroskedastic and correlated errors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):327–355.

[Zhou et al., 2021] Zhou, F., Ren, J., Lu, X., Ma, S., and Wu, C. (2021). Gene–Environment Interaction: A Variable Selection Perspective. In Wong, K.-C., editor, *Epistasis: Methods and Protocols*, pages 191–223. Springer US, New York, NY.

[Zhou and Zhu, 2010] Zhou, N. and Zhu, J. (2010). Group Variable Selection via a Hierarchical Lasso and Its Oracle Property. arXiv:1006.2871 [stat].

[Zhou et al., 2024] Zhou, Y., Xie, J., Zhang, X., Wu, W., and Kwong, S. (2024). Energy-Efficient and Interpretable Multisensor Human Activity Recognition via Deep Fused Lasso Net. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(5):3576–3588. Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence.

[Zou, 2006] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429. Publisher: ASA Website _eprint: https://doi.org/10.1198/016214506000000735.

[Zou et al., 2006] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286. Publisher: ASA Website _eprint: https://doi.org/10.1198/106186006X113430.

[Zou and Zhang, 2009] Zou, H. and Zhang, H. H. (2009). On the Adaptive Elastic-Net with a Diverging Number of Parameters. *The Annals of Statistics*, 37(4):1733–1751. Publisher: Institute of Mathematical Statistics.

Part II

Sparse-group variable selection in the context of booting - theory, implementation, and applications

Chapter 4

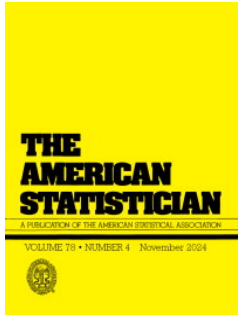# Sparse-group boosting: Unbiased group and variable selection

This chapter introduces methodological extensions for boosting to enable sparse-group variable selection. The method was inspired by the sparse-group lasso and utilizes component-wise and group-component-wise ridge regression combined through a mixing parameter. Theoretical properties of the group/variable selection properties depending on the singular values of the design matrix are studied. Furthermore, the presented method is investigated in simulation studies and real datasets.

Contributing article:

Author contributions:

The manuscript was written by Fabian Obster. Christian Heumann added valuable input and proofread the manuscript.

Sparse-Group Boosting: Unbiased Group and Variable Selection

Fabian Obster & Christian Heumann

View supplementary material ⬀

Published online: 14 Nov 2024.

Submit your article to this journal ⬀

Article views: 185

View related articles ⬀

View Crossmark data ⬀

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Sparse-Group Boosting: Unbiased Group and Variable Selection

Fabian Obster[a] and Christian Heumann[b]

[a]Department of Business Administration, University of the Bundeswehr Munich, Neubiberg, Germany; [b]Department of Statistics, Ludwig Maximilians University Munich, Munich, Germany

ABSTRACT

For grouped covariates, we propose a framework for boosting that allows for sparsity within and between groups. By using component-wise and group-wise gradient ridge boosting simultaneously with adjusted degrees of freedom or penalty parameters, a model with similar properties as the sparse-group lasso can be fitted through boosting. We show that within-group and between-group sparsity can be controlled by a mixing parameter, and discuss similarities and differences to the mixing parameter in the sparse-group lasso. Furthermore, we show under which conditions variable selection on a group or individual variable basis happens and provide selection bounds for the regularization parameters depending solely on the singular values of the design matrix in a boosting iteration of linear Ridge penalized boosting. In special cases, we characterize the selection chance of an individual variable versus a group of variables through a generalized beta prime distribution. With simulations as well as two real datasets from ecological and organizational research data, we show the effectiveness and predictive competitiveness of this novel estimator. The results suggest that in the presence of grouped variables, sparse-group boosting is associated with less biased variable selection and higher predictability compared to component-wise or group-component-wise boosting.

1. Introduction

A key task in empirical science involves the presence of high-dimensional data and the need to perform variable selection, especially if the number of variables is relatively high compared to the number of observations. In biostatistics, this is a common setting, for example, in gene sequencing (Johnstone and Titterington 2009). Two common variable selection strategies are the use of a lasso penalty (Tibshirani 1996) or component-wise boosting (Breiman 1998; Friedman, Hastie, and Tibshirani 2000). Many strategies exist to find a subset of data, including forward selection, backward elimination, or even all-possible subset selection (Chowdhury and Turin 2020), where all possible combinations of variables are considered. Methods differ not only by the selection strategy but also by the metric determining the resulting subset of variables. Some include second-generation p-values (Zuo, Stewart, and Blume 2022) while others use modified loss functions leading to sparsity through shrinkage like the lasso. Often, the variables in the data can be clustered into groups. These could be pathways of genes or items of a construct in a questionnaire, used, for example, in the social sciences or psychology (Agarwal 2011; Gogol et al. 2014). In these cases, it can be of interest to perform variable selection in such a way that this group structure is accounted for. Through the group lasso penalty (Yuan and Lin 2006; Meier, Van De Geer, and Bühlmann 2008) and group-wise boosting (Kneib, Hothorn, and Tutz 2009) this can be achieved. A solution where variable

selection is based on groups, as well as variables, can be of interest if one wants to identify important groups as well as important variables within a group or in addition to a group. This can be achieved by the sparse-group lasso (Simon et al. 2013). Most applications of datasets with sparse-group structures rely on the utilization of the sparse-group lasso penalty in some form, like sparse-group quantile regression (Mendez-Civieta, Aguilera-Morillo, and Lillo 2021), sparse-group neural networks (Yoon and Hwang 2017) and support vector machines (Tang, Adam, and Si 2018). One exception is sparse-group Bayesian regression (Chen et al. 2016). However, to our knowledge, an in-depth analysis of such sparse-group variable selection in the context of boosting has not been conducted. Since boosting is a widely used machine learning algorithm, a boosting variation that can deal with sparse-group structures can offer an alternative modeling approach beyond the sparse-group lasso. Having an alternative to the sparse-group lasso is especially important since many Machine Learning systems use (sparse-group variable) selection methods prior to (Farokhmanesh and Sadeghi 2019) or after (Zhao, Hu, and Wang 2015) fitting another machine learning algorithm, leaving the sparse-group variable selection algorithm a potential bottleneck for predictive power and interpretability. In this article, we show the issues and potential biases, as well as their correction, occurring in the presence of variable selection between and within groups in the context of boosting. In Section 2, we will discuss results from boosting Ridge

CONTACT Fabian Obster ✉ fabian.obster@unibw.de Department of Business Administration, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, 85579 Neubiberg, Germany.

regression, which will be useful for understanding the sparse-group boosting algorithm stated and discussed in Section 3. We also discuss its advantages over alternative definitions. Differences and similarities between the sparse-group lasso and the sparse-group boosting are described with special attention to sparsity. In Section 4, we apply sparse-group boosting to an agricultural dataset and compare its results to component-wise, group component-wise boosting, and sparse-group lasso to showcase its efficacy. The same comparison will be conducted with extensive simulations in Section 5, followed by a discussion and conclusion in Section 6. The code used for the analysis and figure creation and the raw data is available at GitHub (*https:// github.com/FabianObster/sgb*).

1.1. Notation and General Setup

Throughout this article, we consider a (generalized) linear regression framework with outcome $y \in \mathbb{R}^n$ and design matrix X consisting of n observations and p variables. The p variables are grouped in G nonoverlapping groups, where each group $g \in \{1, \ldots, G\}$ consists of p_g variables. We refer to the jth variable as x_j and the sub-matrix containing only the columns belonging to group g is denoted as X_{V_g} where $V_g = \{(v_g)_1, \ldots, (v_g)_{p_g}\} \subseteq \{1, \ldots, p\}$ is the set containing the indices of group g. If groups are not considered, the group index g is omitted. The same notation also applies to the parameter vector $\beta \in \mathbb{R}^p$ and regularization parameters.

1.2. The Sparse-Group Lasso

In a possibly high-dimensional setting, for example $p \gg n$, the sparse-group lasso can fit a model that not only performs variable selection on a variable basis but also on a group basis (Simon et al. 2013). The sparse-group lasso achieves this by combining the group lasso penalty $\sum_{g=1}^{G} \sqrt{p_g} \left\| \beta^{(g)} \right\|_2$ (Yuan and Lin 2006) and the lasso penalty $\|\beta\|_1$ (Tibshirani 1996) with a mixing parameter $\alpha \in [0, 1]$,

$$
\min_\beta \frac{1}{2n} \left\| y - \sum_{g=1}^{G} X_{V_g} \beta^{(g)} \right\|_2^2
$$
$$
+ (1 - \alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \left\| \beta^{(g)} \right\|_2 + \alpha\lambda \|\beta\|_1. \quad (1)
$$

There are two tuning parameters: α and $\lambda \geq 0$. The mixing parameter α determines how much we want to penalize the individual variables (increase α) versus how much we want to penalize groups (decrease α). The special case of $\alpha = 1$ yields the lasso fit, and $\alpha = 0$ yields the group-lasso fit. As in (Simon et al. 2013), we differentiate between the two types of sparsity: "within-group sparsity" refers to the number of nonzero coefficients within each nonzero group, and "group-wise sparsity" refers to the number of groups with at least one nonzero coefficient. Depending on α, both types of sparsity can be balanced. This gives the data scientist the flexibility to include secondary knowledge regarding the two types of sparsity. If α is not known beforehand, it has to be estimated, for example, by using cross-validation on a two-dimensional grid for λ and

α. This has the downside that two hyperparameters have to be tuned.

The sparse-group lasso can also be extended to more general loss functions by replacing the least squares loss with other loss functions. This way, generalized linear models can be fitted by using the negative log-likelihood $l(\beta)$, with group-wise and within-group sparsity

$$
\min_\beta l(\beta) + (1 - \alpha)\lambda \sum_{g=1}^{G} \sqrt{p_l} \left\| \beta^{(g)} \right\|_2 + \alpha\lambda \|\beta\|_1.
$$

1.3. Model-based Boosting

Another way of fitting sparse regression models is through the method of boosting. The fitting strategy is to continuously improve a given model by adding a base-learner to it. Throughout this article, we refer to a base-learner as a subset of columns of the design matrix associated with a real-valued function. To enforce sparsity, each base-learner only considers a subset of the variables at each step (Bühlmann and Hothorn 2007). In the case of component-wise $\mathcal{L}^2$ boosting, each variable will be a base-learner with a linear link function. In the case of a one-dimensional B-Spline, a base-learner is the design matrix representing the basis functions of the B-Spline with a linear link function. The goal of boosting in general is to find a real-valued function that minimizes a typically differentiable and convex loss function $l(\cdot, \cdot)$. Here, we will consider the negative log-likelihood as a loss function to estimate f^* as

$$
f^*(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}[l(y, f)].
$$

General functional gradient descent Algorithm (Friedman 2001)

1. Define base-learners of the structure $h : \mathbb{R}^{n \times p} \to \mathbb{R}$
2. Initialize $m = 0$ and $\widehat{f}^{(0)} \equiv 0$ or $\widehat{f}^{(0)} \equiv \bar{y}$
3. Set $m = m + 1$ and compute the negative gradient $\frac{\partial}{\partial f} l(y, f)$ and evaluate it at $\widehat{f}^{[m-1]}$. Doing this yields the pseudo-residuals $u_1, \ldots, u_n$ with

$$
u_i^{[m]} = \frac{\partial}{\partial f} l(y_i, f)|_{f=\widehat{f}^{[m-1]}},
$$

   for all $i = 1, \ldots, n$
4. Fit the base-learner h with the response $(u_1^{[m]}, \ldots, u_n^{[m]})$ to the data. This yields $\widehat{h}^{[m]}$, which is an approximation of the negative gradient
5. Update

$$
\widehat{f}^{[m]} = \widehat{f}^{[m-1]} + \nu \cdot \widehat{h}^{[m]}
$$

   here ν can be seen as learning rate with $\nu \in]0, 1[$
6. Repeat Steps 2, 3, and 4 until $m = M$

An important case of the general functional gradient descent algorithm is boosting ridge regression which we will use as the framework for defining the sparse-group boosting in a modified form. For L base-learners, denote the lth candidate sets consisting of p_l columns as $V_l = \{(v_l)_1, \ldots, (v_l)_{p_l}\} \subseteq \{1, \ldots, p\}$. We do not require the candidate sets to be disjoint as in Tutz and Binder (2007) leading to possibly overlapping groups, which we will later use for the sparse-group boosting.

Boosting Ridge Regression

1. Initialize $m = 0, \widehat{\beta}^{[0]} = \mathbf{0}_p, \widehat{\mu}^{[0]} = X\widehat{\beta}^{[0]}$,
2. Set $m = m + 1$

   For each candidate set $V_l, l \leq L$, fit Ridge regression to the residuals

$$\widehat{u}^{[m-1]} = y - \widehat{\mu}^{[m-1]},$$

   yielding

$$\widehat{\overline{\beta}}_{V_l}^{[m]} = ((X_{V_l})^T X_{V_l} + \lambda_l I_p)^{-1}(X_{V_l})^T (\widehat{u}^{[m-1]}).$$

3. Select the candidate set which evaluates the lowest residual sum of squares

$$l^* = \underset{l \leq L}{\arg\min}(\widehat{u}^{[m-1]} - X_{V_l}\widehat{\overline{\beta}}_{V_l})^T(\widehat{u}^{[m-1]} - X_{V_l}\widehat{\overline{\beta}}_{V_l}).$$

4. Update for all $l \leq L$

$$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}_{V_l}^{[m-1]} + \nu\widehat{\overline{\beta}}_{V_l} & l = l^*, \\ \widehat{\beta}_{V_l}^{[m-1]} & l \neq l^* \end{cases}$$

   and

$$\widehat{\mu}^{[m]} = X\widehat{\beta}^{[m]}.$$

   Here ν can be seen as learning rate with $\nu \in]0, 1[$.
5. Repeat Steps 3, 4, and 5 until $m = M$ and retrieve $\widehat{\beta}^{[M]}$ as global estimate.

Through early-stopping, or setting M relatively smaller compared to the number of variables in the dataset, and considering the learning rate ν, a sparse overall model can be fitted. The algorithm for boosting generalized linear models can be found in Appendix A.

In the case of component-wise boosting the base procedure is fitted to each variable in the dataset individually by setting the candidate sets $V_l = \{j\}$. For group component-wise boosting, we can set V_l to the indices of the nonoverlapping groups. For the sparse-group boosting, which we will define in Section 3.1, we will combine both candidate sets, leading to overlapping groups. This introduces perfect multicollinearity for each variable in the dataset and a non-identifiable model. To understand variable selection in such cases, and to ensure that both types of base-learners can be selected within the same procedure, we need some results from boosting ridge regression. These include linking the selection criteria, the residual sum of squares/log-likelihood, with the structure of the base-learners, which we will do in the following section.

2. Boosting Ridge Regression and Preliminary Results

The sparse-group boosting as we define it here is based on $\mathcal{L}^2$ regularized regression. Therefore, we first discuss some results for linear Ridge Regression which minimizes $(y - X\beta)^T(y - X\beta)$ with the constraint $\|\beta\|^2 \leq c$, for a positive constant c. Using the Lagrangian form this has an explicit solution $\widehat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^Ty$. We will now discuss results for boosting ridge regression regarding the residual sum of squares (RSS) and the degrees of freedom that will be relevant for the sparse-group boosting. Lemma 1 allows us to characterize the hat matrix in

Ridge Regression using the singular values. The ridge hat matrix will be important to understand the RSS and degrees of freedom, which we need to later define the sparse-group boosting and then understand the variable selection mechanism.

Lemma 1 (Hat matrix in $\mathcal{L}^2$ Ridge Boosting). Consider a design matrix $X \in \mathbb{R}^{n \times p}$ of rank $r \leq p$ with singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n \times p}, V \in \mathbb{R}^{p \times p}$ are unitary matrices and $D = \text{diag}(d_1, \ldots, d_r, 0, \ldots, 0)$ is a diagonal matrix containing the singular values. Let $y \in \mathbb{R}^n$ be the outcome variable and $\widehat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^Ty$ be the Ridge estimate for $\lambda \geq 0$. Then the hat matrix $H^\lambda(m)$ after m boosting steps using a learning rate of $\nu = 1$ is given by

$$H^\lambda(m) = I_n - (I_n - U\tilde{D}U^T)^{m+1} = \sum_{j=1}^r (1 - (1 - \tilde{d}_j)^{m+1})u_ju_j^T,$$

with $\tilde{D} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_r, 0, \ldots, 0) = \text{diag}(\frac{d_1^2}{d_1^2+\lambda}, \ldots, \frac{d_r^2}{d_r^2+\lambda}, 0, \ldots, 0)$.

A derivation can be found in Tutz and Binder (2007). Note that the RSS does not depend on the orthogonal matrix V. Considering the case of only one boosting step, the hat matrix becomes

$$H^\lambda := H^\lambda(0) = U\tilde{D}U^T = \sum_{j=1}^r \frac{d_j^2}{d_j^2 + \lambda}u_ju_j^T.$$

For the residual sum of squares, this means

$$\begin{aligned} RSS(\widehat{\beta}_\lambda) &= (y - X\widehat{\beta}_\lambda)^T(y - X\widehat{\beta}_\lambda) = y^T(I - H^\lambda)^2y \\ &= y^Ty - y^T(2H^\lambda - (H^\lambda)^2)y \\ &= y^Ty - y^T(2U\tilde{D}U^T - U\tilde{D}^2U^T)y \\ &= y^Ty - y^T\Big(\sum_{j=1}^r \Big[2\frac{d_j^2}{d_j^2 + \lambda} - \frac{d_j^4}{(d_j^2 + \lambda)^2}\Big]u_ju_j^T\Big)y. \end{aligned}$$

$$(2)$$

Now, we can introduce the degrees of freedom $\text{df}(\lambda)$, which are either defined as the trace of the hat matrix $\tilde{\text{df}}(\widehat{\beta}_\lambda) = \text{tr}(H^\lambda)$ or as $\text{df}(\lambda) = \text{tr}(2H^\lambda - (H^\lambda)^TH^\lambda)$. As discussed by Hofner et al. (2011) and apparent in (2), df has the advantage over $\tilde{\text{df}}$ of being tailored to the *RSS*. It is worth pointing out that regularizing based on df leads to a greater shrinkage compared to $\tilde{\text{df}}$ for the same base-learner, because for the same base-learner

$$\text{df}(\lambda) = \tilde{\text{df}}(\tilde{\lambda}) \Leftrightarrow \sum_{j=1}^r 2\frac{d_j^2}{d_j^2 + \lambda} - \frac{d_j^4}{(d_j^2 + \lambda)^2}$$

$$= \sum_{j=1}^r \frac{d_j^2}{d_j^2 + \tilde{\lambda}} \Rightarrow \lambda \geq \tilde{\lambda}.$$

This can be seen by contraposition. Assume, that $\lambda < \tilde{\lambda}$. For some j, we substitute $x = \frac{d_j^2}{d_j^2 + \tilde{\lambda}}$ and use $2x - x^2 \geq x$ for $x \in [0, 1]$. Then, $\frac{d_j^2}{d_j^2 + \tilde{\lambda}} \leq 2\frac{d_j^2}{d_j^2 + \tilde{\lambda}} - \frac{d_j^4}{(d_j^2 + \tilde{\lambda})^2} < 2\frac{d_j^2}{d_j^2 + \lambda} - \frac{d_j^4}{(d_j^2 + \lambda)^2}$. The second inequality follows from the assumption $\lambda < \tilde{\lambda}$ and the fact that

$2x - x^2$ is strictly increasing for $x \in [0, 1]$. Since all summands of one sum are strictly smaller, the sums cannot be equal, and the claim follows.

They conclude that the degrees of freedom should be set equal for all base-learner to avoid selection bias, because for two base-learners with design matrices X_γ and X_β and corresponding Ridge estimates $\widehat{\beta}_\lambda$ and $\widehat{\gamma}_\mu$ we have

$$\mathbb{E}[RSS(\widehat{\beta}_\lambda) - RSS(\widehat{\gamma}_\mu)] = 0 \Leftrightarrow df(\lambda) = df(\mu)$$

if the normally distributed outcome random variables $y \sim \mathcal{N}(0, \sigma I)$ are not dependent of the design matrices X_γ and X_β in an ordinary linear regression model. However, we do want to point out that having the same expectation of RSS does not mean, that there is no variable selection bias. The RSS of $\widehat{\beta}_\lambda$ can still have a different variance, shape or different higher order moments than the RSS of $\widehat{\gamma}_\mu$. In the same setting, the RSS is a quadratic form and can be written as $y^T Q_1 y$ and $y^T Q_2 y$ with symmetric and positive definite matrices Q_1 and Q_2 for two base-learners. Such quadratic forms are generally not independent of each other unless $Q_1 Q_2 = 0$ (Craig's theorem). We will later return to the issue of selection bias in the context of sparse-group boosting. We will now look at component-wise Ridge Boosting

Corollary 1. Consider a design matrix vector $x \in \mathbb{R}^{n \times 1}$ of rank one with singular value decomposition $x = ud$, where $u = \frac{x}{\sqrt{x^T x}}$ is the left singular vector and $d = \sqrt{x^T x} \in R^+$ is the singular value. Let $y \in \mathbb{R}^n$ be the outcome variable and $\widehat{\beta}_\lambda = (x^T x + \lambda I)^{-1} x^T y$ be the Ridge estimate for $\lambda \geq 0$. Then,

$$df(\lambda) = df(\widehat{\beta}_\lambda) = \left(2\frac{d^2}{d^2 + \lambda} - \frac{d^4}{(d^2 + \lambda)^2}\right),$$
$$RSS(\widehat{\beta}_\lambda) = y^T y - y^T df(\widehat{\beta}_\lambda) u u^T y,$$

and the Ridge parameter λ in terms of $df(\widehat{\beta}_\lambda)$ is given by

$$\lambda = \frac{-\left(\sqrt{-d^4(df(\lambda) - 1)} + d^2(df(\lambda) - 1)\right)}{df(\lambda)}.$$

This follows directly from Lemma 1 and then solving for λ by finding the zeros and the fact that $df(\lambda)$ is greater than zero. This corollary seems straightforward but has some useful implications. Generally, in model-based boosting, grid search over λ has to be performed to set the degrees of freedom to a fixed value. For individual base-learners one can now compute λ directly with a simple formula without having to try a lot of regularization parameters, which increases speed and accuracy. In addition, one does not have to compute the singular value of the Demmler-Reinsch orthogonalization (App. B.1.Carroll, Ruppert, and Wand (2003)), because the singular values of the design matrix are sufficient in this case. We also see that controlling the variance of a covariate can achieve the same effect as regularization in component-wise boosting. Hence, equalizing the degrees of freedom can be seen as a form of standardization. We will now turn to ridge regression with orthogonal design matrices. In this case, the Ridge estimate is equal to a scaled ordinary least squares estimate $\widehat{\beta}_\lambda = \frac{1}{1+\lambda}\widehat{\beta}_{OLS}$. Orthogonal designs also allow us to characterize the difference between the RSS of Ridge regression and the RSS of the OLS estimate as a Gamma distribution.

Theorem 1 (Distribution of the difference of RSS in orthogonal Ridge regression). Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with orthonormal columns such that $X^T X = I_p$. Let $y \in \mathbb{R}^n$ be the outcome variable, $y = \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ not dependent on the design matrix. Further, assume that the least squares estimate $\widehat{\beta} = X^T y$ exists and $\widehat{\beta}_\lambda$ is the Ridge estimate for some $\lambda > 0$. Define the difference of residual sums of squares as $\Delta = RSS(\widehat{\beta}_\lambda) - RSS(\widehat{\beta}) = (y - \frac{1}{1+\lambda} X\widehat{\beta})^T (y - \frac{1}{1+\lambda} X\widehat{\beta}) - (y - X\widehat{\beta})^T(y - X\widehat{\beta})$. Then if $\left(1 - \frac{df(\lambda)}{p}\right) > 0$, $\frac{\Delta}{\sigma^2}$ follows a gamma distribution with the following shape-scale parameterization

$$\frac{\Delta}{\sigma^2} \sim \Gamma\left(\frac{p}{2}, 2(1 - \frac{2}{p(1 + \lambda)} + \frac{1}{p(1 + \lambda)^2})\right).$$

3. Sparse-Group Boosting

The goal of this article is to adapt the concept of the sparse-group lasso to the boosting such that it is tailored to the boosting framework. One straightforward idea is to use the whole dataset as base-learner equipped with the sparse-group lasso penalty in (1) and update the global model with each boosting step. However, it is not within the scope of this article to fit a sparse-group lasso model through the utilization of boosting. We rather want to build upon the results from boosting Ridge regression within the framework of group-component-wise boosting. With this approach, no Lasso penalty is needed. As proposed by Hofner, Mayr, and Schmid (2014), one can define one base-learner as a group of variables, as well as an individual variable. We define sparse-group boosting using a similar idea as in the sparse-group lasso.

3.1. Definition and Properties of the Sparse-Group Boosting

For the sparse-group boosting we define $p + G$ candidate sets. Of which the first p refer to individual base-learners $l \leq p : V_l = \{l\}$, and the remaining G to the group base-learners of group size $p_l, l > p : V_l = \{(v_l)_1, \ldots, (v_l)_{p_l}\} \subseteq \{1, \ldots, p\}$.

1. Initialize $m = 0, \widehat{\beta}^{[0]} = \mathbf{0}_p, \widehat{\mu}^{[0]} = X\widehat{\beta}^{[0]}$,
2. Set $m = m + 1$
   For each candidate set $V_l, l \leq p + G$, fit Ridge regression to the residuals

$$\widehat{u}^{[m-1]} = y - \widehat{\mu}^{[m-1]},$$

yielding

$$\widehat{\overline{\beta}}_{V_l}^{[m]} = ((X_{V_l})^T X_{V_l} + \lambda_l I_p)^{-1} (X_{V_l})^T (\widehat{u}^{[m-1]}).$$

Regularization parameters λ_l are defined using the Ridge hat matrix $H_{V_l}^\lambda = X_{V_l}((X_{V_l})^T X_{V_l} + \lambda_l I)^{-1}(X_{V_l})^T$ of each base-learner as defined in the previous section, such that

$$\lambda_l = \begin{cases} \lambda_l : df(\lambda_l) = tr(2H_{V_l}^\lambda - (H_{V_l}^\lambda)^2) = \alpha & l \leq p \\ \lambda_l : df(\lambda_l) = tr(2H_{V_l}^\lambda - (H_{V_l}^\lambda)^2) = 1 - \alpha & l > p. \end{cases}$$
$$\tag{3}$$

3. Select the candidate set which evaluates the lowest residual sum of squares

$$l^* = \underset{l \leq L}{\arg\min} (\widehat{u}^{[m-1]} - X_{V_l} \widehat{\overline{\beta}}_{V_l})^T (\widehat{u}^{[m-1]} - X_{V_l} \widehat{\overline{\beta}}_{V_l}).$$

4. Update for all $l \leq p + G$

$$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}^{[m-1]} + \nu \widehat{\overline{\beta}}_{V_{l^*}} & l = l^*, \\ \widehat{\beta}^{[m-1]} & l \neq l^* \end{cases}$$

and

$$\widehat{\mu}^{[m]} = X\widehat{\beta}^{[m]}.$$

Here ν can be seen as learning rate with $\nu \in]0, 1[$.

5. Repeat Steps 2, 3, and 4 until $m = M$ and retrieve $\widehat{\beta}^{[M]}$ as global estimate.

From the derivation in (2) and applying the expectation one can see that in some boosting iteration m,

$$\mathbb{E}[\text{RSS}(\beta_\lambda)] = \mathbb{E}[(\widehat{u}^{[m-1]})^T \widehat{u}^{[m-1]}] - \text{tr}(2H_{V_l}^\lambda - (H_{V_l}^\lambda)^2)$$
$$= \mathbb{E}[(\widehat{u}^{[m-1]})^T \widehat{u}^{[m-1]}] - df(\lambda_l).$$

Therefore, the degrees of freedom and the selection criteria in boosting are directly linked, making the degrees of freedom an excellent choice for changing the chance of a specific base-learner being selected. This means that we can directly change the chance of an individual base-learner being selected over a group base-learner by the choice of α. This is done in step 2 in (3). Since individual base-learners $X_{V_l} \in \mathbb{R}^{n \times 1}, l \leq p$ have degrees of freedom equal to α we will set λ_l such that

$$df(\lambda_l) = \alpha, \tag{4}$$

and for group base-learners $X_{V_l} \in \mathbb{R}^{n \times p_l}, l > p$

$$df(\lambda_l) = (1 - \alpha). \tag{5}$$

$\alpha \in]0, 1[$ is the mixing parameter. Since $df(\lambda) = 0$ means $\lambda \to \infty$, $\alpha = 1$ yields component-wise boosting, and $\alpha = 0$ yields group boosting. This is a similar result as in the sparse-group lasso. With the R package "sgboost" (Obster and Heumann 2024) one can create such a sparse-group boosting formula depending on a given group structure and fit the corresponding model. An alternative definition that looks more like the sparse-group lasso would be to directly regularize the penalty term instead of the degrees of freedom. In this case, the estimate for individual base-learners $l \leq p$ becomes

$$\widehat{\overline{\beta}}_{V_l}^{[m]} = ((X_{V_l})^T X_{V_l} + \alpha \lambda_l I_p)^{-1} (X_{V_l})^T (\widehat{u}^{[m-1]})$$

and for group base-learners $l > p$ with group size p_l

$$\widehat{\overline{\beta}}_{V_l}^{[m]} = ((X_{V_l})^T X_{V_l} + (1 - \alpha)\sqrt{p_l} \lambda_l I_p)^{-1} (X_{V_l})^T (\widehat{u}^{[m-1]})$$

For generalized linear models, the modified loss function $\mathcal{L}$ for individual base-learners $l \leq p$ becomes

$$\mathcal{L}_{V_l}^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l^{[m]}}) + \alpha\lambda(\beta_{V_l})^T \beta_{V_l}, \tag{6}$$

and for group base-learners $l > p$ with group size p_l

$$\mathcal{L}_{V_l}^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l})$$
$$+ (1 - \alpha)\sqrt{p_l}\lambda(\beta_{V_l})^T \beta_{V_l}. \tag{7}$$

The sparse-group boosting algorithms for generalized linear models can be found in Appendix A. However, using this definition does not yield either group boosting or component-wise boosting for $\alpha \in \{0, 1\}$. This is the case because if we compare the loss function of the regularized base-learner with an unregularized base-learner, it is not guaranteed that the unregularized base-learner has a lower loss. We will see this later in Theorem 2. Still, both definitions have their advantages and disadvantages. Using the degrees of freedom allows us to directly control the expectation of the RSS given normal error terms. In this case, α has a natural interpretation and one can set α a priori based on how one wants the RSS of individual base-learners to be compared to the group base-learner. The other advantage of using the degrees of freedom is that one only has to decide on one hyper-parameter, namely α. Based on that choice, all other penalty parameters are already determined. Of course, the optimal stopping parameter and the learning rate have to be set in both definitions. There are also advantages of mixing the penalty term. While more tuning is required, there is a greater flexibility of being able to control two parameters independently of each other which may lead to greater predictive power. Controlling the penalty term directly also has the advantage of seeing which combination of α and λ leads to either only group or individual variables based on the smallest and biggest nonzero singular value of the design matrix and makes the search more efficient, see Theorem 2. In this article, we will mainly focus on the first definition, because of its simplicity, interpretation, and the fact that in boosting the regularization is mainly achieved through the small learning rate and early stopping than finding the optimal regularization parameter as in the sparse-group lasso.

Figure 1 displays a two-variable group example of the evolution of the estimates throughout the sparse-group boosting process for different mixing parameters based on the degrees of freedom. One group base-learner and two individual base-learners were used. All models move toward the least squares estimate indicated with the point at $\alpha = 1$. However, the path they take depends on the mixing parameter. So, in the case of early stopping, different parameter estimates are obtained depending on α. One can see that sometimes only the group base-learner is selected, and in some cases, only the individual base-learners are selected, whenever the path moves only either up or to the right. In some cases, there is an alteration between individual base-learners and the group base-learner. It becomes clear that α has a strong impact on the parameter estimate in the sparse-group boosting even if there is no within-group sparsity or between-group sparsity, as there is only one group and all variables are active in this example. This is especially the case in the early parts of the boosting process. We also see that varying the β-value of one of the variables leaving the other constant does lead to similar jet stretched parameter estimate paths. In this example, the Multi-collinearity of the two predictors seems to only slightly affect the selection process, as
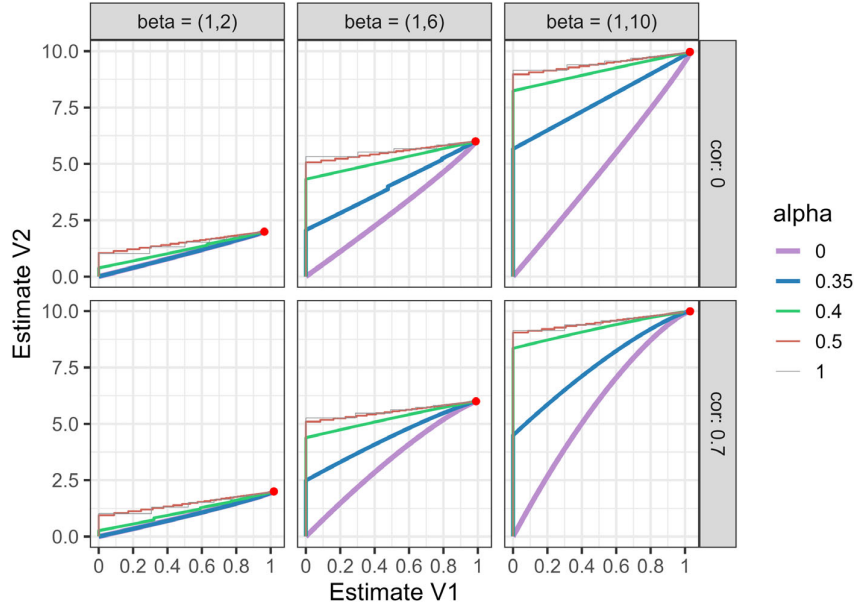
Figure 1. Example: sparse-group boosting parameter estimate paths. Paths throughout 100 boosting iterations for a learning rate of 0.3, depending on the mixing parameter (line-thickness) in an ordinary linear regression model with a normally distributed error term.Independent variables are one group, formed using two variables, where the β-value of the first variable was always set to one, and the β-value of the second variable varied among 1, 6, and 10. The sparse-group boosting model consists of one group base-learner and two individual base-learners. The horizontal axis depicts the estimate for the first variable and the vertical axis the estimate for the second variable within the group. The point at $\alpha = 1$ indicates the least squares estimate. In the case of $\alpha = 0$ group boosting and $\alpha = 1$ component-wise boosting was used.

the upper and lower paths look similar. For the sparse-group boosting to be a useful method for a general design matrix compared to component-wise boosting and group boosting as separate methods, we show that it is more flexible than just using either of the two.

Theorem 2 (Selection intervals of the sparse-group boosting). Consider a design matrix $X \in \mathbb{R}^{n \times p}$ of rank $r \leq p$ with singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$ are unitary matrices and $D = \text{diag}(d_1, \ldots, d_r, 0, \ldots, 0)$ is a diagonal Matrix containing the singular values. Let $y \in \mathbb{R}^n$ be the outcome variable and $\widehat{\beta}_\mu = (X^T X + \mu I)^{-1} X^T y$ be the Ridge estimate for $\mu \geq 0$. For $j \leq p$ let $\widehat{\beta}_{\lambda_j} = (x_j^T x_j + \lambda_j)^{-1} x_j^T y$ be the estimate for the jth individual base-learner, and $\overline{d}_j$ be the singular value of x_j. Denote $\overline{d}^- = \min_{j \leq p} \overline{d}_j^2$ and $\overline{d}^+ = \max_{j \leq p} \overline{d}_j^2$ as well as $d^+ = \max_{j \leq r} d_j^2$ and $d^- = \min_{j \leq r} d_j^2$ accordingly. Then, there are always two mixing parameters $\alpha_1, \alpha_2 \in]0, 1[$ such that

$$(\forall_{j \leq p} : \alpha_1 = \text{df}(\lambda_j) \wedge (1 - \alpha_1) = \text{df}(\mu))$$

$$\Rightarrow \min_{j \leq p} RSS(\lambda_j) \leq RSS(\mu), \text{ and}$$

$$(\forall_{j \leq p} : \alpha_2 = \text{df}(\lambda_j) \wedge (1 - \alpha_2) = \text{df}(\mu))$$

$$\Rightarrow RSS(\mu) \leq \min_{j \leq p} RSS(\lambda_j).$$

Furthermore, the following conditions assure the selection of an individual variable or the whole design matrix

$$\left(\left[\forall_{l \leq k} \frac{(d^- + 2\mu)}{(d^- + \mu)^2} \leq \frac{(\overline{d}_l^2 + 2\lambda_l)}{r(\overline{d}_l^2 + \lambda_l)^2} \right] \vee \left[\text{df}(\mu) \leq \frac{\text{df}(\lambda_l) d^-}{r \overline{d}^+} \right] \right)$$

$$\Rightarrow \min_{j \leq p} RSS(\lambda_j) \leq RSS(\mu), \tag{8}$$

$$\left(\left[\forall_{l \leq k} \frac{(d^+ + 2\mu)}{(d^+ + \mu)^2} \geq \frac{\text{df}(\lambda_l)}{\overline{d}_l^2} \right] \right.$$

$$\left. \vee \left[\forall_{l \leq k} \frac{(d^+ + 2\mu)}{(d^+ + \mu)^2} \geq \frac{(\overline{d}_l^2 + 2\lambda_l)}{(\overline{d}_l^2 + \lambda_l)^2} \right] \right)$$

$$\Rightarrow RSS(\mu) \leq \min_{j \leq p} RSS(\lambda_j). \tag{9}$$

Theorem 2 is useful for both definitions of the sparse-group boosting, as one can either use the bounds by setting $\lambda_j = \alpha\lambda$, $\mu = (1 - \alpha)\lambda$ and further bound (8) and (9) by replacing d_l^2 with either $\overline{d}^+$ or $\overline{d}^-$, respectively. Note that in (9) it is not possible to use the degrees of freedom of the group design matrix as a bound as in (8), because the smallest sum member cannot be expressed in terms of the sum.

We see that there are bounds for the regularization, that always either favor an individual or group base-learner only knowing the largest and smallest nonzero singular values of the group matrix and the column vectors as well as the group size. Especially no assumptions regarding the association between predictors, grouped or individual, and the error term were made. Also, the number of boosting iterations performed and the learning rate play no role in the selection bounds. By restricting the design matrix one can find even tighter bounds in which both individual and group selection can happen for a given α.

Corollary 2. Consider the same setting as in Theorem 2. Setting $X = UD$ meaning $V = I_p$, yields the following bound

$$(\forall_{j \leq p} : \text{df}(\mu) \leq \text{df}(\lambda_j)) \Rightarrow \min_{j \leq p} RSS(\lambda_j) \leq RSS(\mu),$$

and in the case of $X = UdV^T$ with $d \in \mathbb{R}^+$

$$\left(\forall_{j \leq p} : \text{df}(\mu) \geq \frac{1}{p} \text{df}(\lambda_i) \right) \Rightarrow \min_{j \leq p} RSS(\lambda_j) \geq RSS(\mu).$$

Follows directly from the proof of Theorem 2. These bounds have strong implication for setting the mixing parameter α, because set too high or too low one only gets either individual or group base-learners for every design matrix. One example of using Corollary 2 is a categorical variable that is treated as a group base-learners and each dummy variable as an individual base-learner. In this case, $\alpha > 0.5$ should be set. This is also an example where the condition of Hofner et al. (2011) of setting the degrees of freedom equal across all base-learners fails, as in this case only individual base-learners can be selected, which is a clear case of variable selection bias. If the design matrix of one group is a scaled orthogonal matrix then $\alpha \in [\frac{1}{1+p}, \frac{1}{2}]$ should be set. One example of this would be a categorical base-learner with equal number of observations per category. The following Theorem allows us to characterize the pairwise selection probability in one boosting step of two base-learners where one is a sub-matrix of the other for scaled orthogonal matrices.

Theorem 3. Let $X \in \mathbb{R}^{n \times p}$ be a scaled orthogonal design matrix such that $X = dU$ for $d \in \mathbb{R}^+$ and $U^{n \times p}$ orthogonal. Define the sub-matrix $X^{(1)} \in \mathbb{R}^{n \times p_1}, 0 < p_1 < p$. Let $y \in \mathbb{R}^n$ be the outcome variable, $y = \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ not being dependent on the design matrix. Let $\widehat{\beta}_\lambda$ be the Ridge estimate using the design matrix $X^{(1)}$ for some penalty $\lambda > 0$ and $\widehat{\beta}_\mu$ the Ridge using X as design matrix for penalty $\mu > 0$. Let $\text{df}(\lambda)$ and $\text{df}(\mu)$ be the corresponding degrees of freedom. If $\frac{\text{df}(\lambda)}{p_1} \geq \frac{\text{df}(\mu)}{p}$ we can characterize the selection probability based on the residual sum of squares for the two base-learners as

$$P\left(RSS(\widehat{\beta}_\lambda) \geq RSS(\widehat{\beta}_\mu)\right) = F_{\beta'\left(\frac{p_1}{2}, \frac{p-p_1}{2}, 1, \frac{\text{df}(\lambda)p}{\text{df}(\mu)p_1} - 1\right)}(1),$$

where F_β is the distribution function of the beta prime distribution.

Theorem 3 allows us to know the selection probability of a group base-learner versus one individual base-learner for an orthogonal group design matrix. While this is interesting, one would assume that groups are rather defined such that there are dependency structures within a group. However, it is plausible to assume that an individual base-learner from another group is orthogonal to the group design matrix. In that case it is straightforward to see that the selection probability of the individual base-learner versus the group base-learner follows a generalized beta prime distribution $\beta'\left(\frac{1}{2}, \frac{p-1}{2}, 1, \frac{\text{df}(\lambda)p}{\text{df}(\mu)}\right)$, which in the sparse-group boosting becomes $\beta'\left(\frac{1}{2}, \frac{p-1}{2}, 1, \frac{\alpha p}{1-\alpha}\right)$.

3.2. Within-Group and Between-Group Selection

When defining group and individual base-learners at the same time there are two types of variable selection happening at the same time. There is selection between groups: Which group base-learner will be selected? This selection can only be unbiased if there is an equal selection chance for all group base-learners. The sparse-group boosting assures them all to have the same degree of freedom, hence the expected RSS is the same for all base-learners. The same is the case for all individual

base-learners compared to each other as they also have the same degree of freedom. However, in the presence of individual variables, there will always be a challenge at the group level. To illustrate, consider a categorical variable, one containing 3 categories and a linear base-learner, together building an orthogonal system, not being associated with the outcome variable. Then, if one wants the selection chance of any categorical variable versus the linear variable to be equal their degrees of freedom have to be equal, meaning $\alpha = 0.5$ in the sparse-group boosting. But doing this leads to the group base-learner of the categorical variable to be never selected based on Corollary 2. Furthermore, the individual categorical base-learners will have a greater selection chance compared to the linear base-learner, because of the greater group size. To counter this, one could penalize the individual base-learners in bigger groups more than the ones in smaller groups. Doing this could give the categorical variable versus linear variable equal selection chances. But then on the individual variable basis, there would be a bias toward selecting an individual base-learner just because the group it belongs to has a smaller group size. Using the sparse-group boosting, one can decide if one rather want a balance between groups or between individual base-learners. However one should keep in mind, that a perfect balance in the case of unequal group sizes may not be easy to achieve. On which level one wants equal selection chances depends on the research question and interpretation of the data. Generally, varying group sizes impose challenges. If a group contains only one variable then the group and individual base-learner are the same and therefore the greater value of either α or $1 - \alpha$ is used for both, preferring either the group base-learner or the individual base-learner. From Theorem 2 we see that the group size affects the selection bounds. This can also be seen in Figure 2 which compares the selection frequency of the group base-learner in the first boosting iteration for group sizes two and three and different dependency structures. Staying again with the example of two categorical variables of equal number of observations within each category from Corollary 2 we know that the selection interval of the smaller categorical variable is a subset of the selection interval of the bigger categorical variable. This means that for a small enough α one can either have the group or individual variable selected depending on the dataset and for the smaller group only group base-learners regardless of the dataset. One could use a group adjustment by the group size to align the lower bound but then the upper bound would also be affected imposing the same issue.

