My Research: Learning low-dimensional structures

In the last decade machine learning algorithms have shown unprecedented accuracy in a variety of applications and tasks. However, in many applications in-sample prediction itself is of secondary importance and the main or first goal is to get an understanding of the underlying structures. Understanding the underlying structure is for example important when it is employed for decision making where discrimination may play a role, if one wants to make prediction in a different and previously unseen environment, or if the goal is to discover or quantify causal structures; for example understanding the effect of one covariate or a group of covariates on health-outcomes in medicine or on the probability of an accident happening and its magnitude in insurance. Lastly, also in-sample prediction performance may be improved if structure in the data can be found and exploited. The next three paragraphs give a slightly more detailed picture on these points.

Structure & explainability Understanding the effect of one covariate is only possible if it is part of a low-dimensional structure. The alternative is that a covariate effects an outcome only via a complicated interplay with many other covariates while having no direct impact on its own or with a small group of covariates. If this latter is the case then the game may already be lost, since there seems to be a limit on how much complex interplay a human can understand; see also here for attempts of psychologist to quantify this limit `https://en.wikipedia.org/wiki/The_Magical_Number_Seven,_Plus_or_Minus_Two`.

Structure & accuracy Another reason why we may want to seek for low-dimensional structures is the curse of dimensionally. Without structure, models will inevitably suffer from the curse of dimensionality in high dimensional data spaces, i.e., exponentially deteriorating estimation performance in the number of covariates due to the growing distance between data points. The following hyperlink leads to an interactive example of how the number of observations in a neighbourhood decreases when going from one to two to three dimensions: `https://mhiabu.github.io/curse_of_dimensionality.html`.

Structure & robustness If one employs a model in an environment different to the one it has been trained on, this can lead to arbitrarily bad predictions. This happens when the predictions are based on correlation structures not present anymore in the new environment. If the trained model is a composition of low-dimensional structures that are understandable, one can easier add adjustments to the initially trained model that better generalise to the new environment. Often this adjustments are not feasible from the data alone, but may stemm from additional instrumental variables and expert knowledge.

Luckily, in many applications low-dimensional structures seem to play an important role. If data has lower complexity, then learning low-dimensional structures opens up the way for more accurate and robust predictions and global interpretations that may better resemble the underlying causal structure.

Concrete structures

Machine learning methods usually do not assume any structure and have the much celebrated universal approximation property https://en.wikipedia.org/wiki/Universal_approximation_theorem. In a supervised learning task, the model can be described as

$$E[Y|X=x]=f(x).$$

Given independent observations $(X,Y)_{i=1,\dots,n}$, the goal is to estimate the function f while making no structural assumption on f. But having no underlying structure, leads to the problem mentioned in the first page. While the universal approximation property dictates that one can, in principle, approximate any function, this is not the case under finite data conditions, i.e., the universal approximation property says nothing about how the best approximating function can be found. In particular if the chosen machine learning method is performing well, that is probably because the function f has some simple underlying structure. However, this structure is neither directly visible to the user, nor has it been directly employed for better performance.

Going in the other extreme, the probably most simple structure is the linear model:

$$E[Y|X=x]=x^T\beta=\beta_0+x_1\beta_1+\cdots+x_d\beta_d,$$

where the goal is to estimate the slope β. Given independent observations $(X,Y)_{i=1,\dots,n}$ and some strong assumptions on the noise, one can estimate β with an error that is of order $n^{-1/2}$. This simple linear model is well understood and it is usually taught to students in heir first statistics course. However, the model can be made arbitrarily hard when aiming for more realistic or suitable assumptions like assuming an additional time component, measurement errors, correlated noise, unobserved confounders, large number of initial covariates d, and possibly many other little things that pop up in applications. Indeed, much of today's statistical research (for good reason) is still within this linear model.

In my research so far, I have mostly ignored all those little complications (a point I will come back to later at the very end of this document) and considered optimal estimation and inference for example in the more genreal setting of an additive model,

