I work at the intersection of Machine Learning and more traditional areas of Statistics such as *Nonparametric Estimation* and *High-Dimensional Inference*.

My research is driven by the need for statistical insights into the successes and failures of machine learning in emerging applications. In particular, despite its growing success, machine learning technologies still require much human effort to configure and tune to the hidden characteristics of the application data at hand. Typical configuration decisions range from the choice of data representation and preprocessing, to the difficult choice of algorithmic hyperparameters. Adding to these difficulties, typical applications of machine learning involve hard domain constraints, such as time and space constraints in large data and high-dimensional regimes, missing data or costly data, and frequent changes in data distribution requiring expensive reconfiguration efforts.

My general approach to the above problems is to first understand, mathematically, those hidden characteristics, or structure, of application data that should drive configuration and tuning efforts, i.e., which have provable effects on statistical performance. Many such *hidden structures* have been identified that characterize modern machine applications, and often differentiate these from traditional applications domains of Statistics. Namely, much of the high-dimensional data that arise in machine learning applications, are characterized by low-dimensional structures (e.g. data lying close to a submanifold, or is sparse under some unknown dictionary), or can be clustered in ways that might improve prediction performance. Much of my work focuses on such structures, and on understanding how they affect the statistical performance of common machine learning tools, towards automatically configuring such tools to *adapt* and take advantage of such hidden aspects of data.

My research, currently and over the next few years, can be divided into two main directions. The first concerns *supervised learning*, i.e., prediction tools for classification, regression and bandit problems in the presence of high-dimensional data and domain constraints; this work has so far won various honors at leading machine learning venues (i.e. best student paper, and plenary presentations). The second direction is more recent and concerns *unsupervised learning*, i.e., problems such as clustering, density estimation, and causality, where the learner has little information to go by during training (such as labeled data as in prediction problems); this is an area where tuning and configuration efforts are especially difficult due to missing information, and is perhaps the least theoretically understood area to date in data sciences. Next I describe a few representative research results and open questions of interest in these two areas.

Supervised (and Semi-Supervised) Learning

Here we are concerned with typical prediction problems arising in machine learning, i.e., identifying objects in images, classifying documents into topics, deciding which advertisement to present to consumers, or deciding how to route a self-driving car around routine traffic. These are often formalized as predicting some variable Y from a high-dimensional observation X, while the data available to the learner is in the form of pairs $(X_i, Y_i)_{i=1}^n$ of previous observations.

It's been observed that such high-dimensional input $X \in \mathbb{R}^D$ is often more *structured* than it might appear, due to intrinsic correlations between the D predictor variables in X. As such, X might for instance lie close to a low-dimensional manifold, or be sparse under some unknown dictionary, or might be highly clusterable. Such realization at first gave rise to various preprocessing approaches such as *manifold learning*, and *dictionary learning*, which can unfortunately add to the complexity of configuration decisions in practice.

However, as it turns out, much of common prediction procedures (e.g., trees, nearest neighbors) automatically benefit from the presence of such structures in high-dimensional data without the need for expensive preprocessing. Such results were first shown by [Bickel and Li, 2006], in the case of kernel regression methods and data on unknown manifolds. [Scott and Nowak, 2006] showed around the same time that dyadic classification trees are adaptive to *box-dimension*, a fairly general notion of intrinsic complexity.

Following up on this work, I have shown in a series of recent results that adaptivity to intrinsic dimension is a rather general phenomenon. In particular, as shown with various collaborators, common methods such as k-NN regression, kernel regression, and many efficient tree-based regression methods, and their classification counterparts are adaptive to the intrinsic dimension of data, thus require no preprocessing.

My work in this area relies on formalisms of *intrinsic dimension* which tightly capture, in a minimax sense, the complexity of high-dimensional regression and classification. Furthermore, these notions capture, in a unified manner, the various low-dimensional structures usually considered in pattern recognition. It remains open whether other classes of predictors such as support vector machines, and neural networks might also automatically adapt to certain hidden structures in data, although there are reasons to believe so as a general picture is emerging.

