## Machine learning in psychological assessment

Machine learning and artificial intelligence are by now familiar buzzwords in many fields of empirical research. In psychological assessment, we also see an increasing interest and application of methods from these fields. We believe methods from ML and AI can strongly contribute to the field of psychological assessment. At the same time, the buzz around these methods often appears somewhat reminiscent of the tale of the emperor’s new clothes. We believe when it comes to ML and AI, the emperor is in fact wearing clothes, but the clothes are often not new. Many of the powerful techniques used in machine learning (e.g., cross validation, regularization, ensembling) have since long been known and fruitfully applied in more traditional statistical analyses, psychometrics and psychological assessment. Perhaps this view is most poignantly summarized in a popular tweet from bio-statistician Daniella Witten: “When we raise money it’s AI, when we hire it’s machine learning, and when we do the work it’s logistic regression.” (although the original author of the statement is unknown).

In this editorial, we aim to provide a short, general introduction to methods from statistical learning, and to discuss how they relate to more traditional statistical techniques. We hope to underline the strengths of existing works and methods from our field, and to discuss how novel methods and insights from the field of MI may contribute to the continuing improvement of our field. We will use the term statistical learning to refer to both more traditional and more recent tools for data analysis, because there is no clear distinction between statistics, artificial intelligence and machine learning. Newer data-analytic methods can only contribute to our field as long as their use is firmly rooted in existing statistical knowledge about sampling, estimation and uncertainty. For sake of brevity, in this editorial we focus on methods for statistical prediction, a.k.a. supervised learning methods. Thus, unsupervised learning methods (e.g., factor analysis, clustering, correlation networks, topic models from natural language processing) are outside the scope of this editorial.

## Two shifts in predictive modeling

Recent development in statistical learning have lead to two main shifts:

1. An increased focus on prediction.

# Increased flexibility, allowing for capturing non-linear associations and/or large numbers of potential predictor variables.

The field of psychology has traditionally been primarily interested in *explanation* through developing and testing theories of human behavior. This has sometimes led researchers to under appreciate *prediction*, merely because their main aim was not to predict, but to test a specific theory or hypothesis. A theory, however, can only explain real-world phenomena to the extent that it can accurately predict them (Yarkoni and Westfall 2017). The focus on explanation has lead many researchers to focus on effect sizes (*R2*, Cohen’s *d*) computed using observations that were also used to fit the model. At the same time, most researchers in psychology are probably familiar with the phenomenon of *R2* automatically and spuriously increasing when extra explanatory variables are added to the model. When the coefficients of a regression model fitted on one sample are applied to predict the value of the outcome in a different sample, the amount of variance explained tends to be (much) lower. More realistic *R2* values can be obtained, for example, through cross validation: by computing *R2* on a sample of observations not used for computing the regression coefficients (Rooij and Weeda 2020). The importance of cross validation is certainly not new to the field of assessment, where it has been discussed for almost a century (Larson 1931; Mosier 1951).

The interest in explanation may also have resulted in a preference for using *unbiased* estimators. Unbiasedness may sound like a highly favorable property, but all statistical modeling is subject to the bias-variance trade-off: The smaller the bias of our procedure, the higher it’s variance will be. Higher variance means that the fitted model and the predicted values will more strongly differ from one training sample to the next. The smaller the sample size, and the more flexible the fitted model, the higher the variance will be. Often, variance tends to contribute more strongly to generalization error than bias. In general, to obtain a model with reasonable generalizability and predictive power, we need to reduce the variance, either by having a large enough sample, or by introducing some *bias* into our model fitting-procedure. Bias can be introduced in several ways, including:

* Restricting the complexity of the function form (e.g., model only linear associations)
* Careful pre-selection and scaling of predictor variables (e.g., using sum- or factor scores instead of item scores)
* Using regularized estimation procedure (e.g, lasso, ridge, elastic net regression; use of Bayesian priors).

If the bias is well-chosen, this will improve generalizability: We buy predictive power by assumption. If the bias is not well chosen (e.g., we fit a linear model while the associations are in fact highly non-linear; we omit relevant predictors), predictive accuracy and generalizability will suffer.

This brings us to the second shift, towards more flexible modeling. This shift has been made possible through the development of flexible non-parametric methods, combined with increased computational power. A prime example of a highly flexible method, perhaps the most non-parametric method of all, is the method of *k*-nearest neighbours (kNN). In fact, kNN does not even fit a model; it merely remembers the training observations. To compute predictions for a new observation, kNN computes the distance of the new observation to the training observations in the space spanned by the predictor variables, and computes the mean of the response variable over the *k* closest neighbours. kNN provides the greatest flexibility of all prediction methods, because it does not impose any restrictions on the shape of the association between predictors and response. This high flexibility, however, can potentially yield very high variance and even breaks down in high dimensions. In lower dimensions, variance can be reduced by introducting bias through increasing the value of *k*. With larger values of *k*, the predicted value for a new observation averages over a larger number of observations. This is in fact nothing very new: In traditional parametric models, we quantify the variance of an estimated mean by its standard error, which gets smaller as the number of observations increases.

Note however that the great flexibility of kNN comes at the cost of interpretability: the effects of predictor variables on the predicted values cannot be readily computed or explained anymore. If we have more than three predictor variables, it becomes increasingly difficult to visualize which of the training observations are the nearest neighbours. In what follows, we will therefore focus on *interpretable* statistical prediction methods, which allow for quantifying or visualizing the effects of predictor variables. We will see that explanation of the predictions of black-box models brings many pittfals.