3.3. Extensions

The here presented results like the selection bounds were mainly focused on $\mathcal{L}^2$ boosting. However, the definition of the sparse-group boosting can be applied to many cases of grouped datasets. Whenever the degrees of freedom can be computed and modified by a regularization parameter, one can use sgb df as defined in (4), and whenever Ridge regularized regression can be used sgb lambda as defined in (6) can be used. This includes generalized linear and additive models but also regularized regression trees. We also want to highlight that semi-grouped datasets can be analyzed using the sparse-group boost-
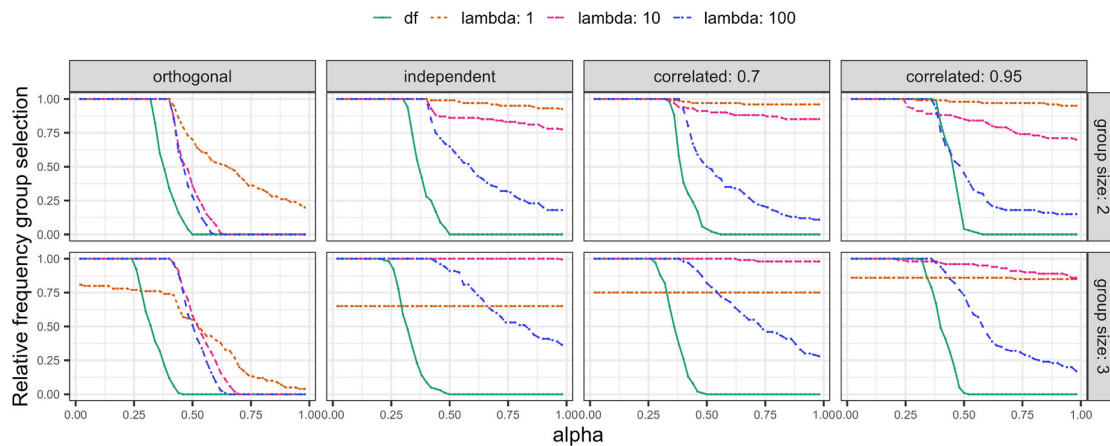
Figure 2. Group selection probability versus individual variables depending on the mixing parameter (alpha) in the sparse-group boosting. The two variables within the group are either orthogonal, independent of each other or correlated.

ing by including additional base-learners which are not split up into groups and individual variables. Examples of this could be random effects, treatment effects, or smoothing splines. The degrees of freedom or regularization parameters of these variables could then either be set to α, $1-\alpha$ or even zero yielding an unregularized base-learner. An extension to generalized additive models for location scale and shape (gamlss) Stasinopoulos and Rigby (2008) and their boosting variant 'gamboostLSS in Hofner et al. (2011) is also possible. This would allow the data analyst to also apply sparse-group penalization to the linear predictor for other moments of the conditional distribution of the outcome given covariates. We believe that (group)—sparsity is especially important for higher-order moments due to the overall model complexity and the number of variables to be interpreted.

4. Empirical Data: Agricultural Dataset

The analysis was performed with R (R Core Team 2022). For visualizations, the R package ggplot was used (Wickham 2016). All computations were conducted on a 3600 MHz Windows machine. Biomedical data are prominent examples of where sparse-group selection can be used. To show the variety of possible applications we analyze an agricultural dataset. Climate change impacts on the agricultural sector are well documented. The type and level of impacts are crop and region-specific. Not surprisingly, exposure to climate change makes many orchard farming communities in Chile and Tunisia vulnerable to climate change impacts. Many susceptibility-related factors may affect farm vulnerability to climatic impacts. Several adaptation resources (measures/tools) are available to directly reduce the impact on farm operations or reduce the number or sensitivity of susceptibility-related factors. The final objective is to increase the resilience of the farming communities.

The dataset (Pechan, Bohle, and Obster 2023) contains 12 binary outcome variables of interest that are related to adaptive measures against climate change impacts. The 147 independent variables can be grouped into 23 groups depending on the construct the variable belongs to. Two group examples are social variables as well as past adaptive measures. 801 farmers have

been included in the study. Further analysis of group variable selection for other outcomes in the dataset can be found in Obster, Bohle, and Pechan (2024) and Obster et al. (2023).

We again use 11 equally spaced α values from zero to one as mixing parameters for sgl, sgb df and sgb lamda. The dataset was split into two, each only containing farmers of one Country. We randomly split 70% of the data into the training data and 30% into the test data. The remaining test data was used for the model evaluation. As in the previous section we used the area under the curve (auc) as an evaluation criterion, since all outcome variables are binary. For the training data, we used a 3-fold cross-validation to estimate the optimal stopping parameter for the boosting models and the optimal λ value for the sgl models. We used 6 values of λ for the sgl and 3000 boosting iterations with a learning rate $\nu = 0.05$ for both sparse-group boosting models. In sgb lambda we used $\lambda = 100$.

Referring to Figure 3 which averages across all 12 outcome variables for each dataset and α value, it becomes apparent, that overall the models performed similarly regarding predictability. In Chile and Tunisia, sgb df had the highest AUC, obtained at $\alpha = 0.4$ in Chile and $\alpha = 0.2$ in Tunisia. At the same α values also the sgl achieved its highest auc. Stronger differences between the models can be found regarding sparsity. For smaller α values all models selected more variables on average, where sgb df and sgb lambda selected more variables for smaller α values than sgl. Component-wise boosting and the lasso yielded roughly the same number of selected variables. The number of partially selected groups, meaning at least one variable within a group gets selected, does on average increase with α. Whereas the number of fully selected groups decreases on average with increasing α values for sgb df and sgl. This effect is less pronounced for sgb lambda. This opens up an interesting discussion on what "between group sparsity" means. If one defines it through the number of fully selected groups, meaning all variables within the group have to be selected, then compared to defining it through the number of partially selected groups one gets an opposing effect of α. The average percentage of selected variables within groups decreases on average with α, corresponding to increasing "within-group-sparsity."
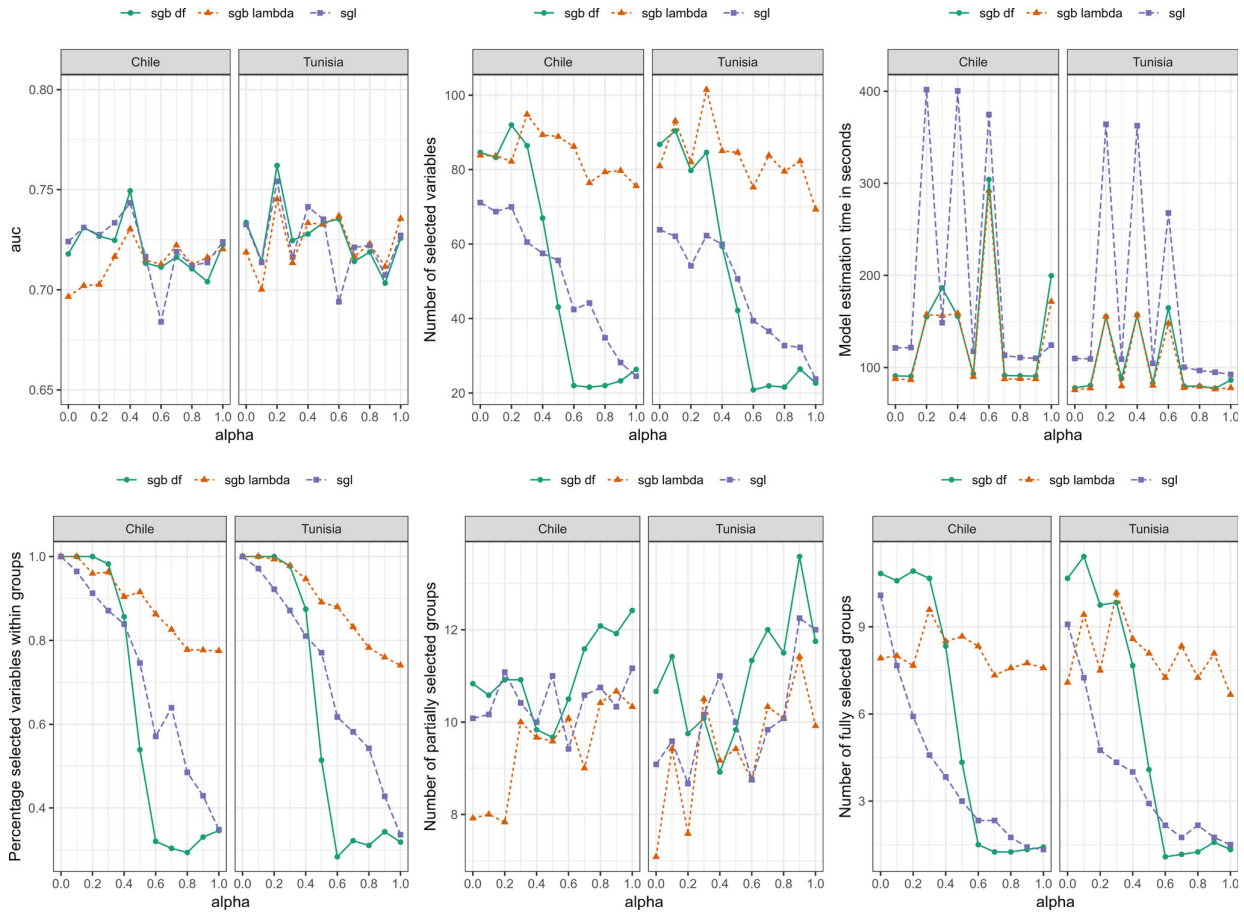
Figure 3. Results of the agricultural dataset. Line-type and point-shape indicate the type model depending on alpha.

The computation time was somewhat volatile, and we had to rerun the models a few times, as the sparse-group lasso cross-validation estimation with the sgl package crashed multiple times. We did not fully optimize for computational speed and there are efforts to improve both the computation time of sgl (Ida, Fujiwara, and Kashima 2019; Zhang et al. 2020) and boosting (Staerk and Mayr 2021). Theoretically, the computation time of the sparse-group boosting should be the sum of the time it takes to fit component-wise boosting and group-component-wise boosting. However, through fitting group base-learners parallel to individual learners, the computation time should be close to either group boosting or component-wise boosting depending on which of the two is slower. Therefore, the speed of the sparse-group boosting depends mostly on the implementation of boosting.

After evaluating the models on the training data, we refit one of the models for the outcome indicating the willingness to invest more than 60% of income in adaptive measures in the future with $\alpha = 0.3$. We can look at the relative variable importance adapted to sparse-group boosting. Let $\hat{\mathcal{L}}_l^{[m]}$ be the reduction of log-likelihood in boosting iteration m and predictor $l \in \{0, \ldots, p, p+1, \ldots, p+G\}$ be the base-learner which was selected in this step. The first p predictors are individual variables and the remaining G are the groups. Then the reduction can be attributed to this predictor. Hence, we can compute the relative

contribution of each group or individual variable to the global model or the relative contribution of all groups $\frac{\sum_{m=1}^{M} \hat{\mathcal{L}}_{\{l:l>p\}}^{[m]}}{\sum_{m=1}^{M} \hat{\mathcal{L}}_l^{[m]}}$. Figure 4(a) and (b) show the variable importance for Chile and Tunisia for each predictor/base-learner contributing at least one percent to the model. The legend shows the relative contribution of groups, indicating that individual variables contribute relatively more to the model than groups. Below in (c) and (d) we see the aggregated coefficient paths for each variable in the dataset by summing up the coefficients from individual selection and group selection. The color indicates whether an update in a particular coefficient came from individual or group variable selection. We can see the alteration between individual and group variable selection depending on the boosting iteration giving us insights into the selection process. The figures were created with the R package "sgboost" (Obster 2024), which implements the sparse-group boosting and contains interpretability tools, explained in Obster and Heumann (2024).

5. Simulated Data

In this section, we will compare the two versions of the sparse-group boosting with the lasso, to see how similar the predictive power and sparsity properties are. Therefore, we consider 11 equally spaced mixing parameters α between zero and one for
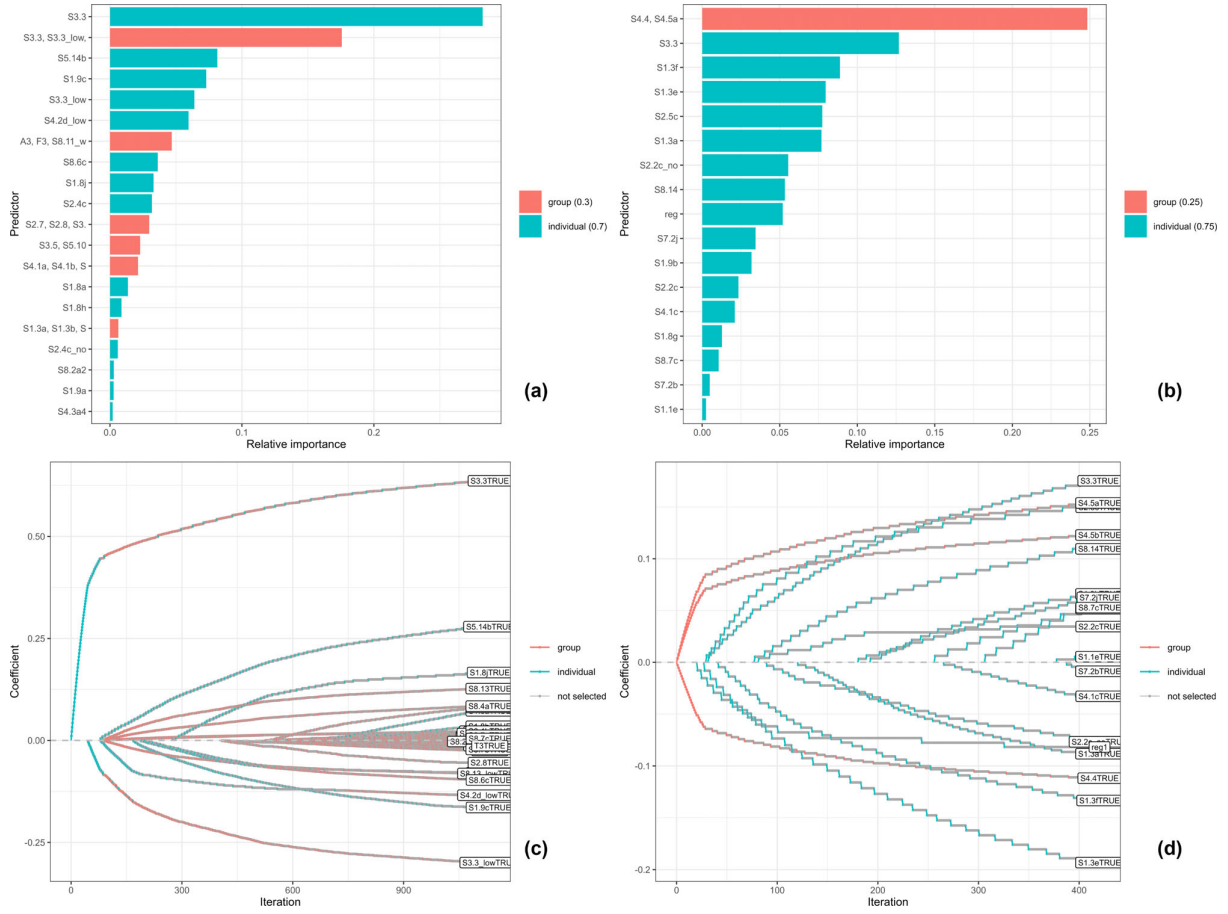
Figure 4. Variable importance for the sparse-group boosting model explaining the willingness to invest in Chile (a) and in Tunisia (b) and coefficient path in Chile (c) and Tunisia (d).

the sparse-group lasso and sparse-group boosting. This way the lasso/boosting and group lasso/group boosting are also covered.

The covariate matrix X was simulated with different numbers of covariates, groups, observations, and covariance structures. The response, y was set to

$$\sum_{g=1}^{G} X_{V_g} \beta^{(g)} + \sigma \epsilon.$$

Here, $\epsilon \sim \mathcal{N}(0, I)$. The signal-to-noise ratio between the nonzero entries of β and $\sigma \epsilon$ was set to 4 through the value of σ. Note that the effective signal-to-noise ratio is additionally altered by setting some elements of β to zero, which additionally increases the noise. In the case of no variables being associated with the outcome, no additional error term ϵ was used.

The tuning of the models was performed with a 3-fold cross-validation performed on the whole simulated data. We used 11 equally spaced mixing parameters α ranging from zero to one. For the sparse-group boosting based on λ and the sparse-group lasso, we chose 10 values for λ. Since no proven method of selecting a good set of λ values in the sparse-group boosting exists yet, we chose $\lambda = 50 \cdot i$ for $i \in \{1, \ldots, 10\}$, as in boosting ridge regression in general bigger values for λ are generally preferable Tutz and Binder (2007). For the boosting models, we used a learning rate of 0.05 and 2500 boosting iterations to

fit the models with early stopping derived from a 3-fold cross-validation. Since the sparse-group boosting using the degrees of freedom has no comparable tuning parameter for λ in the other two models, we used a finer grid of α values. For a given alpha value α in the sparse-group lasso, whenever the model is fitted for λ_i, the sparse-group boosting with the degrees of freedom is fitted with $\alpha + 0.01 \cdot (i - 1)$. This way, for each α, 10 versions of each of the three models are being fitted. Afterward, we compare the estimates with the actual parameter vector β. The parameters used for each scenario are summarized in Table 1. For each scenario, 15 iterations of the data were simulated. As the main evaluation criterion, we used the root mean squared error (RMSE), defined as

$$\text{RMSE} = \frac{1}{p} \sqrt{\sum_{g=1}^{G} \sum_{j=1}^{p_g} \left((\beta_j^{(g)} - \widehat{\beta}_j^{(g)})^2 \right)},$$

where $X_{i\cdot}$ is the ith row of the design matrix.

To make the simulation results comparable across the considered scenarios, we consider the standardized RMSE within each scenario for all iterations $r \in \{1, \ldots, R\}$ defined as $\frac{\text{RMSE}_r - \overline{\text{RMSE}}}{\text{sd(RMSE)}}$ using the sample mean ($\overline{\text{RMSE}}$) and standard deviation (sd(RMSE)). We also computed the proportion of "correct

Table 1. Table with aligned units.

Scenario	full gr.	half gr.	empty gr.	full vars	half vars	empty vars	cor	n
1	5	5	5	15	15	15	0	50
2	5	5	5	5	5	15	0	50
3	5	5	5	5	15	5	0	50
4	5	5	5	15	5	5	0	50
5	2	2	5	15	15	15	0	50
6	5	2	2	15	15	15	0	50
7	2	5	2	15	15	15	0	50
8	0	0	5	0	0	15	0	50
9	5	0	0	15	0	0	0	50
10	5	5	5	15	15	15	0	500
11	5	5	5	15	15	15	0.5	50
12	5	5	5	15	15	15	0.95	50

NOTE: Full gr. refers to the number of groups where each variable inside it has an effect. Half gr. refers to the number of groups that contain exactly five effects and the remaining ones are zero. Empty gr. refers to the number of groups that contain no effects. The number of variables within these groups is described by full vars, half vars, and empty vars. Cor refers to the degree of multicollinearity of the design matrix, measured by the pairwise correlation between the variables in the design matrix.
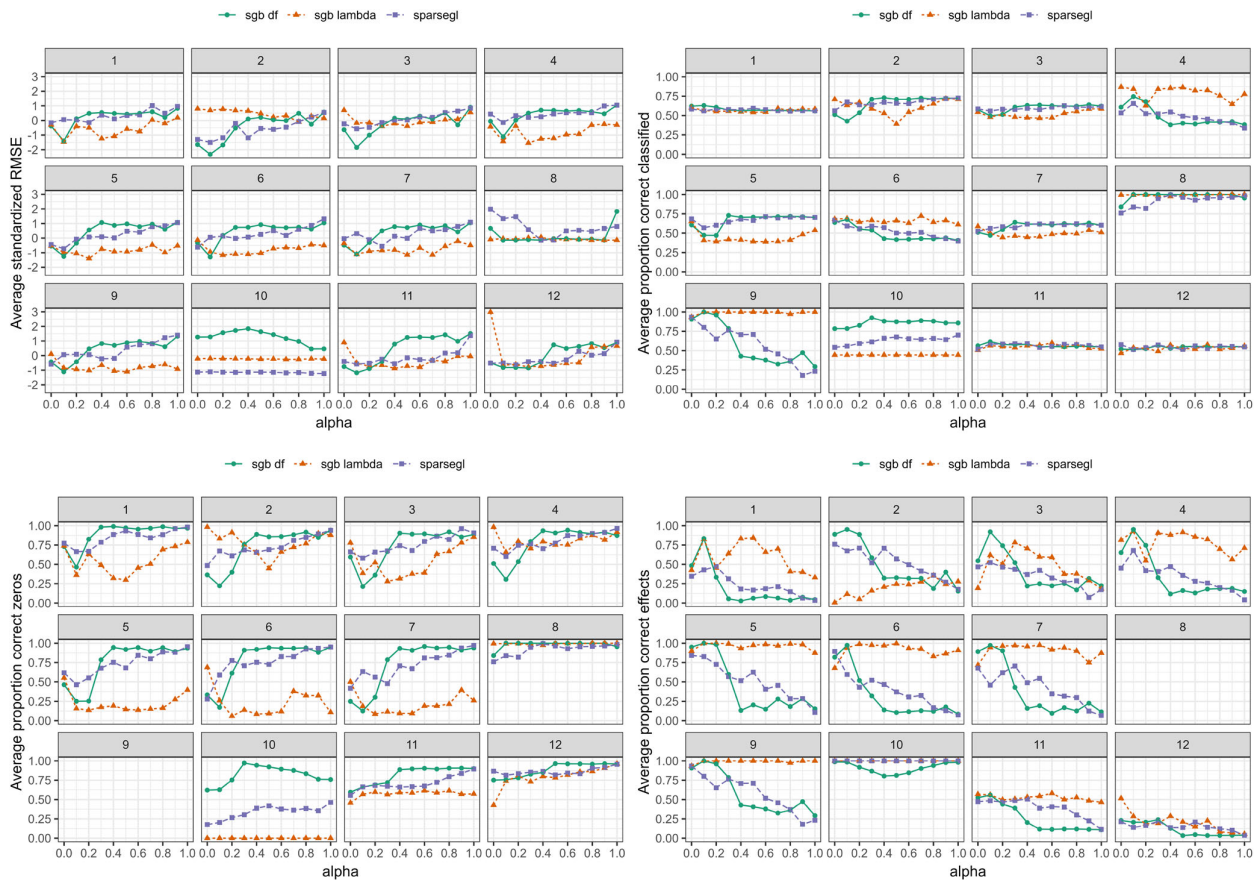
Figure 5. Simulation results for the 12 simulated scenarios averaged across the 15 iterations and 10 hyper-parameter setting for each alpha. Line-type and point-shape indicates the type of model. All metrics compare the model estimates with the true parameter vector.
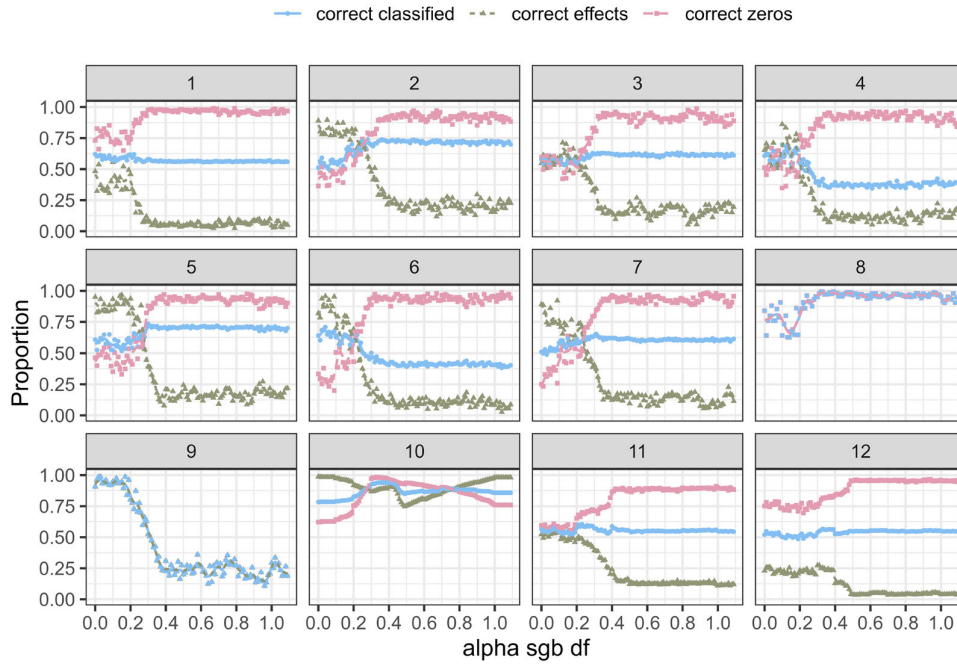
effects"

$$\frac{\sum_{g=1}^{G} \sum_{j=1}^{p_g} \mathbb{1}_{[\beta_j^{(g)} \neq 0 \wedge \widehat{\beta}_j^{(g)} \neq 0]}}{\sum_{g=1}^{G} \sum_{j=1}^{p_g} \mathbb{1}_{[\beta_j^{(g)} \neq 0]}},$$

and the proportion of "correct zeros" and the overall "correct classified" elements of β. In Figure 5, the results of the simulation are displayed. For each model type and α value, out of the 10 hyperparameters, the model with the lowest RMSE was chosen. For each metric considered the values were then

averaged across the 15 iterations. Since sgb df had a finer grid of α values, we removed the second digit of α, for example 0.47 becomes 0.4. The only exceptions were the values between 0.01 and 0.10, which we rounded up to 0.1. This is because $\alpha = 0$ is group boosting and we did not want to mix it with the sparse-group boosting. The results of sgb df on the full scale of α without summarizing are shown in Figure 6.

Generally, in most scenarios and models, sparse-group variable selection improves the fit. For $\alpha \in \{0, 1\}$ (group lasso,

Figure 6. Effect of alpha on the proportion of correct classification of effects and zeros in the sparse-group boosting (sgb df) for the 12 simulated scenarios. Each point represents an average of the 15 iterations. Line-type and point-shape indicates the type of detection rate comparing the estimates with the actual parameter vector β. Smoothed lines were computed by the LOESS with a span of 0.1.

group boosting, lasso and component-wise boosting), sgb df and sgl yield similar estimates, except scenarios 8 and 10, as the RMSE and also the detection rates are close together. This is in line with Hepp et al. (2016). However, the effect of α on the evaluated metrics differs. This is in line with the results from Theorem 2 and Corollary 2, as for $\alpha \geq 0.5$, sgb df will be close to component-wise boosting with varying degrees of freedom. This is a difference to sgl, in which the resulting model changes through the whole range of α. Generally, there is a tradeoff between the correct detection of effects and zeros, which is affected by α through the selection of either groups or single variables. This is the case for all covered models. The range of the correct detection of zeros and effects is greater for sgb df than all other models, which can partly be explained by a finer grid of α values. In this regard, it is important to note again, that sgb lambda is not guaranteed to yield only group-wise selection as seen in Figure 2. In scenario 10 sgl outperformed both variations of the sparse-group boosting. However, in this scenario, sgb df was almost always stopped out meaning the number of boosting iterations was set too small for this dataset. A similar issue happens with the models fitted with the "SGL" package, as the vector of lambda values contained too small values leading to severe over-fitting (compare with Figure 1 in Appendix C). Therefore, we used the 'sparsegl' package for the simulations, which chooses more sensible values for lambda in the cross-validation.

As in many cases, there is no "the best model", only the best model for a given metric and dataset, especially if one looks at opposing metrics. For sgb df and sgl, whenever there are more full groups or the number of variables in the full groups is greater compared to the half and empty groups (Scenarios 4,6, and 9), α decreases the correct detection of β elements and increases RMSE, meaning group-wise selection is more

important than individual variable selection. The opposite is true when the number of half groups and the group size of the half groups is greater (Scenarios 3 and 7). We observe that the correct classification rate of active variables is higher for sgb lambda than for sgb df, especially for higher values of α. In Scenario 2, however, this is not the case, as the group size of empty groups is greater than that of non-empty groups. On the other hand, the correct classification of non-active variables is higher for sgb df. This observation is in line with the results from the agricultural data, where within-group sparsity and overall sparsity are higher for sgb df compared to sgb lambda. Hence, sgb lambda may be more prone to over-selection (overfitting) and sgb df to under-selection. Over the whole range of α the range of all considered metrics is greater for sgb df than for the other methods in most scenarios. This makes it more likely to find a good tradeoff between the correct detection of active and inactive variables/groups by tuning α using sgb df. This was also observed in the agricultural data. In the case of Multicollinearity (Scenario 1 vs. Scenarios 11 vs. 12), the correct detection rates are similar for the lasso and component-wise boosting. It seems to affect the models for smaller values of α more. As also apparent in Figures 2 and 6, the bounds of the interval in which group selection and individual variable selection happen together in the sgb df is shifted to the right in the case of multicollinearity.

6. Conclusion, Limitations, Future Work, and Discussion

In this article, we presented a framework for adapting sparse-group variable selection to boosting. Combining group and individual base-learners in the same model, by weighting the

degrees of freedom as mixing parameter α, one can fit a model with similar characteristics as the sparse-group lasso. However, the effect of α is different in the sparse-group lasso compared to α in the sparse-group boosting. Even though both models yield only group selection for $\alpha = 0$ and only individual variable selection for $\alpha = 1$, there is a greater range where the sparse-group boosting only selects individual variables or groups. We found theoretical bounds for this range depending on the singular values of the group design matrix in $\mathcal{L}^2$ boosting. This implies that for model tuning one should focus on the α values within the theoretical bounds. A good proxy without computing the singular values a priori is to use $\alpha \in [\frac{1}{p_{\max}+1}, 0.6]$. This interval is the one from Corollary 2 and gives some room for multicollinearity and strongly varying singular values of a group design matrix. In the simulations as well as the two real datasets this range was sufficient. While finding the right value for α in the sparse-group boosting, it has a natural interpretation in addition to just being a mixing parameter, as it corresponds to the degrees of freedom.

Mixing the ridge regularization parameter directly (sgb lambda) one can also fit a sparse-group boosting. However, the fitting becomes harder as one has an additional hyper-parameter which has a strong effect on the selection bounds and one loses the interpretability of α. Further research would be required to make this formulation a useful competitor.

We want to note that there are also other possible ways to fit a similar model through boosting. One idea is boosting the sparse-group lasso, or using group boosting with an elastic net penalty within each group-base-learner. The main difference between the sparse-group boosting and the sparse-group lasso is the fitting philosophy. When thinking about the sparse-group lasso one thinks of shrinking effects and making them vanish either on a group level or on an individual variable level. When thinking about boosting, one rather think about adding individual variables or groups of variables to the global model. Boosting the sparse-group lasso or group boosting the elastic net would combine both approaches iteratively while adding and shrinking at the same time. This could have an interesting and different selection behavior over the here proposed sparse-group boosting, as one does not have to update full groups if a group base-learner is selected. A distinct advantage of the philosophy of adding rather than shrinking of the sparse-group boosting is that one can have a group being selected, updating all variables equally and also additional individual variables on top of the group, which are more important than the other variables within the group. This way one can compare the variable importance of individual variables versus group variables through the proportion of explained variance/loss function which cannot be done by the sparse-group lasso and boosted variants of it. An example of this can be found in Obster, Bohle, and Pechan (2024). This way, sparse-group boosting can facilitate new research questions and provide additional insights and interpretability. The group variable importance as defined in Section 4 can also be used for construct validation or an exploratory identification of latent variables, as often performed in the social sciences and psychology. Suppose variables are selected through a group rather than through individual variables. In that case, the group construct is likely responsible for explaining the outcome, rather than some specific aspects of it, underpinning construct validity.

Supplementary Materials

Acknowledgments

Disclosure Statement

Funding

ORCID

Fabian Obster 🔵 http://orcid.org/0000-0002-6951-9869

References

Agarwal, N. K. (2011), "Verifying Survey Items for Construct Validity: A Two-Stage Sorting Procedure for Questionnaire Design in Information Behavior Research," *Proceedings of the American Society for Information Science and Technology*, 48, 1–8. DOI:10.1002/meet.2011.14504801166. *https://onlinelibrary.wiley.com/doi/abs/10.1002/meet.2011.14504801166*. [1]

Breiman, L. (1998), "Arcing Classifier," (with discussion and a rejoinder by the author), *The Annals of Statistics*, 26, 801–849. DOI:10.1214/aos/1024691079. Available at *https://projecteuclid.org/journals/annalsof-statistics/volume-26/issue-3/Arcing-classifier-with-discussion-and-a-rejoinder-by-the-author/10.1214/aos/1024691079.full*. [1]

Bühlmann, P., and Hothorn, T. (2007), "Boosting Algorithms: Regularization, Prediction and Model Fitting," *Statistical Science*, 22, 477–505. DOI:10.1214/07-STS242. *https://projecteuclid.org/journals/statisticalscience/volume-22/issue-4/Boosting-Algorithms-Regularization-Prediction-and-Model-Fitting/10.1214/07-STS242.full*. [2]

Carroll, R. J., Ruppert, D., and Wand, M. P. (2003), "Computational Issues," in *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, eds. R. J. Carroll, D. Ruppert, and M. P. Wand, pp. 336–360, Cambridge: Cambridge University Press. DOI:10.1017/CBO9780511755453.023. *%5Curl %7Bhttps://www.cambridge.org/core/books/semiparametric-regression/02FC9A9435232CA67532B4D31874412C%7D*. [4]

Chen, R.-B., Chu, C.-H., Yuan, S., and Nian, Y. (2016). "Bayesian Sparse Group Selection," *Journal of Computational and Graphical Statistics*, 25, 665–683. *https://www.jstor.org/stable/44861885*. [1]

Chowdhury, M. Z. I., and Turin, T. C. (2020), "Variable Selection Strategies and its Importance in Clinical Prediction Modelling," *Family Medicine and Community Health*, 8, e000262. DOI:10.1136/fmch-2019-000262. Available at *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7032893/*. [1]

Farokhmanesh, F., and Sadeghi, M. T. (2019), "Deep Feature Selection using an Enhanced Sparse Group Lasso Algorithm,". in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1549–1552. DOI:10.1109/IranianCEE.2019.8786386. *https://ieeexplore.ieee.org/abstract/document/8786386?casa_token=0QiOzmn2ObkAAAAA:WQ8uemI_FBoJvMiy8vWzdBG7RX-kk_bLE2I0o4dJw2M0SaGhFAg0Y_iDN8CtaZ9oYclmo4J5jv1q*. [1]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," (with discussion and a rejoinder by the authors), *The Annals of Statistics*, 28, 337–407. DOI:10.1214/aos/1016218223. *https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-2/Additive-logistic-regression--a-statistical-view-of-boosting-With/10.1214/aos/1016218223.full*. [1]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. DOI:10.1214/aos/1013203451. *https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full*. [2]

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., and Preckel, F. (2014), ""My Questionnaire is Too Long!" The Assessments of Motivational-Affective Constructs with Three-Item and Single-Item Measures," *Contemporary Educational Psychology*, 39, 188–205. DOI:10.1016/j.cedpsych.2014.04.002. Available at *%5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0361476X14000204%7D*. [1]

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., Mayr, A. (2016), "Approaches to Regularized Regression - A Comparison between Gradient Boosting and the Lasso," *Methods of Information in Medicine*, 55, 422–430. DOI:10.3414/ME16-01-0033. *http://www.thiemeconnect.de/DOI/DOI?10.3414/ME16-01-0033*. [12]

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011), "A Framework for Unbiased Model Selection Based on Boosting," *Journal of Computational and Graphical Statistics*, 20, 956–971. DOI:10.1198/jcgs.2011.09220. *http://www.tandfonline.com/doi/abs/10.1198/jcgs.2011.09220*. [3,7,8]

Hofner, B., Mayr, A., and Schmid, M. (2014), "gamboostLSS: An R Package for Model Building and Variable Selection in the GAMLSS Framework," arXiv:1407.1774 [stat]. arXiv: 1407.1774. Available at *http://arxiv.org/abs/1407.1774*. [4]

Ida, Y., Fujiwara, Y., and Kashima, H. (2019), "Fast Sparse Group Lasso," in *Advances in Neural Information Processing Systems* (Vol. 32), Curran Associates, Inc. Available at *https://proceedings.neurips.cc/paper/2019/hash/d240e3d38a8882ecad8633c8f9c78c9b-Abstract.html*. [9]

Johnstone, I. M., and Titterington, D. M. (2009), "Statistical Challenges of Highdimensional Data," *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 367, 4237–4253. DOI:10.1098/rsta.2009.0159. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2865881/*. [1]

Kneib, T., Hothorn, T., and Tutz, G. (2009), "Variable Selection and Model Choice in Geoadditive Regression Models," *Biometrics*, 65, 626–634. *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2008.01112.x*. [1]

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, Series B, 70, 53–71. DOI:10.1111/j.1467-9868.2007.00627.x. [1]

Mendez-Civieta, A., Carmen Aguilera-Morillo, M., and Lillo, R. E. (2021), "Adaptive Sparse Group LASSO in Quantile Regression," *Advances in Data Analysis and Classification*, 15, 547–573. DOI:10.1007/s11634-020-00413-8. [1]

Obster, F. (2024), *sgboost: Sparse-Group Boosting*, available at *https://CRAN.R-project.org/package=sgboost*. [9]

Obster, F., Bohle, H., and Pechan, P. M. (2024), "The Financial Well-Being of Fruit Farmers in Chile and Tunisia Depends More on Social and Geographical Factors than on Climate Change," *Communications Earth & Environment*, 5, 1–12. DOI:10.1038/s43247-023-01128-2. *https://www.nature.com/articles/s43247-023-01128-2*. [8,13]

Obster, F., and Heumann, C. (2024), "Introducing sgboost: A Practical Guide and Implementation of sparse-group boosting in R," arXiv:2405.21037 [stat]. *http://arxiv.org/abs/2405.21037*. [5,9]

Obster, F., Heumann, C., Bohle, H., and Pechan, P. (2023), "Using Interpretable Boosting Algorithms for Modeling Environmental and Agricultural Data," *Scientific Reports* 13, 12767. DOI:10.1038/s41598-023-39918-5. *https://www.nature.com/articles/s41598-023-39918-5*. [8]

Pechan, P. M., Bohle, H., and Obster, F. (2023), "Reducing Vulnerability of Fruit Orchards to Climate Change," *Agricultural Systems*, 210, 103713. DOI:10.1016/j.agsy.2023.103713. *https://www.sciencedirect.com/science/article/pii/S0308521X2300118X*. [8]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. DOI:10.1080/10618600.2012.681250. *http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250*. [1,2]

Staerk, C., and Mayr, A. (2021), "Randomized Boosting with Multivariable Baselearners for High-Dimensional Variable Selection and Prediction," *BMC Bioinformatics*, 22, 441. DOI:10.1186/s12859-021-04340-z. [9]

Stasinopoulos, D. M., and Rigby, R. A. (2008), "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R," *Journal of Statistical Software*, 23, 1–46. DOI:10.18637/jss.v023.i07. [8]

Tang, F., Adam, L., and Si, B. (2018), "Group Feature Selection with Multiclass Support Vector Machine," *Neurocomputing*, 317, 42–49. DOI:10.1016/j.neucom.2018.07.012. *https://www.sciencedirect.com/science/article/pii/S0925231218308403*. [1]

R Core Team. (2022), "R: A Language and Environment for Statistical Computing," available at *https://www.R-project.org/*. [8]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. *https://www.jstor.org/stable/2346178*. [1,2]

Tutz, G., and Binder, H. (2007), "Boosting Ridge Regression," *Computational Statistics & Data Analysis*, 51, 6044–6059. DOI:10.1016/j.csda.2006.11.041. *https://www.sciencedirect.com/science/article/pii/S0167947306004749*. [2,3,10]

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag. Available at *https://ggplot2.tidyverse.org*. [8]

Yoon, J., and Hwang, S. J. (2017), "Combined Group and Exclusive Sparsity for Deep Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3958–3966, PMLR. Available at *https://proceedings.mlr.press/v70/yoon17a.html*. [1]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. DOI:10.1111/j.1467-9868.2005.00532.x. *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x*. [1,2]

Zhang, Y., Zhang, N., Sun, D., and Toh, K.-C. (2020), "An Efficient Hessian based Algorithm for Solving Large-Scale Sparse Group Lasso Problems," *Mathematical Programming*, 179, 223–263. DOI:10.1007/s10107-018-1329-6. [9]

Zhao, L., Hu, Q., and Wang, W. (2015), "Heterogeneous Feature Selection With Multi-Modal Deep Neural Networks and Sparse Group LASSO," *IEEE Transactions on Multimedia*, 17, 1936–1948. DOI:10.1109/TMM.2015.2477058. *https://ieeexplore.ieee.org/abstract/document/7244241?casa_token=A2MMpqF5kWQAAAAA:8agVi67iudlCIFVWXUI7FyMC4CXglX5c0TH_cTPmyF3I9f_WXf_19rIkr3mTFmY7BLSGEmEsagHo*. [1]

Zuo, Y., Stewart, T. G., and Blume, J. D. (2022), "Variable Selection With Second- Generation P-Values," *The American Statistician*, 76, 91–101. DOI:10.1080/00031305.2021.1946150. [1]

Appendix
Sparse-group boosting:
Unbiased group and variable selection

Fabian Obster*

Department of Business Administration,

University of the Bundeswehr Munich, Germany

and

Christian Heumann

Department of Statistics,

Ludwig Maximilians University Munich, Germany

September 19, 2024

A Generalizations

Let $h : \mathbb{R}^{n \times p} \mapsto \mathbb{R}^n$ be the invertible strictly increasing response function of a generalized linear model, with $\mathbb{E}[y|X] = \mu = h(\eta)$. Where $y|x$ is a member of the simple exponential family in canonical form depending on the linear predictor $\eta = X\beta$. For L base-learners, denote the l-th candidate sets consisting of p_l columns as $V_l = \{(v_l)_1, ..., (v_l)_{p_l}\} \subseteq \{1, ..., p\}$. We do not require the candidate sets to be disjoint as in Tutz and Binder 2007 leading to possibly overlapping groups, which we will later utilize for the sparse-group boosting.

Generalized Ridge Boosting

1. Initialize $m = 0, \widehat{\beta}^{[0]} = \mathbf{0}_p, \widehat{\eta}^{[0]} = X\widehat{\beta}^{[0]}$, and $\widehat{\mu}^{[0]} = h(\widehat{\eta}^{[0]})$

2. Set $m = m + 1$

   For each candidate set V_l, $l \leq L$, fit the model

   $$\mu = h(\widehat{\eta}^{[m-1]} + X_{V_l}\beta_{V_l}),$$

   by minimizing the penalized negative log-likelihood

   $$\mathscr{L}_l^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l}) + \lambda(\beta_{V_l})^T \beta_{V_l},$$

   with offset $X\widehat{\beta}^{[m-1]}$ derived from the previous iteration. This can be done by Fisher scoring or iterative weighted least squares. For the l-th base-learner denote the estimate of β_{V_l} as $\widehat{\overline{\beta}}_{V_l}$ and the estimate of the negative log-likelihood as $\widehat{\mathscr{L}}_l^{[m]}$.

3. Select the candidate set which evaluates the lowest negative log-likelihood $l^* = \arg\min_{l \leq L} \widehat{\mathscr{L}}_l^{[m]}$.

4. Update for all $l \leq L$

   $$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}_{V_l}^{[m-1]} + \nu\widehat{\overline{\beta}}_{V_l} & l = l^*, \\ \widehat{\beta}_{V_l}^{[m-1]} & l \neq l^* \end{cases}$$

   and

   $$\widehat{\eta}^{[m]} = X\widehat{\beta}^{[m]},$$

   $$\widehat{\mu}^{[m]} = h(X\widehat{\beta}^{[m]}).$$

   Here ν can be seen as learning rate with $\nu \in \,]0, 1[$.

5. Repeat steps 3,4 and 5 until $m = M$ and retrieve $\widehat{\beta}^{[M]}$ as global estimate.

Generalized Sparse-group Boosting

For the sparse-group boosting we define $p+G$ candidate sets. Of which the first p refer to individual base-learners $l \leq p : V_l = V_l = \{l\}$, and the remaining G to the group base-learners $l > p : V_l = \{(v_l)_1, ..., (v_l)_{p_l}\} \subseteq \{1, ..., p\}$.

1. Initialize $m = 0, \widehat{\beta}^{[0]} = \mathbf{0}_p, \widehat{\eta}^{[0]} = X\widehat{\beta}^{[0]}$, and $\widehat{\mu}^{[0]} = h(\widehat{\eta}^{[0]})$.

2. Set $m = m + 1$

   For each candidate set V_l, $l \leq G + p$, fit the model

   $$\mu = h(\widehat{\eta}^{[m-1]} + X_{V_l}\beta_{V_l}),$$

   by minimizing the penalized negative log-likelihood

   $$\mathcal{L}_l^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l}) + \lambda(\beta_{V_l})^T \beta_{V_l},$$

   with offset $X\widehat{\beta}^{[m-1]}$ derived from the previous iteration. Regularization parameters are defined using the Ridge hat matrix $H_{V_l}^{\lambda} = W^{1/2} X_{V_l}((X_{V_l})^T W X_{V_l} + \lambda_l I)^{-1}(X_{V_l})^T W^{1/2}$ of each base-learner as defined in the previous section. W is the weighting matrix from generalized linear models (Tutz and Binder 2007).

   $$\lambda_l = \begin{cases} \lambda_l : \mathrm{df}(\lambda_l) = \mathrm{tr}(2H_{V_l}^{\lambda} - (H_{V_l}^{\lambda})^2) = \alpha & l \leq p \\ \lambda_l : \mathrm{df}(\lambda_l) = \mathrm{tr}(2H_{V_l}^{\lambda} - (H_{V_l}^{\lambda})^2) = 1 - \alpha & l > p. \end{cases} \quad (1)$$

   This can be done by Fischer scoring or iterative weighted least squares. For the l-th base-learner denote the estimate of β_{V_l} as $\widehat{\overline{\beta}}_{V_l}$ and the estimate of the negative log-likelihood as $\widehat{\mathcal{L}}_l^{[m]}$.

3. Select the candidate set which evaluates the lowest negative log-likelihood $l^* = \arg\min_{\{l \leq G+p\}} \widehat{\mathcal{L}}_l^{[m]}$.

   $$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}^{[m-1]} + \nu\widehat{\overline{\beta}}_{V_{l^*}} & l = l^*, \\ \widehat{\beta}^{[m-1]} & l \neq l^* \end{cases}$$

3

and

$$\widehat{\eta}^{[m]} = X\widehat{\beta}^{[m]},$$

$$\widehat{\mu}^{[m]} = h(X\widehat{\beta}^{[m]}).$$

Here ν can be seen as learning rate with $\nu \in]0,1[$.