While these initial insights on adaptivity assume ideal scenarios with complete data and no particular constraint, the practical reality is otherwise in modern applications of machine learning. Data is often incomplete with missing or costly labels, the distribution of data often changes overtime, and practitioners often have to compromise on ideal predictive approaches in order to meet various real-world time and space constraints. My current and future research efforts is therefore focused on how to best meet such constraints while taking advantage of, and automatically tuning to hidden structure in application data.

In particular, we have recently shown that the mere presence of low-dimensional structure in data can help meet time and space constraints without trading-off the statistical performance of *local* predictive tools (e.g. k-NN, trees, and kernel regressors and classifiers) [Kpotufe and Verma, 2017]. Many interesting questions remain concerning other families of predictive tools, and their limitations under practical constraints.

Furthermore, tradeoffs under changing data distributions and missing data remain largely unclear. We have recently shown that, in *active learning* (where the learner can actively choose which data to label at a cost), there is a richer set of regimes than previously thought, where good statistical rates are achievable with few label requests, and with many interesting rate-transitions that were previously unknown despite over a decade of work on the subject [Locatelli et al., 2017].

My main aim for this research direction is to generate new theoretical understanding that informs practice.

Unsupervised Learning

Tuning and configuration decisions are particularly difficult in unsupervised learning problems such as clustering data $\{X_i\}_{i=1}^n$ into unknown groups of similar points (e.g. documents without topic information, speech signals without tags or delimiters), or estimating salient structures in high-dimensional data (e.g., in medical imaging, or in astronomy. Unlike in prediction problems, it is hard to validate choices of algorithmic hyperparameters such as number of clusters, shape or dimension of topological structures.

I have recently started work in this direction, under established statistical formalisms for the structures of interest. For instance, *clusters* might be viewed as regions of high-density in data (defined by level-sets of an unknown density) which make it possible, at least theoretically, to automatically identify such structures in data without a priori knowledge of the number of structures, nor their shape or dimension.

In particular, we have recently proposed some new clustering approaches that are provably robust to choices of tuning parameters, attain minimax rates of estimation under general situations, while competing favorably against state-of-the-art heuristics on a range on real-world data [Jiang and Kpotufe, 2016]. This work is the culmination of years of research effort, that aimed at understanding how particular geometric graphs on data encode the level-sets of the unknown data-generating density [Kpotufe and von Luxburg, 2011, Chaudhuri et al., 2014, Dasgupta and Kpotufe, 2014].

My current and future research here aim at addressing the following question: it remains unclear how to properly scale clustering approaches in general (including ours) to big data and high-dimensional regimes without loosing performance. One idea is to consider scenarios where the underlying data density might depend weakly on most dimensions, or where the data itself happens to lie close to low-dimensional structures (and where in fact a proper *density* might not exist). The aim is then to first understand how such structure might reduce the statistical difficulty of the problem, and then to design tools that might automatically take advantage of such situations while requiring little tuning effort from practitioners.

References

[Bickel and Li, 2006] Bickel, P. and Li, B. (2006). Local polynomial regression on unknown manifolds. *Tech. Re. Dep. of Stats. UC Berkley*.

[Chaudhuri et al., 2014] Chaudhuri, K., Dasgupta, S., Kpotufe, S., and von Luxburg, U. (2014). Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12):7900–7912.

[Dasgupta and Kpotufe, 2014] Dasgupta, S. and Kpotufe, S. (2014). Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, pages 2555–2563.

[Jiang and Kpotufe, 2016] Jiang, H. and Kpotufe, S. (2016). Modal-set estimation with an application to clustering. *arXiv preprint arXiv:1606.04166*.

[Kpotufe and Verma, 2017] Kpotufe, S. and Verma, N. (2017). Time-accuracy tradeoffs in kernel prediction: controlling prediction quality. *Journal of Machine Learning Research*, 18(44):1–29.

[Kpotufe and von Luxburg, 2011] Kpotufe, S. and von Luxburg, U. (2011). Pruning nearest neighbor cluster trees. *International Conference on Machine Learning*.

[Locatelli et al., 2017] Locatelli, A., Carpentier, A., and Kpotufe, S. (2017). Adaptivity to noise parameters in nonparametric active learning. In *Conference on Learning Theory*, pages 1383–1416.

[Scott and Nowak, 2006] Scott, C. and Nowak, R. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52.