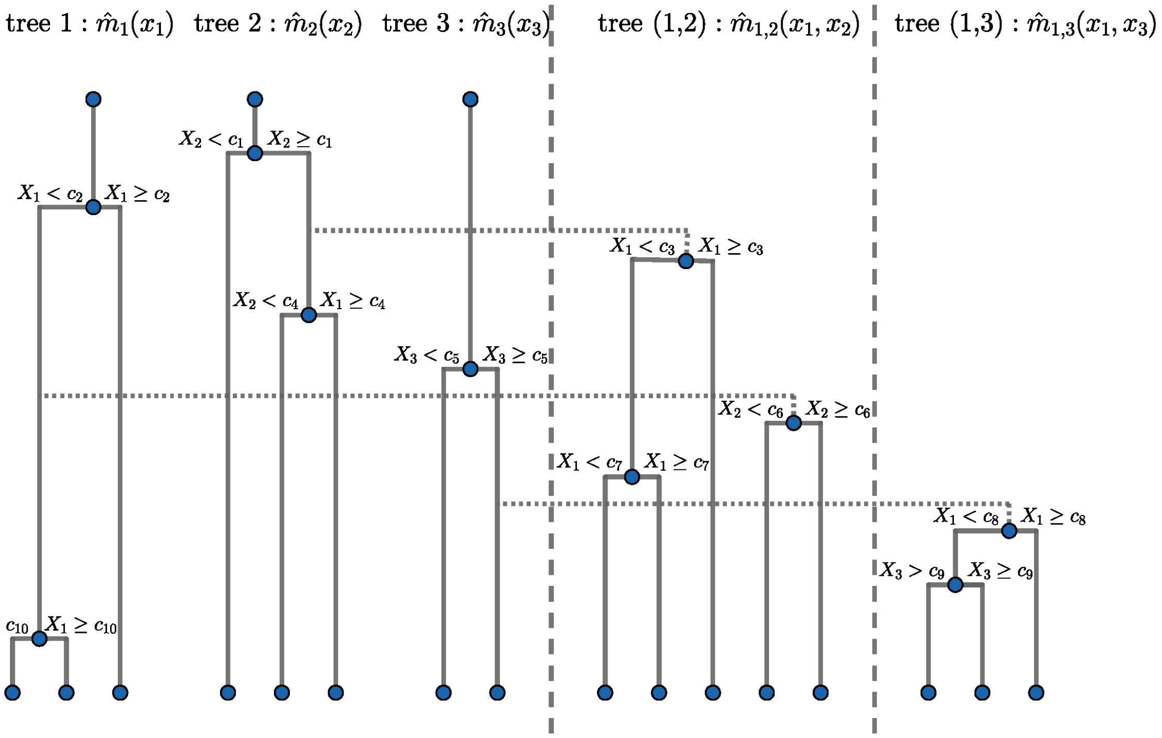
$$E[Y|X=x]=m_0+m_1(x_1)+\cdots+m_d(x_d),$$

with no structural assumption on the components $m_j, j=1,\dots,d$. Compared to the linear model it relaxes the linearity assumption while still assuming that covariates do not interact when affecting the response Y. It is well known, that if the components are twice differentiable, that they can be estimated with an error rate of $n^{-2/5}$, which is also the optimal rate in the corresponding one-dimensional problem.

More recently, I have been working on more general structures, that in principle have the universal approximation property. One example is the decomposition of the regression function into a sum of main effects, bivariate interactions, etc., up to a d-variate interaction term.

$$E[Y|X=x]=\sum_{S\subseteq\{1,\dots,d\}}m_S(x_S)=m_0+\sum_{k=1}^{d}m_k(x_k)+\sum_{k<l}m_{kl}(x_k,x_l)+\cdots+m_{1,\dots,d}(x). \qquad (1)$$

The heuristic of the decomposition is that if the underlying function $m(x)$ only lives on low-dimensional structures, then m_S should be zero for most covariate subsets S and the order of maximal interaction $q = \max\{|S| : m_S \neq 0\}$ should be much smaller than the number of covariates: $q << d$. Together with Enno Mammen and his PhD student Jospeh Meyer from Heidelberg, we have developed a machine learning algorithm that estimates the non-zero components m_S. We have coined the method Random Planted Forest, as it is a tree based method that follows a hierarchical structure along a planted path and only estimates m_S if it has already found signal in m_U with $U \subsetneq S, |S \setminus U| = 1$. Experiments conducted so far hint hat our algorithm seems to be very competitive compared to state-of-the-art machine learning methods in many data set.



In one current stream of research, we are looking into the behaviour of tree based methods in general (which function classes can a random forest consistently estimate?), and we are also trying to show that a simplified version of Random Planted Forest can estimate the component m_S with $|S|$ −dimensional rate of $n^{|S|/(2|S|+1)}$. Interestingly, possibly due to the proof techniques, proving this turns out to be already quite challenging for dimensions ≥ 3.

Other streams of research I am interested in or working on is how the structure (1) can be further exploited via Random Planted Forest or related algorithms. I am also considering other structures than (1), and how structures like (1) can be employed for different targets and in more complicated situations. In the latter case I am in particular thinking of specific biomedical or insurance applications with time-dependent covariates and censored observations. Other open questions are how expert knowledge and structures like in (1) can be optimally combined for optimal predictions and to uncover causal structures.