4. Repeat steps 2, 3, and 4 until $m = M$ and retrieve $\widehat{\beta}^{[M]}$ as global estimate.

B   Theorms with proofs

Theorem 1 (Distribution of the difference of RSS in orthogonal Ridge regression). *Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with orthonormal columns such that $X^T X = I_p$. Let $y \in \mathbb{R}^n$ be the outcome variable, $y = \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ not dependent on the design matrix. Further, assume that the least squares estimate $\widehat{\beta} = X^T y$ exists and $\widehat{\beta}_\lambda$ is the Ridge estimate for some $\lambda > 0$. Define the difference of residual sums of squares as $\Delta = RSS(\widehat{\beta}_\lambda) - RSS(\widehat{\beta}) = (y - \frac{1}{1+\lambda}X\widehat{\beta})^T(y - \frac{1}{1+\lambda}X\widehat{\beta}) - (y - X\widehat{\beta})^T(y - X\widehat{\beta})$. Then if $\left(1 - \frac{df(\lambda)}{p}\right) > 0$, $\frac{\Delta}{\sigma^2}$ follows a gamma distribution with the following shape-scale parametrization*

$$\frac{\Delta}{\sigma^2} \sim \Gamma\left(\frac{p}{2}, 2(1 - \frac{2}{p(1+\lambda)} + \frac{1}{p(1+\lambda)^2})\right).$$

Proof.

$$\begin{aligned}
\Delta &= y^T\left(I_p - \frac{1}{1+\lambda}XX^T\right)^2 y - y^T(I_p - XX^T)y \\
&= y^T\left(-\frac{2}{1+\lambda}XX^T + \frac{1}{(1+\lambda)^2}XX^T + XX^T\right)y \\
&= \left(1 - \frac{2}{1+\lambda} + \frac{1}{(1+\lambda)^2}\right)y^T XX^T y \\
&= \left(1 - \frac{df(\lambda)}{p}\right)y^T XX^T y.
\end{aligned}$$

Since $1 - \frac{df(\lambda)}{p} > 0$ and $\frac{y^T XX^T y}{\sigma^2} \sim \chi^2(p)$ we end up with $\frac{\Delta}{\sigma^2} \sim \Gamma\left(\frac{p}{2}, 2(1 - \frac{df(\lambda)}{p})\right)$. $\square$

Theorem 2 (Selection intervals of the sparse-group boosting). *Consider a design matrix $X \in \mathbb{R}^{n \times p}$ of rank $r \leq p$ with singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n \times p}, V \in \mathbb{R}^{p \times p}$ are unitary matrices and $D = diag(d_1, ..., d_r, 0, ..., 0)$ is a diagonal Matrix containing the singular values. Let*

4

$y \in \mathbb{R}^n$ be the outcome variable and $\widehat{\beta}_\mu = (X^T X + \mu I)^{-1} X^T y$ be the Ridge estimate for $\mu \geq 0$. For $j \leq p$ let $\widehat{\beta}_{\lambda_j} = (x_j^T x_j + \lambda_j)^{-1} x_j^T y$ be the estimate for the j-th individual base-learner, and $\overline{d}_j$ be the singular value of x_j. Denote $\overline{d}^{\,-} = \min_{j \leq p} \overline{d}_j^2$ and $\overline{d}^{\,+} = \max_{j \leq p} \overline{d}_j^2$ as well as and $d^+ = \max_{j \leq r} d_j^2$ and $d^- = \min_{j \leq r} d_j^2$ accordingly. Then, there are always two mixing parameters $\alpha_1, \alpha_2 \in]0, 1[$ such that

$$(\forall_{j \leq p} : \alpha_1 = df(\lambda_j) \wedge (1 - \alpha_1) = df(\mu)) \Rightarrow \min_{j \leq p} RSS(\lambda_j) \leq RSS(\mu), \text{ and}$$

$$(\forall_{j \leq p} : \alpha_2 = df(\lambda_j) \wedge (1 - \alpha_2) = df(\mu)) \Rightarrow RSS(\mu) \leq \min_{j \leq p} RSS(\lambda_j).$$

Furthermore, the following conditions assure the selection of an individual variable or the whole design matrix

$$\left(\left[\forall_{l \leq k} \frac{(d^- + 2\mu)}{(d^- + \mu)^2} \leq \frac{(\overline{d}_l^2 + 2\lambda_l)}{r(\overline{d}_l^2 + \lambda_l)^2} \right] \vee \left[df(\mu) \leq \frac{df(\lambda_l) d^-}{r \overline{d}^{\,+}} \right] \right) \Rightarrow \min_{j \leq p} RSS(\lambda_j) \leq RSS(\mu), \quad (2)$$

$$\left(\left[\forall_{l \leq k} \frac{(d^+ + 2\mu)}{(d^+ + \mu)^2} \geq \frac{df(\lambda_l)}{\overline{d}_l^2} \right] \vee \left[\forall_{l \leq k} \frac{(d^+ + 2\mu)}{(d^+ + \mu)^2} \geq \frac{(\overline{d}_l^2 + 2\lambda_l)}{(\overline{d}_l^2 + \lambda_l)^2} \right] \right) \Rightarrow RSS(\mu) \leq \min_{j \leq p} RSS(\lambda_j). \quad (3)$$

Proof. First, choose a base-learner with a design matrix vector denoted as x_l. By using the singular value decomposition of $x_l = \overline{u}_l \overline{d}_l$ and $X = UDV^T$, we can rewrite $RSS(\lambda_l)$ and $RSS(\mu)$ as

$$RSS(\mu) = y^T y - y^T \left(\sum_{j=1}^r \left[2 \frac{d_j^2}{d_j^2 + \mu} - \frac{d_j^4}{(d_j^2 + \mu)^2} \right] u_j u_j^T \right) y$$

and

$$RSS(\lambda_l) = y^T y - y^T \left(\left[2 \frac{\overline{d}_l^2}{\overline{d}_l^2 + \lambda_l} - \frac{\overline{d}_l^4}{(\overline{d}_l^2 + \lambda_l)^2} \right] \overline{u}_l \overline{u}_l^T \right) y = y^T y - y^T df(\lambda_l) \overline{u}_l \overline{u}_l^T y.$$

Denote the diagonal elements $\tilde{d}_j = 2 \frac{d_j^2}{d_j^2 + \mu} - \frac{d_j^4}{(d_j^2 + \mu)^2}$ of $\tilde{D}$ and note that $\overline{d}_j^2 \in [d^-, d^+]$, because x_l is a sub-matrix of X.

Then for some $l \leq p$

$$RSS(\mu) - RSS(\lambda_l) = y^T y - y^T \left(\sum_{j=1}^r \tilde{d}_j u_j u_j^T \right) y - \left(y^T y - y^T df(\lambda_l) \overline{u}_l \overline{u}_l^T y \right)$$

$$= y^T \left[- \sum_{j=1}^r \tilde{d}_j u_j u_j^T + df(\lambda_l) \overline{u}_l \overline{u}_l^T \right] y$$

$$= y^T \left(- U \tilde{D} U^T + \frac{df(\lambda_l)}{\overline{d}_l^2} U D (V^T)_l (V^T)_l^T D U^T \right) y$$

$$= y^T U \left(- \tilde{D} + \frac{df(\lambda_l)}{\overline{d}_l^2} D (V^T)_l (V^T)_l^T D \right) U^T y. \quad (4)$$

5

Here we have used the fact that $\overline{u}_l$ is the left singular value of $x_l = UD(V^T)_l$.

We will now consider the first claim (2).

Using the norm inequality $\forall z \in \mathbb{R}^r : \|z\|_\infty^2 \geq \frac{1}{r}\|z\|_2^2$ we get

$$RSS(\mu) - \min_{l \leq k} RSS(\lambda_l) = \max_{l \leq k} y^T U\Big(-\tilde{D} + \frac{\mathrm{df}(\lambda_l)}{\overline{d}_l^2}D(V^T)_l(V^T)_l^{\,T}D\Big)U^T y$$

$$\geq y^T U(-\tilde{D} + \frac{\mathrm{df}(\lambda_l)}{r\overline{d}_l}D^2)U^T y. \tag{5}$$

Here we have also used that $\frac{\mathrm{df}(\lambda_l)}{\overline{d}_l^+}$ is minimal for l^*: $\overline{d}_{l^*} = \overline{d}^+$, because $d \mapsto \frac{1}{d^2}\Big(\frac{2d^2}{d^2+\lambda} - \frac{(d^2)^2}{(d^2+\lambda)^2}\Big)$ is a decreasing function which can be seen by taking the derivative with respect to d. In the case of $\mathrm{df}(\mu) \leq \frac{\mathrm{df}(\lambda)d^-}{r\overline{d}^+}$ all diagonal elements in (5) are greater zero because $\mathrm{df}(\mu) \geq \tilde{d}_j$. To see the other part of (2) we continue with (5), look at diagonal element j and observe that

$$\left[\frac{2d_j^2}{d_j^2+\mu} - \frac{d_j^4}{(d_j^2+\mu)^2}\right] \leq \frac{\frac{2\overline{d}_l^2}{\overline{d}_l^2+\lambda} - \frac{\overline{d}_l^4}{(\overline{d}_l^2+\lambda_l)^2}}{r\overline{d}_l^2}d_j^2$$

$$\Leftrightarrow \frac{(d_j^2+2\mu)}{(d_j^2+\mu)^2} \leq \frac{(\overline{d}_l^2+2\lambda_l)}{r(\overline{d}_l^2+\lambda_l)^2}.$$

Therefore, all diagonal elements are greater zero if (2) holds and $d_j^2 = d^-$, since $d \mapsto \frac{-(d^2+2\mu)}{(d^2+\mu)^2}$ is a decreasing function, which can be easily seen by taking the derivative with respect to d.

For the second claim in (3), we return to(4), which can be bounded by

$$RSS(\mu) - RSS(\lambda_l)$$

$$= y^T U(-\tilde{D} + \frac{\mathrm{df}(\lambda_l)}{\overline{d}_l^2}D(V^T)_l(V^T)_l^{\,T}D)U^T y.$$

$$\leq y^T U\left(-\mathrm{diag}\left[\frac{2d_j^2}{d_j^2+\mu} - \frac{(d_j^2)^2}{(d_j^2+\mu)^2}\right] + \left[\frac{\mathrm{df}(\lambda_l)}{\overline{d}_l^2}\right]\mathrm{diag}(d_j^2)\right)U^T y.$$

$$= y^T U\left(-\mathrm{diag}\left[\frac{d_j^2(d_j^2+2\mu)}{(d_j^2+\mu)^2}\right] + \left[\frac{(\overline{d}_l^2+2\lambda_l)}{(\overline{d}_l^2+\lambda_l)^2}\right]\mathrm{diag}(d_j^2)\right)U^T y$$

$$\leq y^T U\left(\mathrm{diag}(d_j^2)\left[\frac{-(d^++2\mu)}{(d^++\mu)^2} + \frac{(\overline{d}_l^2+2\lambda_l)}{(\overline{d}_l^2+\lambda_l)^2}\right]\right)U^T y \leq 0.$$

In the last step we have used that $d \mapsto \frac{(d^2+2\mu)}{(d^2+\mu)^2}$ is a decreasing function.

The first part follows from (2) and (3) $\qquad\square$

6

Theorem 3. *Let $X \in \mathbb{R}^{n \times p}$ be a scaled orthogonal design matrix such that $X = dU$ for $d \in \mathbb{R}^{+}$ and $U^{n \times p}$ orthogonal. Define the sub-matrix $X^{(1)} \in \mathbb{R}^{n \times p_1}, 0 < p_1 < p$. Let $y \in \mathbb{R}^n$ be the outcome variable, $y = \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ not being dependent on the design matrix. Let $\widehat{\beta}_\lambda$ be the Ridge estimate using the design matrix $X^{(1)}$ for some penalty $\lambda > 0$ and $\widehat{\beta}_\mu$ the Ridge using X as design matrix for penalty $\mu > 0$. Let $df(\lambda)$ and $df(\mu)$ be the corresponding degrees of freedom. If $\frac{df(\lambda)}{p_1} \geq \frac{df(\mu)}{p}$ we can characterize the selection probability based on the residual sum of squares for the two base-learners as*

$$P\big(RSS(\widehat{\beta}_\lambda) \geq RSS(\widehat{\beta}_\mu)\big) = F_{\beta'\left(\frac{p_1}{2}, \frac{p-p_1}{2}, 1, \frac{df(\lambda)p}{df(\mu)p_1} - 1\right)}(1),$$

where F_β is the distribution function of the beta prime distribution.

Proof.

$$\mathrm{RSS}(\widehat{\beta}_\lambda) - \mathrm{RSS}(\widehat{\beta}_\mu) \geq 0 \Leftrightarrow$$

$$y^T\left(I_p - \frac{1}{1+\lambda} X^{(1)} X^{(1)T}\right)^2 y - y^T\left(1 - \frac{1}{1+\mu} XX^T\right)^2 y \geq 0 \Leftrightarrow$$

$$-\frac{df(\lambda)}{p_1} y^T U^{(1)} U^{(1)T} y + \frac{df(\mu)}{p} y^T UU^T y \geq 0 \Leftrightarrow$$

$$df(\mu) y^T U_{\{p_1+1,\dots,p\}} U_{\{p_1+1,\dots,p\}}^T y - \left(\frac{df(\lambda)}{p_1} - \frac{df(\mu)}{p}\right) y^T U^{(1)} U^{(1)T} y \geq 0$$

$$\frac{\left(\frac{df(\lambda)}{p_1} - \frac{df(\mu)}{p}\right) y^T U^{(1)} U^{(1)T} y}{\frac{df(\mu)}{p} y^T U_{\{p_1+1,\dots,p\}} U_{\{p_1+1,\dots,p\}}^T y} \leq 1$$

In the 8-th line $\frac{df(\lambda)}{p_1} - \frac{df(\mu)}{p}$ was used. From the derivations in Lemma 1 we know that

$$\frac{\left(\frac{df(\lambda)}{p_1} - \frac{df(\mu)}{p}\right) y^T U^{(1)} U^{(1)T} y}{\frac{df(\mu)}{p} y^T U_{\{p_1+1,\dots,p\}} U_{\{p_1+1,\dots,p\}}^T y} \sim \frac{\Gamma\left(\frac{p_1}{2}, 2\left(\frac{df(\lambda)}{p_1} - \frac{df(\mu)}{p}\right)\right)}{\Gamma\left(\frac{p-p_1}{2}, \frac{2df(\mu)}{p}\right)} \sim \beta'\left(\frac{p_1}{2}, \frac{p-p_1}{2}, 1, \frac{df(\lambda)p}{df(\mu)p_1} - 1\right).$$

For the last step, the independence of the two gammas was used which follows from the orthogonality of X and therefore the independence of the two quadratic forms. $\square$

C Organisational research data

As a second application, we use a dataset in the field of innovation research in the public sector conducted in 2020 (https://github.com/FabianObster/sgb). A number of 208 soldiers have been

interviewed with a focus on organizational empowerment and its determining factors within the German armed forces. We use 10 groups of variables each containing 4 variables associated with the individual innovation potential (Schießl 2015) and one group containing 20 variables describing the organizational innovation (Intrapreneurship) potential (Moghaddas, Tajafari, and Nowkarizi 2020) to explain the numeric outcome variable "the organizational empowerment scale" (Matthews, Diaz, and Cole 2003). A common way to analyze these types of datasets in the social sciences is to average the variables (items) belonging to a group (construct), as the number of items is relatively high and they are in many cases correlated within a group. However, with this approach, within-group comparisons and sparsity are not obtainable. Models performing sparse-group variable selection allow for more flexibility. We compare the sparse-group lasso with the here proposed versions of the sparse-group boosting. We are interested in two properties, the predictive performance on held-out data and the sparsity property depending on the mixing parameter α. To do this, we fit the three models to half of the data and compare the predictive performance measured by the mean squared error (MSE) on the other half of the data. We also compare the total number of selected variables as a sparsity measure. All variables were standardized. For all models, we used 11 equally spaced mixing parameters α ranging from zero to one. For the sparse-group boosting based on λ and the sparse-group lasso, we chose 10 values for λ. For the sparse-group lasso, we used a 5-fold cross-validation with the function 'cvSGL' from the R package 'SGL' which determines its own values for λ. Since no proven method of selecting a good set of λ values in the sparse-group boosting exists yet, we chose $\lambda = 50 \cdot i$ for $i \in \{1, ..., 10\}$, as in boosting ridge regression in general bigger values for λ are generally preferable Tutz and Binder 2007. For the boosting models, we used a learning rate of 0.01 and 2000 boosting iterations to fit the models with early stopping derived from a 5-fold cross-validation. Since the sparse-group boosting using the degrees of freedom has no comparable tuning parameter for λ in the other two models, we used a finer grid of α values. For a given alpha value α in the sparse-group lasso, whenever the model is fitted for λ_i, the sparse-group boosting with the degrees of freedom is fitted with $\alpha + 0.01 \cdot (i - 1)$. This way, for each α, 10 versions of each of the three models are being fitted. We always chose the model with the lowest MSE for each α evaluated on the training data and plotted the results in Figure 1. We see that for all α values both versions of the sparse-group boosting are competitive comparing the MSE and yield a sparser set of selected variables at the same time. However, in this dataset, it looks

8

like the utilization of group structure decreases the predictive power. the lasso outperformed the group lasso and sparse-group lasso, and the MSE of both sparse-group boosting versions is lower for $\alpha \geq 0.6$ compared to smaller values.
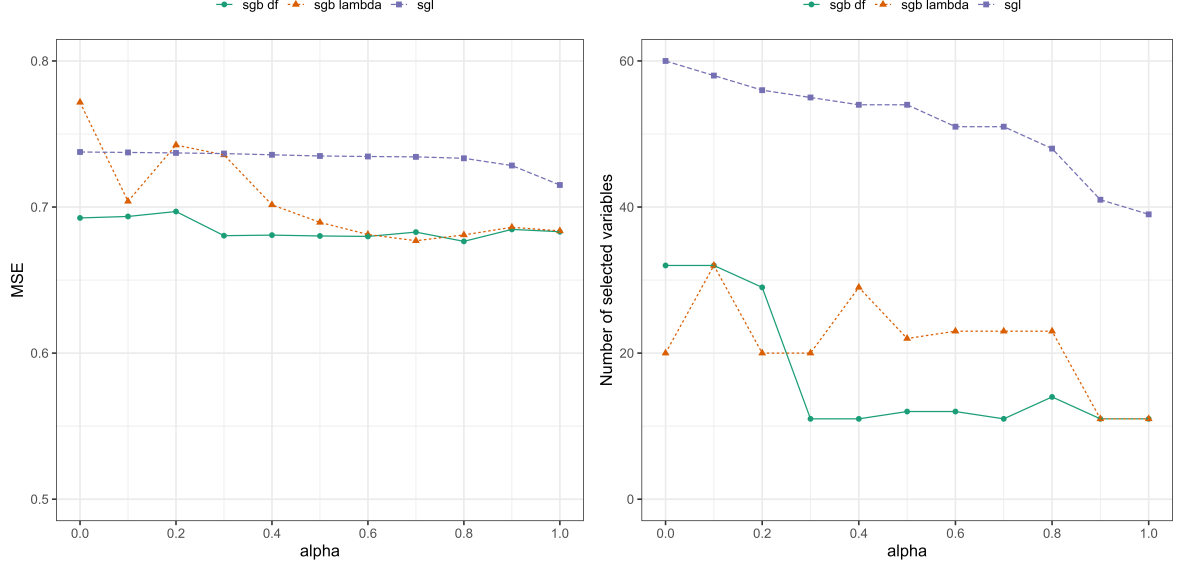

Figure 1: Out of sample MSE (left) and the number of selected variables out of the 60 variables in the dataset (right) for various mixing parameters alpha on the x-axis. Line-type and point-shape indicate the model sparse-group boosting mixing the degrees of freedom (sgb df), sparse-group boosting mixing the ridge regularization parameter (sgb lambda), and the sparse-group lasso (sgl).
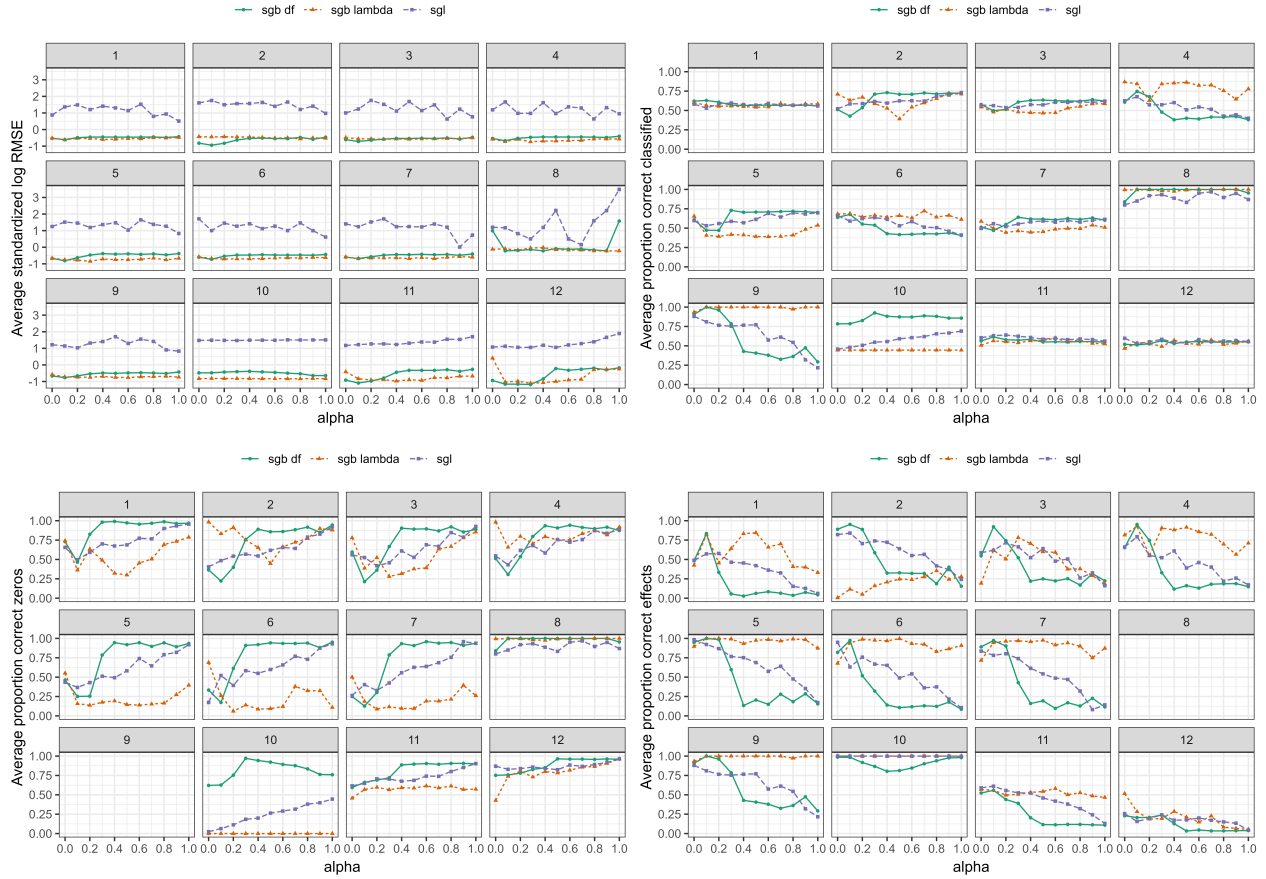
D   Further simulation results

Figure 2: Simulation results for the 12 simulated scenarios averaged across the 15 iterations and 10 hyperparameter setting for each alpha. Colour indicates the type of model. All metrics compare the model estimates with the true parameter vector. Sparse-group lasso fitted via the R-package 'SGL'

References

Matthews, Russell, Wendy Diaz, and Steven Cole (June 2003). "The organizational empowerment scale". In: *Personnel Review* 32, pp. 297–318. DOI: 10.1108/00483480310467624.

Moghaddas, Seyedeh Zeinab, Masoumeh Tajafari, and Mohsen Nowkarizi (June 2020). "Organizational empowerment: A vital step toward intrapreneurship". en. In: *Journal of Librarianship and Information Science* 52.2. Publisher: SAGE Publications Ltd, pp. 529–540. ISSN: 0961-0006. DOI: 10.1177/0961000619841658. URL: https://doi.org/10.1177/0961000619841658.

Schießl, Nina (2015). "Erstellung des Befragungsinstruments". de. In: *Intrapreneurship-Potenziale bei Mitarbeitern: Entwicklung, Optimierung und Validierung eines Diagnoseinstruments*. Ed. by Nina Schießl. Innovation und Entrepreneurship. Wiesbaden: Springer Fachmedien, pp. 63–86. ISBN: 978-3-658-09428-7. DOI: `10.1007/978-3-658-09428-7_6`. URL: `https://doi.org/10.1007/978-3-658-09428-7_6`.

Tutz, Gerhard and Harald Binder (Aug. 2007). "Boosting ridge regression". In: *Computational Statistics & Data Analysis* 51.12, pp. 6044–6059. ISSN: 0167-9473. DOI: `10.1016/j.csda.2006.11.041`. URL: `https://www.sciencedirect.com/science/article/pii/S0167947306004749`.

Chapter 5

Sparse-Group Boosting with Balanced Selection Frequencies: A Simulation-Based Approach and R Implementation

This chapter introduces the CRAN package 'sgboost', which implements sparse-group boosting alongside tools for model interpretation. Key features include group-wise coefficient path visualization and a sparse group-variable importance measure. The package also provides practical guidance and R code for applying sparse-group boosting in real-world settings. Additionally, it presents and implements the group balancing algorithm, designed to ensure fair group selection in scenarios with unequal group sizes and structures. This method enhances interpretability and reduces bias in grouped variable selection

Contributing article:

Obster, F., & Heumann, C. (2024). "Sparse-Group Boosting with Balanced Selection Frequencies: A Simulation-Based Approach and R Implementation". *ArXiv e-prints* *arXiv: 2405.21037*

Author contributions:

The manuscript and R package "sgboost" were written by Fabian Obster. Christian Heumann added valuable input and proofread the manuscript.

Sparse-Group Boosting with Balanced Selection Frequencies: A Simulation-Based Approach and R Implementation

Fabian Obster [a] [b] and Christian Heumann[b]

aDepartment of Business Administration, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, Bavaria, Germany
bDepartment of Statistics, LMU Munich, Ludwigstr. 33, Bavaria, Germany

ABSTRACT
This paper introduces a novel framework for reducing variable selection bias by balancing selection frequencies of base-learners in boosting and introduces the `sgboost` package in R, which implements this framework combined with sparse-group boosting. The group bias reduction algorithm employs a simulation-based approach to iteratively adjust the degrees of freedom for both individual and group base-learners, ensuring balanced selection probabilities and mitigating the tendency to over-select more complex groups. The efficacy of the group balancing algorithm is demonstrated through simulations. Sparse-group boosting offers a flexible approach for both group and individual variable selection, reducing overfitting and enhancing model interpretability for modeling high-dimensional data with natural groupings in covariates. The package uses regularization techniques based on the degrees of freedom of individual and group base-learners. Through comparisons with existing methods and demonstration of its unique functionalities, this paper provides a practical guide on utilizing sparse-group boosting in R, accompanied by code examples to facilitate its application in various research domains.

1. Introduction

Regularized regression is used to model high-dimensional data to reduce the risk of overfitting and to perform variable selection. In many cases, covariates have natural groupings, such as with gene data or categorical data often found in survey data. In such cases, one may want to select whole groups of variables or just individual parts. Sparse-group boosting is a powerful statistical method that extends classical boosting methods through the incorporation of structured sparsity. This structure is particularly useful in high-dimensional settings where predictors exhibit natural groupings. 'sgboost' implements the sparse-group boosting in R and other useful functions for sparse group model interpretation unique to boosting, and visualization for group and individual variable selection. The package is available on CRAN [13]. Despite their utility and predictive performance, existing variable selection techniques

---

often lack explicit mechanisms to balance sparsity within and between groups, making sgboost as an implementation of sparse-group boosting a valuable contribution to the field of variable selection.

To address a well-known limitation in boosting methods—namely, the tendency to favor larger or more flexible groups due to inherent selection bias - we introduce a group balancing algorithm. This algorithm employs a simulation-based approach to iteratively adjust the degrees of freedom assigned to each group, thereby equalizing the selection probabilities under the null hypothesis of no association. By using the shrinkage of the RSS through the degrees of freedom in ridge regression, the algorithm ensures that groups of differing sizes are penalized appropriately. This adjustment mitigates over-selection of complex groups and enhances the overall interpretability and fairness of the model. The resulting group balancing can be combined with sparse-group boosting and is integrated into sgboost via the `balance()` function, representing a significant methodological advancement in achieving balanced variable selection across heterogeneous groups.

The increasing availability of high-dimensional datasets such as in economics, climate research, and bioinformatics, where variable selection plays a crucial role and group structures are relevant, motivates the implementation of sgboost. Many real-world datasets contain naturally predefined groups, such as geographical regions in climate modeling, questionnaire sections in survey-based research or gene data. Traditional boosting methods struggle with these structures, either selecting too many irrelevant variables or failing to capture group-level effects. By explicitly incorporating the group structure through two-level sparsity, sgboost addresses these challenges while also enhancing predictive accuracy and model interpretability.

Sparse-group boosting [15] is an alternative method to sparse-group lasso [19], employing boosted Ridge regression. Although there are many methods of variable selection, most focus on group selection only, e.g. [12], [9] and [23], or individual variable selection e.g. [1], [22] and [6]. However, it should be noted, that in some cases of group variable selection with overlapping groups, one could also end up with sparse-group variable selection. There are not many R packages implementing sparse-group variable selection methods. There is 'SGL' [18] implementing [19], 'sparsegl' [10] with a faster implementation of the sparse-group lasso as well as "grpreg" [4] implementing the group exponential least absolute shrinkage and selection operator (GEL) [2], the Composite minimax concave penalty (cMCP) [3] and the group bridge [9].

The goal of this paper is to provide a practical guide including the code on how to use sparse-group boosting in R and get the most out of the method. The code is presented within this manuscript and can also be found also on GitHub together with the used dataset (https://github.com/FabianObster/sgboost-introduction).

While the mboost package [8] already allows for structured regularization, sgboost is specifically designed for structural sparsity and simplifies the application of sparse-group boosting through a dedicated formula constructor, enhanced visualization, and interpretability tools.

To demonstrate the relevance of sgboost, we focus on two types of datasets: simulated data designed to reflect real-world sparsity structures and real-world data where sparse-group boosting provides clear advantages. In climate economics, the willingness of farmers to take adaptive measures against climate change is affected by multiple interdependent factors such as climatic/weather patterns, market conditions, and agronomic factors. These independent variables can be sorted into natural groups (e.g., climatic

variables, economic indicators, and agronomic environment), making such data ideal for sparse-group boosting. This paper explores how sgboost can be applied and performs in these settings, highlighting its practical benefits.

2. Methods

Throughout this paper, we consider a grouped dataset $X \in \mathbb{R}^{n \times p}$ with G groups and each group g contains p_g variables. We also consider a generalized linear regression setting, such as explained in [21] and [11] and their extended variations for ridge regression with a similar notation as in [20]. For simplicity, we primarily illustrate the method using linear ridge regression; however, the approach can be readily extended to generalized linear models. We refer to the parameter vector as β and to its estimate as $\widehat{\beta}$. The linear predictor with the response function $g^{-1}(\cdot)$ together form the conditional expectation of the response variable $\mathbb{E}[y|X] = \mu = g^{-1}(X\beta)$. As we can subset the design matrix, we can estimate a model using a subset of the design matrix, which we denote as $\widehat{\beta}_{V_g}$. Note that with this notation $\widehat{\beta}_{V_g}$ is not the same as the subset of $\widehat{\beta}$ using the index set V_g, but rather separate estimates. We will also use the notion of a Ridge hat matrix as defined in [20] using the penalty term $\lambda \geq 0$: $H^\lambda = X(X^T X)^{-1}X^T$ and the degrees of freedom for which we use the definition of $\mathrm{df}(\lambda) = \mathrm{tr}(2H^\lambda - (H^\lambda)^2)$.

2.1. Sparse-Group Boosting Framework

sgboost extends traditional boosting by incorporating structured sparsity, combining component-wise and group-wise selection. This approach is particularly beneficial in high-dimensional settings such as correlated independent variables. Unlike standard boosting, which selects individual variables in isolation, sgboost allows the selection of groups or individual variables. This way, entire groups of predictors can be included or excluded, improving interpretability and predictive performance. The same holds for individual variables.

We define $p + G$ candidate sets denoted as $(V_l \subseteq \{1, ..., p\})_{l \leq p+G}$. Each candidate set describes the indices of the variables to be considered as one group. This yields $p + G$ submatrices of the design matrix X_{V_l} only containing the columns corresponding to the index set.

- The first p are individual base-learners only containing one variable:

$$V_l = \{l\} \text{ for } l \leq p,$$

- and the remaining G are group base-learners with group size p_l:

$$V_l = \{(v_l)_1, ..., (v_l)_{p_l}\} \subseteq \{1, ..., p\}, \text{ for } l > p:$$

Through V_l, $l \geq p$, the group structure is defined using no overlapping groups. Through this bi-level structure, within-group sparsity and between-group sparsity can be balanced. By using Ridge Regression, regularization is controlled through the degrees of freedom constraint, regulating sparsity levels. The optimal base-learner is selected based on the residual sum of squares (RSS), yielding the most informative structure either via a group or individual variable at each step.

3

Algorithm 1 Sparse-Group L^2 Boosting Algorithm

1: **Initialize:** $m \leftarrow 0$, $\widehat{\beta}^{[0]} \leftarrow \mathbf{0}_p$, $\widehat{\mu}^{[0]} \leftarrow X\widehat{\beta}^{[0]}$

2: **while** $m < M$ **do**

3: $\quad m \leftarrow m + 1$

4: $\quad$ **for** each candidate set V_l, $l \leq p + G$ **do**

5: $\quad\quad$ Compute residuals: $\widehat{u}^{[m-1]} \leftarrow y - \widehat{\mu}^{[m-1]}$

6: $\quad\quad$ Fit Ridge regression:

$$\widehat{\overline{\beta}}_{V_l}^{[m]} = ((X_{V_l})^T X_{V_l} + \lambda_l I_p)^{-1} (X_{V_l})^T (\widehat{u}^{[m-1]})$$

7: $\quad\quad$ Set regularization parameter λ_l:

$$\lambda_l = \begin{cases} \lambda_l : \mathrm{df}(\lambda_l) = \mathrm{tr}(2H_{V_l}^\lambda - (H_{V_l}^\lambda)^2) = \alpha, & l \leq p \\ \lambda_l : \mathrm{df}(\lambda_l) = \mathrm{tr}(2H_{V_l}^\lambda - (H_{V_l}^\lambda)^2) = 1 - \alpha, & l > p \end{cases} \tag{1}$$

8: $\quad$ **end for**

9: $\quad$ Select candidate set:

$$l^* = \arg\min_{l \leq L} (\widehat{u}^{[m-1]} - X_{V_l}\widehat{\overline{\beta}}_{V_l})^T (\widehat{u}^{[m-1]} - X_{V_l}\widehat{\overline{\beta}}_{V_l})$$

10: $\quad$ **for** all $l \leq p + G$ **do**

11: $\quad\quad$ Update coefficients:

$$\widehat{\beta}_{V_l}^{[m]} = \begin{cases} \widehat{\beta}^{[m-1]} + \nu\widehat{\overline{\beta}}_{V_{l^*}}, & l = l^* \\ \widehat{\beta}^{[m-1]}, & l \neq l^* \end{cases}$$

12: $\quad$ **end for**

13: $\quad$ Update estimate:

$$\widehat{\mu}^{[m]} = X\widehat{\beta}^{[m]}$$

14: **end while**

15: **Output:** $\widehat{\beta}^{[M]}$

4

The same algorithm can be used to fit (generalized) linear models by replacing the L^2 loss function with the modified loss function $\mathcal{L}$. This yields for individual base-learners $l \leq p$

$$\mathcal{L}_{V_l}^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l}) + \alpha \lambda_l (\beta_{V_l})^T \beta_{V_l},$$

and for group base-learners $l > p$

$$\mathcal{L}_{V_l}^{[m]} = -\sum_{i=1}^{n} \ell_i(\widehat{\beta}^{[m-1]} + \beta_{V_l}) + (1-\alpha) \lambda_l (\beta_{V_l})^T \beta_{V_l}.$$

2.2. Group adjustment

Variable selection bias can occur in the presence of grouped variables, such as categorical or functional data, making the definition of the degrees of freedom $df(\lambda) = \text{tr}(2H^\lambda - (H^\lambda)^2)$ preferable [7]. However, group selection bias can still occur because of the group size. The same issue occurs in the sparse group lasso of the group bridge, which is met by using group standardization depending on the type of regularization, such as $\sqrt{p_g}$ [19] [3]. Many algorithms use such an adjustment, which is also referred to as outer adjustment [5]. This is to prevent an over-selection of groups with larger group sizes. Unlike traditional methods that use fixed penalties or shrinkage parameters, this algorithm dynamically adjusts selection probabilities through repeated sampling, enabling data-driven balancing. To overcome this issue, we introduce a simulation-based algorithm that balances the selection chance of one group over another by using the degrees of freedom.

Assume we have G groups, described by the index sets $V_1, ..., V_G$, with group sizes $p_1, ..., p_g$. Denote the scaling vector for the degrees of freedom as $d = (d_1, ..., d_G)$, where each group g has its own value for the degrees of freedom d_g for $g \leq G$.

Algorithm 2 Group balancing algorithm

Initialize: Set $r = 0$ and $d_g^* = d_g^{[1]} \equiv c$ with constant $c \in]0, 1[$ for all $g \leq G$. A reasonable starting value is $c = 0.5$.

2: **for** $r \leq R$ **do**

$\quad r \leftarrow r + 1$

4: $\quad$ Simulate K versions of the outcome variable $y^{(k)}$, e.g., $y^{(k)} \sim \mathcal{N}(0_n, I_n)$ for $k \leq K$.

$\quad$ **for** $k \leq K$ **do**

6: $\quad\quad$ Fit the learning algorithm $f : X \to y^{(k)}$ with the degrees of freedom d^* to obtain the fitted model $\widehat{f}^{(k)}$.

$\quad$ **end for**

8: $\quad$ Retrieve the activation vector $\left(s_1^{(k)}, ..., s_G^{(k)}\right) \in \{0, 1\}^G$ for each $\widehat{f}^{(k)}$, indicating selected groups. If a group is selected, the value one is assigned; if not, then the value zero.

$\quad$ Compute the average selection frequency vector:

$$\bar{s} = \left(\bar{s}_1, ..., \bar{s}_G\right) = \left(\frac{1}{K} \sum_{k=1}^{K} s_1^{(k)}, ..., \frac{1}{K} \sum_{k=1}^{K} s_G^{(k)}\right)^T.$$

10: $\quad$ Compute the error vector:

$$c^{[r]} = \left(\frac{1}{G}, ..., \frac{1}{G}\right)^T - \bar{s}.$$

$\quad$ **for** $g \leq G$ **do**

12: $\quad\quad$ Update:

$\quad\quad$ **if** $\sum_{g=1}^{G} (c_g^{[r]})^2 < \sum_{g=1}^{G} (c_g^*)^2$ **then**

14:

$$d_g^* = d_g^{[r-1]}$$
$$d_g^{[r]} = d_g^* + \nu c^{[r]}$$

$\quad\quad$ **else**

16:

$$\nu = \gamma \nu$$
$$d_g^{[r]} = (1 - \eta)d_g^* + \eta(d_g^{[r-1]} + \nu c^{[r]}).$$

$\quad\quad$ **end if**

18: $\quad$ **end for**

$\quad$ **end for**

20: **Return:** d^* as the degrees of freedom scaling vector.

Note that it is sufficient to run the boosting algorithm for only one step instead of fitting the whole algorithm, to achieve the balance, as the algorithm does not depend on the actual outcome variable. This allows for a highly efficient estimation of group preference without fully fitting the model. This compensates for simulating many repetitions of the outcome variable and refitting for each sample

The general idea is to decrease the degrees of freedom for over-selected groups and increase the degrees of freedom for under-selected groups. The step size is proportional to the imbalance, meaning strongly imbalanced groups are adjusted more than slightly imbalanced groups, and is multiplied by the learning rate ν, which impacts how far the update goes away from the current estimate d^*. A larger ν leads to larger corrections, hence fewer necessary iterations, but may cause oscillations or overshooting, especially in small-sample or highly collinear settings. Choosing an appropriate learning rate is therefore a trade-off between speed of convergence and stability. Also, K, the number of samples of the outcome variable, increases stability and should also affect the choice of the learning rate. If the algorithm overshoots and the overall imbalance increases, a convex combination between the current best estimate and the updated parameter is used, where η is the mixing parameter. Also, the learning rate is reduced by $\gamma \in]0, 1[$, e.g. 0.9 to avoid overshooting in future steps. This approach incorporates new information while preserving the information from the previous best estimate, balancing exploration and robustness. The algorithm can be stopped after a fixed number of iterations R or if $\sum_{g=1}^{G}(c_g^{[r]})^2$ is smaller than some predefined value. The algorithm does not yield a unique solution, which depends strongly on the initialization of c. The existence of a solution that balances the selection frequencies is guaranteed by the mean-value theorem and the law of large numbers if $K \to \infty$. One can verify that $\mathrm{df}(\lambda_g) \to 0$ implies $\bar{s}_g \to 0$ and $\mathrm{df}(\lambda_g) \to 1$ implies $\bar{s}_g \to z > 0$. Furthermore, the residual sum of squares is monotonous and continuous in $df(\lambda)$. Therefore, $\frac{1}{K}\sum_{k=1}^{K} s_g^{(k)} = P(RSS(\lambda_g) = max_{l \leq G}(RSS(\lambda_l)))$ is also monotonous and continuous in $\mathrm{df}(\lambda_g)$. A unique solution could be achieved by fixing the degrees of freedom for one base-learner and only updating the others. Another variation of the algorithm is to perform the algorithm by updating the ridge regularization parameter λ for each group instead of the degrees of freedom. The algorithm can be used for group boosting, but also for sparse-group boosting by expanding the group index set to include also $\tilde{V} = 1, ..., p, V_1, ...V_G$, leading to overlapping groups. In this case, one could also update the error vector to $c^{[r]} = \left(\frac{\alpha}{p+G}, \overset{p}{\cdots}, \frac{\alpha}{p+G}, \frac{1-\alpha}{p+G}, \overset{G}{\cdots}, \frac{1-\alpha}{p+G}\right)^T - \bar{s}$, instead of mixing the degrees of freedom. In this case, α would have a natural interpretation, though the odds of an individual base-learner being selected over a group base-learner as $\frac{\alpha}{1-\alpha}$. This would make the choice of α much easier. Then, $\alpha = 0$ would still correspond to group boosting, $\alpha = 1$ to componentwise boosting. The case of $\alpha = 0.5$, would lead to equal selection frequencies of each base-learner, regardless of the group size and type of base-learner.

3. Results

We first simulate the sample data and corresponding group structure with 40 equal-sized groups to show the sparse-group boosting workflow. Based on a linear regression model we simulate the response variable y as part of the data.frame with $n = 100$ observations and $p = 200$ predictor variables (each group is formed by 5 predictors).

```
beta <- c(
  rep(5, 5), c(5, -5, 2, 0, 0), rep(-5, 5),
```

```
    c(2, -3, 8, 0, 0), rep(0, (200 - 20))
)
X <- matrix(data = rnorm(20000, mean = 0, sd = 1), 100, 200)
df <- data.frame(X) %>%
    mutate(y = X %*% beta+rnorm(100, mean = 0, sd = 1)) %>%
    mutate_all(function(x){as.numeric(scale(x))})
group_df <- data.frame(
    group_name = rep(1:40, each = 5),
    variable_name = colnames(df)[1:200]
)
```

3.1. Defining the model

Now we use the group structure to describe the sparse group boosting formula with the function `create_formula()`. We only need the data.frame() describing the group structure. It should contain two variables, one indicating the name of the variable in the modeling data (`var_name`), and one indicating the group it belongs to (`group_name`). Additionally, we need to pass the mixing parameter `alpha` and the name of the outcome variable.

```
sgb_formula <- create_formula(
    alpha = 0.4, group_df = group_df, outcome_name = "y",
    group_name = "group_name", var_name = "variable_name")
```

This function returns an R-formula consisting of p model terms defining the individual base-learners and G group base-learners.

```
labels(terms(sgb_formula))[[1]]
## bols(X1, df = 0.4, intercept = FALSE)
labels(terms(sgb_formula))[[201]]
## bols(X1, X2, X3, X4, X5, df = 0.6, intercept = FALSE)
```

3.2. Fitting and tuning the model

sgboost is to be used in conjunction with the **mboost** package, which provides many useful functions and methods that can also be used for sparse-group boosting models. Now we pass the formula to `mboost()` and use the arguments as seems appropriate. The main hyperparameters are `nu` and `mstop`. For model tuning, the function `cvrisk` can be used and plotted. Running the cross-validation/bootstrap in parallel can speed up the process.

```
sgb_model <- mboost(
    formula = sgb_formula, data = df,
    control = boost_control(nu = 1, mstop = 600)
)
cv_sgb_model <- cvrisk(sgb_model)
mstop(cv_sgb_model)
## 204
plot(cv_sgb_model
```
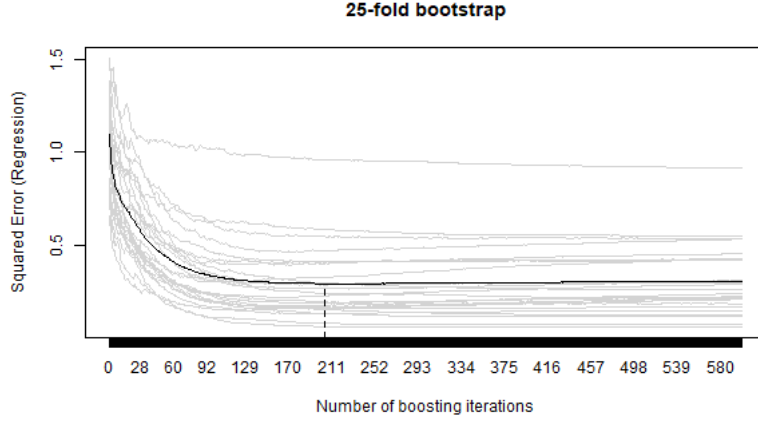
Figure 1. Out of sample error depending on the boosting iteration

In this example, the lowest out-of-sample risk is obtained at 204 boosting iterations, so we only use the first 204 updates for the final model.

3.3. Interpreting and plotting a sparse-group boosting model

`sgboost` has useful functions to understand sparse-group boosting models, and reflects that the final model estimates of a specific variable in the dataset can be attributed to group base-learners as well as individual base-learners depending on the boosting iteration.

3.3.1. Variable importance

A good starting point for understanding a sparse-group boosting model is the variable importance. In the context of boosting, the variable importance can be defined as the relative contribution of each predictor to the overall reduction of the loss function (negative log-likelihood). `get_varimp()` returns the variable importance of each base-learner/predictor selected throughout the boosting process. In the case of the selection of an individual variable - call it x_1 - as well as the group it belongs to -$x_1, x_2, ...x_p$ -, both base-learners (predictors) will have an associated variable importance as defined before. This allows us to differentiate between the individual contribution of x_1 as its own variable and the contribution of the group x_1 belongs to. It is impossible to compute the aggregated variable importance of x_1 as it is unclear how much x_1 contributes to the group. However, the aggregated coefficients can be computed using `get_coef()`, which also returns the aggregated importance of all groups vs. all individual variables in a separate data.frame. With `plot_varimp()` one can visualize the variable importance as a barplot. Since group sizes can be large, the function allows for cutting of the name of a predictor after `max_char_length` characters. One can indicate the maximum number of predictors to be printed through `n_predictors` or through the minimal variable importance a predictor has to have through `prop`. Through both parameters, the number of printed entries can be reduced. Note, that in this case, the relative importance of groups in the legend is based only on the plotted variables and not the ones removed. Adding information about the direction of effect sizes, one could add arrows behind the bars [14]. For groups, one can use the aggregated coefficients from `get_coef()`.

9

```
slice(get_varimp(sgb_model =sgb_model_linear)$varimp,1:5)
```

```
# A tibble: 5  6
  reduction blearner       predictor   selfreq type  relative_
                                                     importance
      <dbl> <chr>          <chr>         <dbl> <chr>      <dbl>
1    0.297  bols(X1, X2,... X1, X2, ... 0.206  group      0.301
2    0.288  bols(X18, in... X18         0.0196 indi...    0.292
3    0.230  bols(X11, X1... X11, X12... 0.25   group      0.233
4    0.0414 bols(X7, int... X7          0.0784 indi...    0.0419
5    0.0392 bols(X6, int... X6          0.0833 indi...    0.0397
```
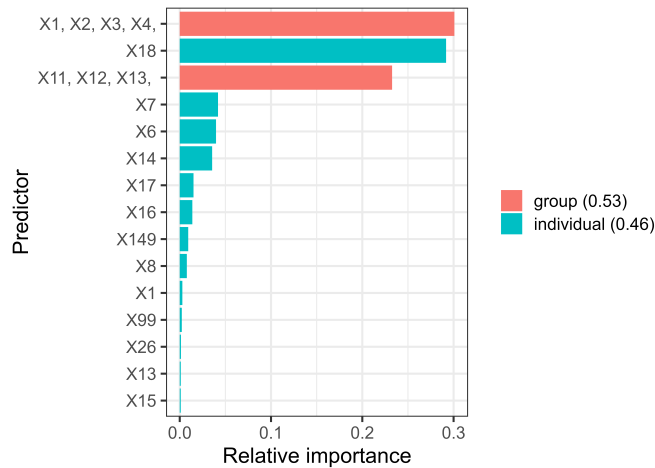
```
get_varimp(sgb_model = sgb_model_linear)$group_importance
```

```
    # A tibble: 2 × 2
  type        importance
  <chr>            <dbl>
1 group           0.534
2 individual      0.466
```

```
plot_varimp(sgb_model = sgb_model_linear, n_predictors = 15)
```


Figure 2. Variable importance of the sparse-group boosting model for simulated data. The variable labels in the groups are cut off after 15 characters by defalut.

In this example, we see that both individual variables and groups were selected and contributed to the reduction of the loss function. The most important predictor is the first group, followed by variable 18, and then by group three. This is in line with what was simulated, as variable 18 has the biggest beta value, and groups one and three are full groups, meaning all variables within the groups have a non-zero beta coefficient. Groups two and four have within-group sparsity, therefore, they were selected as individual variables rather than groups.

3.3.2. Model coefficients

The resulting coefficients can be retrieved through `get_coef()` In sparse-group boosting models, a variable in a dataset can be selected as an individual variable or through a group. Therefore, there can be two associated effect sizes for the same variable. This function aggregates both and returns them in a data.frame sorted by the effect size `'effect'`.

```
slice(get_coef(sgb_model = sgb_model)$raw,1:5)
```

```
# A tibble: 5 × 5
  variable effect blearner        predictor                 type
  <chr>     <dbl> <chr>           <chr>                     <chr>
1 X18       0.364 bols(X18, int... X18                      individual
2 X5        0.250 bols(X1, X2, ... X1, X2, X3, X4, X5       group
3 X15      -0.249 bols(X11, X12... X11, X12, X13, X14, X15  group
4 X4        0.234 bols(X1, X2, ... X1, X2, X3, X4, X5       group
5 X11      -0.228 bols(X11, X12... X11, X12, X13, X14, X15  group
```

```
slice(get_coef(sgb_model = sgb_model)$aggregate,1:5)
```

```
# A tibble: 5 × 4
 variable effect learner          predictor
  <chr>     <dbl> <chr>            <chr>
1 X18       0.364 bols(X18, inte...  X18
2 X15      -0.272 bols(X11, X12,...;  X11, X12, X13, X14, X15; X15
                  bols(X15, inte...
3 X5        0.250 bols(X1, X2,...    X1, X2, X3, X4, X5
4 X4        0.234 bols(X1, X2,...    X1, X2, X3, X4, X5
5 X13      -0.230 bols(X11, X12,...;  X11, X12, X13, X14, X15; X13
                  bols(X13, inte...
```

We see that the effect sizes differ between the two perspectives. The variable X15, for example, has a more extreme model coefficient of -0.272 in the aggregated case compared to the coefficient of -0.249 derived only from the group base-learner. Consequently, the ordering also differs. X11 has a greater absolute model coefficient from the group than X13, but in the aggregated version, the absolute model coefficient of X13 exceeds the one of X11.

3.3.3. Plotting model coefficients and importance

With `plot_effects()` we can plot the effect sizes of the sparse-group boosting model in relation to the relative importance to get an overall picture of the model. Through the parameter `'plot_type'` one can choose the type of visualization. `'radar'` refers to a radar plot using polar coordinates. Here, the angle is relative to the cumulative relative importance of predictors, and the radius is proportional to the effect size. `'clock'` does the same as `'radar'` but uses clock coordinates instead of polar coordinates. `'scatter'` uses the effect size as the y-coordinate and the cumulative relative importance as the x-axis in a classical Scatter plot.
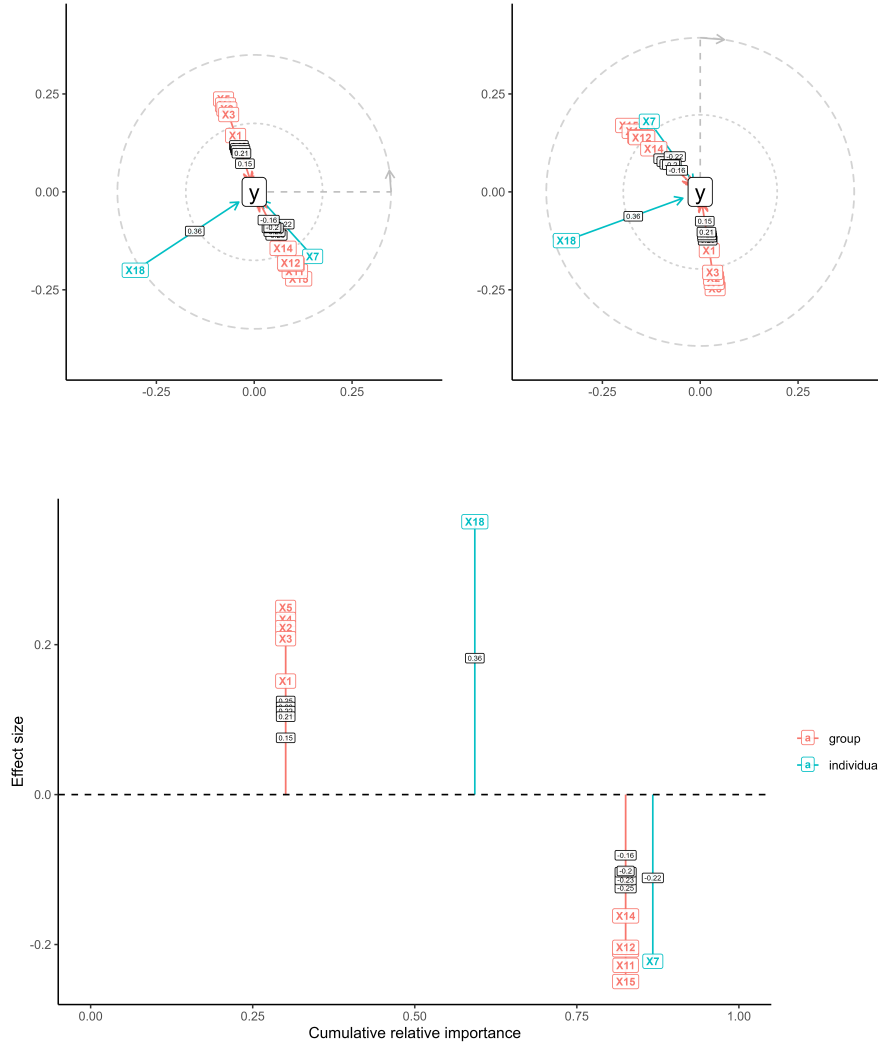
```
plot_effects(sgb_model = sgb_model, n_predictors = 5,
             base_size = 10)
```

```
plot_effects(sgb_model = sgb_model, n_predictors = 5,
             plot_type = "clock", base_size = 10)
plot_effects(sgb_model = sgb_model, n_predictors = 5,
             plot_type = "scatter", base_size = 10)
```


Figure 3. Three visualizations of Effect size vs. relative importance of individual and group base-learners

3.3.4. Coefficient path

`plot_path` calls `get_coef_path()` to retrieve the aggregated coefficients from a `mboost` object for each boosting iteration and plots it, indicating if a coefficient was updated by an individual variable or group.

```
plot_path(sgb_model = sgb_model)
```
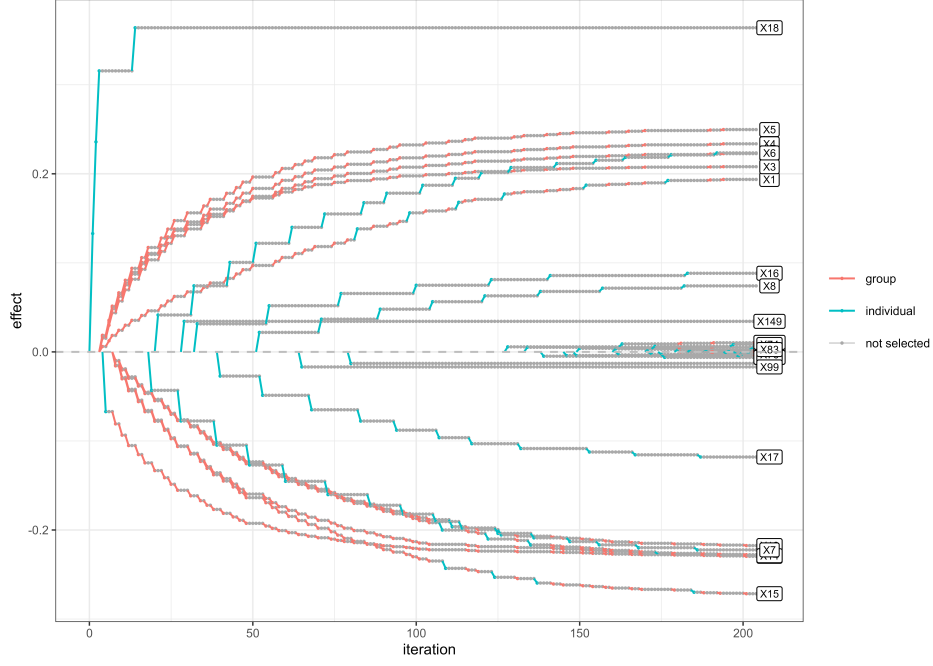
Figure 4. Coefficient path of a sparse-group boosting model with simulated data

In the coefficient path shown in Figure 4, we see the change in model coefficients. Since the path shows the aggregated model coefficients, the path of one variable in the dataset may have both colors. This is the case with variable X1 which was first updated through the group and then also as an individual variable or with variable X15 in reverse order.

3.4. Real data

In this section, we will fit a sparse-group boosting model with **sgboost** to a real dataset. We will use behavioral ecological data and an associated group structure [16] to explain whether farmers in Chile and Tunisia are planning to take adaptive measures against climate change in the following years. We will use a logistic regression model for this binary decision. The data consists of 14 groups and 84 variables for the 801 farmers. Groups include vulnerability to climate change [17], social, biophysical, and economic assets, as well as perceptions of the farmers. After loading the data and group structure, we create the formula with mixing parameter $\alpha = 0.3$. Then, we pass the formula to mboost() with 1000 boosting iterations and a learning rate of 0.3.

```
model_df <- readRDS('model_df.RDS') %>%
  mutate_at(index_df$col_names, factor)
index_df <- readRDS('index_df.RDS')
sgb_formula <- create_formula(
  group_df = index_df, var_name = 'col_names',
  group_name = 'index', outcome_name = 'S5.4'
)
model <- mboost(
  sgb_formula, data = model_df,
  family = Binomial(link = 'logit'),
  control = boost_control(mstop = 1000, nu = 0.3)
```

13

```
)
cv_model <- cvrisk(model)
model <- model[mstop(cv_model)]
```

The model is stopped early after 466 boosting iterations. We examine the coefficient path and see that in the early stage, individual base-learners were dominantly selected like the variable 'S1.8b or 'S8.11 river' which indicates whether river irrigation is used. Many of the variables were first included as individual variables and later also through group base-learners like 'S8.1b' or 'S2.5c proximity' (Proximity to extreme weather events), which we also saw in the simulated data.

```
plot_path(model)
```
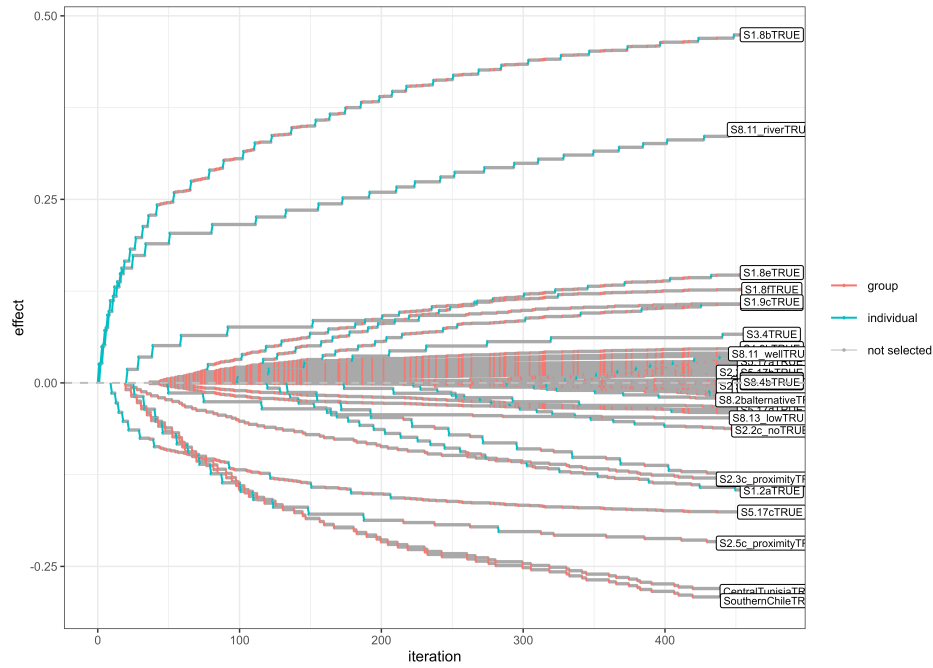

Figure 5. Coefficient path using the ecological dataset

In figure 6, we look at the variable importance with the default values, plotting all 27 selected predictors of which 8 are groups, the latter having a relative variable importance of 22 percent. The most important base-learner is the individual variable 'S1.8b', indicating whether farming journals are being used and the most important group is the social asset group, followed by the group consisting of the four considered regions.

```
plot_varimp(model)
```

14
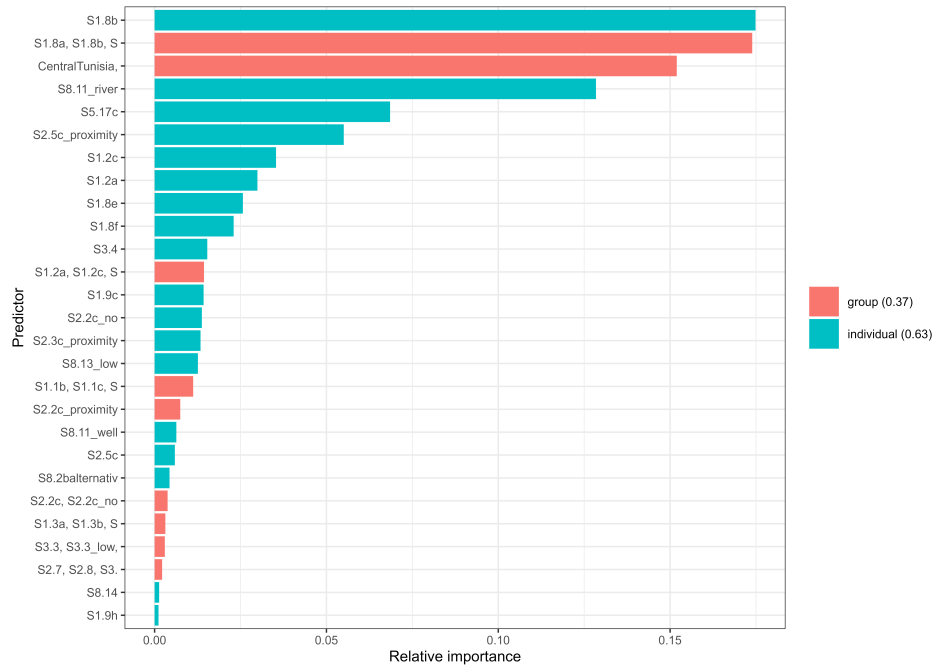
Figure 6. Variable importance using the ecological dataset

Plotting the effect sizes of all predictors having a relative importance of greater than 1.5 percent shows the tendency for more important variables to have greater absolute effect sizes. For readability, we set the number of printed characters per variable to 6 and use the 'scatter' version of the plot.

```
plot_effects(
    model, plot_type = 'scatter',
    prop = 0.015, max_char_length = 6
)
```
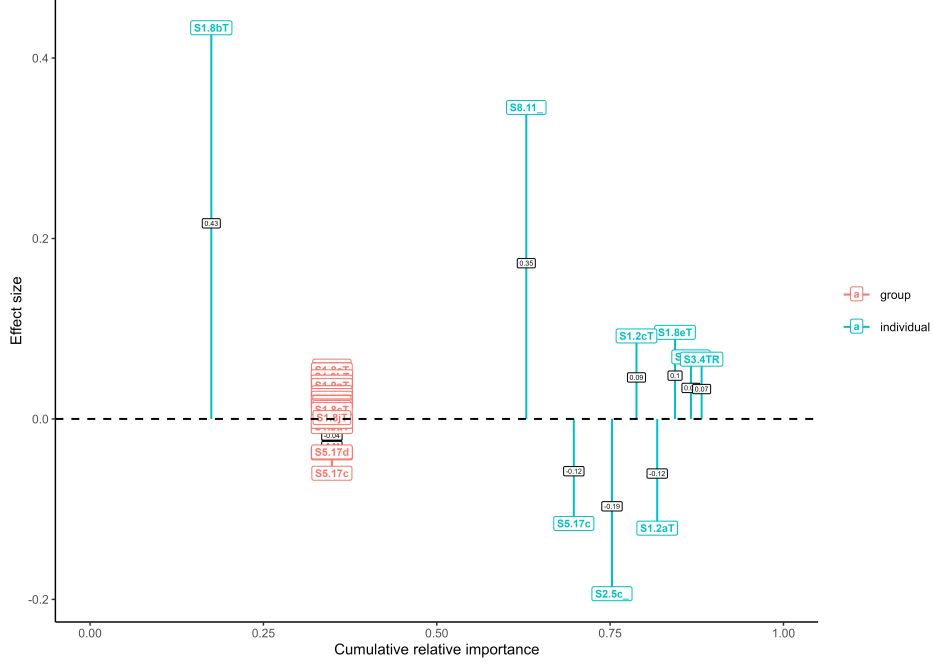
Figure 7. Coefficient plot using the ecological dataset

3.5. *Group selection bias and balance*

The function `balance()` returns the optimal degrees of freedom for each baselearner, such that all groups have equal selection chances if the outcome is not associated with any group. This is especially important in genetic research as group sizes based on genes may vary strongly. Not controlling for group sizes may lead to strong false detection because of a bias towards more complex groups eg. genes. To illustrate this problem, four scenarios are considered. The first three scenarios have three groups, where the first group is a categorical predictor with three categories, the second group is a categorical predictor with two categories, and the third group is a numerical variable simulated with a standard normal distribution. The fourth scenario consists of two groups, the first having 46 members and the second having 4. In the first scenario, the sample size is 50; in the second and third scenarios, the sample size is 500; in the fourth scenario, the sample size is 30, leading to $p > n$. In scenario 3 the outcome variable is i.i.d gamma distributed with shape one and rate 1, Scenarios one, two, and four have standard normal outcomes. Table 1 shows the selection frequencies in these scenarios for each group, for group boosting with three versions of group boosting: one with equal penalties ($\lambda = 0.1$), one with equal degrees of freedom df(λ) = 0.5, and one with the degrees of freedom based on the `balance()` function which implements Algorithm 2. The default settings of 3000 repetitions of i.i.d standard normal outcomes, 20 iterations, a learning rate of 0.5, and a reduction factor of 0.9.

The results suggest, that generally ridge regression with equal penalties leads to the greatest group imbalance, and equal degrees of freedom lead to relatively lower imbalances, especially when two categorical predictors are compared, and the only group adjustment-based group boosting balances the the selection frequencies in all scenarios. In scenario 4, equal lambda only selects the larger group, and in equal degrees of freedom, the chance of the larger group being selected is 2.9 times higher than the smaller

16

one. equal lambda is better at balancing a binary variable with a numerical variable compared to equal df, where equal df is better at balancing group sizes compared to equal lambda. The distribution of the outcome variable seems to not play a great role as the results in scenarios two and three are quite comparable for all three models, which makes the balancing algorithm robust, even if the distribution of the error is not known. Comparing scenario one with two, the balancing algorithm works better with a greater sample size.

Table 1. Selection frequencies in group boosting with degrees of freedom adjustment (group adjustment) compared to ridge regression with equal degrees of freedom (equal df) and equal penalty term (equal lambda). The degrees of freedom used as group adjustment are shown in brackets.

Scenario	group	equal lambda	equal df	group adjustment
1	1	0.699	0.453	0.345 (df=0.377)
1	2	0.157	0.364	0.341 (df=0.406)
1	3	0.144	0.183	0.314 (df=0.717)
2	1	0.701	0.407	0.337 (df=0.397)
2	2	0.15	0.419	0.339 (df=0.352)
2	3	0.149	0.174	0.324 (df=0.751)
3	1	0.695	0.417	0.338 (df=0.397)
3	2	0.155	0.408	0.326 (df=0.352)
3	3	0.15	0.175	0.336 (df=0.751)
4	1	1	0.744	0.518 (df=0.394)
4	2	0	0.256	0.482 (df=0.606)

4. Discussion

4.1. Sparse-group boosting

'sgboost' when applied to high-dimensional grouped data such as ecological data on climate adaptation, reveals meaningful patterns, such as socio-economic and biophysical variables, thereby providing actionable insights for policy and practice. Moreover, the integration of comprehensive visualization tools—such as variable importance plots, coefficient paths, and effect size charts—enhances the interpretability of the models, making it easier for practitioners to understand the contributions of different predictors.

4.2. Group bias

Attempts to reduce selection bias in boosting have been made through equal degrees of freedom using the definition $\mathrm{df}(\lambda) = \mathrm{tr}(2H^\lambda - (H^\lambda)^2)$ [7] The results of our simulations highlight a fundamental issue in group boosting: a systematic selection bias towards more complex base-learners. This phenomenon follows the principle that "whoever shouts the loudest is rarely right," meaning that larger or more flexible groups are favored in the selection process. The proposed group balancing function `balance()` in sgboost implementing the group balancing algorithm effectively eliminating this bias by using a simulation-based approach to equalize selection probabilities across different groups.
While this issue is particularly evident in group boosting, it is not limited to this setting.

A similar bias can arise in standard boosting when base-learners differ significantly in scale or distribution. This occurs, for example, when comparing binary and numerical variables or in the context of functional regression. The proposed adjustment method provides a systematic way to address these imbalances across various modeling settings and is robust against small sample sizes and varying outcome variable distributions. Even if one assumes a wrong error distribution in the simulation e.g. standard normal, the group adjustment seems to still balance the group selection frequencies if the actual error distribution differs, e.g. gamma errors.

4.3. Limitations of the group balancing algorithm

Despite its benefits, the balancing approach has several challenges: The resampling and iterative adjustments make the method time-intensive. Different combinations of degrees of freedom can yield similar selection frequencies. One way to address this is by fixing the degrees of freedom of a reference base-learner and adjusting only the others. If the learning rate is too high or the number of resamples is too low, the algorithm may fail to converge. To mitigate this, a reduction factor is applied. The method can overshoot adjustments, leading to situations where the ridge regression problem becomes ill-posed (e.g., degrees of freedom approaching zero or leading to negative. This is controlled via predefined bounds (`max_df` and `min_df`) which can be used instead of the too-extreme solution.

While the algorithm increases computational cost, it is important to compare this to alternative approaches. Standard group boosting with equal penalties often requires extensive tuning, including thousands of boosting iterations and 25-fold cross-validation to determine optimal stopping. In contrast, the balancing approach achieves the same computational efficiency when considering this tuning overhead. Furthermore, parallelization can significantly reduce runtime. Additionally, equal degrees of freedom approaches also require multiple λ values to equalize the selection behavior, making the balancing method computationally more efficient in comparison if the λ values are optimized and not the degrees of freedom.

Funding

References

[1] K.N. Berk, *Forward and backward stepping in variable selection*, Journal of Statistical Computation and Simulation 10 (1980), pp. 177–185. Available at https://doi.org/10.1080/00949658008810367, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00949658008810367.

[2] P. Breheny, *The group exponential lasso for bi-level variable selection*, Biometrics 71 (2015), pp. 731–740.

[3] P. Breheny and J. Huang, *Penalized methods for bi-level variable selection*, Statistics and

its interface 2 (2009), pp. 369–380. Available at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904563/`.

[4] P. Breheny, Y. Zeng, and R. Kurth, *grpreg: Regularization Paths for Regression Models with Grouped Covariates* (2024). Available at `https://CRAN.R-project.org/package=grpreg`.

[5] G. Buch, A. Schulz, I. Schmidtmann, K. Strauch, and P.S. Wild, *A systematic review and evaluation of statistical methods for group variable selection*, Statistics in Medicine 42 (2023), pp. 331–352. Available at `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9620`, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9620.

[6] P. Bühlmann and T. Hothorn, *Boosting Algorithms: Regularization, Prediction and Model Fitting*, Statistical Science 22 (2007), pp. 477–505. Available at `https://projecteuclid.org/journals/statistical-science/volume-22/issue-4/Boosting-Algorithms-Regularization-Prediction-and-Model-Fitting/10.1214/07-STS242.full`, Publisher: Institute of Mathematical Statistics.

[7] B. Hofner, T. Hothorn, T. Kneib, and M. Schmid, *A Framework for Unbiased Model Selection Based on Boosting*, Journal of Computational and Graphical Statistics 20 (2011), pp. 956–971. Available at `http://www.tandfonline.com/doi/abs/10.1198/jcgs.2011.09220`.

[8] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, B. Hofner, F. Otto-Sobotka, F. Scheipl, and A. Mayr, *mboost: Model-Based Boosting* (2023). Available at `https://CRAN.Rproject.org/package=mboost`.

[9] J. Huang, S. Ma, H. Xie, and C.H. Zhang, *A group bridge approach for variable selection*, Biometrika 96 (2009), pp. 339–355.

[10] X. Liang, A. Cohen, A.S. Heinsfeld, F. Pestilli, and D.J. McDonald, *sparsegl: An R Package for Estimating Sparse Group Lasso* (2023). Available at `http://arxiv.org/abs/2208.02942`, arXiv:2208.02942 [stat].

[11] P. McCullagh and J.A. Nelder, *Generalized Linear Models (2nd ed.).*, Journal of the American Statistical Association 88 (1993), p. 698. Available at `https://www.jstor.org/stable/2290358?origin=crossref`.

[12] L. Meier, S. Van De Geer, and P. Bühlmann, *The Group Lasso for Logistic Regression*, Journal of the Royal Statistical Society Series B: Statistical Methodology 70 (2008), pp. 53–71. Available at `https://doi.org/10.1111/j.1467-9868.2007.00627.x`.

[13] F. Obster, *sgboost: Sparse-Group Boosting* (2024). Available at `https://CRAN.R-project.org/package=sgboost`.

[14] F. Obster, H. Bohle, and P.M. Pechan, *The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change*, Communications Earth & Environment 5 (2024), pp. 1–12. Available at `https://www.nature.com/articles/s43247-023-01128-2`, Number: 1 Publisher: Nature Publishing Group.

[15] F. Obster and C. Heumann, *Sparse-Group Boosting: Unbiased Group and Variable Selection*, The American Statistician 0 (2024), pp. 1–14. Available at `https://doi.org/10.1080/00031305.2024.2408007`, Publisher: ASA Website _eprint: https://doi.org/10.1080/00031305.2024.2408007.

[16] F. Obster, C. Heumann, H. Bohle, and P. Pechan, *Using interpretable boosting algorithms for modeling environmental and agricultural data*, Scientific Reports 13 (2023), p. 12767. Available at `https://www.nature.com/articles/s41598-023-39918-5`, Number: 1 Publisher: Nature Publishing Group.

[17] P.M. Pechan, H. Bohle, and F. Obster, *Reducing vulnerability of fruit orchards to climate change*, Agricultural Systems 210 (2023), p. 103713. Available at `https://www.sciencedirect.com/science/article/pii/S0308521X2300118X`.

[18] N. Simon, J. Friedman, T. Hastie, and a.R. Tibshirani, *SGL: Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization* (2019). Available at `https://CRAN.R-project.org/package=SGL`.

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *A Sparse-Group Lasso*, Journal

of Computational and Graphical Statistics 22 (2013), pp. 231–245. Available at `http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250`.

[20] W.N.v. Wieringen, *Lecture notes on ridge regression* (2023). Available at `http://arxiv.org/abs/1509.09169`, arXiv:1509.09169 [stat].

[21] S.N. Wood, *Generalized Additive Models: An Introduction with R, Second Edition*, 2nd ed., Chapman and Hall/CRC, New York, 2017 May.

[22] Z. Zhang, *Variable selection with stepwise and best subset approaches*, Annals of Translational Medicine 4 (2016), p. 136. Available at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/`.

[23] N. Zhou and J. Zhu, *Group Variable Selection via a Hierarchical Lasso and Its Oracle Property* (2010). Available at `http://arxiv.org/abs/1006.2871`, arXiv:1006.2871 [stat].

Chapter 6

The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change

This chapter illustrates the applicability of sparse-group boosting in economic and environmental data analysis. The predictive power of the method is compared to other Machine Learning algorithms and the group variable analysis unique to the sparse-group boosting is utilized to answer novel research questions relevant to the applied research field.

Contributing article:

Author contributions:

Fabian Obster and Paul Pechan wrote the manuscript. Fabian Obster conducted the statistical analysis and Machine Learning and wrote parts related to the statistical analysis and its results. Paul Pechan and Heidi Bohle curated the data, and Paul Pechan wrote parts related to the climate change literature, theory, and how the results relate to it. Heidi Bohle added valuable input and proofread the manuscript.

ARTICLE

OPEN

Check for updates

The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change

Fabian Obster [1,2✉], Heidi Bohle[3] & Paul M. Pechan [3✉]

Climate change has significant implications for economically important crops, yet understanding its specific impact on farm financial wellbeing remains a challenging task. In this study we present self-reported perceptions of fruit farmers about their financial well-being when confronted with different climate change factors. We employed a combination of supervised machine learning and statistical modelling methods to analyze the data. The data collection was conducted through face-to-face interviews with 801 randomly selected cherry and peach farmers in Tunisia and Chile. Specific climate change factors, namely increases in temperature and reductions in precipitation, can have a regionally discernible effect on the self-perceived financial wellbeing of fruit farmers. This effect is less pronounced in Tunisia than in Chile. However, climate change is of lessor importance in predicting farm financial wellbeing, particularly for farms already doing well financially. Social assets, which include reliance on and trust in information sources, community and science, play an important role in increasing the probability of fruit farm financial wellbeing in both Tunisia and Chile. However, the most influential predictive factors differ between the two countries. In Chile, the location of the farm is the primary determinant of financial wellbeing, while in Tunisia it was the presence of social assets.

[1] Department of Business Administration, University of the Bundeswehr, Munich 85577 Neubiberg, Germany. [2] Department of Statistics, LMU Munich, 80539 Munich, Germany. [3] Department of Media and Communication, LMU Munich, 80539 Munich, Germany. ✉email: fabian.obster@unibw.de; paul.pechan@ifkw.lmu.de

Climate change impact on fruit tree yields and farm economic wellbeing. Climate change can impact crops, with yields of many important crops projected to decline in the future[1,2]. Increases in temperature, in particular, can reduce yields of major crops worldwide[3]. Such climatic impacts can and will have a detrimental effect on food availability and its nutritional value[4]. Because of its nature, much of climate change agricultural research is crop, region or country-specific. While there have been numerous investigations into the effects of climate change on various crops, the studies have tended to focus on wheat, rice, corn, and soybean, primarily grown in Asia, Europe and North America[3]. Unfortunately, there has been a lack of assessments regarding the vulnerabilities of fruit crops in the regions that we are interested in, namely North Africa and South America. In particular, no information is available on the extent climate change factors impact not only the fruit yields but also the overall financial wellbeing of farms.

In this paper we focus on the effects of climate change on crops that have an important nutritional and monetary value in Chile and Tunisia: cherry and peach fruit tree[5,6]. Both crops are sensitive to climate change damage, with reproductive organs being particularly vulnerable to climatic impacts, leading to a reduction in the quantity and quality of harvestable fruit[7–9]. Increases in winter temperatures can affect fruit tree chill requirements resulting in changes of bud, flower and fruit set[10–14]. Similarly, elevated temperatures during fruit set and development can lead to changes in fruit growth and maturation[9,15,16]. Combined with reduced water availability, high temperatures can affect both fruit yield and quality[7,17,18]. These effects can vary between fruit tree cultivars and species. Additionally, extreme events (hail, wind, frost) have also been observed to impact the physical environment and cause fruit crop damage[7,19–21]. These climate events are region-specific, affecting food production the crops to varying extents. While climate change impacts on fruit crop quality and yield can be estimated, evaluating climate change impacts on farm financial wellbeing is much more difficult and uncertain[22]. Yet the ability to predict the impacts of factors on the farm financial well-being is crucial for the development of appropriate policy measures that target factors with the highest monetary impacts.

Use of a hybrid approach to predict farm financial wellbeing. In this paper, we introduce a novel hybrid approach that combines machine learning and generalized linear models to address this challenge of predicting farm financial well-being. Traditional economic climate change impact models typically estimate effects of climate change on crop yields using climate and crop simulation models, and then translate this information into likely farm financial performance. However, these analyses are based on a number of assumptions that seldom take a combination of adaptive measures, socio-economic and other factors, such as regional differences, into consideration[23]. One of the most often used economic models measuring impacts of climate change on agriculture is the Ricardian approach that focuses on the land value and agricultural revenue[24,25], with cross-sectional and panel regression analysis as the analytical tools of choice[26]. Whatever the approach and type of analysis performed, the omission of variables that may directly or indirectly affect crop/farm incomes/ revenue makes climate change financial impact assessments highly uncertain. In classical statistics, regression analysis can have predictive powers. But there are situations where regression analysis is not sufficient to handle the generated datasets or the specific questions to be answered or where the assumption of the existence of a linear function between independent and dependent variables doesn't hold. This is especially the case when complex variable interactions are present in the dataset. And this is where machine learning becomes a useful tool that complement traditional statistical analysis[27–29].

Machine learning offers the ability to analyse large datasets and many variables simultaneously, reducing the chance that important variables are left out of the data analysis process. It comes thus as no surprise that machine learning has vast potential to analyse big data in agriculture[30–37], especially when considered in combination with other research domains, such as climate change[38–41]. However, tackling agricultural problems is complex. For example, whether a new crop variety actually provides better yield and farm income under certain climatic conditions is potentially dependent not only on its genetic traits but also on many other factors, such as those related to biophysical and farm management issues[42]. This means that complex and deep interactions could exist in the datasets. Such data can become quickly difficult to properly analyze using classical statistical approaches. The resulting datasets, just for one farm, could encompass millions of data point combinations. Importantly, analysis of such data can provide answers as to which variables, from the millions of possible combinations, are associated and important for the outcome variable - in our case financial well-being of a farm. This is where the power of machine learning can be explored to its full potential[36,43]. By including not only biophysical variables such as microclimate effects, soil structure and quality, but also socio-economic variables, such as land use, urban-farm water accessibility, farm size, demographic data and access to markets, machine learning enables analysis at every step of the agricultural value chain[32,44,45]. Thus the usefulness of machine learning is evident not only when considering ultimate outcome variables, such as the financial well-being of a farm, but also to assess whether adaptive measures were effective in maintaining or increasing crop yield under certain climatic conditions, provide information on the relative importance of an intervention for a desired outcome and generally help with future predictions and strategies[38,40,46–48]. However, the interpretability of machine learning models, especially complex algorithms like

support vector machines, deep neural networks, and random forest or boosted trees, can be limited. Although there are post hoc interpretability methods to approximate the functioning of such black box models, there is no straightforward way of understanding and interpreting the exact processes leading to the outcome. This is potentially a major drawback for research questions that aim to deepen the understanding of the processes or factors associated with the desired outcome.

To overcome this problem, we introduce herein a hybrid method that combines analysis of datasets based on generalized linear models combined with strategies from machine learning, such as cross-validation and boosting and group-variable selection. The output of this approach preserves interpretability, respects the group structure of the data and is still competitive with state-of-the-art machine learning algorithms. Detail information about this strategy can be found in the data analysis section of this paper. Use of this hybrid model has allowed us to effectively address our research objectives.

Research objectives. The primary objective of this paper is to assess the potential impact of climate change on the financial well-being of fruit farms. To achieve this, we relied on farmer self-reporting about the past experiences with climate change and examined whether these experiences had any bearing on the financial performance of their farms. The information was collected through face-to-face interviews. It is important to note that because it was the farmers who provided the information for subsequent data analysis, we are in effect, reporting herein on
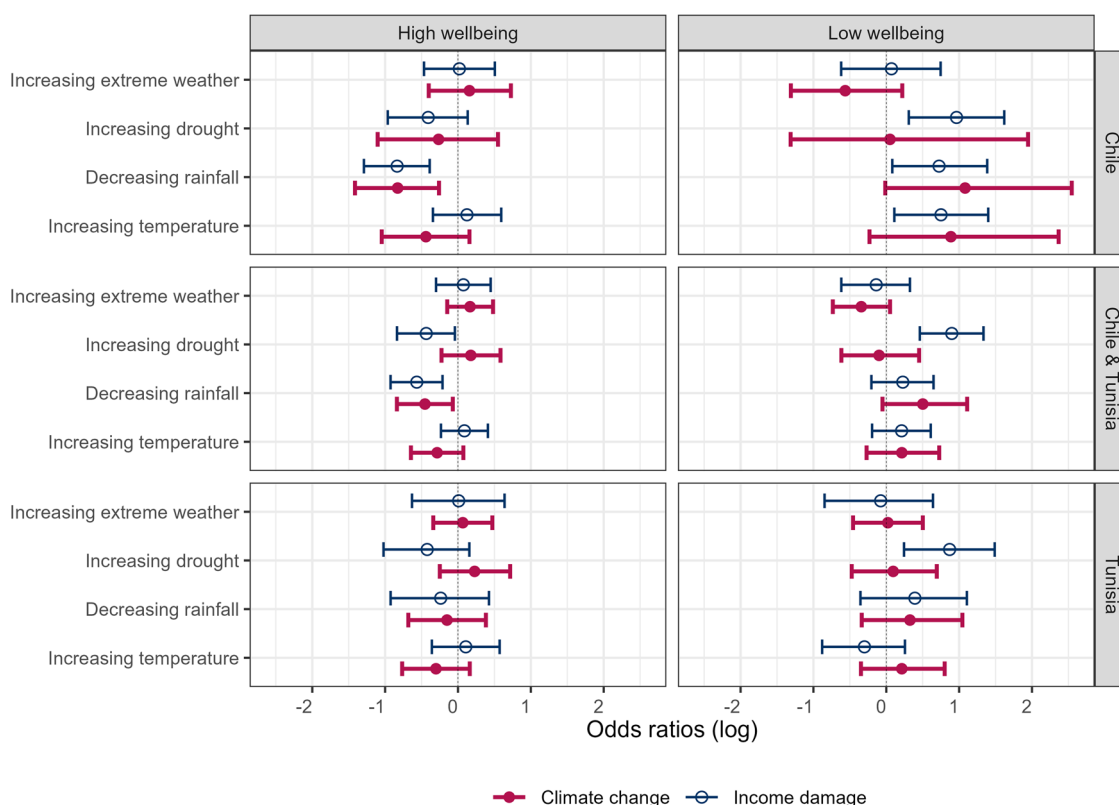
Fig. 1 Effect of experiences with climate change and crop financial damage on financial wellbeing of a farm in Chile and Tunisia. Confidence intervals of the Odds-ratios (OR), based on logistic regression (see Supplementary Tables 4 and 5 for more data).

farmer´s perceived financial well-being. Perception in the context of this paper refers to how individual farmers interpret, assess and experience climate change information. Their perceptions may be influenced by sensory observations as well as their previous memories, knowledge and expectations of climate change.

To address the complex nature of the datasets, which includes many grouped and single independent variables, we employed a combination of classical statistical analysis and machine learning techniques. This approach allowed us to consider the high dimensionality of the data and determine the relative importance and predictive power of both individual and grouped independent variables in relation to the outcome variable.

In this paper, we aim to answer three basic research questions based on farmer self-reporting. First, we investigate whether climate change has a discernible impact on how well fruit farmers are doing financially. Second, in cases where climate change is not important for the farm financial well-being, we explore what other factors may influence this outcome variable. And third, we examine the potential effects of factor interactions on farm financial well-being. By addressing these questions, we seek to enhance our understanding of the relationship between climate change and farm financial well-being.

Results

Climate change effects on farm financial well-being. First, we evaluated whether experiencing climate change had any impact on farm financial wellbeing. Decreasing rainfall and increasing temperatures were associated with reduced farm financial well-being (Fig. 1). Combined, farmers in Chile and Tunisia, who have experienced reduced rainfalls, were significantly less likely to do financially well than farmers who did not experience reduced rainfalls (0.635, $p = 0.020$). To a lesser extent, increases in temperature in the two countries also resulted in the likelihood farms

to do financially well (0.751, $p = 0.11$). Increasing drought frequencies and extreme weather experiences had no significant impact on farm financial well-being in any of the regions studied. The effects of increasing temperatures or decreasing rainfall were more discernible in Chile than Tunisia. Thus negative experiences with certain climatic factors lowered in some cases farm financial well-being, with the provision that the effects of the negative experiences may be country-specific.

Second, we investigated to what extent financial damage to crops, caused by specific climate change impacts, is associated with overall financial farm well-being. The results indicate that farms that performed financially well, the odds were that only decreasing rainfall-associated income impacts were significantly associated with farm-high well-being (0.568, $p = 0.002$ for Chile and Tunisia combined, 0.434, $p < 0.001$ for Chile). Farms that were not doing financially well, the odds were that higher temperature-associated income impacts were significantly associated with farms low wellbeing (2.119, $p = 0.021$ for Chile) and more frequent drought (2.457, $p < 0.001$ for Chile and Tunisia combined, 2.623, $p = 0.003$ for Chile and 2.385, $p = 0.006$ for Tunisia). Decreasing rainfall, especially in Chile, seemed to be somewhat relevant for explaining low well-being farms. It is noteworthy that although experiencing drought was not significantly associated with low or high financial well-being, the financial impacts of drought tended to be significantly associated with farm financial well-being.

Variables important for farm financial wellbeing. The sparse group boosting (sgb) algorism allowed the model to choose between individual and grouped independent variables for the predictive modeling (Fig. 2). Arrow directions indicate the added effect size (log odds) of all variables within one group on the farm financial wellbeing, resulting in a latent variable. For high
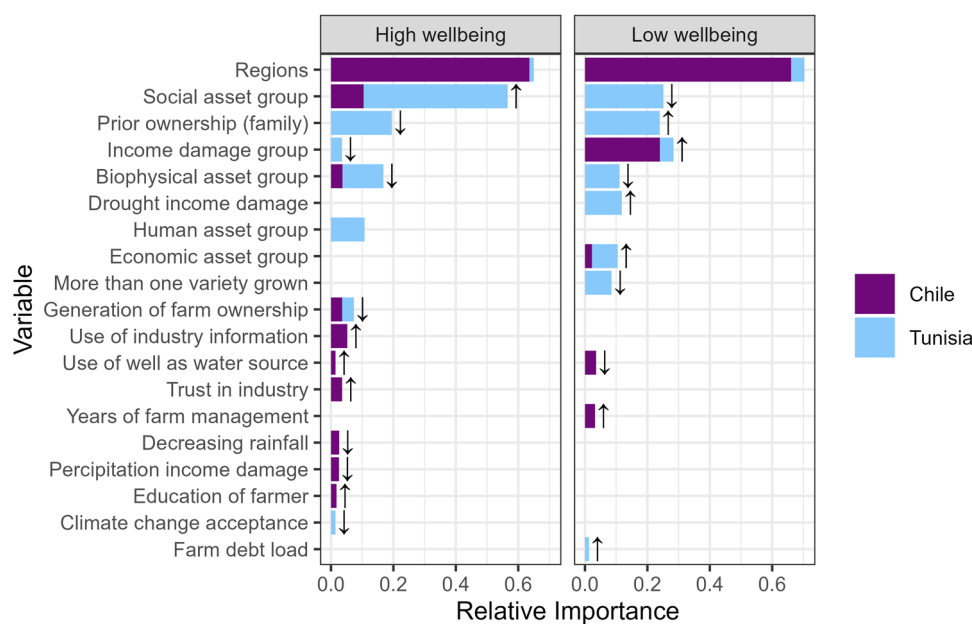
Fig. 2 Most important variables contributing to farm financial well-being. Sparse group boosting model for Chile and Tunisia and high and low financially performing farms separately. Central Chile was associated with higher financial wellbeing compared Southern Chile and Northern Tunisia slightly higher than Central Tunisia.

financial well-being, upward pointing arrows indicate that an overall increase of group variable values lead to an increased probability for high financial well-being, while downward pointing arrows indicate a decreased probability of high well-being. Similarly, for low financial well-being, an arrow pointing upward means that increases in the group variable values increase the probability of low well-being. Thus higher/increasing social assets will increase the probability of farm high well-being. Note that no arrows were added for nonordinal variables or groups of variables.

Generally, variables not related to the climate change factors were comparatively more important for predicting farm financial well-being. Thus the most important predictors of farm high financial well-being, common both to Chile and Tunisia, are social (reliance on/use of information, trust in information sources, community, science or religion) and biophysical (farm size, water management systems used on the farm, diversity of crops used) assets, as well as one individual variable, years of owning the farm (Fig. 2). The latter two tend to have a negative effect on farm financial wellbeing. Natural assets (regional differences) are important predictors almost exclusively only for Chile, where farms in Central Chile tend to exhibit higher financial well-being. Prior farm ownership and the human asset group (including education, age, gender, and knowledge) are important factors specific for Tunisia only. The most important predictors of farm low financial well-being, common both to Chile and Tunisia, are regional differences, income impact and economic asset groups, where for example increasing farm debt and reliance on orchard income increase the likelihood of farm low well-being. A number of factors are associated with the likelihood of farm low well-being in Tunisia only: these are the length of farm ownership, drought, social and biophysical assets groups, and varieties grown. The latter three are associated with increased likelihood of reducing low financial well-being. For Chile only, the important individual factors are use of a well and years of farm management. The more farms use wells, the less likely will they exhibit low financial well-being, whereas longer the farmer is managing the farm, higher the likelihood of low

financial well-being. Factors unique to Chile are not very important variables.

Note that some factors are important predictors of both high and low financial well-being, just with oposing effect. For example, increased well usage in Chile increases the likelihood of high well-being while decreasing the likelihood of low well-being. In Tunisia, prior family ownership decreases the likelihood of high well-being while increasing the likelihood of low well-being. The exception are biophysical assets, that decrease the odds for high wellbeing and also decrease the odds for low wellbeing, indicating using biophysical assets, like adaptive measures, are only useful to help farmers with low financial wellbeing.

Variable interactions affecting farm financial wellbeing. We have examined whether interactions between independent variables may change the model outcomes vis a vie financial well-being of a farm (Fig. 3). Even though the model that included variable interactions was not as predictive as the model including only additive effects (Table 1), the importance of each interaction still showcases interesting and important inter-dependencies in the datasets. One outcome is that the region variable seems to be less important when other interactions are considered. Interactions within and between social and human assets seem to be relevant for the farm's financial well-being, especially those related to use of information and trust. Interactions that involve adaptive measures, current assessment of climate change as well as education are also of relative importance. Such interactions point to inter-dependencies between variables and to likely confounding and mediating effects of certain variables.

Figure 4 provides information about some noteworthy interactions that can affect farm financial well-being. Without the use of newspapers as a source on information the probability of high well-being drops in Chile and Tunisia when temperature increases or precipitation decreases (Fig. 4, top left). However, when farmers used newspapers, financial well-being in Chile and Tunisia is not markedly reduced by increasing temperatures or decreasing precipitation (Fig. 4, bottom left). Indeed, use of
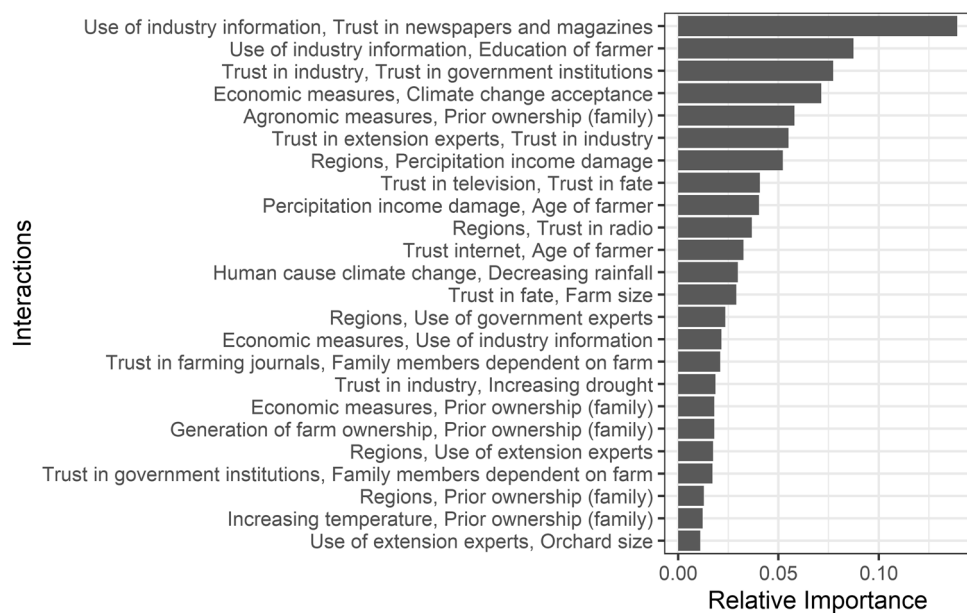
Fig. 3 The most important interacting variables for farm financial wellbeing. Component-wise boosting model for Chile and Tunisia combined.

Table 1 Predictive power for farm financial high and low well-being.

Accuracy	wellbeing	Sgb	mb	Mb int	glm	rf	nn	gbm
Chile	High	0.65	0.675		0.642	0.733	0.575	0.683
Chile	Low	0.833	0.833		0.833	0.842	0.833	0.817
Chile & Tunisia	High	0.71	0.693	0.685	0.618	0.734	0.556	0.705
Chile & Tunisia	Low	0.809	0.809		0.822	0.817	0.822	0.793
Tunisia	High	0.595	0.579		0.545	0.645	0.529	0.57
Tunisia	Low	0.76	0.744		0.736	0.769	0.529	0.727
AUC	wellbeing	Sgb	mb	Mb int	glm	rf	nn	gbm
Chile	High	0.655	0.717		0.687	0.757	0.603	0.727
Chile	Low	0.830	0.802		0.676	0.837	0.642	0.773
Chile & Tunisia	High	0.733	0.723	0.731	0.627	0.796	0.619	0.758
Chile & Tunisia	Low	0.763	0.733		0.637	0.746	0.721	0.735
Tunisia	High	0.663	0.661		0.537	0.710	0.616	0.658
Tunisia	Low	0.596	0.562		0.579	0.614	0.492	0.579

The accuracy and Area Under the Curve (AUC) of all fitted models was evaluated on the test data from Chile and Tunisia. For corresponding receiver operator curves (see Supplementary Fig. 2). For abbreviation explanation, see Methods - Choice of predictive models for data analysis.

newspapers increased the probability of farm financial well-being irrespective whether or not temperature increases or precipitation decreases: the use of newspapers eliminated any negative effect of reduction in precipitation or increases in temperature on doing financially well. A similar effect was observed for trust in industry (Fig. 4, top right and bottom right). Farmers, especially in Tunisia, who trusted industry as a source of information, were more likely to do financially well than farmers who did not trust industry, regardless whether or not they experienced a reduction of precipitation. However, the effect of increasing temperatures on high wellbeing seems to be unchanged by trust in industry in Tunisia while in Chile, trust in industry, compared to no trust in industry, intensified the negative effect of temperature increases on financial farm wellbeing.

Trust in media, use of industry information and farm financial well-being indicate that farmers, regardless of their country of origin, who did not trust media and did not use information from industry had the lowest probability of doing financially well (Fig. 5, top left). Farmers who did trust media sources but still did not use industry information, performed financially substantially better. Farmers with the highest probability of doing financially

well were those that trusted the media and used industry information, where the trust factor acted synergistically with the use of information. The importance of trust for financial well-being can be illustrated with the effect

of trust in industry, experts and government. Thus, trust in industry acted synergistically with trust in experts (Fig. 5, top right) as did trust in government and trust in industry (Fig. 5, bottom left).

In all cases, farmers that trusted industry, experts or the government were more likely to be financial well off than farmers who had no trust in their information sources. Other interactions, for example, education and use of media also have a positive modifying effect on farm financial well-being in Chile but not in Tunisia: educated farmers who used media tended to be more likely to do well financially than farmers with low education (Fig. 5, bottom right).

Discussion

Previous studies have highlighted the detrimental effect of individual climate change factors on crop yields and farm income[3,47,49–53]. Our research contributes to these findings by
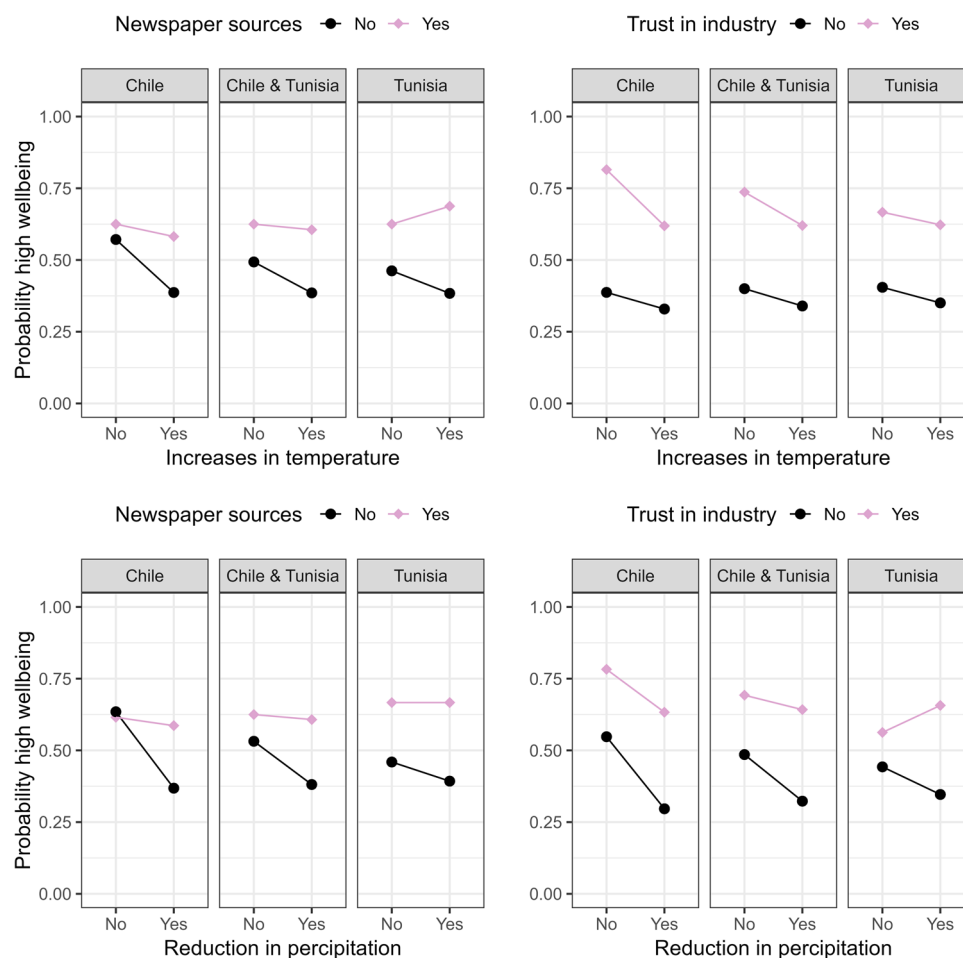
Fig. 4 Probability for high financial well-being of the farm. Comparisons based on an interaction between country, climate change factors, use of newspapers and trust in industry.

showing that climate change factors, when analyzed concurrently, impact fruit farm financial well-being to different extents. Whereas odds are that increasing drought and reduction in the amount of rain will negatively affect fruit farm financial well-being, especially in Chile, extreme climatic events do not seem to play such a role. Thus, while farmers have discussed possible fruit damage due to frost or hail events[54], such events do not appear to affect the mid to long-term farm income prospects. Indeed, fruit farmers are more likely to be concerned about drought issues (and consequently future water availability)[54], reflecting findings herein showing that the increasing frequency of droughts had a negative effect on farm income and farm financial wellbeing.

Contrary to expectations, our analysis reveals that climate change is, compared to other factors we investigated, not the most important factor for predicting fruit farm financial wellbeing. In Chile, farm location emerged as the strongest indicator of farm financial well-being, with farms in central Chile doing better than farms in Southern Chile. In Tunisia, farms that have been in family possession for multiple generations, did worse financially. Chile and Tunisia also shared a number of important predictors. In both Chile and Tunisia, access to information and trust of information sources are more important than climate change in predicting farm financial well-being. These shared factors are useful to predict both financially high and low-performing farms: better the information access and more trust there is in information sources, better the farm financial performance and vice versa. On the other hand, climate change-related factors do play a more important role for farms not doing financially well.

As predictive factors differ between farms doing financially well and those experiencing financial hardship, policymakers or farmers need to employ different strategies depending whether they wish to focus on maintaining or improving fruit farm financial performance. An argument can be made to focus on factors important for improving farm financial well-being as financially healthier farms are more likely to be resilient against climatic impacts[48,55]. Furthermore, synergistic effects and interactions between factors can affect their individual or combined importance for farm financial wellbeing. It is important to note that inter-dependencies between factors can motivate farmers to respond to climate change[56]. In this respect, the specificity of some factors implicated in fruit farm financial wellbeing advocates for collecting extensive regional rather than country-wide datasets.

Although our findings presented herein indicate that climate change currently is not important for predicting fruit farm financial wellbeing, the situation may change in the future. This is evident from climate change trends analyzed in this paper: the odds are that with higher temperatures and less precipitation fruit farm financial performance may decrease. Temperature predictions indicate continuing increases of winter night temperatures in Tunisia and Chile in the future[14,57]. This will lead to winter chill deficits and potential problems with fruit tree phenology necessitating changes in the types of fruit trees grown. Similarly, reduced precipitation and water availability in Tunisia has serious implications for the future of fruit trees in that country[58]. Much of the irrigation water for fruit trees comes from underground
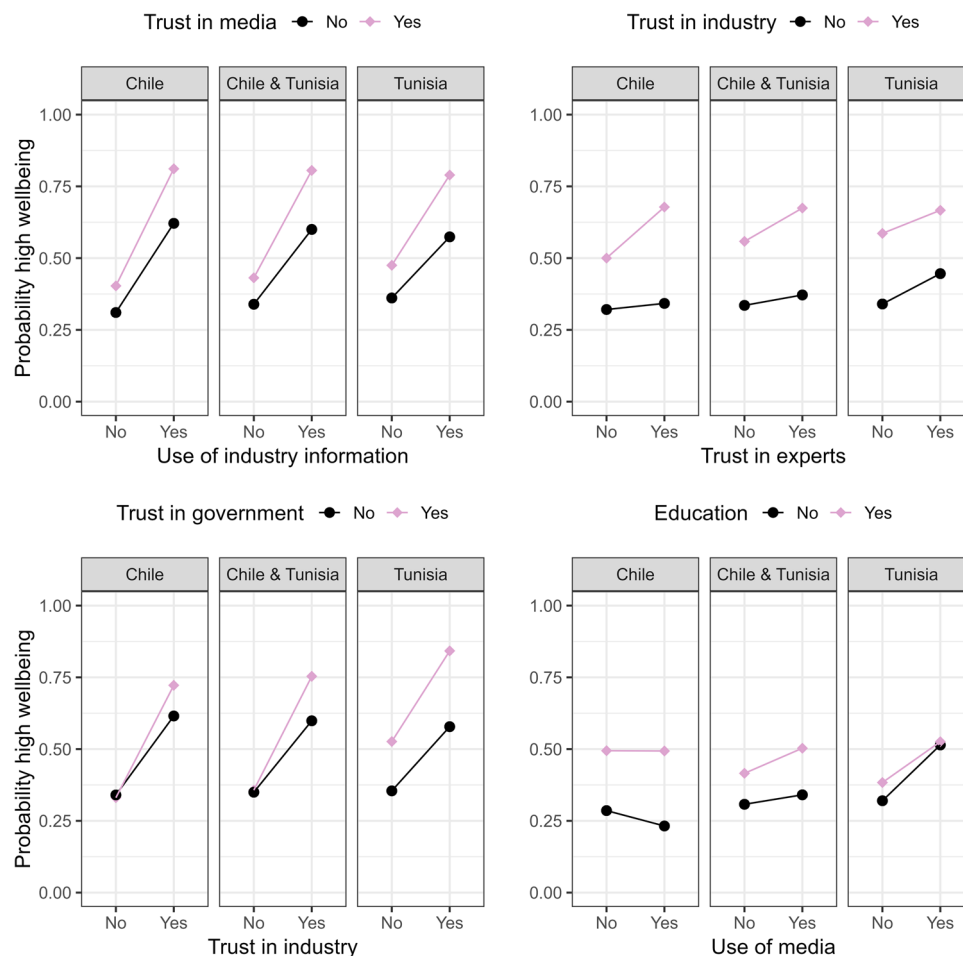
Fig. 5 Probability for high financial wellbeing of the farm. Comparisons based on an interaction between country, trust, use of information sources and education.

aqua ducts. If they are depleted or become saline, farmers in Tunisia may face major crop yield loses. In Chile, reduced water flow from the Andes, increasing urbanization and inappropriate crop use could create water distribution bottlenecks[59,60]. These latter predictions are in agreement with farmer climate predictions for the future: most worry about the effects of drought and water availability (63 > 54). Ensuring general access to water, beyond relying solely on rainfall, is critical for adequate irrigation and reducing drought exposure. Resolution of these problems will require the implementation of specific adaptive measures that will reduce the future vulnerability of fruit farms to climate change, measures that farmers and governments need to be willing to pay[55,61]. Policymakers must enact regulations to guarantee fair and sufficient access, distribution, and use of limited water resources among all stakeholders. Furthermore, policymakers should provide fruit farmers with effective, affordable, and accessible resources and tools to enhance farm adaptive capacity and reduce vulnerability to drought, such as sustainable irrigation systems, insurance schemes, crop alternatives, and farm management training[54]. When it comes to communication efforts to convince stakeholders to adapt the necessary protective measures, policy makes must keep in mind that climatic impacts may not be a primary risk to farm financial wellbeing. Indeed, due to the current conflict in the Ukraine and Covid aftermath, costs associated with adaptive measures are likely to become a dominant concern of many farmers around the world. It is also worthwhile to remember that more media coverage does not necessarily

influence farmer perception of climate change: we found no substantial association between use or trust in media and farmer perceptions of climate change. Similarly, if farmers trust or use media as their source of news, they don't necessarily think that precipitation decline is bad for farm financial well-being.

To our knowledge, this is the first instance of using a hybrid modeling approach combining statistical models with machine learning techniques to analyze data in a much more complex and integrated manner. Using this approach, our aim was to improve predictability while maintaining interpretability. We found that statistical models, utilizing limited datasets that reflect the requirements of relevant theories, can be used to make adequate predictions about the relationships between climate change, intervening variables, and the outcome variable. However, we have also shown that by combining statistical models with specific machine learning methods, such as boosting and cross-validation, we were able to substantially improve the predictability of the (generalized linear) statistical model. This hybrid model can still be interpreted through variable importance and odds ratios, but classical inference based on F and t statistics is not valid for variables selected through a data-driven process[62]. Predictive modeling provided new insights into data relationships that can serve to generate and test new hypotheses by classical statistical means. Even though the random forest (a typical black box model) outperformed the sparse group boosting model, we believe that this improvement generally does not compensate for the loss of interpretability. With a similar analysis methodology,

neural networks marginally outperformed logistic regression[63]. The predictive sparse group boosting and component-wise boosting models were ultimately chosen for the current data analysis. The former model provided evidence of regional or supra-regional variables that are important for predicting whether fruit farms will perform financially well. The latter model revealed that the interaction between various variables and farm financial well-being, as the outcome variable, is not a simple one-to-one relationship. Rather, certain variables, such as trust and use of specific information sources, appear to have a modulating effect on variables that may directly affect the outcome variable.

Conclusions and future considerations. Our research underlines the usefulness of the hybrid analytical approach and highlights specific climate change factors that impact fruit farm financial wellbeing while emphasizing the significance of other influential variables. Policymakers, stakeholders, and researchers can utilize these findings to develop targeted strategies and adaptive measures to support fruit farmers, reduce their vulnerability to climate change while enhancing the financial stability.

Our experience with the hybrid model indicates that, especially when it is necessary to balance predictive improvements (usually requiring larger datasets) with loss of model interpretability, the research questions and the modeling tools available will dictate the extent and complexity of the data to be collected, whether the focus should be on regional or supra-regional datasets and the type and depth of analysis that can be performed. Machine learning provided the opportunity to include a broader range of independent variables with substantially better predictability of farm's financial well-being and clarity of data presentation than offered by traditional regression analysis. We believe that, through group-component-wise boosting of generalized linear models, our hybrid approach can generate useful predictions in high dimensional settings, while still preserving basic interpretability, like variable importance and odds ratios. This way, new hypotheses and models can be generated, left to be validated or rejected by future research. The key challenge for future studies will be to find the correct balance between a theory-based approach, where a limited number of likely relevant variables are included in the survey design and resulting datasets, and a black-box approach that relies on deep mining of the largest possible number of data points.

Our results indicate that self-reporting of changes in temperatures and precipitation within the last ten years generally reflect the meteorological observations over the past 30 years. Farmer's perceptions and self-assessment are thus a valid tool to investigate the linkage between climate change and other factors, such as farmers' perception of financial well-being as an outcome variable and ultimately allows investigation into the influence of perceived financial well-being on farmer behavior. It is, however, important to note that the perception of farm well-being is not the same as using actual financial performance data from farms or regions to assess its impact on farmer behaviour. Future research should consider collecting actual farm financial data and conducting comparative studies with self-assessment data collected from face-to-face interviews with farmers. The relatively small size of the resulting dataset, based on face-to-face interviews with 800 farmers, restricted subnational comparisons and increased the possibility of false selections due to the large number of influencing variables. However, the project size, the complexity of the survey and the length of the interview (ca. one hour), precluded a larger sample size and the number of variables and items to be investigated.

Fruit farming is an important sector for the economy, particularly due to high export potential. It is essential to develop policies that support fruit farmers in improving their financial well-being and achieving financial stability as the climate changes. Farmer experiences with climate change is reflected in perception of their financial well-being but it is factors other than climate change that are deemed to be more important for farm financial well-being. Policymakers should thus prioritize supporting and strengthening farmers' financial well-being beyond climate change considerations. Addressing issues such as trust, information sharing and targeted communications can contribute to these goals.

Methods

General agricultural attributes of the study areas. According to FAO statistical yearbook for 2022, the world value of primary agricultural production reached USD 2.7 trillion, of which fruits represented 17%[64]. World Food and Agriculture- Statistical yearbook 2022. Rome. doi.org/10.4060/cc2211en). More specifically, the agriculture and related sectors in Chile represent 24.4% of total exports, 9% of total GDP, and employs around 10% of Chile's labor force[65]. In Tunisia, agriculture represents 12% of the country's GDP, employing 16% of the country's workforce[66]. It is, however, very difficult to obtain up-to-date and reliable information on the importance of cherry and peach crops for the economies of Chile and Tunisia. Chile 2022 cherry production was estimated at 255 711 metric tons, ranking number 6 in the world[67]. Majority of the production is exported to China, valued at over USD 2 billion[68]. Tunisia 2022 peach production was estimated at 123 000 metric tons, ranking number 20 in the world[69]. Majority of the exported production is destined for the Gulf states[70].

Environmental attributes of the study areas. Four contrasting geographical and climatic regions were selected for the study, two regions in Tunisia and two in Chile. In Tunisia, these were the Mornag and Reueb peach-growing regions. In Chile, these were the Rengo and Chillán cherry-growing regions.

Tunisia. Mornag, Tunisia, hereafter referred to as Northern Tunisia, has an elevation of 110 meters and is located approximately 20 km east of the capital Tunis. The region has a Mediterranean climate. Precipitation in Mornag is characterized by a rainy fall-winter season spanning October and March (ca. 400 mm) and a relatively dry spring and summer (ca. 130 mm). The coldest month is February with minimum and maximum average temperatures of 5.5 °C and 16 °C, respectively. The warmest month is August with average minimum and maximum temperatures of 22 °C and 34 °C respectively.

Regueb, Tunisia, hereafter referred to as Central Tunisia, has an elevation of 160 meters and is located approximately 230 km south of Tunis. It is a semi-arid region characterized by low rainfall and high temperatures. Most of the rainfall is between October and the end of March (ca. 210 mm). Spring and summer are dry (ca.80 mm). The coldest month is January with minimum and maximum average temperatures of 5 °C and 15 °C, respectively. The warmest month is July with minimum and maximum average temperatures of 21.5 °C and 36°C respectively.

Chile. Rengo, Chile, hereafter referred to as Central Chile, has an elevation of 570 m and is located approximately 110 km south of Santiago de Chile. The Mediterranean climate in this region is characterized by rainy, cool, wet winters and hot, dry summers. Rainfall is concentrated in the winter months between May and September (ca. 500 mm). Spring and summer tend to be dry (ca. 60 mm). The coldest month is July with minimum and maximum average temperatures of 0 °C and 10 °C, respectively. The

warmest month is January with minimum and maximum average temperatures of 10 ºC and 24 ºC, respectively.

Chillán, Chile, hereafter referred to as Southern Chile, has an elevation of 120 to 150 meters and is located approximately 380 km south of Santiago de Chile. The climate of the region is Mediterranean, with the rainy season occurring primarily during the winter months. Summers are relatively dry. Most of the rainfall occurs in the winter between May and September (ca.700 mm). Rainfall in the spring and summer is ca. 200 mm. July is the coldest month with minimum and maximum average temperatures of 0.5 ºC and 11 ºC. The warmest month is January with minimum and maximum average temperatures of 10.5 ºC and 25 ºC, respectively.

In order to place farmer perceptions in the context of climate change, we analysed regional Chile and Tunisia climatic data for the last 30 years (see Supplementary Fig. 1). Changes in temperatures and precipitation within the last 10 years generally reflect the meteorological observations over the past 30 years. Farmer´s perceptions are thus a valid tool to investigate the linkage between climate change and other factors, such as farmer´s perception of financial wellbeing.

Data collection: survey methodology and sampling. The data collection instrument used in this study was a face-to-face survey with cherry farmers in Chile and peach farmers in Tunisia. A total of 801 farmers were interviewed, 401 in Tunisia and 400 in Chile in the fall of 2018 and spring 2019, respectively.

Survey methodology. The questionnaire for the survey was prepared in English and translated into Tunisian Arabic and Chilean Spanish. The translated documents were back-translated into English to check for inconsistencies. The survey was pre-tested with 12 farmers in consultation with Qualitas Agro-Consultores in Chile and Elka Consulting in Tunisia. Based on their feedback, and that of our research colleagues in Tunisia and Chile, some questions were removed while others were refor-mulated. The same consultants carried out the face-to-face interviews. Farmers were asked to answer a combination of multiple-choice, open, Likert Scale and Yes / No questions related to climate change and climate impacts on their farms between the years 2009 and 2018 and to their past, present and planned adaptive measures. The relevant survey questions and analysed variables are presented in the Supplementary Tables 1 and 2.

We analysed threat to fruit farms from four different climate change factors: temperature, precipitation, extreme weather and drought. Farmers were asked whether any of the factors over the past 10 years were increasing, decreasing, staying the same or became unpredictable.

In addition to climate change, there may be groups or individual farm-related variables that may, by themselves or in interaction with climate threat, affect farm financial well-being.

- **Groups of variables**. We have focused our analysis on groups of farm variables (assets) that may be important for farm financial well-being. These were:

  Natural (geographical regions)
  human (education, age, gender, knowledge)
  social (reliance on/use of information, trust in information sources, community, science or religion)
  biophysical/manufactured (farm size, water management systems used on the farm, diversity of crops used, adaptive measures)
  economic (farm debt, farm performance, reliance on orchard income)

  climate experience
  income damage

The choice of the above variables was made on the basis of the five resource/capital sustainability model that addresses the concept of sustainable wealth creation[71,72].

- **Individual variables**. Above listed grouped variables were also assessed individually. In addition, other variables were examined that may, by themselves or in interaction with climate threat, affect farm financial wellbeing.

- **Dependent variable**. The question given to farmers that defines the dependent variable was: "When it comes to financial matters of your farm operation, how well is your farm doing?" The variable consists of three categories. Doing well and very well, neither doing or not doing well ("neutral"), and not doing well or not well at all. Throughout the analysis, the financial well-being variable is coded as two separate variables. We refer to the first variable as "high well-being" comparing farmers who are doing well and very well financially with farmers who are doing neutral or not well (reference category) and the second one as "low well-being" differentiating between farmers who are not doing well financially with farmers who are doing neutral, well or very well (reference category). This enabled us to differentiate between the process leading to farmers not doing well and the process leading to farmers doing well, as the farmers who are neither doing or not doing well are always part of the reference category.

Sampling. A list of individual fruit farms in regions of interest were obtained from respective Ministries of Agriculture. Farms from these lists were randomly selected for the survey if they fulfilled the following criteria: farmers had to own the farm, manage and work on the farm and derive over 70% of their income from their farming activities. A total of 801 face-to-face interviews were subsequently conducted with farmers who ful-filled the preselection criteria – 401 peach farmers in Tunisia (201 in Mornag and 200 in Regueb regions) and 400 cherry farmers in Chile (200 in Rengo and 200 in Chillán regions). The approxi-mately one-hour-long interviews were carried out with farmers directly on their farms. The interviews were carried out after harvest completion in the fall of 2018 by Elka Consulting in Tunisia and in the spring 2019 by Qualitas AgroConsultores in Chile. Guidance was sought from the Department of Commu-nication and Media Research, University of Munich about the participation of human subjects in the survey research and sub-sequent data use. The farm data was collected according to data collection procedures applicable in each country. Informed con-sent for the data collection was provided by the survey partici-pants. No personal identifiable data was collected, assuring full anonymity. After compiling the data from farmer interviews, the resultant datasets were checked for errors and integrated into excel formats for further data analysis.

Data analysis strategy
Research question one: does climate change have an effect on how well the farm is doing financially?. We used a statistical approach to determine the effect of independent variables on the farm financial well-being. As the two outcome variable "high well-being" and "low well-being " are binary, we used logistic regres-sion and analysed the odds ratios as well as associated p-values and confidence intervals of adaptive measures and past experi-ence for the outcome.

Research question two: what factors, other than climate change, may be important for the financial well-being of a farm?. This research question imposes a major challenge. There are many possible influencing variables in the dataset. Some may be relevant for the outcome variable, but others may not. Variables not related to the outcome variable create unnecessary background "noise" because generalized linear models tend to over-adapt to the data (the so-called overfitting) in high-dimensional cases. In the extreme case, where the number of independent variables is higher than the number of observations, linear models cannot be fitted. The solution to this problem is to perform variable selection, and then include only these variables in the model. The current practice is to perform this selection based on literature and expert knowledge. In fact, there is always an implicit variable selection process based on which such data is collected. However, one may still end up with a large number of possible influencing variables. In this situation, the combination of statistics and machine learning can be used to perform the variable selection. We used model-based boosting[73], but other strategies, such as the Lasso[74] can be utilized. The model-based boosting strategy is to improve a given model by only adding variables that improve the overall model the most. The process of adding variables is stopped if a further update would not result in a "better" model.

Importantly, in some instances, grouped variables may be more important for the model than individual variables. We used sparse group boosting for this purpose[75]. In sparse group boosting, the model can decide between individual variables and groups of variables. New hypotheses can be generated about the association of selected variables or groups of variables and the farm's financial well-being. Being able to differentiate between the importance of groups and individual variables may help in designing questionnaires because if individual variables are more important than the group, only the important individual variables need to be included in the questionnaire. This may greatly shorten the questionnaire without loss of information. Conversely, variable groups may provide information about variable interactions.

Research question three: Are observed effects on the financial well-being of a farm the result of moderating effects and/or more complex relationships between variable?. We analysed (pairwise) interaction effects of all variables on the financial well-being of the farm. Interactions of variables were evaluated with the help of model-based boosting, allowing comparisons of their relative importance for the outcome variable. Note that if there are p variables in the dataset, then there are $0.5 * p*(p-1)$ possible interactions in the dataset, leading to an even higher dimensional noise problem. However, this brute force method has the potential to identify important moderation or additive variable effects, and thus increase our understanding of the processes leading up to the outcome.

Depending on the research question being asked, the complexity of data analysis, as described above, may still not be sufficient. In such situations, noninterpretable black-box machine learning models should be used. Comparing the predictive performance of these machine learning models with the interpretable hybrid and statistical models gives an indication of the necessary analytical complexity. If the hybrid model outperforms the black-box model regarding the predictive power (i.e. delivers better AUC), then further complexities are not necessary. If the converse is true, the goal of future research should be to understand how these complexities can be explained, for example, by using highly nonlinear relationships or higher-order interactions.

Models used for data evaluation

Statistical models. We used generalized linear models[76] to answer whether interventions had an impact on the outcome of interest.

As the outcome variables were binary, logistic regression was used to provide odds ratios, the corresponding *p*-values, and confidence intervals.

Machine learning. We have compared different popular machine learning models to ensure that the models used for our analysis were competitive in their predictability. A list of all models used is given in Supplementary Table 3. In contrast to the model-based boosting models and the logistic regression, these machine-learning models do not allow insight into the data.

Hybrid statistical - machine learning-based predictive models. We decided to use model-based boosting as means to select variables for the predictive models. The number of boosting iterations was controlled by 25-fold cross-validation using the training data. This hyper-parameter controls effect penalization (smoothness) and regularization (variable selection)[73]. Variable selection was completed in under 4000 iterations. The effect sizes, in our cases the odds ratios, were shrunken to zero through ridge regularization. This makes it easier to interpret the results since only the most important variables for the outcome must be analyzed and irrelevant variables are not considered by the model. Since the influencing variables can be clustered into groups, as described in the contextual definitions, we used sparse group boosting[75] as an extension of model-based boosting. The chosen approach allows the resulting model and variables to be interpreted similarly to generalized linear models[77]. A possible alternative for this approach is to use the lasso and the sparse group lasso[78].

Model evaluation. 70 percent of the observations in the data were randomly assigned to the training dataset and the remaining 30 percent were assigned to the test data set for the final evaluation.

Model evaluation was based on the area under the receiver operator curve, as evaluated on the test data. For the binary outcome variables, two major performance metrics were evaluated at every threshold of probability. First, the rate of correctly identified farms doing well financially, and second, the rate of correctly identified farms not doing well financially yielding the receiver operator curve (ROC). The area under the ROC (AUC) takes both rates into account by considering all possible thresholds of probabilities computed by a prediction model. We also computed the Accuracy as additional metric, which is the percentage of all correctly identified/predicted farmers in the test data set by a classification model. Even though this metric does not balance the true positive and true negative rate in unbalanced data like the AUC, it is used because of its intuitive interpretation property.

All data analyses were performed using the statistical programming environment R, visualizations were created with the R package ggplot2[79].

Choice of predictive models for data evaluation. We compared different predictive models to ascertain which model has the best predictive power and should therefore be used for the data analysis (Table 1). Except for Chile and Tunisia combined low financial wellness, the random forest (rf) tended to outperform all other models for Chile and Tunisia combined as well as for Chile and Tunisia separately. The overview of ROC curves for selected models can be found in Extended Data Fig. 2 Boosted decision trees (gbm) performed similarly to sparse group boosting (sqb) and model-based boosting (mb). In all cases, neural networks (nn) performed worse than sgb and mb. Generalized linear model (glm), which consisted only of experiences with climate change and its financial impact, had lower predictive properties than sgb and mb. However, when the glm was fitted with boosting (model-based boosting-mb), which included more variables related to the

farm vulnerability to climate change and geographical location, the accuracy and AUC tended to improve compared to the glm-only. Including interactions between all independent variables (mb-int) did not improve the predictive outcomes of model-based boosting. The results imply that only considering experiences with climate change and its financial impact as in the glm is not enough to explain both financial well-being variables. Thus, additional variables had to be considered. When compared to the interpretable models, accounting for deep

interactions and complex relationships like the random forest could, in some cases, result in marginal improvements in accuracy and AUC predicting high well-being, but for predicting low well-being the simpler models seem to suffice. Since our investigation necessitated data interpretation, sgb was chosen for subsequent data analysis.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The data for conducting the analysis can be found in the supplement or on github (https://github.com/FabianObster/pasit_financial_wellbeing).

Code availability
Code for conducting the analysis can be found in the supplement or on github (https://github.com/FabianObster/pasit_financial_wellbeing).

References

1. IPCC. Summary for Policymakers. *In: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems* (eds. P. R. Shukla, et al.). (Cambridge University Press, Cambridge, 2019).
2. IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability.* Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (eds. Pörtner, H.-O; et al.). (Cambridge University Press, 2022).
3. Zhao, C. et al. Temperature increase reduces global yields of major crops in four independent estimates. *PNAS* **114**, 9326–9331 (2017).
4. Fanzo, J., McLaren, R., Davis, C. & Choufani, J. How to ensure nutrition for everyone under climate change and variability. GCAN policy notes 1. (International Food Policy Research Institute (IFPRI) 2017).
5. Quero-García, J., Iezzoni, A., Pulawska,J., Lang, G. A. (Eds.). Cherries: Botany, Production and Uses. (CABI, 2017).
6. Manganaris, G. A. et al. Peach for the future: A specialty crop revisited. *Sci. Horticult.* **305**, 111390 (2022).
7. Predieri, S., Dris, R., Sekse, L. & Rapparini, F. Influence of environmental factors and orchard management on yield and quality of sweet cherry. *J. Food Agric. Environ.* **1**, 263–266 (2003).
8. Jackson, D. I. Climate and Fruit Plants. In *Temperate and subtropical Fruit production*. Jackson,D. I., Looney, N. F., Morley-Bunker, M. Ed. 3rd Edition. CAB International. pp. 11–17 (2011).
9. Ghrab, M., BenMimoun, M., Masmoudi, M. M. & BenMechlia, N. The behaviour of peach cultivars under warm climatic conditions in the Mediterranean area. *Int. J. Environ. Stud.* **7**, 3–14 (2014).
10. Measham, P. F., Quentin, A. G. & MacNair, N. Climate, winter chill, and decision-making in sweet cherry production. *HortScience* **49**, 254–259 (2014).
11. Zhang, L., Ferguson, L. & Whiting, M. D. Temperature effects on pistil viability and fruit set in sweet cherry. *Sci. Horticult.* **241**, 8–17 (2018).
12. Sønsteby, A. & Heide, O. M. Temperature effects on growth and floral initiation in sweet cherry (Prunus avium L.). *Sci. Horticult.* **257**, 108762 (2019).
13. Penso, G. A. et al. Development of Peach Flower Buds under Low Winter Chilling Conditions. *Agronomy* **10**, 428–448 (2020).
14. Fernandez, E., Whitney, C., Cuneo, I. F. & Luedeling, E. 2020. Prospects of decreasing winter chill for deciduous fruit production in Chile throughout the 21st century. *Clim. Chan.* **159**, 423–439 (2020).
15. Lopez, G. & DeJong, T. M. Spring temperatures have a major effect on early peach fruit growth. *J. Hort. Sci. Biotech.* **82**, 507–512 (2007).
16. Usenik, V. & Stampar, F. The effect of environmental temperature on sweet cherry phenology. *Euro. J. Horticult. Sci.* **76**, 1–5 (2011).
17. Syvertsen, J. P. Integration of water stress in fruit trees. *HortScience* **20**, 1039–1043 (1985).
18. Alae-Carew, C. et al. The impact of environmental changes on the yield and nutritional quality of fruits, nuts and seeds: a systematic review. *Environ. Res. Lett.* **15**, 023002 (2020).
19. Botzen, W. J. W., Bouwer, L. M. & van den Bergh, J. C. J. M. Climate change and hailstorm damage: Empirical evidence and implications for agriculture and insurance. *Res. En. Econ.* **32**, 341–362 (2010).
20. Seneviratne, S. I. et al. Changes in climate extremes and their impacts on the natural physical environment. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation [eds. Field, C. B. et al.]. A Special Report of Working Groups I and II of the Intergovernmenta l Panel on Climate Change (IPCC). 109–230 (Cambridge University Press, Cambridge, UK, and New York, NY, USA, 2012).
21. Lauren, E., Parker, A., McElrone, J., Ostoja, S. M. & Forrestel, E. J. Extreme heat effects on perennial crops and strategies for sustaining future production. *Plant Sci.* **295**, 110397 (2020).
22. Nelson, G. et al. Agriculture and climate change in global scenarios: why don't the models agree. *Agric. Econ.* **45**, 85–101 (2014).
23. Lampe, M. et al. Why do global long-term scenarios for agriculture differ? An overview of the AgMIP Global Economic Model Inter-comparison. *Agric. Econ.* **45**, 3–20 (2014).
24. Mendelsohn, R., R., Nordhaus, W. D. & Shaw, D. The impact of global warming on agriculture: a Ricardian analysis. *Am. Econ. Rev.* **84**, 753–771 (1994).
25. Mendelsohn, R. O. & Massetti, E. The use of cross-sectional analysis to measure climate impacts on agriculture: Theory and evidence. *Rev. Environ. Econ. Policy* **11**, 280–298 (2017).
26. Carter, C., Cui, X., Ghanem, D. & Mérel, P. Identifying the economic impacts of climate change on agriculture. *Ann. Rev. Res. Econ.* **10**, 361–380 (2018).
27. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intel. Med.* **23**, 89–109 (2001).
28. Bishop, C. M. *Pattern Recognition and Machine Learning.* (Springer 2006).
29. Bzdok, D., Altman, A. & Krzywinski, M. Statistics versus Machine Learning. *Nat. Methods* **15**, 233–234 (2018).
30. McQueen, R. J., Garner, S. R., Nevill-Manning, C. G. & Witten, I. H. Applying machine learning to agricultural data. *Comput. Electron. Agric.* **12**, 275–293 (1995).
31. González-Recio, O., Guilherme, J. M., Rosa, G. J. M. & Gianola, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Sci.* **166**, 217–231 (2014).
32. Coble, K. H., Mishra, A. K., Ferrell, S. & Griffin, T. Big Data in Agriculture: A Challenge for the Future. *Appl. Econ. Perspect. Policy* **40**, 79–96 (2018).
33. Kamilaris, A. & Prenafeta-Boldú, F. X. Deep learning in agriculture: A survey. *Comput. Electron. Agric* **147**, 70–90 (2018).
34. Liakos, K. G., Busato, P., Moshou, D., Pearson, S. & Bochtis, D. Machine learning in agriculture: A review. *Sensors* **18**, 2674 (2018).
35. Zhu, N. et al. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *Int. J. Agric. Biol. Engineer.* **11**, 32–44 (2018).
36. Ansarifar, J., Wang, L. & Archontoulis, S. V. An interaction regression model for crop yield prediction. *Sci. Rep.* **11**, 17754 (2021).
37. Tong, H. & Nikoloski, Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* **257**, 153354 (2021).
38. Jakariya, M. et al. Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. *Sci. Total Environ.* **742**, 140255 (2020).
39. Avand, M. & Moradi, H. Using machine learning models, remote sensing, and GIS to investigate the effects of changing climates and land uses on flood probability. *J. Hydrol.* **595**, 125663 (2021).
40. Guo, Y. et al. Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecol. Indicators* **120**, 106935 (2021).
41. Rolnick, D. et al. Tackling climate change with machine learning. *ACM Comput. Surveys.* **55**, Article 42 (2022).
42. IPCC. Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Pörtner, H.O. et al. (eds.)]. Cambridge University Press. In Press (2022).

43. Crane-Droesch, A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* **13**, 114003 (2018).

44. Van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **177**, 105709 (2020).

45. Meshram, V. et al. Machine learning in the agricultural domain: state-of-art survey. *Artif. Intell. Life Sci.* **1**, 100010 (2021).

46. Mark, H. S. et al. Adapting agriculture to climate change. *PNAS* **104**, 19691–19696 (2007).

47. Khan, T., Sherazi, H. H. R., Ali, M., Letchmunan, S. & Butt, U. M. Deep Learning-Based Growth Prediction System: A Use Case of China Agriculture. *Agronomy* **11**, 1551 (2021).

48. Shahhosseini, M., Hu, G., Huber, I. & Archontoulis, S. V. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci. Rep.* **11**, 1606 (2021).

49. Lobell, D. B. & Gourdji, S. M. The Influence of Climate Change on Global Crop Productivity. *Plant Physiol.* **160**, 1686–1697 (2012).

50. Bobojonov, I. & Aw-Hassan, A. Impacts of climate change on farm income security in Central Asia: An integrated modeling approach. *Agr. Ecosyst. Environ.* **188**, 245–255 (2014).

51. Abraham, T. W. & Fonta, W. M. Climate change and financing adaptation by farmers in northern Nigeria. *Financ. Innov.* **4**, 11 (2018).

52. Dalhaus, T. et al. The Effects of Extreme Weather on Apple Quality. *Sci. Rep.* **10**, 7919 (2020).

53. El Yaacoubi, A. et al. Potential vulnerability of Moroccan apple orchard to climate change–induced phenological perturbations: effects on yields and fruit quality. *Int. J. Biometeorol.* **64**, 377–387 (2020).

54. Pechan, P., Bohle, H. & Obster, F. Reducing vulnerability of orchards to climate change impacts. *Agri. Syst.* **210**, 103713 (2023).

55. Pechan, P. Obster, F. Marchioro, L. & Bohle, H. Climate change impact on fruit farm operations in Chile and Tunisia. Preprint at https://doi.org/10.31220/agriRxiv.2023.00171 (2023).

56. van Valkengoed, A. M. & Steg, L. Meta-analyses of factors motivating climate change adaptation behaviour. *Nat. Clim. Chan.* **9**, 158–163 (2019).

57. Benmoussa, L., Luedeling, E., Ghrab, M. & Ben Mimoun, M. Severe winter chill decline impacts Tunisian fruit and nut orchards. *Climatic Change* **162**, 1249–1267 (2020).

58. Verner, D. et al. Climate Variability, Drought, and Drought Management in Tunisia's Agricultural Sector. World bank Group, 114 pp, (2018).

59. Meza, F. J., Wilks, D. S., Gurovich, L. & Bambach, N. Impacts of Climate Change on Irrigated Agriculture in the Maipo Basin, Chile: Reliability of Water Rights and Changes in the Demand for Irrigation. *J. Water Resour. Plann. Manag.* **138**, 421–430 (2012).

60. Novoa, V. et al. Understanding agricultural water footprint variability to improve water management in Chile. *Sci. Total Environ.* **670**, 188–199 (2019).

61. OECD. Water and Climate Change Adaptation: Policies to Navigate Uncharted Waters. Studies on Water, OECD Publishing (2013).

62. Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. Valid Post-Selection Inference. *Ann. Stat.* **41**, 802–837 (2013).

63. Raval, M. et al. Automated predictive analytic tool for rainfall forecasting. *Sci. Rep.* **11**, 17704 (2021).

64. FAO 2022. World Food and Agriculture- Statistical yearbook 2022. Rome. https://doi.org/10.4060/cc2211en (2022).

65. International Trade Administration (US). www.trade.gov/country-commercial-guides/chile-agricultural-sector. Accessed June 2023.

66. International Trade Administration (US). www.trade.gov/country-commercial-guides/tunisia-agricultural-sectors. Accessed June 2023.

67. FAO. www.fao.org/faostat/en/#data/QCL. Accessed June 2023.

68. Fresh fruit portal. www.freshfruitportal.com/news/2023/04/10/chilean-cherries-hit-new-exports-record. Accessed June 2023.

69. FAO. www.fao.org/faostat/en/#data/QCL. Accessed June 2023.

70. Fresh plaza. www.freshplaza.com/north-america/article/9516186/strong-demand-for-tunisian-peaches-and-apricots/. Accessed June 2023.

71. Porritt, J. The World in Context: beyond the business case for sustainable development. University of Cambridge Programme for Industry. (2003). www.cisl.cam.ac.uk/publications/the-world-in-context. Accessed 12/09/14.

72. Ivory, S. & Brooks, S. B. An updated conceptualisation of corporate sustainability: Five resources sustainability. In Proceedings. British Academy of Management (BAM), British Academy of Management Annual Conference 2018, Bristol, United Kingdom (2018).

73. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. Mboost: Model-Based Boosting. R package version 2.9-7 (2022). https://CRAN.R-project.org/package=mboost.

74. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. Royal Stat. Society: Series B (Methodological)* **58**, 267–288 (1996).

75. Obster, F. & Heumann, C. Sparse-group boosting -- Unbiased group and variable selection. Preprint at https://arxiv.org/abs/2206.06344 (2022).

76. Nelder, J. A. R. & Wedderburn, W. M. Generalized Linear Models. *J. Royal Stat. Society. Series A (General)* **135**, 370–384 (1972).

77. Hofner, B. et al. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput. Sta.t* **29**, 3–35 (2014).

78. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).

79. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. (2016). https://ggplot2.tidyverse.org.

Acknowledgements

Author contributions

Conceptualisation: P.M.P. and H.B. Investigation: P.M.P. and H.B. Data curation: P.M.P. and H.B. Formal analysis: F.O. and P.M.P. Writing— F.O. and P.M.P. All authors reviewed and approved the manuscript.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-023-01128-2.

Correspondence and requests for materials should be addressed to Fabian Obster or Paul M. Pechan.

Peer review information *Communications Earth & Environment* thanks Ariel Soto-Caro, Nyong Princely Awazi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Martina Grecequet and Alienor Lavergne. A peer review file is available.

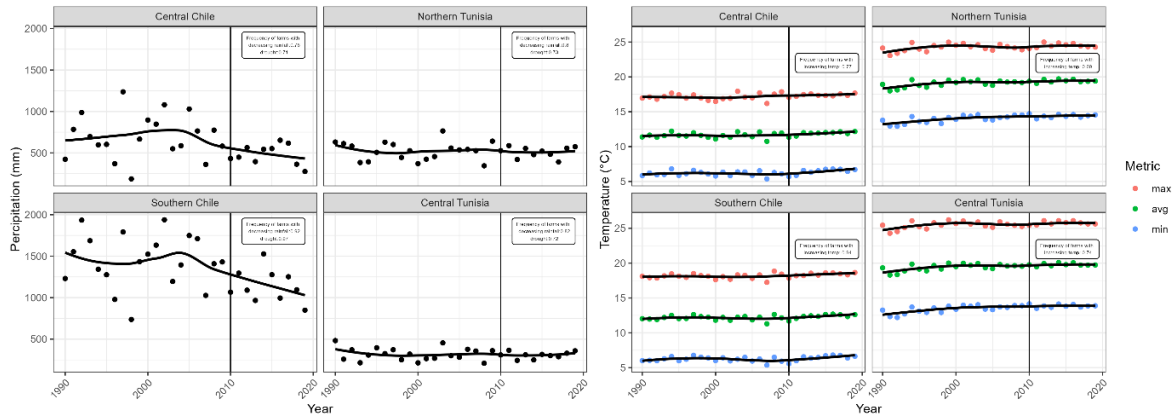**Reprints and permission information** is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

Supplementary Figure 1. Yearly average daily precipitation and temperature in the four study areas over a 30 year period until 2019. Local smoothing lines were computed using the LOESS estimator from the R package ggplot2.

Supplementary Figure 1 presents yearly average precipitation and temperature within each of the four covered regions over a 30 year period (World Bank, 2022). Calculations were based on the annual observed data for the two countries. Overall in Tunisia, the precipitation remained relatively constant compared to both regions in Chile where the variability is much larger. The largest decrease in precipitation within the covered 10-year period was witnessed in Southern Chile, where 97% of farmers reported decreasing rainfall. A similar but reverse trend can be found regarding temperature changes, with the largest increase in Southern Chile within the covered 10 year period, where 94% of farmers reported increasing temperatures.

Variable name	Category*	n	Group	Original Questionnaire scale
Agronomic measures	yes (no)	647	biophysical asset	Dichotomous
Economic measures	yes (no)	464	biophysical asset	dichotomous
Technological measures	Yes (no)	721	biophysical asset	dichotomous
Use of well as water source	Yes (no)	231	biophysical asset	dichotomous
Farm size	>7ha (≤7ha)	283	biophysical asset	Interval
Orchard size	>2ha (≤2ha)	318	biophysical asset	Interval
More than one variety grown	Yes (no)	508	biophysical asset	dichotomous
Other products	Yes (no)	571	biophysical asset	dichotomous
Regions	CentralChile	200	natural asset	Nominal
Regions	CentralTunisia	200	natural asset	Nominal
Regions	NorthernTunisia	201	natural asset	Nominal
Regions	SouthernChile	200	natural asset	Nominal
Percentage of income invested	≥80% (<80%)	137	economic asset	Dichotomous
Farm debt load	Heavy in debt (rest)	96	economic asset	Dichotomous
Family members dependent on farm	>2 (≤ 2)	528	economic asset	Count
Family farm engagement	>2 (≤2)	203	economic asset	Count
Climate change acceptance	Yes (no)	676	human asset	dichotomous
Human cause climate change	Yes (no)	685	human asset	dichotomous
Climate change causes extremes	Yes (no)	755	human asset	dichotomous
Age of farmer	>50 (≤50)	438	human asset	Count
Gender of farmer	M (F)	680	human asset	dichotomous
Education of farmer	≥ primary (no primary)	632	human asset	dichotomous
Generations of farm ownership	≥3 (0)	218	human asset	Count
Generations of farm ownership	2 (0)	130	human asset	Count
Generations of farm ownership	1 (0)	229	human asset	Count
Prior ownership	Family (other)	399	human asset	dichotomous
Years of farm management	>10 (≤10)	437	human asset	Count
Use of newspapers and magazines	4,5 (1,2,3)	95	social asset	Likert 5 point
Use of farming journals	4,5 (1,2,3)	161	social asset	Likert 5 point
Use of television	4,5 (1,2,3)	415	social asset	Likert 5 point
Use of radio	4,5 (1,2,3)	219	social asset	Likert 5 point
Use of internet	4,5 (1,2,3)	319	social asset	Likert 5 point
Use of extension experts	4,5 (1,2,3)	346	social asset	Likert 5 point
Use of government experts	4,5 (1,2,3)	166	social asset	Likert 5 point
Use of neighbours	4,5 (1,2,3)	313	social asset	Likert 5 point
Use of industry information	4,5 (1,2,3)	192	social asset	Likert 5 point
Use of farm associations	4,5 (1,2,3)	97	social asset	Likert 5 point
Trust in newspapers and magazines	4,5 (1,2,3)	174	social asset	Likert 5 point
Trust in farming journals	4,5 (1,2,3)	291	social asset	Likert 5 point
Trust in television	4,5 (1,2,3)	329	social asset	Likert 5 point
Trust in radio	4,5 (1,2,3)	241	social asset	Likert 5 point
Trust in internet	4,5 (1,2,3)	319	social asset	Likert 5 point
Trust in extension experts	4,5 (1,2,3)	433	social asset	Likert 5 point
Trust in government workers	4,5 (1,2,3)	268	social asset	Likert 5 point
Trust in neighbours	4,5 (1,2,3)	319	social asset	Likert 5 point
Trust in industry	4,5 (1,2,3)	215	social asset	Likert 5 point
Trust in farm associations	4,5 (1,2,3)	184	social asset	Likert 5 point
Trust in government institutions	4,5 (1,2,3)	213	social asset	Likert 5 point
Trust in other farmers	4,5 (1,2,3)	168	social asset	Likert 5 point
Trust in my religion	4,5 (1,2,3)	236	social asset	Likert 5 point
Trust in fate	4,5 (1,2,3)	268	social asset	Likert 5 point
Temperature increase	Yes (no)	629	climate experience	Dichotomous
Rainfall decrease	Yes (no)	659	climate experience	Dichotomous
Drought increase	Yes (no)	671	climate experience	Dichotomous
Extreme weather increase	Yes (no)	542	climate experience	Dichotomous
Temperature income damage	4,5 (1,2,3)	294	income damage	Likert 5 point
Precipitation income damage	4,5 (1,2,3)	219	income damage	Likert 5 point
Drought income damage	4,5 (1,2,3)	161	income damage	Likert 5 point
Extreme weather income damage	4,5 (1,2,3)	187	income damage	Likert 5 point

*numbers in bracket indicate the reference category

Supplementary Table 1. Overview of all independent variables and corresponding groups used in the analysis

Variable name	Question
Agronomic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *a.Changed tree thinning and pruning practices to reduce hail and rain damage b. Used chemical treatments for bud breaking and control flowering time c. Planted/re-grafted new varieties that have low chilling requirements d. Planted/re-grafted drought tolerant varieties e. Planted/re-grafted early or late maturing varieties*
Technologic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *b. Used chemical treatments for bud breaking and control flowering time f. Installed canopy (nets) against hail, sun and heat damage g. Installed water irrigation systems (for example tree drip irrigation) h. Improved the efficiency of irrigation systems to reduce water use and energy costs*
Economic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *i. Took land out of production j. Sold or rented part or all of the farm property k. Purchased crop damage insurance l. Got an off-farm job to supplement farm income (you and/or your spouse)*
Use of well as water source	Which water sources are used on your farm for agriculture (for example well, river water)
Farm size	Farm size (ha)
Orchard size	Size of your peach/cherry orchard (ha)
More than one variety grown	Varieties grown
Other products	Other farm products
Regions	Central Chile
Regions	Southern Chile
Regions	Northern Tunisia
Regions	Central Tunisia
Percentage of income invested	What percentage of your annual farm brutto income have you invested into adaptive measures in recent years to reduce the impact of climate change on your farm operations? *5 point Likert scale from 0-20% to 80-100%*
Farm debt load	How much in debt is your farm business? No debt, lightly in debt, moderately in debt, heavily in debt
Family members dependent on farm	Number of family members dependent on the farm activities
Family farm engagement	Number of family members working on your farm
Climate change acceptance	Global climate is not changing *Correct, incorrect.*
Human cause climate change	Human activities, such as burning of fossil fuels, are an important cause of climate change Correct, incorrect
Climate change causes extremes	Climatic changes can lead to an increased intensity and frequency of extreme weather events, such as hail, floods, frost and high winds Correct, incorrect
Age of farmer	Age
Gender of farmer	Gender *Male, Female*
Education of farmer	Education *1.None, 2. Incomplete primary, 3. Primary, 4. Incomplete secondary, 5. Complete secondary, 6. Technical education, (specify subject area), 7. University (specify subject area), 8. University postgraduate (specify subject area), 9. Other*
Generations of farm ownership	How many generations does your family own this farm?
Prior ownership	Who owned the farm before you?
Years of farm management	How many years have you been managing this farm?
Use of newspapers and magazines	How often do you use the following information sources on how to deal with climate change impacts on agricultural production? National newspapers or magazines *5 point Likert scale from not at all to very often*
Use of farming journals	Farming journals *5 point Likert scale from not at all to very often*
Use of television	TV *5 point Likert scale from not at all to very often*
Use of radio	Radio *5 point Likert scale from not at all to very often*
Use of internet	Internet *5 point Likert scale from not at all to very often*
Use of extension experts	Farm extension workers *5 point Likert scale from not at all to very often*
Use of government experts	Government workers/experts *5 point Likert scale from not at all to very often*
Use of neighbours	Neighbours/communities *5 point Likert scale from not at all to very often*
Use of industry information	Agriculture industry/export industry *5 point Likert scale from not at all to very often*
Use of farm associations	Farmer associations/cooperatives *5 point Likert scale from not at all to very often*
Trust in newspapers and magazines	How much do you trust the following information sources to provide you with reliable information on how protect your farm against possible climate change impacts? National newspapers or magazines *5 point Likert scale from not at all to very often*
Trust in farming journals	Farming journals *5 point Likert scale from not at all to very often*
Trust in television	TV *5 point Likert scale from not at all to very often*
Trust in radio	Radio *5 point Likert scale from not at all to very often*
Trust in internet	Internet *5 point Likert scale from not at all to very often*
Trust in extension experts	Farm extension workers *5 point Likert scale from not at all to very often*
Trust in government workers	Government workers/*experts 5 point Likert scale from not at all to very often*
Trust in neighbours	Neighbours/communities *5 point Likert scale from not at all to very often*
Trust in industry	Agriculture industry/export industry *5 point Likert scale from not at all to very often*
Trust in farm associations	Farmer associations/cooperatives *5 point Likert scale from not at all to very often*
Trust in government institutions	I trust government institutions to help me to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Trust in other farmers	I trust other farmers to advise me on what adaptive measures I should select to reduce future impacts of climate change on my farm *5 point Likert scale from strongly disagree to strongly agree*
Trust in my religion	I trust my religion more than science to guide me how to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Trust in fate	I trust in fate to guide me how to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Temperature increase	In recent years, I have observed that the temperature on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Rainfall decrease	In recent years, I have observed that the rainfall on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Drought increase	In recent years, I have observed that the dry periods and drought on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Extreme weather increase	In recent years, I have observed that the extreme weather events on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Temperature income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in temperature during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Winter temperature income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in temperature during the winter tree dormancy period, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Precipitation income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in rainfall during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Drought income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in dry periods and droughts during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Extreme weather income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in extreme weather events during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*

Variable name	Question
Agronomic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *a.Changed tree thinning and pruning practices to reduce hail and rain damage b. Used chemical treatments for bud breaking and control flowering time c. Planted/re-grafted new varieties that have low chilling requirements d. Planted/re-grafted drought tolerant varieties e. Planted/re-grafted early or late maturing varieties*
Technologic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *b. Used chemical treatments for bud breaking and control flowering time f. Installed canopy (nets) against hail, sun and heat damage g. Installed water irrigation systems (for example tree drip irrigation) h. Improved the efficiency of irrigation systems to reduce water use and energy costs*
Economic measures	From the following list, please select the adaptive measures you have already undertaken in the past 10 years to reduce the impact of climatic changes on your farming operations. More than one answer is possible. *i. Took land out of production j. Sold or rented part or all of the farm property k. Purchased crop damage insurance l. Got an off-farm job to supplement farm income (you and/or your spouse)*
Use of well as water source	Which water sources are used on your farm for agriculture (for example well, river water)
Farm size	Farm size (ha)
Orchard size	Size of your peach/cherry orchard (ha)
More than one variety grown	Varieties grown
Other products	Other farm products
Regions	Central Chile
Regions	Southern Chile
Regions	Northern Tunisia
Regions	Central Tunisia
Percentage of income invested	What percentage of your annual farm brutto income have you invested into adaptive measures in recent years to reduce the impact of climate change on your farm operations? *5 point Likert scale from 0-20% to 80-100%*
Farm debt load	How much in debt is your farm business? No debt, lightly in debt, moderately in debt, heavily in debt
Family members dependent on farm	Number of family members dependent on the farm activities
Family farm engagement	Number of family members working on your farm
Climate change acceptance	Global climate is not changing *Correct, incorrect.*
Human cause climate change	Human activities, such as burning of fossil fuels, are an important cause of climate change Correct, incorrect
Climate change causes extremes	Climatic changes can lead to an increased intensity and frequency of extreme weather events, such as hail, floods, frost and high winds Correct, incorrect
Age of farmer	Age
Gender of farmer	Gender *Male, Female*
Education of farmer	Education *1.None, 2. Incomplete primary, 3. Primary, 4. Incomplete secondary, 5. Complete secondary, 6. Technical education, (specify subject area), 7. University (specify subject area), 8. University postgraduate (specify subject area), 9. Other*
Generations of farm ownership	How many generations does your family own this farm?
Prior ownership	Who owned the farm before you?
Years of farm management	How many years have you been managing this farm?
Use of newspapers and magazines	How often do you use the following information sources on how to deal with climate change impacts on agricultural production? National newspapers or magazines *5 point Likert scale from not at all to very often*
Use of farming journals	Farming journals *5 point Likert scale from not at all to very often*
Use of television	TV *5 point Likert scale from not at all to very often*
Use of radio	Radio *5 point Likert scale from not at all to very often*
Use of internet	Internet *5 point Likert scale from not at all to very often*
Use of extension experts	Farm extension workers *5 point Likert scale from not at all to very often*
Use of government experts	Government workers/experts *5 point Likert scale from not at all to very often*
Use of neighbours	Neighbours/communities *5 point Likert scale from not at all to very often*
Use of industry information	Agriculture industry/export industry *5 point Likert scale from not at all to very often*
Use of farm associations	Farmer associations/cooperatives *5 point Likert scale from not at all to very often*
Trust in newspapers and magazines	How much do you trust the following information sources to provide you with reliable information on how protect your farm against possible climate change impacts? National newspapers or magazines *5 point Likert scale from not at all to very often*
Trust in farming journals	Farming journals *5 point Likert scale from not at all to very often*
Trust in television	TV *5 point Likert scale from not at all to very often*
Trust in radio	Radio *5 point Likert scale from not at all to very often*
Trust in internet	Internet *5 point Likert scale from not at all to very often*
Trust in extension experts	Farm extension workers *5 point Likert scale from not at all to very often*
Trust in government workers	Government workers/*experts 5 point Likert scale from not at all to very often*
Trust in neighbours	Neighbours/communities *5 point Likert scale from not at all to very often*
Trust in industry	Agriculture industry/export industry *5 point Likert scale from not at all to very often*
Trust in farm associations	Farmer associations/cooperatives *5 point Likert scale from not at all to very often*
Trust in government institutions	I trust government institutions to help me to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Trust in other farmers	I trust other farmers to advise me on what adaptive measures I should select to reduce future impacts of climate change on my farm *5 point Likert scale from strongly disagree to strongly agree*
Trust in my religion	I trust my religion more than science to guide me how to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Trust in fate	I trust in fate to guide me how to protect my farm against future impacts of climate change *5 point Likert scale from strongly disagree to strongly agree*
Temperature increase	In recent years, I have observed that the temperature on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Rainfall decrease	In recent years, I have observed that the rainfall on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Drought increase	In recent years, I have observed that the dry periods and drought on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Extreme weather increase	In recent years, I have observed that the extreme weather events on my farm *1.has increased, 2. has not changed, 3. has decreased, 4. has become unpredictable*
Temperature income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in temperature during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Winter temperature income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in temperature during the winter tree dormancy period, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Precipitation income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in rainfall during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Drought income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in dry periods and droughts during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*
Extreme weather income damage	If you deal with damage to peaches/cherries on your farm in recent years due to changes in extreme weather events during the fruit growing season, how serious would you rate the impact of the crop damage on your farm income that year(s)? *5 point Likert scale from not at all serious to very serious*

Supplementary Table 2. Relevant survey questions utilized for this manuscript

Model	Short name	package	Main hyperparameters	Interpretability
Logistic regression	glm	base	-	interpretable
Model-based boosting without interactions	mb	mboost	Nu=0.3, mstop = 3000	interpretable
Sparse group boosting	sgb	mboost	Nu=0.3, alpha = 0.5, mstop = 3000	Interpretable
Model-based boosting with interactions	mb int	mboost	Nu= 0.3, mstop = 3000	Interpretable
Random forest	Rf	randomForest	Ntree=500, mtry=7	Not interpretable
Gradient boosting machines	gbm	gbm	Trees=100, interaction.depth=3	Not interpretable
Neural Network	nn	neuralnet	hidden layers: 1, logistic activation	Not interpretable

Supplementary Table 3. Overview of the models used in the data analysis
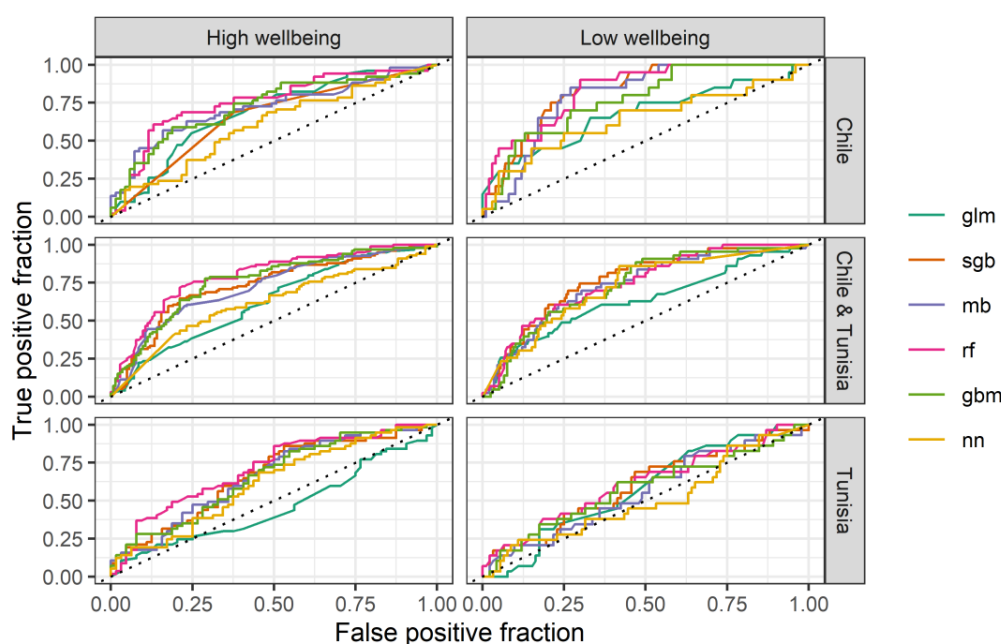
Supplementary Figure 2. Overview of ROC curves for selected models using the Chile and Tunisia datasets

	Chile & Tunisia	Chile & Tunisia	Chile	Chile	Tunisia	Tunisia
	Odds ratio (p-value)	Odds ratio (p-value)	Odds ratio (p-value)	Odds ratio (p-value)	Odds ratio (p-value)	Odds ratio (p-value)
Climate Change experience	High wellbeing	Low wellbeing	High wellbeing	Low wellbeing	High wellbeing	Low wellbeing
Increasing temperatures	0.751 (0.11)	1.238 (0.400)	0.644 (0.150)	2.428 (0.167)	0.739 (0.201)	1.237 (0.466)
Decreasing rainfall	0.635 (0.020)	1.650 (0.089)	0.437 (0.004)	2.950 (0.086)	0.860 (0.577)	1.384 (0.353)
Increasing drought frequency	1.29 (0.394)	0.906 (0.715)	0.766 (0.523)	1.052 (0.949)	1.257 (0.353)	1.099 (0.749)
Increasing extreme weather frequency	1.18 (0.302)	0.709 (0.086)	1.170 (0.584)	0.569 (0.146)	1.068 (0.751)	1.022 (0.930)
Income impact						
Increasing temperatures	1.092 (0.592)	1.234 (0.305)	1.131 (0.606)	2.119 (0.021)	1.113 (0.652)	0.741 (0.299)
Decreasing rainfall	0.568 (0.002)	1.254 (0.297)	0.434 (<0.001)	2.064 (0.028)	0.789 (0.489)	1.481 (0.289)
Increasing drought frequency	0.647 (0.031)	2.457 (<0.001)	0.664 (0.143)	2.623 (0.003)	0.656 (0.158)	2.385 (0.006)
Increasing extreme weather frequncy	1.08 (0.699)	0.871 (0.565)	1.019 (0.940)	1.074 (0.837)	1.009 (0.976)	0.926 (0.838)

Supplementary Table 4. Effect of climate change on the financial wellbeing of a farm in Chile and Tunisia. Odds ratios and corresponding p-values based on logistic regression

	Central Chile	Central Chile	Southern Chile	Southern Chile	Northern Tunisia	Northern Tunisia	Central Tunisia	Central Tunisia
	Odds ratio (p-value) High wellbeing	Odds ratio (p-value) Low wellbeing	Odds ratio (p-value) High wellbeing	Odds ratio (p-value) Low wellbeing	Odds ratio (p-value) High wellbeing	Odds ratio (p-value) Low wellbeing	Odds ratio (p-value) High wellbeing	Odds ratio (p-value) Low wellbeing
Climate Change experience								
Increasing temperatures	0.993 (0.986)	1.606 (0.676)	0.588 (0.436)	1.682 (0.537)	1.545 (0.221)	0.833 (0.643)	0.368 (0.004)	2.073 (0.139)
Decreasing rainfall	0.480 (0.059)	1.338 (0.796)	0.704 (0.549)	3.006 (0.176)	0.579 (0.171)	1.984 (0.169)	1.072 (0.863)	0.992 (0.987)
Increasing drought frequency	0.856 (0.738)	0.658 (0.717)	-*	0.511 (0.617)	1.400 (0.357)	0.912 (0.822)	1.003 (0.994)	1.364 (0.497)
Increasing extreme weather frequency	0.907 (0.848)	-*	0.793 (0.568)	0.617 (0.272)	1.187 (0.566)	1.117 (0.745)	0.998 (0.996)	1.005 (0.990)
Income impact								
Increasing temperatures	1.316 (0.455)	0.686 (0.715)	0.790 (0.528)	3.083 (0.002)	1.219 (0.590)	0.903 (0.803)	0.869 (0.677)	0.757 (0.511)
Decreasing rainfall	0.896 (0.775)	1.343 (0.762)	0.415 (0.014)	1.487 (0.307)	0.896 (0.812)	1.543 (0.362)	0.831 (0.745)	0.877 (0.842)
Increasing drought frequency	2.102 (0.191)	1.499 (0.745)	0.622 (0.229)	1.924 (0.076)	0.923 (0.837)	1.955 (0.103)	0.308 (0.044)	3.305 (0.033)
Increasing extreme weather frequncy	0.742 (0.389)	2.037 (0.445)	1.271 (0.554)	1.186 (0.670)	0.710 (0.537)	1.452 (0.496)	1.459 (0.385)	0.657 (0.455)

Supplementary Table 5. Effects of climate change on the financial wellbeing of a farm in the subregions of Chile and Tunisia. Odds ratios and corresponding p-values based on logistic regression. *Only 3 farmers did not experience extreme weather in central Chile and 17 did not experience drought in southern Chile. In these cases, no odds ratios were computed.

Supplementary Table 5 provides additional information about the effects of climate change experience and income damage on region-specific basis. Notable differences are found in Chile, where in Central Chile the effect of experiencing decreasing temperatures on high wellbeing is more pronounced than in Southern Chile and the effect of the income damage associated with decreasing rainfall is more pronounced in Southern Chile. For low wellbeing there is a strong negative effect (high odds ratios) of the financial damage associated with increasing temperatures which is not the case in Central Chile. In Central Tunisia, the effect of income damage associated with increasing drought frequency is stronger compared to Northern Tunisia for both high-wellbeing and low-wellbeing. There is also a strong negative effect of experiencing increasing temperatures on high wellbeing in Central Tunisia which is not the case in Northern Tunisia.

Supplementary references

1. World Bank 2022 "Climate change knowledge portal for development practitioners and policy makers". https://climateknowledgeportal.worldbank.org/download-data.

Part III

Variable selection biases and k-step boosting

Chapter 7

Using interpretable boosting algorithms for modeling environmental and agricultural data

This chapter deals with the problem of identifying interactions in high-dimensional data while preserving a stable selection of the main effects using a two-step boosting approach. The method uses componentwise boosting, only considering the main effects. After the first model is stopped, the base-learners are changed such that only interaction effects are boosted starting with the negative gradient of the first model in the first iteration. The method is compared to parallel estimation of the main effects, and interaction effects for the prediction of the vulnerability of farmers against five climate hazards.

Contributing article:

Author contributions:

The manuscript was written by Fabian Obster. Heidi Bohle, Christian Heumann, and Paul Pechan added valuable input and proofread the manuscript.

scientific reports

Check for updates

OPEN

Using interpretable boosting algorithms for modeling environmental and agricultural data

Fabian Obster[1,2][✉], Christian Heumann[2], Heidi Bohle[3] & Paul Pechan[3][✉]

We describe how interpretable boosting algorithms based on ridge-regularized generalized linear models can be used to analyze high-dimensional environmental data. We illustrate this by using environmental, social, human and biophysical data to predict the financial vulnerability of farmers in Chile and Tunisia against climate hazards. We show how group structures can be considered and how interactions can be found in high-dimensional datasets using a novel 2-step boosting approach. The advantages and efficacy of the proposed method are shown and discussed. Results indicate that the presence of interaction effects only improves predictive power when included in two-step boosting. The most important variable in predicting all types of vulnerabilities are natural assets. Other important variables are the type of irrigation, economic assets and the presence of crop damage of near farms.

In this work, we show how interpretable boosting algorithms can be used to predict financial vulnerabilities against multiple hazards based on environmental factors but also based on human, social, and biophysical factors as well as their interactions. For finding interactions we propose a new method based on two-step boosting, which is still interpretable and blends together with component-wise boosting. Interpretability tools like variable importance, effect sizes, and partial effects are utilized to better understand the underlying factors that may cause these vulnerabilities against climatic changes.

Model-based boosting algorithms have been used in environmental sciences for multiple purposes. For example for quantifying several soil parameters based on soil samples[1], predicting the financial wellbeing of farmers based on environmental factors[2], and predicting the number of zoo visitors based on climatic variables[3]. Also non-interpretable boosting algorithms based on classification or regression trees like Adaboost[4] have been used for environmental predictions based on environmental data because of their high predictive power. Applications include landslide susceptibility[5] and predicting the presence of juvenile sea-trouts based on environmental factors[6].

Through the proposed boosting models we want to achieve the following goals:

- *Predictive Power* The model should not only have a good fit for the analyzed data but also for unseen data from the same domain assuming a similar distribution of the variables.
- *Interpretability* We are interested in the question of which variables are associated with the outcome. But we also want to know how the associations look like. In the agronomic case, we want to derive actions to reduce vulnerability against adverse environmental changes. This is only possible if the effect of adaptive measures is known. Only if the associations are known, one can state causal hypotheses and test them with new specific experiments. We also want the effects to be modeled as simply as possible while retaining the power of the model. Linear effects should be prioritized over nonlinear effects and over interaction effects. Black-Boxes should be avoided in this case.
- *Sparsity* We consider high dimensional data sets where the number of variables p is relatively large compared to the number of observations n or even possibly higher if we consider the case with interactions. Out of the many possible variables, we want to know which ones are actually associated with the outcome and which

[1]Department of Business Administration, University of the Bundeswehr Munich, 85577 Neubiberg, Germany. [2]Department of Statistics, LMU Munich, 80539 Munich, Germany. [3]Department of Media and Communication, LMU Munich, 80539 Munich, Germany. [✉]email: fabian.obster@unibw.de; paul.pechan@ifkw.lmu.de

1

ones are not. Therefore, the model should perform variable selection to enforce sparsity. The goal is to find the smallest subset of variables that still has high predictive power. Sparsity also increases interpretability because the scientist and stakeholders only have to look at the truly relevant variables and can disregard the unimportant ones. In the vulnerability setting this could mean that farmers focus on selected variables like the type of irrigation systems rather than not selected variables like financial adaptive measures.

- *Complexity* The model should be as complex as necessary and as simple as possible. Complexity is the characteristic that balances all previously stated points. Out of two explanations with the same predictive power the model should pick the one that is simpler. By simpler, we mean sparser, more interpretable, and without interactions. On the other hand, we do not want to neglect important complexities like non-linearity and interactions. It is important to identify if some variables are modified by others. There could also be non-hierarchical interactions, where a variable has by itself no effect on the outcome, but may have a positive effect in one subset of the data and a negative one in the other. One example could be, that in one region a high variety of crops has a positive effect on vulnerability and in another region a negative effect.
- *Group structure* The variables in the data can be clustered into groups. "Climate change experience" is one example and contains the binary variables "increasing temperature", "increasing drought", "increasing extreme weather" and "decreasing rain". The question is whether the outcome is influenced by each or only by some of the individual variables or if they act as a group due to the similarity. Group structures also increase interpretability, because it is often easier for humans to comprehend the overall effect of an abstract concept than to look at all its facets.

There are many approaches to deal with each of the above specifications. For example, sparsity can be achieved through Lasso Regression[7] or boosted Lasso[8], predictiveness can be achieved through a big variety of models and group structures can be incorporated with the sparse group lasso[9].

In this work we focus on how these goals can be met using boosting algorithms, namely componentwise boosting (mb), componentwise boosting with interactions (mb int), sparse group boosting (sgb), and two-step boosting for interactions (2-boost). We compare their predictive power, effect sizes, and the relative importance of variables/groups. In the following, we describe the used methods for the analysis and discuss how they help to achieve the stated goals using modifications of the generic boosting algorithm.

Methods

Introduction of the data.
Randomly selected cherry and peach farmers in the selected regions of Tunisia and Chile. In order to be selected for the survey, farmers had to own the farm, manage and work on the farm and derive the majority of their income from their farming activities. A total of 801 face-to-face interviews were subsequently conducted with farmers who fulfilled the selection criteria—401 peach farmers in Tunisia (201 in Mornag and 200 in Regueb regions) and 400 cherry farmers in Chile (200 in Rengo and 200 in Chillán regions). Mornag, Tunisia (longitude: 10.28805, latitude: 36.68529, altitude: 110 m), hereafter referred to as Northern Tunisia, is located approximately 20 km east of the capital Tunis. Regueb (longitude: 9.78654, latitude: 34.85932; altitude: 230 m), Tunisia, hereafter referred to as Central Tunisia, is located approximately 230 km south of Tunis. Rengo (longitude: −70.86744, latitude: −34.40237, altitude: 570 m), Chile, hereafter referred to as Central Chile, is located approximately 110 km south of Santiago de Chile. Chillán (longitude: −72.10233, latitude: −36.60626, altitude: 120–150 m), Chile, hereafter referred to as Southern Chile, is located approximately 380 km south of Santiago de Chile. The approximately one-hour-long interviews were carried out with farmers directly on their farms. The interviews were carried out after harvest completion in the fall of 2018 by Elka Consulting in Tunisia and in the spring 2019 by Qualitas AgroConsultores in Chile. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent for the data collection was provided by the survey participants. No personality-identifiable data was collected, assuring full anonymity. Department of Communication and Media Research, University of Munich had been consulted about the participation of human subjects in the survey research. Guidance was sought from our institute about the survey implementation and data use that included participation of human subjects. Experimental protocol was approved by University of Munich. A descriptive description of the data[10] and further mixed methods analysis on vulnerability[11] with similar data was performed.

Code availability. The R code of the analysis can be found at https://github.com/FabianObster/boostingEcology.

Independent variables. The analyzed variables can be clustered into groups, including

- Climate experience group (Increasing temperature, decreasing rain, increasing drought, increasing extreme weather)
- Natural asset group (geographical regions)
- Social asset group (reliance on/use of information, trust in information sources, community, science or religion)
- Human asset group (age, gender, education)
- Biophysical asset group (farm size, water management systems used on the farm, diversity of crops used, adaptive measures)
- Economic asset group (farm debt, farm performance, reliance on orchard income)
- Goals group (Keep tradition alive, work independently)

- Harm group (Climate threatens farm, Optimism)
- Spatial group (Crop damage near farms, Crop damage of farms in Country)

An overview of all variables and the belonging groups can be found in Tables 4 and 5. There, also the number of farmers in each category can be found (Tables 1, 2).

Outcome variables. The outcome variables measure financial vulnerability against the 5 climate hazards, increasing winter temperatures, increasing summer temperatures, decreasing rainfall, increasing drought, and increasing extreme weather based on self-assessment of the farmers. For each of the hazards, a binary variable indicating if a farmer is vulnerable to the hazard is defined as the outcome variable. The main category includes farmers, who are not financially vulnerable and the reference category includes farmers who are financially vulnerable. The number of farmers in each category can be found in Table 3.

Interaction variables. 22 variables were used as variables that may have an interaction effect with the other variables on the outcome. The interaction variables include regions as well as socio-demografic variables amongst others and are indicated in bold in Tables 4 and 5. Together with all other variables, this yields 1366 interaction terms and over 4000 possible model parameters to estimate. Since there are 801 farmers in the data,

AUC sgb	AUC mb	AUC 2-boost	AUC mb int	Outcome vulnerability
0.656	0.619	0.608	0.587	Summer temperature
0.707	0.708	0.713	0.705	Winter temperature
0.852	0.852	0.852	0.500	Decreasing rainfall
0.768	0.768	0.768	0.500	Drought
0.776	0.778	0.783	0.773	Extreme weather

Table 1. AUC values for the sparse group boosting (sgb), component-wise boosting (mb), parallel boosting with interaction (mb int) and two-step boosting with interactions (2-boost) for all vulnerability outcomes evaluated on the test data.

Model	Number selected interaction terms	1-Sparsity in percent	Outcome vulnerability
mb int	13	0.95	Summer temperature
2-boost	0	0	Summer temperature
mb int	38	2.78	Winter temperature
2-boost	12	0.88	Winter temperature
mb int	48	3.51	Decreasing rainfall
2-boost	1	0.07	Decreasing rainfall
mb int	27	1.98	Drought
2-boost	16	1.17	Drought
mb int	32	2.34	Extreme weather
2-boost	10	0.73	Extreme weather

Table 2. Comparison of the number of selected interaction terms based on two-step estimation (2-boost) and the parallel estimation (mb int) and the percentage of selected interactions (1-Sparsity) of the 1366 interaction terms.

Variable	Category	n
No summer temperature vulnerability	Yes	358
No winter temperature vulnerability	Yes	579
No decreasing rainfall vulnerability	Yes	451
No drought vulnerability	Yes	492
No extreme weather vulnerability	Yes	453

Table 3. Overview over outcome variables. Financial vulnerability against climate hazards. The "n" column gives the number of farmers who are not financially vulnerable to each of the hazards.

Variable name	Category	n	Group name
Agronomic measures	Yes	647	Biophysical asset group
Economic measures	Yes	464	Biophysical asset group
Use of river irrigation	Yes	138	Biophysical asset group
Use of well irrigation	Yes	231	Biophysical asset group
Farm size	Yes	283	Biophysical asset group
Orchard size	Yes	318	Biophysical asset group
More than one variety grown	Yes	508	Biophysical asset group
Other products	Yes	571	Biophysical asset group
Technological measures	Yes	721	Biophysical asset group
Increasing temperature	Yes	629	Climate experience group
Decreasing rainfall	Yes	659	Climate experience group
Increasing drought	Yes	671	Climate experience group
Increasing extreme weather	Yes	542	Climate experience group
Income invested > 80 Percent	Yes	137	Economic asset group
Income invested <40 percent	Yes	358	Economic asset group
High financial wellbeing	Yes	346	Economic asset group
Low financial wellbeing	Yes	148	Economic asset group
Farm debt load	High	96	Economic asset group
Dependent on farm	Yes	528	Economic asset group
Family farm engagement	Yes	203	Economic asset group
Adaptive measures efficacy	High	490	Efficacy group
Work independent	Yes	635	Goals group
Keep tradition alive	Yes	460	Goals group
Provide good living environment	Yes	466	Goals group
Be in profitable business	Yes	320	Goals group
Climate change is harmful	Yes	258	Harm group
High optimism	Yes	446	Harm group
High certainty	Yes	470	Harm group
Climate threatens farm	Yes	629	Harm group
Climate risks > benefits	Yes	648	Harm group
Climate change acceptance	Yes	676	Human asset group
Human cause climate change	Yes	685	Human asset group
Climate extremes	Yes	755	Human asset group
Age > 50	Yes	438	Human asset group
Gender	F	121	Human asset group
Gender	M	680	Human asset group
Education	Yes	459	Human asset group
Years of farm possession	Yes	577	Human asset group
Prior ownership (family)	Yes	399	Human asset group
Years of farm managing	Yes	437	Human asset group
Natural assets	CentralChile	200	Natural asset group
Natural assets	CentralTunisia	200	Natural asset group
Natural assets	NorthernTunisia	201	Natural asset group
Natural assets	SouthernChile	200	Natural asset group
Adapive measures near farms	1	424	Norms group
Adapive measures near farms	2	151	Norms group
Adapive measures near farms	3	226	Norms group
High optimism	Yes	446	Perception group

Table 4. Overview over variables and groups. The 22 variables used as interaction variables (potential moderators) are bold. The number of observations within each category of each variable is in the n column. For binary variables, only one category is presented and the remaining category is "no" if the shown category is "yes" and "low" if the shown category is "high".

finding interactions results in a "p > n" problem, where the number of variables in the design matrix is greater than the number of observations.

Variable	Category	n	Group
Use of newspapers	Yes	95	Social asset group
Use of farming journals	Yes	161	Social asset group
Use of TV	Yes	415	Social asset group
Use of radio	Yes	219	Social asset group
Use of internet	Yes	319	Social asset group
Use of extension workers	Yes	346	Social asset group
Use of government workers	Yes	166	Social asset group
Use of neighbours	Yes	313	Social asset group
Use of industry	Yes	192	Social asset group
Use of farm associations	Yes	97	Social asset group
Trust in newspapers	Yes	174	Social asset group
Trust in farming journals	Yes	291	Social asset group
Trust in TV	Yes	329	Social asset group
Trust in radio	Yes	241	Social asset group
Trust in internet	Yes	319	Social asset group
Trust in extension workers	Yes	433	Social asset group
Trust in government workers	Yes	268	Social asset group
Trust in neighbours	Yes	319	Social asset group
Trust in industry	Yes	215	Social asset group
Trust in farm associations	Yes	184	Social asset group
Trust in government institutions	Yes	312	Social asset group
Trust in other farmers	Yes	351	Social asset group
Trust in religion	Yes	317	Social asset group
Trust in fate	Yes	360	Social asset group
Crop damage near farms	Yes	643	Spatial group
Crop damage farms in Country	Yes	673	Spatial group
Climate change occurs	Yes	592	Spatial group

Table 5. Overview over variables and groups continued. The 22 variables used as interaction variables (potential moderators) are bold. The number of observations within each category of each variable is in the n column. For binary variables, only one category is presented and the remaining category is "no" if the shown category is "yes" and "low" if the shown category is "high".

General setup, model formulation and evaluation.

All analyses were performed with R[12] and the boosting models were fitted with the package "mboost"[13].

Since all outcome variables are binary, we use the Ridge penalized negative log-likelihood of the binomial distribution as a loss function and a logit link, which yields

$$h(\beta, X_i) = P(y_i == 1) = \frac{1}{1 + \exp{(-X_i^T \beta)}},$$

$$l(y, h) = - \left[\sum_{i=1}^{n} y_i \log{(h(\beta, X_i))} + (1 - y_i) \log{(1 - h(\beta, X_i))} \right] + \lambda \|\beta\|_2^2.$$

Before performing any analysis the data was split into 70 percent training data and 30 percent test data, which was only used for the final evaluation. Variable importance and partial effects were computed using the whole data after the predictive analysis. Model evaluation was based on the area under the receiver operator curve (ROC) and computed using the test data. The area under the ROC (AUC) takes both the true positive and the false positive rate into account by considering all possible thresholds of predicted probabilities computed by a prediction model. While the AUC is often used for discriminatory performance, it is also limited by not assessing calibration and in the presence of strong class imbalances.

In the analysis, we use multiple boosting models for multiple purposes. All boosting models were fitted with the R package "mboost"[14]. For early stopping, the stopping parameter was determined using a 10-fold cross-validation performed at every boosting step. The first and most simple one is component-wise model-based boosting (mb) with ridge-regularized linear effects of all variables, such that the degrees of freedom are all equal to one. This model allows us to perform variable selection and allows for a comparison between all variables regarding their relative importance. For the second model, we used sparse group boosting with a mixing parameter $\alpha = 0.5$, which balances group selection and individual variable selection. This way it is possible to see if variables are important on their own for the outcome, or if they rather act as groups of variables.

To find interactions in the data we use two approaches. The first one is the standard approach by defining linear effects and interaction effects at the same time in each iteration. Then the model can decide weather it selects the main effects or the interaction effects. In the second approach we use a two-stage boosting model. As the first step we use the already fitted mb model, which only uses individual linear base-learners. The second step uses solely interactions. This way linear base-learners are prioritized over interaction base-learners since they are fitted first.

This remaining part of the methods section is more technical and may be skipped by the application-oriented reader.

Generic boosting algorithm. We will start with the general formulation of the boosting algorithm which can also be described as a functional gradient descent algorithm. The goal is to find a function f^* that minimizes some Loss function $l(y, f)$. Here, we only consider differentiable convex loss functions. The loss function has two arguments. The first argument $y \in \{1, ..., n\}$ is the outcome variable with n observations. The second argument f is a real-valued function $f : \mathbb{R}^{n \times p} \mapsto \mathbb{R}$, which is a function of the data $X \in \mathbb{R}^{n \times p}$.

Another way of fitting sparse regression models is through the method of boosting. The fitting strategy is to consecutively improve a given model by adding a base-learner to it. Throughout this article, we refer to a base-learner as a subset of columns of the design matrix associated with a real-valued function. To enforce sparsity, each base-learner only considers a subset of the variables at each step[15]. In the case of component-wise $\mathcal{L}^2$ boosting, each variable will be a base-learner. In the case of a one-dimensional B-Spline, a base-learner is the design matrix representing the basis functions of the B-Spline. The goal of boosting in general is to find a real valued function that minimizes a typically differentiable and convex loss function $l(\cdot, \cdot)$. Here we will consider the negative log-likelihood as a loss function to estimate f^* as

$$f^*(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}[l(y, f)].$$

General functional gradient descent Algorithm[16].

1. Define base-learners of the structure $h : \mathbb{R}^{n \times p} \to \mathbb{R}$
2. Initialize $m = 0$ and $\widehat{f}^{(0)} \equiv 0$ or $\widehat{f}^{(0)} \equiv \bar{y}$
3. Set $m = m + 1$ and compute the negative gradient $\frac{\partial}{\partial f} l(y, f)$ and evaluate it at $\widehat{f}^{[m-1]}$. Doing this yields the pseudo-residuals $u_1, ..., u_n$ with

$$u_i^{[m]} = \frac{\partial}{\partial f} l(y_i, f)|_{f = \widehat{f}^{[m-1]}},$$

   for all $i = 1, ..., n$
4. Fit the base-learner h with the response $(u_1^{[m]}, ..., u_n^{[m]})$ to the data. This yields $\widehat{h}^{[m]}$, which is an approximation of the negative gradient
5. Update

$$\widehat{f}^{[m]} = \widehat{f}^{[m-1]} + \eta \cdot \widehat{h}^{[m]}$$

   here η can be seen as learning rate with $\eta \in]0, 1[$
6. Repeat steps 3,4 and 5 until $m = M$

Boosted ridge regression. The loss function $l(\cdot, \cdot)$ can be set to any function. In the case of interpretable boosting, the negative log-likelihood is a reasonable choice. The log-likelihood can also be modified using a Ridge penalty. By introducing the hyperparameter $\lambda > 0$, one can modify the loss function l. Let h be a function of a parameter vector $\beta \in \mathbb{R}^p$ and the design matrix $X \in \mathbb{R}^{n \times p}$, then

$$l_{\text{Rige}}(u, h) = l(u, h) + \lambda \|\beta\|_2^2$$

is the Ridge penalized loss function. By increasing λ, the parameter vector β can be shrunken towards zero. Closely related to λ are the degrees of freedom. Let S be the approximated generalized ridge hat matrix as in Proposition 3 in[17]. We remark that in the special case of ordinary least squares ridge regression we have $S = X(X^T X + \lambda I)^{-1} X^T$. Generally, the degrees of freedom can be defined as

$$\text{df}(\lambda)) = \text{tr}(2S - (S)^T S)).$$

It is recommended to set the regularization parameter for each base-learner, such that the degrees of freedom are equal for all base-learners. Thus, the regularization parameter enables using complex base-learners like polynomial effects and simple effects like linear effects at the same time. Since the more complex base-learners are regularized more than the simpler ones it is possible to prioritize simple and more interpretable base-learners over complex ones, introducing an inductive bias towards interpretability, as we demanded in the problem statement.

Component-wise and group component-wise boosting. In step 4 of the general functional gradient descent algorithm, the function h is applied. Instead of just one function, one can also use a set of R functions

6

$\{(h_r)_{r \leq R}\}$. Then the update in step 5 is only performed with the function that has the lowest loss function applied to the data, meaning $r^* = \arg\min_{r \leq R} \mathbb{E}[l(u, h_r)]$. In the case of component-wise boosting, for each variable in the dataset, a function is used that is only a function of this variable and not the others. This way in each step only one variable is selected. Then through early-stopping, or setting M relatively smaller compared to the number of variables in the dataset, a sparse overall model can be fitted. This addresses the sparsity requirement in the problem statement section. In the case of grouped variables, one can also define base-learners as groups of variables, which are a function of only the variables belonging to one group. These could be all item variables that belong to a specific construct like in sociological data[18] or all climate change-related variables in agricultural and environmental data[2]. This allows group variable selection, where only a subset of groups is selected, yielding a group/construct-centric analysis rather than on an individual-variable basis. This way, the group structure can be taken into account.

Sparse group boosting.   It is also possible to use individual and group-based base-learners at the same time. Then at each step, either an individual variable or a group of variables is selected. Using a similar idea as in the sparse group lasso[9], the sparse group boosting can be defined[19]. We do this again by modifying the degrees of freedom. Each variable will get its own base-learner, and each group of variables will get one base-learner, containing all variables of the group. Let G be the number of groups and p_g the number of variables in group g. Then, for the degrees of freedom of an individual base-learner $x_j^{(g)} \in \mathbb{R}^{n \times 1}$ we will use

$$\mathrm{df}\left(\lambda_j^{(g)}\right) = \frac{1}{p_g} \cdot \alpha.$$

For the group base-learner we use

$$\mathrm{df}\left(\tilde{\lambda}^{(g)}\right) = \frac{1}{p_g} \cdot (1 - \alpha).$$

The mixing parameter $\alpha \in [0, 1]$ allows to change the prioritization of groups versus individual variables in the selection process. If $\mathrm{df}(\lambda) = 0$ means $\lambda \to \infty$, $\alpha = 1$ yields component-wise boosting, and $\alpha = 0$ yields group boosting.

Two-step boosting.   In the generic boosting algorithm, a single set of functions is applied sequentially to the data. While there is variable selection within the set of functions, the set itself does not change during the boosting procedure. We describe a modification of the general that allows more flexibility, namely a two-step version of boosting. A similar idea of two-step boosting, called hierarchical boosting has been used in genetic research[20] in transfer learning[21], and also deep learning applications[22]. In most cases, hierarchical boosting is used, if the outcome variable consists of a hierarchical class structure[23]. In contrast to the data analyzed in the literature, the data we analyze here does not contain hierarchical class structures. Hence, we do not use hierarchical boosting as in most cases presented in the literature, but for hierarchical and non-hierarchical interaction detection.

We formulate and generalize the two-step boosting. Let K be the number of steps and for every step $k \leq K$ let H_k be the set of base-learners.

K-step boosting algorithm.

1. Set K as the number of steps
2. For every step $k \leq K$ define the set of base-learners H_k to be used and set M_k to the number of boosting iterations
3. Initialize $m_0 = 0$ and $\widehat{f}^{(0)} \equiv 0$ or $\widehat{f}^{(0)}$
4. For $k \leq K$ repeat:
5. For $m_k \leq M_k$ perform steps 2-6 of the general boosting algorithm
6. Set Initialization $m_k = 0$ and $u^{[0]} = u^{[M_{k-1}]}$

One may ask why it is necessary to run multiple boosting algorithms after each other if it is possible to just use more base-learners in parallel in the original method. Previous research has shown high predictive powers in some combinations of steps. However, as described in the problem statement for us predictive power is only one part of the requirements and not necessarily desirable if it comes at the cost of interpretability and understanding of the data. Also, the sequential nature of the algorithm reduces computational improvements through parallelization, as not all base-learners can be fitted in the same boosting iteration in parallel. The k-step boosting algorithm can also be seen as a special case of the general boosting algorithm, where the base-learners themselves are boosting algorithms.

Variable importance.   For each of the previously described boosting methods, it is possible to compute a variable importance measure. In each step, the log-likelihood is computed, which means that one can compute the reduction of log-likelihood attributed to the base-learner being selected in the step. After the fitting for each base-learner the total reduction of likelihood can be computed. This way, one can compute the percentage of reduction in the negative log-likelihood attributed to each base-learner, regardless of the type of base-learner. The variable importance allows us to compare the relative importance of variables compared to each other and

is distinct from the concept of significance or *p* values which tests a hypothesis of a parameter not being zero based on a set of assumptions. Hence a variable can be important in boosting while not being significant base on classical regression and vice versa.

Partial effect and effect sizes.   For boosted generalized linear models, partial effects can be computed[13]. Similar to classical logistic regression, odds ratios for all base-learners can be computed by first summing up all linear predictors for one base-learner. These odds ratios can then be interpreted similarly to effect sizes in logistic regression. Based on the linear predictor one can also compute predicted probabilities for categories of variables if all other base-learners are set to average values. This way partial effects can be plotted, both for individual variable base-learners and for interaction-base-learners. Thus model-based boosting models are by themselves interpretable compared to other machine learning models where only post-hoc explanations can be derived. One can also track which variable was selected in each boosting iteration and thus understand how the model works internally.

Applications
Predictability.   Referring to Table 1 and Fig. 1 we can see that the AUC values are comparable between the boosting models except for the interaction model with parallel estimation. Averaging the AUC values across the five vulnerabilities, sgb yields 0.752, mb and 2-boost yield 0.745, and mb int 0.613. For precipitation and drought vulnerability, the parallel estimation of interactions resulted in no variables being selected and therefore the AUC takes a value of 0.5. In 2-boost, also no variables were selected in the second estimation resulting in the same model as mb, which had the highest AUC for these outcome vulnerabilities. For summer temperature vulnerability, sgb had the highest AUC, and for winter temperature and extreme weather 2-boost had the highest AUC. Comparing the predictability of the individual vulnerabilities with each other, we see, that vulnerability against decreasing rainfall can be predicted better with the given variables, followed by vulnerability against increasing extreme weather, decreasing drought, increasing winter temperature, and summer temperature.

Importance of individual variables and groups.   Comparing the variable importance of the sparse group boosting (sgb) and componentwise boosting (mb) in Fig. 2, it becomes apparent, that while there is some overlap, also some variables differ. The single most important variable for all outcomes is "Natural assets" indicating the four regions of the farm. However, the relative importance of the natural assets is higher for sgb than for mb for all five vulnerabilities. Groups seem to be more important in explaining increasing temperature vulnerability than the other vulnerabilities, as the economic asset group is the second most important variable for summer temperature vulnerability and the goals group is the second most important variable for winter temperature vulnerability. The spatial group is the third most important variable for decreasing rainfall vulnerability but the relative importance is minor compared to the most important variable.
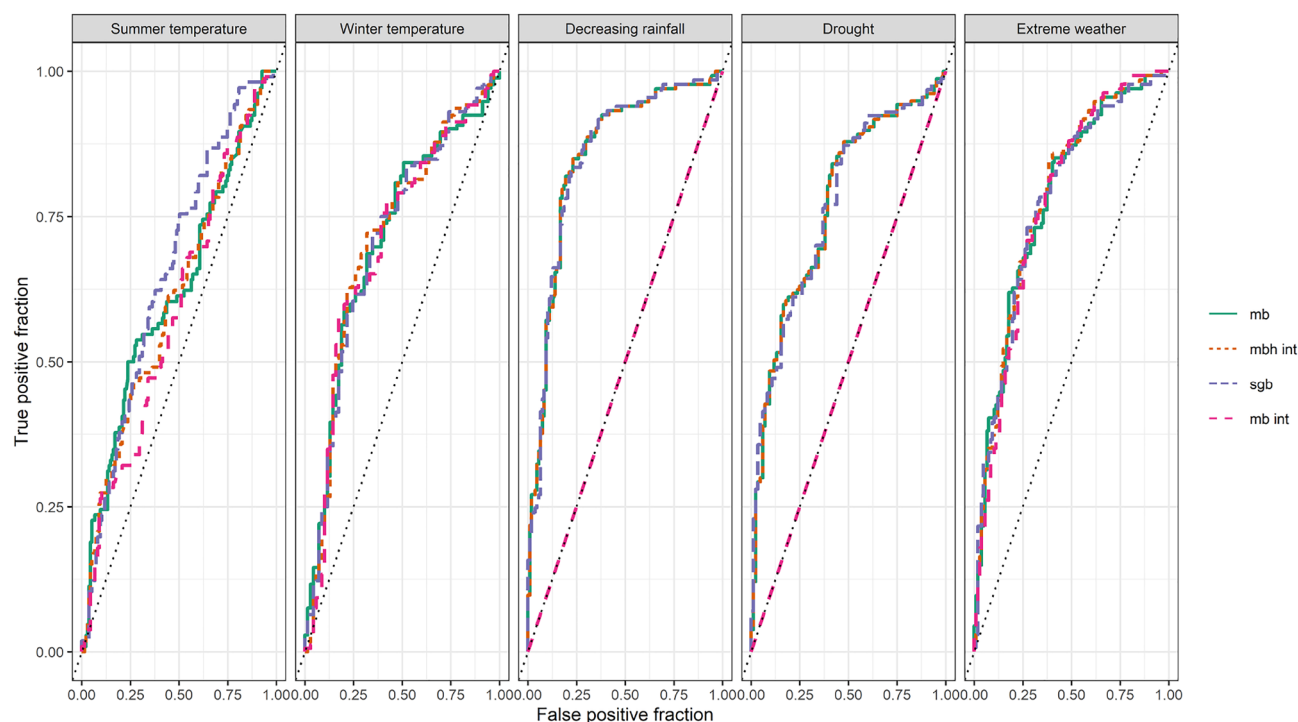
Figure 1. ROC-curves for the sparse group boosting (sgb), component-wise boosting (mb), parallel boosting with interaction (mb int) and two-step boosting with interactions (2-boost) for all vulnerability outcomes evaluated on the test data.
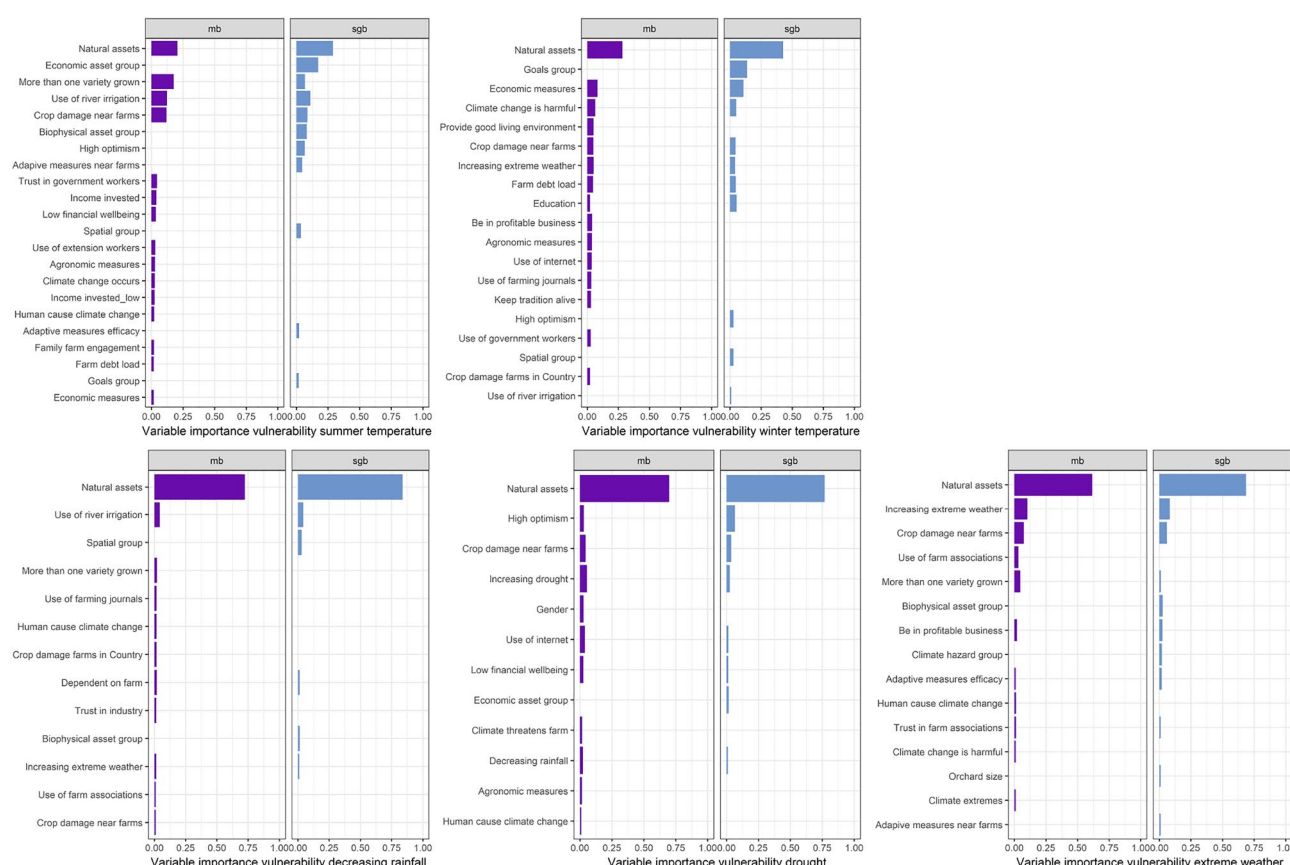
Figure 2. Comparison of variable importance based on component-wise boosting (mb) and sparse group boosting (sgb) for each vulnerability. The ordering of variables is based on the sum of relative importance for both models. only variables with a relative contribution of at least one percent and at most 15 variables per model are shown.

Importance of interactions.   In the predictability section, we have already seen some differences between the two-step and the parallel estimation for interaction effects. For predictions, only models trained on the training data were used for model evaluation on the test data. For the variable importance in Fig. 3 and Sparsity in Table 2 the whole data was used. The parallel estimation selected only interaction effects and no main effects (individual variables), whereas the two-step estimation selected both.

Referring to Table 2 it becomes apparent that the selection of variables differs substantially. Overall, the two-step estimation in 2-boost yields much fewer interactions. For summer temperature vulnerability, no interaction term was selected, whereas for the parallel estimation, 13 interaction effects were selected. For decreasing rainfall vulnerability the differences are also substantial. The two-step estimation selected only one interaction, namely the one between Agronomic measures and trust in TV was selected and mb int selected 48. For drought vulnerability, the difference was the smallest with 27 interactions for the parallel and 16 for the two-step estimation. The percentage of selected interactions was four out of five times below one percent for 2-boost and for mb int it was above one percent four out of five times.

Not only does the sparsity differ, but also the selected interactions themselves. Referring to Fig. 3, for winter temperature vulnerability the two interactions "Natural assets"-"Be profitable business" and "Country"-"Farm debt load" have high relative importance based on both models. But apart from those two, there is almost no overlap. For example for decreasing rainfall vulnerability, the only selected interaction between "Agronomic measures"-"Trust in TV" has a relative importance of 1 based on 2-boost and is not selected based on mb int, which in turn selected 48 other interactions.

In Figs. 4, 5, 6, 7 and 8 we plotted the four most important interaction effects for each of the vulnerabilities found in mb int and 2-boost based on a classical logistic regression only using one interaction term at a time. There, the probability of no vulnerability is plotted based on the joint categories of the interaction. This is done once for the data in Chile, Tunisia, and the whole data. Exemplary, we interpret the two common interaction effects "Natural assets"-"Be profitable business" and "Country"-"Farm debt load" for winter temperature vulnerability, which was selected by both models. In the northern region of Chile, having compared to not having the goal of being a profitable business is associated with a higher probability of not being vulnerable to increasing winter temperatures. In the southern Region of Chile, the association is reversed, meaning that having compared to not having the goal of being a profitable business is associated with a lower probability of not being vulnerable against increasing winter temperatures. In Tunisia, in both regions, the association of having the goal of being a profitable business is negative but more negative in the Southern region compared to the Northern
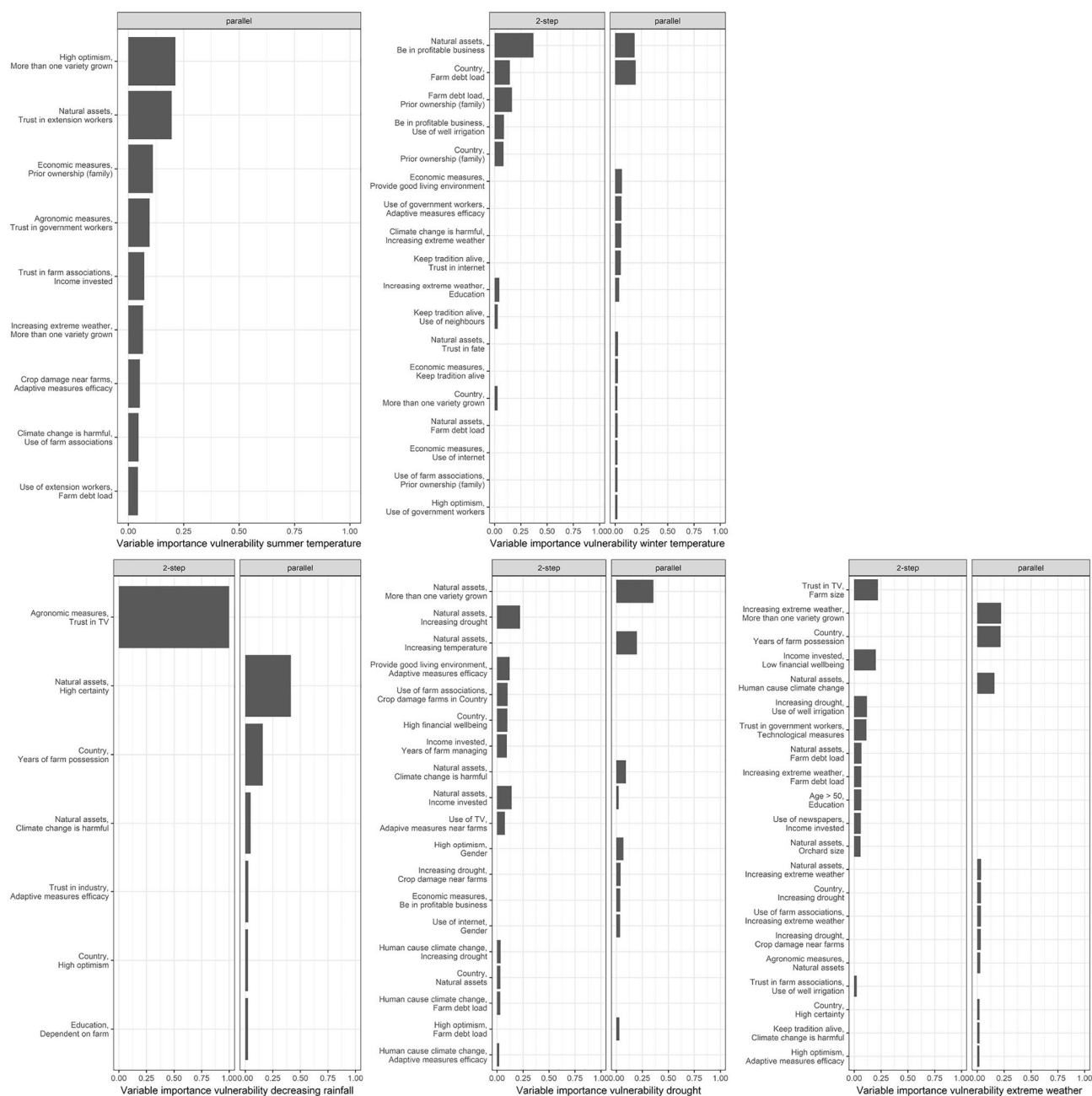
Figure 3. Variable importance of interaction terms in two-step estimation (2-boost) and parallel estimation (2-boost) for each vulnerabilities. The ordering of variables is based on the sum of relative importance for both models Only variables with a relative contribution of at least two percent and at most 15 variables per model are shown.

region. Based on the interaction term "Country"-"Farm debt load", high farm debt load has a positive association with the probability of not being vulnerable to increasing winter temperature, where the association is negative in Tunisia. The positive association in Chile is stronger in the northern region and the negative association in Tunisia is stronger in the southern region.

Discussion

The results indicate that the vulnerability of farmers in Chile and Tunisia against climate hazards can be predicted with the interpretable boosting algorithms and their variations by the variables and groups of variables used in the analysis. All models performed variable selection. The highest predictive power measured in AUC was achieved for vulnerability against decreasing rainfall and the lowest for summer temperature increases regardless of the type of boosting approach. For predicting summer temperature vulnerability the sparse group boosting outperformed all other models indicating that there may be underlying latent variables that cause the effects rather than the individual variables. The group variable importance mainly points to economic and
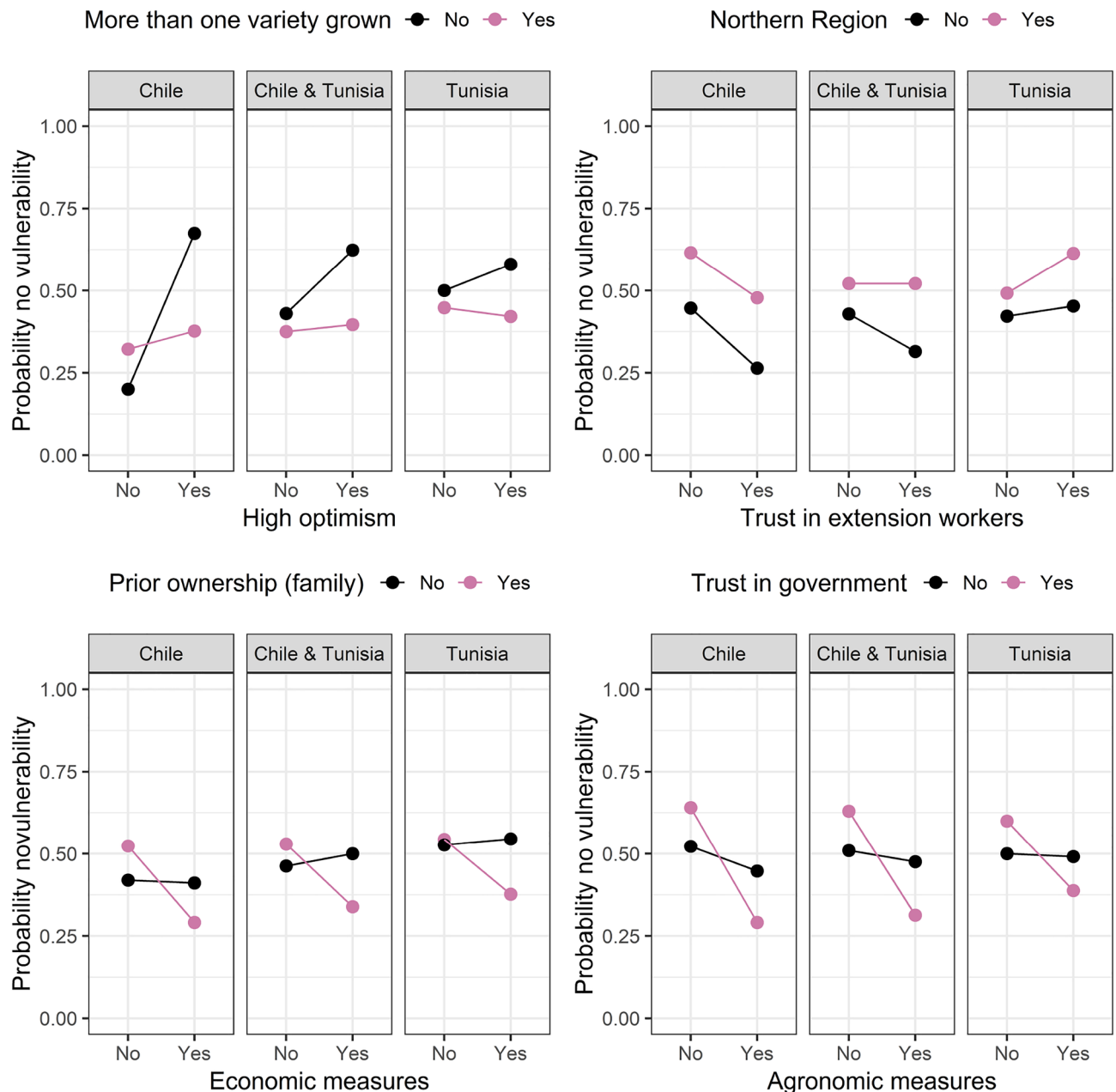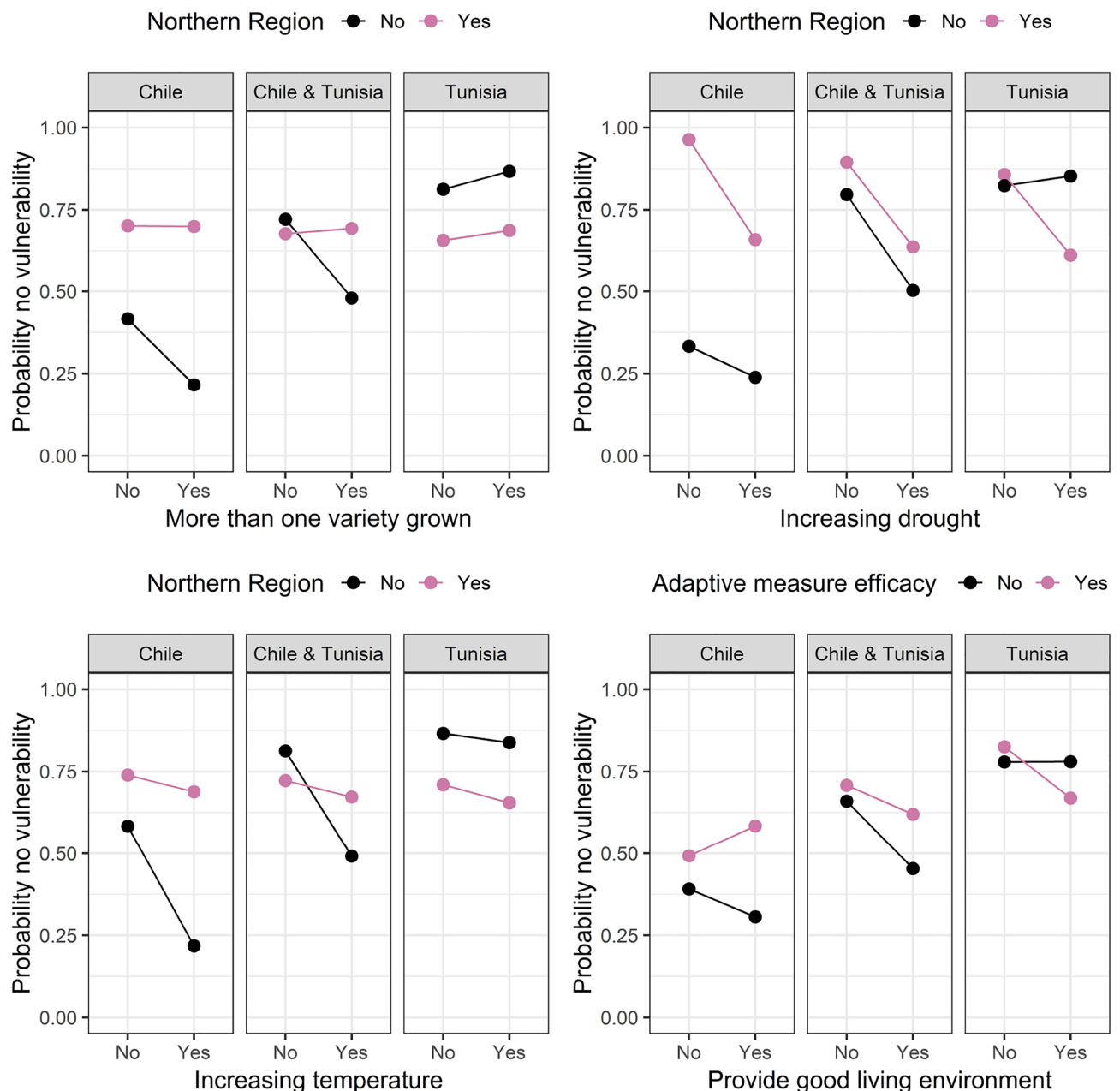
Figure 4. Probability of not being vulnerable against increasing summer temperature based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

biophysical assets including adaptive measures which may be an underlying determinant for summer temperature vulnerability. The variable importance strongly points to Natural assets consisting of the four different regions in Chile and Tunisia, which are a main determinant of all types of vulnerability. This indicates strong within and between country differences. The interaction analyses also confirm the importance of regionality, as some effects are strongly modulated by Country and North-South comparisons. The modulated effect of debt load by region may be an indication of economic differences between regions and closeness to bigger cities or could be a result of the different climatic zones.

Even though there are strong interaction effects present in the data as seen in the univariate interaction analysis, it is not a simple task to transfer their presence into higher predictive power in a high-dimensional setting. This becomes apparent since the model including interactions base-learners additionally to the main effects performed worse than the same model without interactions in all cases. One of the reasons is probably overfitting, as the number of parameters to estimate exceeds the number of variables by a factor of over four. The result was that the interaction model did not include any main effects and only interactions. We believe that this issue of overfitting becomes more systematic in high-dimensional data than purely random because there
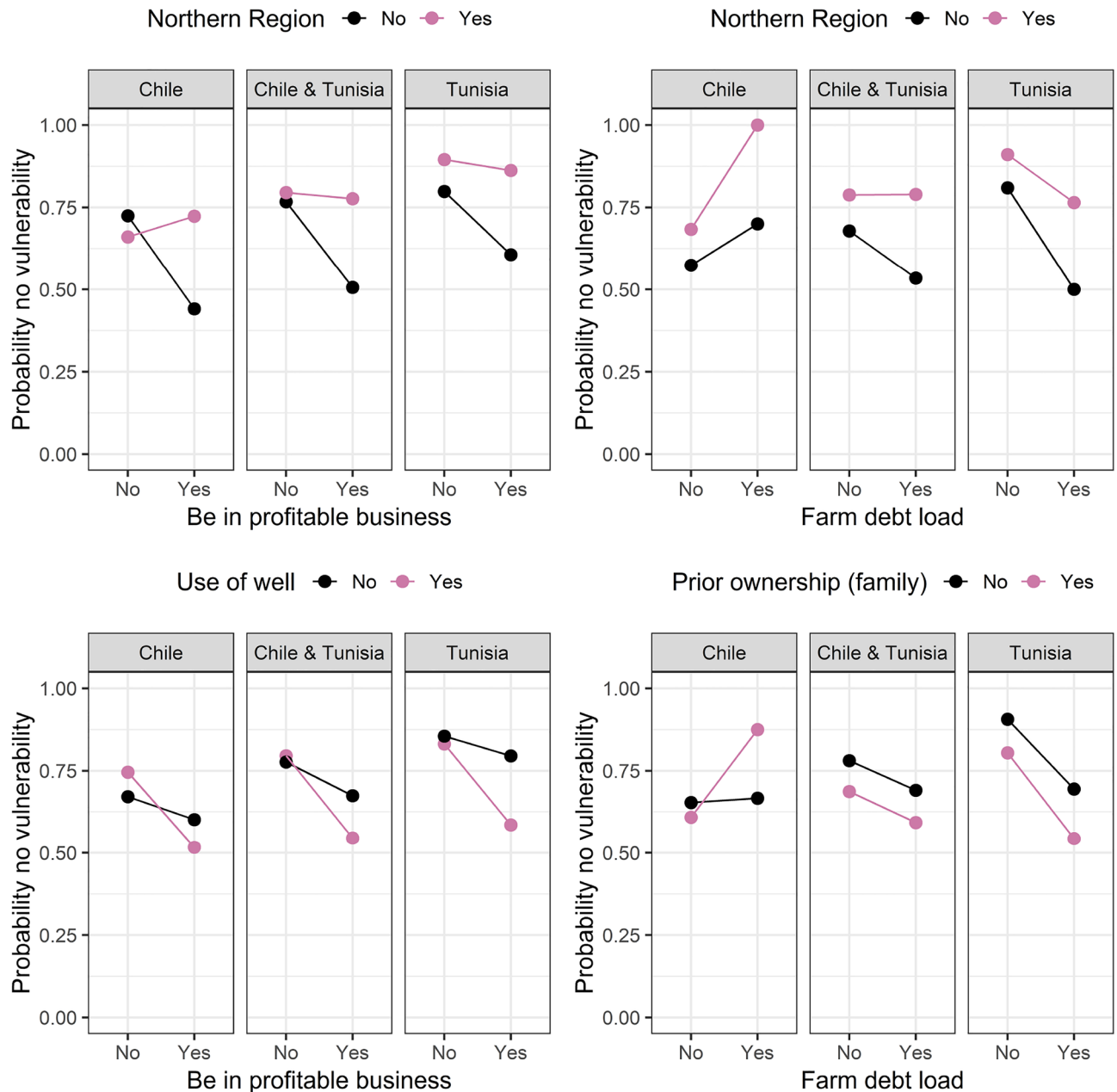
Figure 5. Probability of not being vulnerable against increasing winter temperature based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

if there are p variables in the dataset, then there are $\mathcal{O}(p^2)$ possible interaction terms. So, with increasing p, the chance of selecting an interaction term over a main effect increases with regardless of the actual effect sizes. This implicit interaction selection bias could be addressed successfully by the proposed two-step boosting approach. The two-step boosting yielded higher predictive power and a higher degree of sparsity with fewer interactions being present in the resulting model. This leads us to believe that this approach is superior to the "classical" parallel estimation by including interaction terms in the main model formula in boosting. The only drawback we see is, that one has to estimate two models instead of just one which slightly increased the programming effort and reduces the potential for further parallelization as the models are fitted sequentially and not in parallel. However, it is common practice and in line with the principle of sparsity to always fit one model that contains only individual variables if one wants to do an interaction analysis[24]. In this case, the two-step boosting is also computationally more efficient because one can build upon the first model and avoid having to refit the main effect.
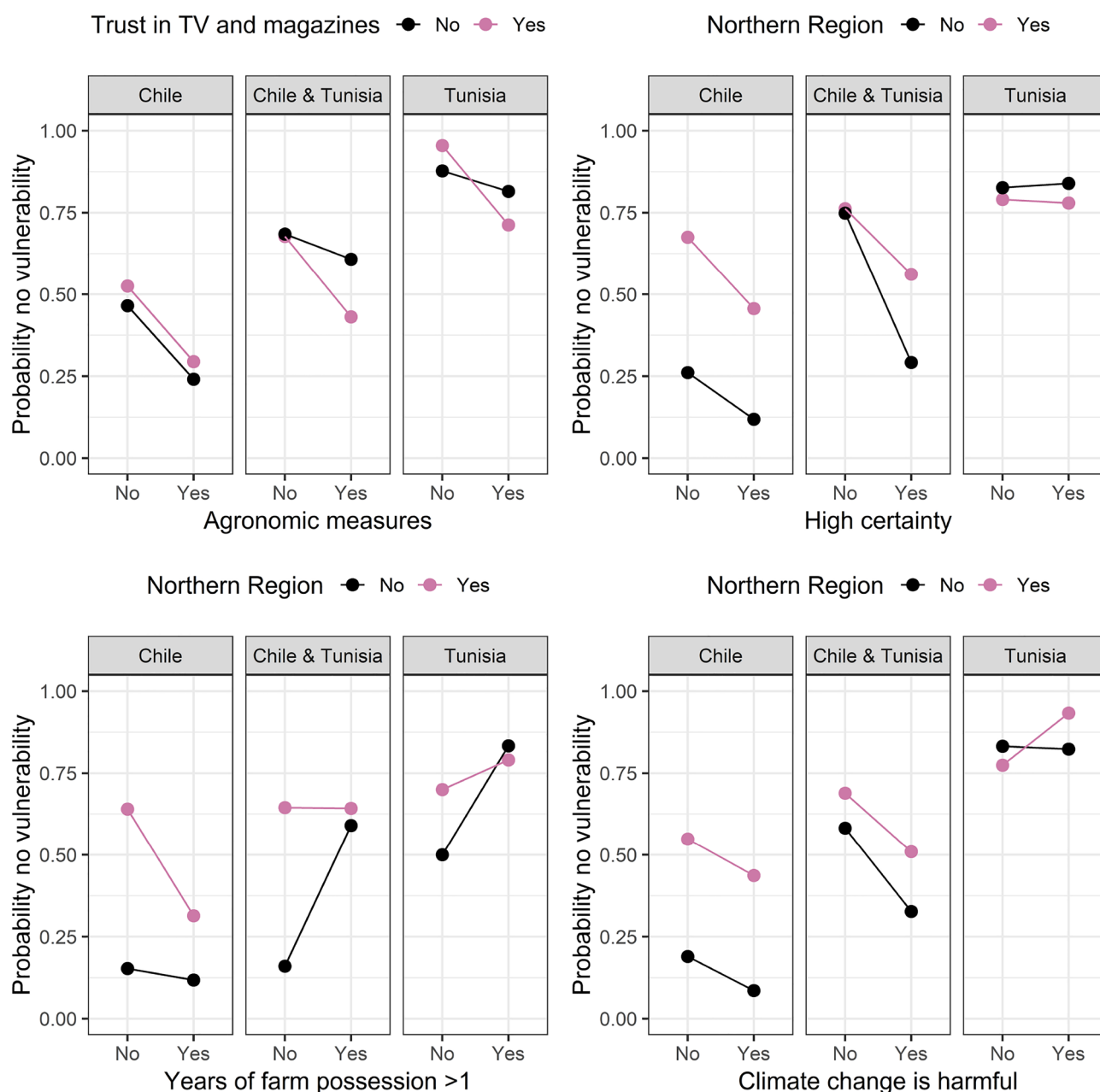
Figure 6. Probability of not being vulnerable against decreasing rainfall based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

In environmental research, consistently finding associations in high-dimensional datasets requires new methods to advance knowledge. These new methods allow more flexibility but often come at the cost of classical statistical inference, including *p* values and estimations of standard errors as in the case of boosting[25].

Figure 7. Probability of not being vulnerable against drought based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

Often, there are multiple plausible explanations for a phenomenon. The here proposed methods can enable direct comparison of a large number of explanations, estimating their explanatory importance for the outcome. This approach can accelerate understanding, particularly for newer phenomena like climate change, by gathering all variables that may be associated with the outcome and sampling observations for them. Starting with a relatively small sample size, one can estimate the relative importance of hypotheses and prioritize future research based on the results.

Using an apriori interpretable method, such as those previously described, provides the great advantage of being able to assess the predictability of a given set of explanations for an outcome. In contrast, post-hoc interpretability tools applied to a black box provide only a simplified explanation of how black-box predictions may be derived, without being able to assess how good the explanations themselves are at predicting the outcome.
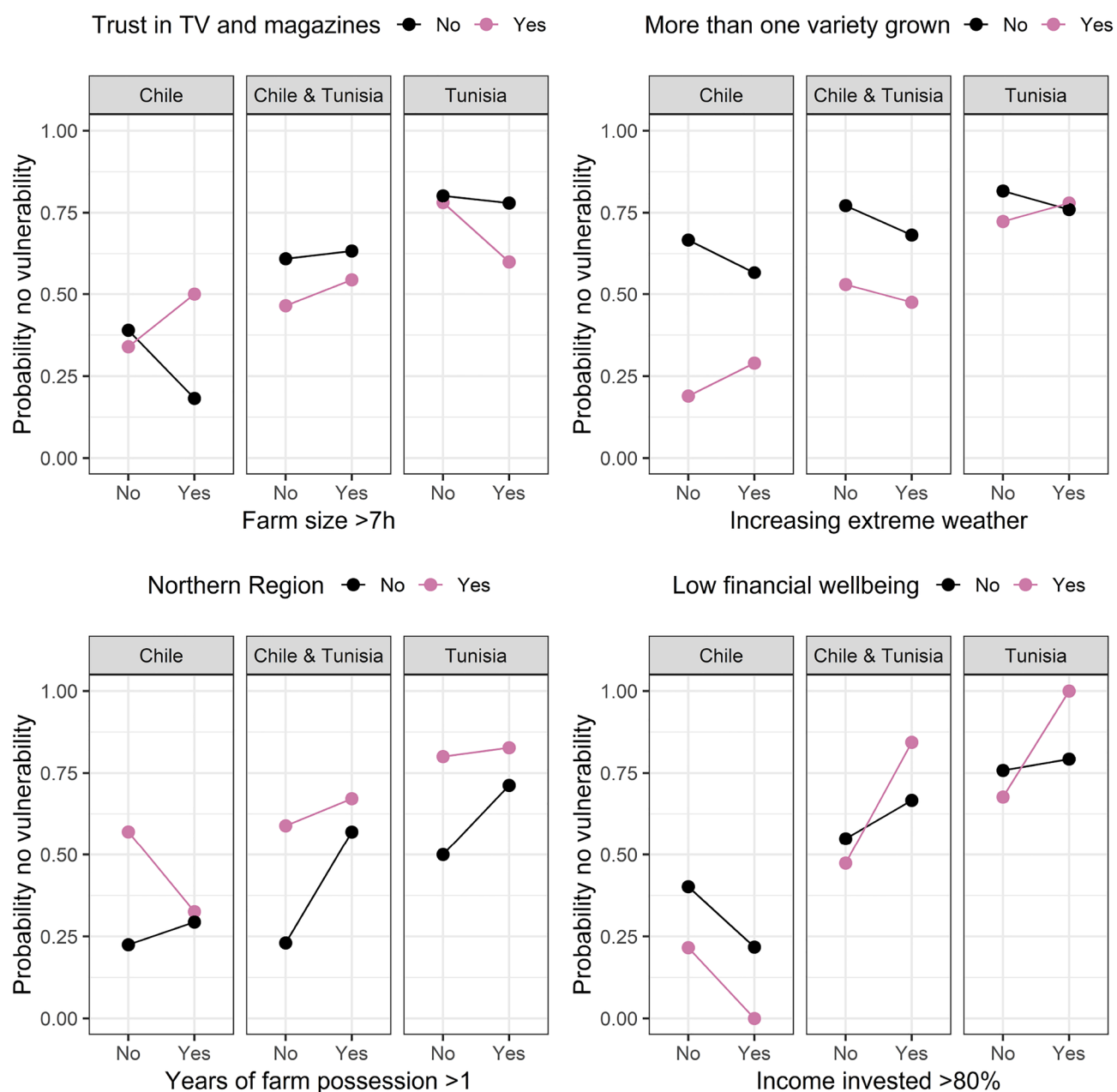
Figure 8. Probability of not being vulnerable against extreme weather based on the categories of the four most important interaction effects found in mb int and 2-boost. Probabilities are based on classical logistic regression only using one interaction term at a time. The results are once stratified by country (Chile, Tunisia) and once estimated on the whole data.

Data availability
The dataset used and analysed during the current study is available from the corresponding author on reasonable request.

References
1. Li, B., Chakraborty, S., Weindorf, D. C. & Yu, Q. Data integration using model-based boosting. *SN Comput. Sci.* **2**, 400. https://doi.org/10.1007/s42979-021-00797-0 (2021).
2. Obster, F., Bohle, H. & Pechan, P. M. Factors other than climate change are currently more important in predicting how well fruit farms are doing financially. arXiv:2301.07685 [cs, stat] (2023).
3. Obster, F., Brand, J., Ciolacu, M. & Humpe, A. Improving boosted generalized additive models with random forests: a zoo visitor case study for smart tourism. *Procedia Comput. Sci.* **217**, 187–197. https://doi.org/10.1016/j.procs.2022.12.214 (2023).
4. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139. https://doi.org/10.1006/jcss.1997.1504 (1997).

5. Jennifer, J. J. Feature elimination and comparison of machine learning algorithms in landslide susceptibility mapping. *Environ. Earth Sci.* **81**, 489. https://doi.org/10.1007/s12665-022-10620-5 (2022).
6. Froeschke, J. T. & Froeschke, B. F. Spatio-temporal predictive model based on environmental factors for juvenile spotted seatrout in Texas estuaries using boosted regression trees. *Fish. Res.* **111**, 131–138. https://doi.org/10.1016/j.fishres.2011.07.008 (2011).
7. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**, 267–288 (1996).
8. Zhao, P. & Yu, B. Boosted Lasso. Tech. Rep., California Univ Berkeley Dept of Statistics. Section: Technical Reports (2004).
9. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
10. Pechan, P. M., Obster, F., Marchioro, L. & Bohle, H. Climate change impact on fruit farm operations in Chile and Tunisia. *agriRxiv* **2023**, 20230025166. https://doi.org/10.31220/agriRxiv.2023.00171 (2023).
11. Pechan, P. M., Bohle, H. & Obster, F. Reducing vulnerability of fruit orchards to climate change. *Agric. Syst.* **210**, 103713. https://doi.org/10.1016/j.agsy.2023.103713 (2023).
12. Team, R. RStudio: Integrated Development Environment for R (2020).
13. Hofner, B., Mayr, A., Robinzonov, N. & Schmid, M. Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput. Stat.* **29**, 3–35. https://doi.org/10.1007/s00180-012-0382-5 (2014).
14. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. mboost: Model-based boosting. *CRAN* (2022).
15. Bühlmann, P. & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* **22**, 477–505. https://doi.org/10.1214/07-STS242 (2007).
16. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
17. Tutz, G. & Binder, H. Boosting ridge regression. *Comput. Stat. Data Anal.* **51**, 6044–6059. https://doi.org/10.1016/j.csda.2006.11.041 (2007).
18. Agarwal, N. K. Verifying survey items for construct validity: A two-stage sorting procedure for questionnaire design in information behavior research. *Proc. Am. Soc. Inf. Sci. Technol.* **48**, 1–8. https://doi.org/10.1002/meet.2011.14504801166 (2011).
19. Obster, F. & Heumann, C. Sparse-group boosting—Unbiased group and variable selection. arXiv:2206.06344 [stat] (2022).
20. Pybus, M. *et al.* Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31**, 3946–3952. https://doi.org/10.1093/bioinformatics/btv493 (2015).
21. Wang, C., Wu, Y. & Liu, Z. Hierarchical boosting for transfer learning with multi-source. In *Proceedings of the International Conference on Artificial Intelligence and Robotics and the International Conference on Automation, Control and Robotics Engineering*, ICAIR-CACRE '16, 1–5. https://doi.org/10.1145/2952744.2952756 (Association for Computing Machinery, New York, 2016).
22. Yang, F. *et al.* Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. arXiv:1901.06340 [cs] (2019).
23. Valentini, G. Hierarchical ensemble methods for protein function prediction. *ISRN Bioinform.* **2014**, 901419. https://doi.org/10.1155/2014/901419 (2014).
24. Aguinis, H. & Gottfredson, R. K. Best-practice recommendations for estimating interaction effects using moderated multiple regression. *J. Organ. Behav.* **31**, 776–786. https://doi.org/10.1002/job.686 (2010).
25. Hofner, B., Mayr, A. & Schmid, M. gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. arXiv:1407.1774 [stat] (2014).

Author contributions

P.P. and H.B. accumulated the data, F.O. performed machine learning and statistical modeling, and F.O. analysed the results. All authors reviewed the manuscript.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.O. or P.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 8

Improving Boosted Generalized Additive Models with Random Forests: A Zoo Visitor Case Study for Smart Tourism

In Chapter 7, a more complex model is used to improve the predictions of a simpler model. This chapter uses the somewhat counterintuitive idea of utilizing a complex model to improve the simpler model. Generalized additive models (GAMs) can be sensitive to outliers or unstable in sparse areas of the feature space. Instead of fitting a boosted GAM to the observed outcome, the same model is fitted to the predictions of a random forest. In a case study of predicting zoo visitors, the idea is tested. Both versions of the boosted GAMs are compared in terms of predictive performance, sparsity, and variable importance.

Contributing article:

Author contributions:

The manuscript was written by Fabian Obster. Josephine Brand, Monica Ciolacu, and Andreas Humpe added valuable input and proofread the manuscript

4th International Conference on Industry 4.0 and Smart Manufacturing

Improving Boosted Generalized Additive Models with Random Forests: A Zoo Visitor Case Study for Smart Tourism

Fabian Obster[a]*, Josephine Brand[b], Monica Ciolacu[c], Andreas Humpe[b]

aUniversity of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
bMunich University of Applied Sciences, Schachenmeierstrasse 35, 80636 Munich, Germany
cUniversity of Passau, Dr. Hans-Karpfinger-Str. 14d, 94032 Passau, Germany

Abstract

Smart Tourism for the Industry 4.0 and post Covid-19 challenge needs explainable AI Algorithms adapted for the Volatility, Uncertainty, Complexity and Ambiguity (VUCA) World with smart (physical components, algorithms, and IoT/mobile connectivity) elements. This paper shows how boosted generalized additives models (GAM) and random forest can be used in conjunction to improve the prediction and model explainability at the same time. This is achieved by using the predictions of the random forest as an outcome of the boosted GAM. Boosted GAMs can not only improve the explainability of random forest, but the random forest can also improve the predictability of boosted GAMs for modeling zoo visitors. This approach also has desirable regularization properties, such as model sparsity of the boosted GAMs. In addition, the current state of the art is provided and a detailed description with descriptive analysis of a case study for zoo visitors. The procedure with integrated XAI techniques, like variable importance measures and partial effects, is explained. In the future, the proposed concept can be implemented also for other industries or as a general method of XAI

Keywords: Boosted Generalised Additive Models; Random Forests; Zoo Visitor Prediction; Explainable AI in Smart Tourism; Partial Effects and Variable Importance

* Corresponding author.
E-mail address: fabian.obster@unibw.de

1. Introduction

Explainable AI Although COVID-19 has changed the way many people work and interact, people are still at the center of business - and they need digitized processes to operate in today's environment. For example, sensors/RFID tags were used to determine whether employees wash their hands regularly. Computer Vision determined whether employees were adhering to mask protocol, and loudspeakers were used to alert people to protocol violations. In addition, these behavioral data were collected and analyzed by the organizations to influence employee behavior on the job. The collection and use of such data to drive behavior is referred to as the Internet of Behavior (IoB). As organizations improve not only the amount of data they collect but also the way they combine and use data from multiple sources, IoB will continue to influence the way organizations to interact with people [1].

The past decade has seen rapid advancements and increasing use of Artificial Intelligence (AI) in all industry sectors. However, this leap in performance has often been achieved through high model complexity with "black-box" approaches. These, in turn, lead to uncertainty about how they operate and how they arrive at decisions. This is highly problematic, especially in sensitive and critical areas such as autonomous driving or healthcare. As a result, scientific interest in the field of explainable artificial intelligence (XAI) has grown tremendously. This, in contrast to established methods, is a field concerned with the development of new methods that explain and interpret machine learning models. An overview of the development of XAI methods offers inter alias Linardatos et al [2] and Confalonieri et al. [4]. Figure 1 illustrates the major milestones of AI (left) and XAI (right). For example, local explanation methods are used in XAI. Here, the individual predictions of a black-box model can be approximated by generating local surrogate models as well as interpreted intrinsically. This method is implemented, for example, in the LIME (Local Interpretable Model-agnostic Explanations) algorithm by Ribeiro et al. [3]. Here, the LIME approach exploits the fact that the trained black-box model can be queried multiple times for the predictions of specific instances. By changing the data used for training, LIME generates a new data set. After the black-box model is fed the modified data, it creates a new interpretable model from the predictions generated over the new data set. When the XAI method provides an explanation for only a particular instance, it is called a local approach, and when the method explains the entire model, it is called global. In contrast to LIME as a local approach, methods like PDPbox or Shapley Additive Explanations (SHAP) are global XAI approaches that try to explain the whole black-box model [2].

Fig. 1.: Major milestones of AI (left) and XAI (right).

The US Defense Advanced Research Project Agency (DARPA) is working on the Explainable Artificial Intelligence project: an open-source framework that makes procedures and methods available for the explainability and comprehensibility of AI-based recommendations for action [31]. The Harvard project Visual Analysis for

Recurrent Neural Networks (LSTMVis) focuses on the explainability of neural sequence models that can be generated by so-called recurrent neural networks, for example for text translation, generation, or interpretation. The aim of this article is to investigate the differences between an interpretable machine learning model, that are fit to the data, compared to the same model fit to the same data with the predictions of a black-box machine learning model as an outcome. We will specifically analyze the differences regarding predictability, quantified by the mean squared error using independent data, variable importance, and partial effects. Here we define the variable importance as the reduction of the loss function, that can be contributed to each individual variable. As black-box model, we use a random forest (rf) and as the interpretable model we use boosted generalized additive models (GAM). A comparison of post hoc interpretability methods explaining random forests, is given in [30], where also GAMs as interpretable surrogate model are discussed. We want to shed light on two important questions. First, can the GAM fit to the predictions of the black-box model to identify bias of the black-box predictions? The comparison might reveal biases of the machine learning model and help to understand the relationship of variables and the prediction of the black-box model compared to the actual relationship of the variables with the outcome estimated by the GAM. As a result, we do not only try to understand the behavior of the black-box but also detect any biases of the black-box. This might help the modeler to decide between two black-box models, based on which yields the most accurate interpretation. Further, it could help to adjust the modeling, based on the detected bias. Second, can the black-box model help to reduce the variance or bias of the interpretable model? Generally, GAMs are prone to overfitting if the raw data set is used.

2. Industrial Application

The tourism industry has been growing for years and continues to be considered an increasingly important sector of the economy [5, 6, 7]. However, the growth has led to increased competition and the industry must face additional challenges, such as the Covid-19 pandemic or climate change. Due to these developments, it is more important than ever for each individual company to optimize and analyze its business activities, resources, and planning. [8] For efficient forward planning, it is important for tourism companies to be able to predict future visitor numbers as accurately as possible [8]. Two of the factors that influence visitor numbers are climate and weather. The climate describes the weather pattern of a place. The period of the weather data must be sufficient to provide significant mean values. Weather, on the other hand, reflects the state of the atmosphere at a given location within a brief time span. [9] Consequently, climate influences the decision of global tourists who prefer a certain weather pattern for their vacation. Weather, on the other hand, has an impact on spontaneous decisions of domestic tourists and thus outbound and domestic tourism, as weather depicts a brief moment of climate [5, 7]. To date, research has mostly focused on the effects of weather and the resulting behavior of tourists in non-urban destinations [10]. For example, Ploner and Brandenburg [11] modeled visitor volume as a function of weather and days of the week in an Australian national park. Álvarez-Díaz and Rosselló-Nadal [12] focused on improving forecasts of British tourist arrivals to the Balearic Islands. Becken [13] measured the effect of different weather variables on the number of flights and visits from a tour operator and visitor center in the community of Franz Josef in Westland, Australia. Other research addresses the monitoring of the Castelldefels urban beach in Barcelona, Spain, to provide the best possible service for the corresponding demand [14] and the analysis of climate sensitivity and the impact of climate change on outdoor recreational activities of two beach resorts in the city of Zurich [15]. Furthermore, literature exists on the influence of weather on the Coachella Valley in California [16] and visitor forecasts for the Museum of New Zealand "Te papa tongarewa" [10]. With the topic of forecasting attendance at zoological parks, Aylen, Albertson, and Cavan [17], Perkins and Debbage [18], Perkins [19], and Álvarez and Barquín [20] have already written research papers in different destinations such as Spain, England, and the United States. Aylen, Albertson, and Cavan [17] and Álvarez and Barquín [20] conclude that visitor numbers are influenced by the weather but are also very dependent on seasonal rhythms, such as school vacations and public holidays. This is because the weather did not seem to be the determining variable in summer [20]. Perkins and Debbage [18] and Perkins [19] noted that it is equally important to consider in which climate zone the zoo is located, as locals consider different weather conditions to be comfortable. Zoo visitors are predominantly day visitors and thus highly dependent on the weather, as they often make their travel decisions on short notice and adjust their plans to the short-term weather forecast [17]. In addition, short-term weather forecasts

have improved in recent years and are now considered to be very accurate. Therefore, it can be assumed that the actual weather data can be used as a good substitute for the originally predicted weather. For the prediction of visitor numbers based on weather data, the following different methods have been used and partially compared so far. Table 1 shows which methods have been used in the literature to predict visitor numbers.

Table 1. Methods used in the existing literature.

Authors\ Methods	Time series model	Decision tree	Bayesian model	Gradient boosting	Generalized linear model	Neuronal network	Random forest
Lise & Tol (2002)					X		
Ploner & Brandenburg (2003)		X			X		
Álvarez-Díaz & Rosselló-Nadal (2010)						X	
Bergmeir & Benitez (2011)	X						
Finger & Lehmann (2012)					X		
Akin (2014)						X	
Aylen, Albertson & Cavan (2014)	X						
Clark et al. (2019)			X				
Yap et al. (2020)				X	X	X	X
Domingo (2021)		X				X	
Lionetti et al. (2021)		X		X	X	X	

3. Data & descriptive analysis

According to Porter & Heppelmann [21] smart products have three elements: physical components, smart components, and connectivity components. The analysis used the daily data from the Helsinki Zoo visitor count as a dependent variable (physical component). The dataset includes the data from the 1st of January in 2010 until the 31st of December in 2021. As independent variables, the weather data of the weather station Kaisaniemi in Helsinki and the public holidays were collected (physical components). The weather data consists of precipitation, snow, average air temperature, minimum temperature, and maximum temperature. Table 2 gives an overview of the sources for the Korkeasaari Zoo visitor numbers, weather data, and public holidays. The zoo visitor prediction through machine learning algorithms (smart components) can be used for smart tourism applications (connectivity components) like visitor management or resource planning.

Table 2. Overview of the sources for the data.

Sources	Visitor Numbers of Korkeasaari Zoo	Weather data	Corona cases	Public holidays
Helsinki, Finland	https://hri.fi/data/en_GB/dataset/korkeasaaren-kavijamaarat	https://cdn.fmi.fi/fmiodata-convert-api/preview/5d7ea17c-52c4-4c8e-9045-4770ec434e57/?locale=en	https://sampo.thl.fi/pivot/prod/en/epirapo/covid19case/summary_tshcddaily?alue_0=445222&alue_1=445193	https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2010&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2011&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2012&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2013&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2014&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2015&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2016&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2017&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2018&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2019&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2020&klasse=0&hl=en https://www.feiertagskalender.ch/ferien.php?geo=3295&jahr=2021&klasse=0&hl=en

The outcome of interest is number of daily zoo visitors. Figure 2 shows that this count variable is right skewed. However, on a logarithmic scale this is not the case anymore, but there are two modes. Referring to Figure 3, for instance the univariate relationship between the air temperature and the number of visitors is nonlinearly increasing and there is a degree of rising heteroscedasticity. That means that the number of visitors scatters more for lower air temperatures than for higher air temperatures. In contrast, logarithmic transformation of the number of visitors, yields a "more linear" relationship and less heteroscedasticity. This relationship could also be found in other numeric variables in the dataset.
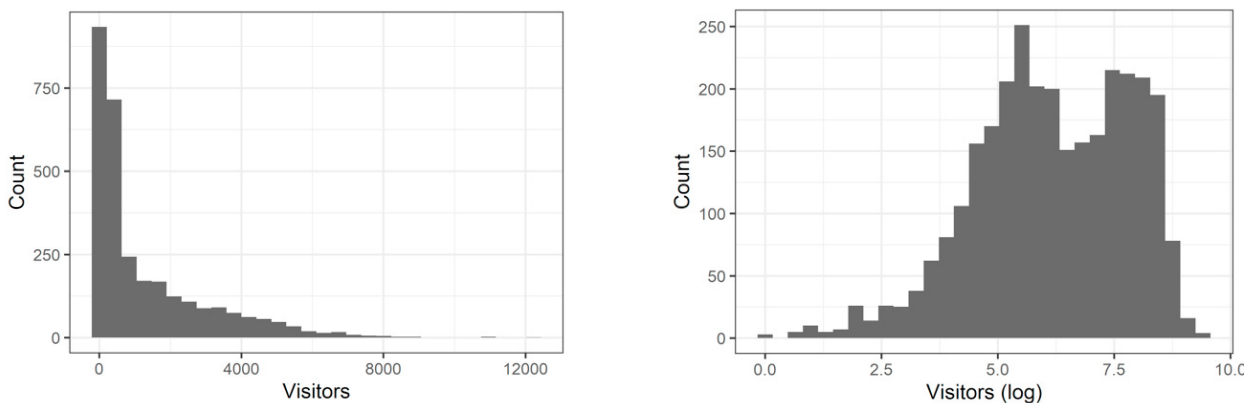

Fig. 2.: Histogram showing the distribution of the number of zoo visitors (left) and the number of zoo visitors on a logarithmic scale (right).
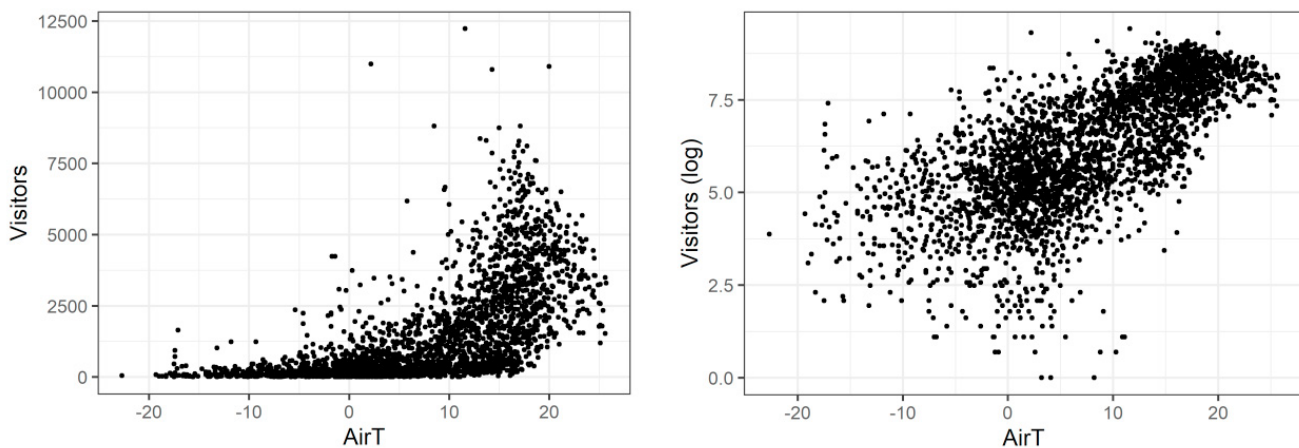
Fig. 3.: Scatterplot showing the relationship between the air temperature (AirT) and the number of zoo visitors (left) as well as the number of zoo visitors on a logarithmic scale (right).

4. Methods & Setup

For our analysis we use stratified sampling by year, with a 70 percent training and a 30 percent test sample. For all models the loss function is the negative sum of least squares. We start by first modelling the training data with generalized additive models (GAM) that were fitted through model-based boosting [22], as well as the random forest (rf). Rf is an ensemble of regression trees through the method of bagging. For the random forest we used 1000 trees and three randomly selected variables at each split. GAMs are a generalization of generalized linear models, that also incorporate (penalized) splines for smooth effects. We consider two versions of GAMs. The first model takes each variable in the dataset as a smooth effect and the second one additionally incorporates interactions. All interactions of variables are considered. If two variables for an interaction are numerical, two-dimensional penalized splines are fitted. The difference to one-dimensional P Splines is that the knots are two-dimensional and span the space of the cartesian product of two variables. If one variable is numerical and the other one is categorical, then for each level of the categorical variable a P Spline is being fitted for the numerical variable. Since interpretability is of interest, we want the GAM to enforce sparsity. This is especially important, since we consider many interactions. To achieve that, we use model-based boosting, where the final model is fitted iteratively, by only updating it by one base learner in each step, selected on the basis of reduction of the loss function. The model is then stopped early using a 25-fold cross-validation, which leads not only to a sparse solution, but also to one where the effect sizes are slightly shrunk towards zero. All boosted GAM models were stopped early before the maximum number of boosting iterations of 2000 and the learning rate was set to 0.1. Next, we apply the resulting models to the training data set to obtain predicted values (e.g., prediction of the random forest and predictions of the GAM). Furthermore, we fit the GAM to the rf predictions using the same covariates as used in the models obtained from the training data. The following steps illustrate how the GAM – rf model is being fitted:

1. Fit the random forest regression model with the k features $x^{(1)}, \dots, x^{(k)}$ to the outcome vector $y = (y_1, \dots, y_n)^T$
2. Use the model form step 1 to generate the predictions $p_1, \dots, p_n$
3. Fit the GAM with the same features $x^{(1)}, \dots, x^{(k)}$ to the new outcome vector consisting of the predictions of the random forest $p = (p_1, \dots, p_n)^T$

From that we compute the variable importance for the GAM as well as GAM interpreting the random forest. Further, we plot the partial effects of both methods for comparison. For the partial effects, all other variables are held constant. Now, the random forest, the GAM and the GAM based on the random forest model are applied to the test data. To evaluate the performance of the different models, mean squared errors (MSE) are calculated and plots of predicted versus true outcomes of the independent variable for all models are plotted.

For the analysis we used the statistical software R [23]. All visualizations were created with the R package "ggplot2" [24]. The random forest was fitted using the package "randomForest" [25] and the GAM was fitted using the package "mboost" [26].

5. Results

Figure 4, Tables 4 and 5 show the importance of the variables for modelling the dependent variable of the relevant models. The plot indicates that the importance ranking of the different variables is the same for the GAM and the GAM explaining the random forest. This means that both GAMs yield basically the same interpretation regarding the variable importance. Figure 5 on the left shows the partial effect of the maximum temperature, the most important variable for the outcome. We see that both curves look the same with only small differences at the boundary knots of the curve. The only variable that showed substantial differences between the GAM and the GAM explaining the random forest is the air temperature, the least important variable. The GAM explaining the outcome yields a much more regularized effect for this variable than the one explaining the random forest. But since this variable reduces the MSE less than the most important variable by the factor of less than 0.001, the effect of this difference for the whole model is only very marginal. However, the plot on the right in Figure 5 shows the results for the GAM and GAM explaining the random forest when variable interactions are included. Here we get a different result as the order of importance is not the same anymore (e.g. the interaction between air temperature and maximum temperature) and also the size of the variable importance show differences. It is important to note that interactions of two-dimensional P-Splines are not identifiable, if one variable is used in two distinct of such P-Splines. This means that the effect of one individual variable can be captured by one of the interactions and not by the other. Regardless of this issue, there are still substantial differences. For example, all interactions containing the variable workdays are more important in the GAM than in the GAM explaining the random forest. Also, there are interactions that were selected by the GAM and not selected by the GAM explaining the random forest, namely the interaction between Precipitation and Snow, as well as the interaction between Precipitation and MinT. This makes the GAM explaining the random forest sparser than the GAM, since fewer variables are selected, leading to an easier interpretation of the overall model. Next, Table 3 shows the mean squared error for the different models evaluated using the test data. The comparison of all GAM models - including the ones with interaction - with the random forest points to the black-box model being superior in predicting the outcome variable. Overall, both GAM with interactions were associated with a lower MSE than the corresponding GAM without interactions. However, the GAM explaining the random forest sows the same MSE as the GAM explaining the outcome and the GAM with interactions explaining the random forest even outperformed the GAM with interaction explaining the outcome. In Figure 6 we also displayed the global fitted/predicted values against the actual values of the outcome, for all models, including the models explaining the random forest. But instead of plotting all points (over 15000 points) on the scatterplot, we only plotted a smoothed line using the LOESS estimator provided by the "ggplot" package. We did this to showcase if there is any bias in the prediction. Between the range of 4 and 8 on the x-axis all smoothing curves lie on the bisector, which indicates no structural bias in that area. However, at the boundary deviations can be detected. There are also differences regarding the degree of deviation depending on the type of model. The largest deviation can be seen in the GAM with interactions, indicating the severest misfit for very low or very high numbers of visitors. In contrast to this, the GAM with interactions explaining the random forest does not show such a misfit for these high and low visitor numbers. The same tendency for a low number of visitors can be seen in the comparison of the GAM explaining the random forest with the GAM, even though the differences are not as apparent as with the GAM including interactions.
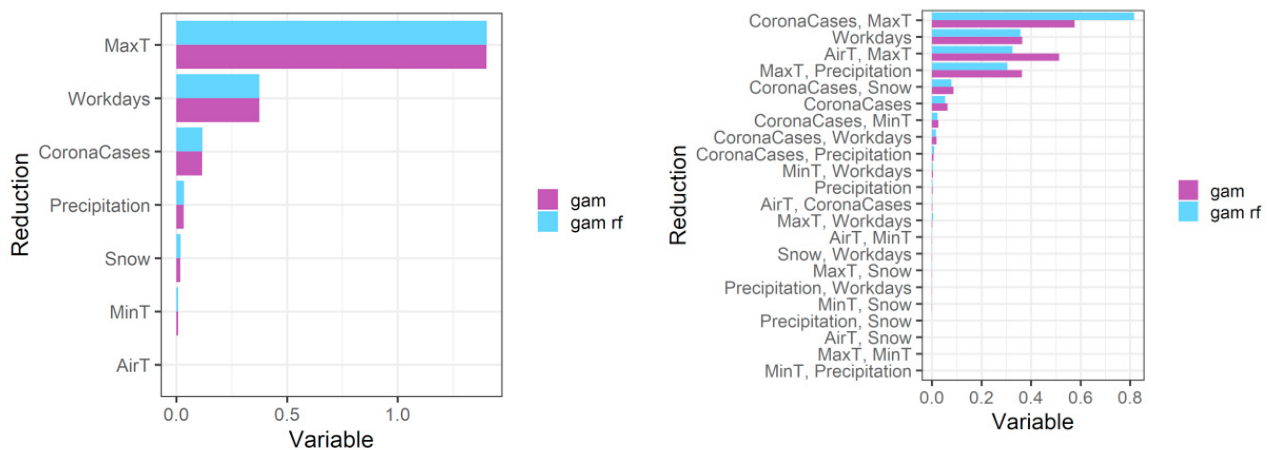
Fig. 4.: Variable importance, as absolute reduction of loss function attributed to each variable, of the GAM (left) and the GAM with interactions (right). The color indicates if the model was fitted to outcomes (gam) or to the predictions of the random forest (gam rf)

Table 3. Mean squared error (MSE) of the models.

RF	GAM	rf GAM	GAM interaction	RF GAM interaction
0,651	0,819	0,819	0,771	0,760

Table 4. Variable importance of the fitted to the outcome (GAM) and the one fitted to the predictions of the random forest (GAM rf), as absolute reduction of loss function attributed to each variable corresponding to Figure 4.

Variable	GAM	RF GAM
MaxT	1,4013	1,4022
Workdays	0,3752	0,3752
CoronaCases	0,1173	0,1183
Precipitation	0,0344	0,0346
Snow	0,0190	0,0198
MinT	0,0083	0,0093
AirT	<0,0001	0,0006

Table 5. Variable importance of the GAM fitted to the outcome (GAM) and the one fitted to the predictions of the random forest (GAM rf), as absolute reduction of loss function attributed to each variable corresponding to Figure 5.

Variable	GAM	RF GAM
CoronaCases, MaxT	0,5758	0,8167
AirT, MaxT	0,5132	0,3246
Workdays	0,3645	0,3571
MaxT, Precipitation	0,3635	0,3046
CoronaCases, Snow	0,0878	0,0791
CoronaCases	0,0629	0,0543
CoronaCases, MinT	0,0263	0,0229
CoronaCases, Workdays	0,0187	0,0168
CoronaCases, Precipitation	0,0067	0,0090

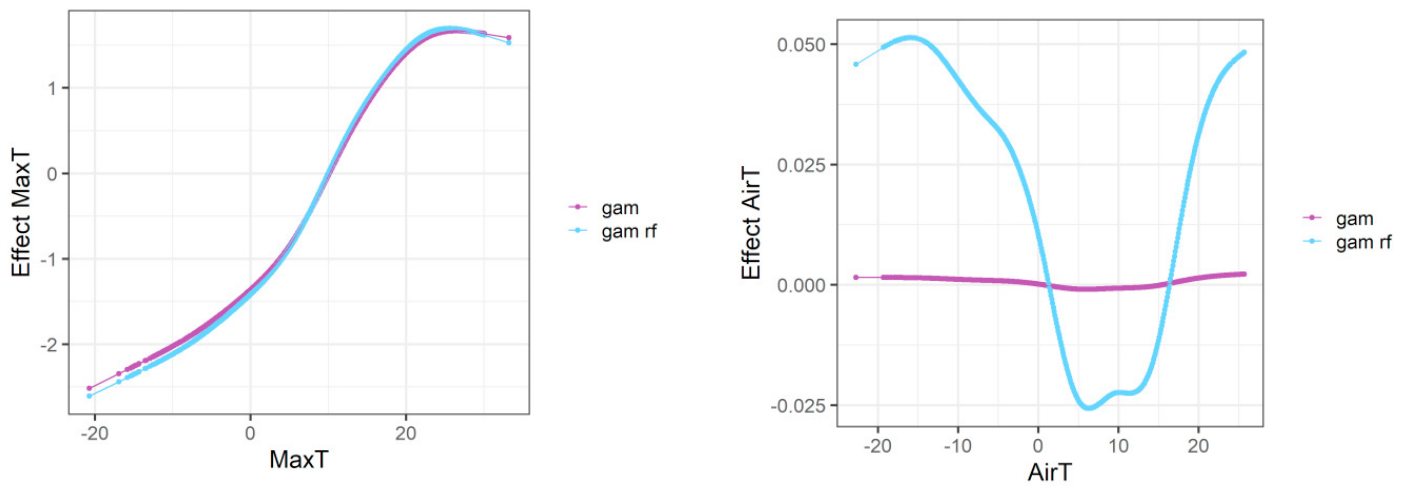|		
MinT, Workdays	0,0050	0,0046
Precipitation	0,0034	0,0031
AirT, CoronaCases	0,0027	0,0021
MaxT, Workdays	0,0018	0,0057
Precipitation, Workdays	0,0015	0,0009
AirT, MinT	0,0013	0,0013
Snow, Workdays	0,0013	0,0011
MaxT, Snow	0,0010	0,0010
MinT, Snow	0,0008	0,0005
AirT, Snow	0,0001	<0,0001
Precipitation, Snow	0,0001	
MaxT, MinT	<0,0001	<0,0001
MinT, Precipitation	<0,0001	

Fig. 5.: Partial effect of the maximum temperature (MaxT) on the expected number of visitors on a logarithmic scale (Effect). The color indicates, if the model was fitted to the outcome (gam) or the random forest (gam rf).
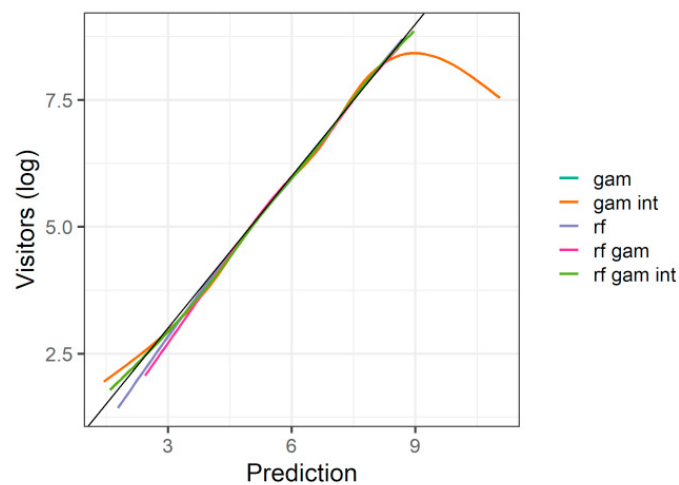

Fig. 6.: Global prediction plot showing a smooth line using the LOESS estimator for the predicted values on the x-axis against the actual values on the y-axis for all fitted models.

6. Discussion

A popular modeling approach for complex data is to fit several models to the same data. Then the best performing models or ensemble of models are used for prediction and the interpretable models are used to gain insight into the data. For this approach random forest and GAMs are common choices, as in Arnold [27]. The unsatisfying aspect of this approach is that it is unclear whether the interpretation of the data provided by the interpretable model matches with how the black-box model derives predictions. Therefore, the field of XAI has mainly focused on explaining the behavior of well predicting but complex models by simple models that are interpretable but not as predictive. This way the interpretation matches the black-box model to a certain degree. But it remains unclear how the explanation of the data matches the interpretation of the black-box. Our findings suggest that for modeling the number of zoo visitors, the interpretation of the GAM itself matches the interpretation of the GAM explaining the random forest. This is a strong statement, as there is a high degree of multicollinearity present in the data. For example, the minimum and maximum temperature were strongly correlated with a Pearson correlation coefficient of 0.95. Still, both GAMs captured similar effects and yielded similar variable importance. As this holds true for the simpler case, the interpretation of the GAM with interactions differs from the GAM with interactions explaining the random forest. Interestingly, not only the variable importance of both models differ but also the sparsity and predictability differed in favor of the model explaining the random forest. GAMs are known to be prone to overfitting [28]. Even though we used boosting to fit the GAM, which imposes further regularization compared to the standard maximum likelihood estimation, we still believe that a certain degree of overfitting is present. Other regularization techniques like bagging or random subspace methods have been proposed to combat this problem [29]. We believe that the pre-processing of the data by the random forest poses further regularization on the outcome and filters out noise, as the random forest itself utilizes bagging as regularization. This regularization led to fewer predictors being selected by the GAM with interactions, and therefore a higher degree of sparsity. This not only reduces the variance but also simplifies the interpretation of the model. The reduction of variance was most apparent at the bounds of the fitted values in the global prediction plot. In the case of zoo visitors, predictions are especially important for very high and low numbers of visitors as the demand has to be planned in advance. So, instead of only trying to make black-box models more interpretable through statistical models, we should also think about improving the predictability of statistical models through the black-box, while retaining the interpretability. The improvement of the GAM through the random forest should be investigated with more data and experiments. Understanding the change of interpretable models, such as GAMs, before and after fitting a black-box algorithm as the random forest, may help to both improve the interpretable model as well as the interpretability of the black-box. This may lead to a more refined scale of finding a good trade-off between interpretability and predictability or, as in this analysis, to a solution that improves both at the same time. Using GAMs or boosted GAMs as a surrogate model to explain another model imposes some limitations. Apart from well-known limitations of GAMs, such as the tendency for overfitting [28], the most limiting factor of using GAMs as in this analysis is probably its lack of capturing deep interactions, especially if the black-box model captures these interactions. In this case, the explanation provided by the GAM may be too simplistic. This may lead to wrong conclusions about the functioning of the black-box model.

Acknowledgements

References

[1] Panetta, Kasey. "Gartner Top Strategic Technology Trends for 2021." Gartner. https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021 (accessed Jan. 06, 2022).

[2] Linardatos, Pantelis, Papastefanopoulos, Vasilis and Kotsiantis, Sotiris. (2021, January). "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* **23 (1):** 18. doi: 10.3390/e23010018.

[3] Marco T. Ribeiro, Singh, Sameer and Guestrin, Carlos. (2016, August). "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* New York, USA, 1135–1144. doi: 10.1145/2939672.2939778.

[4] Confalonieri, Roberto, Coba, Ludovik, Wagner, Benedikt and Tarek R. Besold. (2021). "A historical perspective of explainable Artificial Intelligence." *WIREs Data Mining and Knowledge Discovery* **11 (1)**: e1391. doi: 10.1002/widm.1391

[5] Scott, Daniel and Lemieux, Christopher. (2010). "Weather and Climate Information for Tourism," *Procedia Environmental Sciences* **1**: 146–183.

[6] Christopher R. De Freitas. (2003). "Tourism climatology: Evaluating environmental information for decision making and business planning in the recreation and tourism sector." *International Journal of Biometeorology* **48 (1)**: 45–54.

[7] Lise, Wietze and Richard S. J. Tol. (2002). "Impact of climate on tourist demand." *Climatic Change* **55 (4)**: 429–449.

[8] Lionetti, Simone, Pfäffli, Daniel, Pouly, Marc, Tim vor der Brück, and Wegelin, Philipp. (2021). "Tourism Forecast with Weather, Event, and Cross-industry Data." *in Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, 1097–1104.

[9] Sabine L. Perch-Nielsen, Amelung, Bas and Knutti, Reto. (2010) "Future climate resources for tourism in Europe based on the daily Tourism Climatic Index." *Climatic Change* **103 (3)**: 363–381.

[10] Yap, Norman, Gong, Mingwei, Ranesh K. Naha, and Mahanti, Anaket. (2020). "Machine learning-based modelling for museum visitations prediction." *presented at the 2020 International Symposium on Networks, Computers and Communications*, ISNCC 2020.

[11] Ploner, Alexander and Brandenburg, Christiane. (2003). "Modelling visitor attendance levels subject to day of the week and weather: A comparison between linear regression models and regression trees." *Journal for Nature Conservation* **11 (4):** 297–308.

[12] Álvarez-Díaz, Marcos and Rosselló-Nadal, Jaume. (2010). "Forecasting British tourist arrivals in the Balearic Islands using meteorological variables." *Tourism Economics* **16 (1)**: 153–168.

[13] Becken, Susanne. (2013). "Measuring the Effect of Weather on Tourism: A Destination- and Activity-Based Analysis." *Journal of Travel Research* **52 (2):** 156–167.

[14] Mari C. Domingo. (2021). "Deep learning and internet of things for beach monitoring: An experimental study of beach attendance prediction at Castelldefels beach." *Applied Sciences (Switzerland)* **11 (22)**.

[15] Finger, Robert and Lehmann, Niklaus. (2012). "Modeling the sensitivity of outdoor recreation activities to climate change." *Climate Research* **51 (3)**: 229–236.

[16] Cindy C. Yañez, Francesca M. Hopkins, and William C. Porter. (2020). "Projected impacts of climate change on tourism in the Coachella Valley, California." *Climatic Change* **162 (2)**: 707–721.

[17] Aylen, Jonathan, Albertson, Kevin and Cavan, Gina. (2014). "The impact of weather and climate on tourist demand: the case of Chester Zoo." *Climatic Change* **127 (2)**: 183–197.

[18] David R. Perkins IV and Keith G. Debbage. (2016). "Weather and tourism: Thermal comfort and zoological park visitor attendance." *Atmosphere* **7 (3)**.

[19] David R. Perkins IV. (2018). "Using synoptic weather types to predict visitor attendance at Atlanta and Indianapolis zoological parks." *International Journal of Biometeorology* **62 (1):** 127–137.

[20] Domingo F. Rasilla Álvarez and Sonia Crespo Barquín. (2021). "Weather influences on zoo visitation (Cabárceno, Northern Spain)." *International Journal of Biometeorology* **65 (8):** 1357–1366.

[21] Michael E. Porter and James E. Heppelmann. (2015). "How Smart, Connected Products Are Transforming Companies." *Harvard Business Review* **92 (11)**: 64–88.

[22] Hothorn, Torsten, Bühlmann, Peter, Kneib, Thomas, Schmid, Matthias, and Hofner, Benjamin. (2010) "Model-based Boosting 2.0." *J. Mach. Learn. Res.* **11**: 2109–2113.

[23] R Core Team (2021). "R: A language and environment for statistical computing. R Foundation for Statistical Computing" Vienna, Austria. https://www.R-project.org/.

[24] Wickham, Hadley (2016). "ggplot2: Elegant Graphics for Data Analysis." *Springer-Verlag New York* ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

[25] Liaw, Andy, Wiener Matthew (2002). "Classification and Regression by randomForest." *R News*, **2 (3)**: 18-22. https://CRAN.R-project.org/doc/Rnews/.

[26] Hothorn, Torsten, Buehlmann Peter, Kneib Thomas, Schmid Matthias and Hofner, Benjamin. (2022). "mboost: Model-Based Boosting." R package version 2.9-7, https://CRAN.R-project.org/package=mboost.

[27] Arnold, Denis, Wagner, Petra, and Baayen, Harald. (2013) "Using generalized additive models and random forests to model German prosodic prominence", *Proceedings of Interspeech 2013*: 272-276.

[28] Larsen, Kim. (2015). "GAM: The Predictive Modeling Silver Bullet."

[29] Koen W De Bock and Dirk Van Den Poel, (2012). "Reconciling Performance and Interpretability in Customer Churn Prediction using Ensemble Learning based on Generalized Additive Models" *Working Papers of Faculty of Economics and Business Administration*, Ghent University, Belgium 12/805, Ghent University, Faculty of Economics and Business Administration

[30] Aria, Massimo, Cuccurullo, Corrado and Gnasso, Agostino, (2021). "A comparison among interpretative proposals for Random Forests" *Machine Learning with Applications* **6**

[31] Gunning, David and David W. Aha. (2019). "DARPA's Explainable Artificial Intelligence (XAI) Program" *AI Magazine*, **40 (2)**, 44-58.