My research investigates how to identify subspaces from incomplete data. Subspace models lie at the heart of data analysis. From biologists studying genes to astronomers studying galaxies, scientists often want to find subspaces that explain their data, for example using linear regression, principal component analysis (PCA) or mixture models. But in many modern applications data is missing. For example, in computer vision, occlusions produce missing data; in surveys and recommender systems, subjects do not know or do not want to provide all information. Fortunately, dependencies and redundancies in large systems can be exploited so that very accurate inferences can be made based on incomplete data. The key idea is that the measurements of such systems tend to lie mostly in low-dimensional subspaces. If one can estimate the subspace, then one can infer the missing values. Therefore it is of great interest to understand how to estimate subspaces from limited or incomplete measurements.

Current Work

Subspaces in Big [incomplete] data

We love subspaces. As soon as we get our hands on some data, finding a subspace that explains it is one of the first things we try. But with big data come big challenges. Identifying subspaces from incomplete datasets is one of them, often known as low-rank matrix completion (LRMC). While this is a flourishing topic, our knowledge is still limited. For instance, previous theory mostly focused on uniformly distributed observation patterns. However, in many situations missing patterns are not uniform. For example, in vision, occlusion of objects can produce missing data in very non-uniform random patterns. My work focuses on arbitrary observation patterns. In [1], we give a characterization of deterministic sampling conditions to identify subspaces.

Another limitation of previous theory is that due to coherence assumptions, it could only guarantee to identify subspaces that are unaligned with the canonical axes. In contrast, our theory holds for all subspaces in general position. Our findings have important implications regarding lower bounds for identifiability, sample and computational complexity, the role of coherence, adaptive settings and validation conditions.

Subspaces in Big [corrupted] data

The task of identifying subspaces from severely corrupted data is often known as Robust PCA. One typical example is background segmentation, as the background in a video lies approximately in a low-dimensional subspace (because it has small, structured variations), and the foreground (e.g., people passing by) can be modeled as sparse, gross errors (because they typically take only a small fraction of each frame). Similar to LRMC, existing theory and methods usually require bounded coherence and uniform distribution of errors. However, the foreground is hardly ever uniformly distributed. For example, the location of the head of a person is highly correlated to the location of his body. Furthermore, in many cases, like microscopy and astronomy videos, the background can be mostly dark, whence it lies in a highly coherent subspace. So existing theory and methods do not cover scenarios like these. In contrast, we extended our theory for missing data to derive theory and algorithms for Robust PCA that do not rely on these assumptions and outperform state of the art methods in a broad range of settings [2].

Is there really a subspace?

Scientists want to find subspaces. For example, in medical applications, subspaces can be used to predict the effects of drug compounds. This allows to create medicines that attack the source of illness while minimizing secondary effects. But finding a subspace is not enough. Scientists want to make sure they find the right subspace. But in most cases, they do not know a priori whether there even is a subspace — as is typically assumed in previous theory. Our theory gives conditions to guarantee that if we find a subspace, it is because there really is a subspace [3].

This, together with our theory to identify one subspace from incomplete data provides the tools to explore a more general and challenging problem: identifying mixtures of subspaces from incomplete data.

If one subspace is good, more must be better.

Mixture models assume that each data point lies in one of several underlying subspaces, and the goal is to find such subspaces. Mixtures arise in a wide variety of problems where data is missing. Examples range from classical mixture
regression to face clustering, object tracking and network topology estimation. However, the requirements to identify a mixture of subspaces from incomplete data remained a fundamental open question, answered by our theory. [4]

On the other hand, if identifying one subspace from incomplete data is challenging enough, identifying a mixture is all the more. In fact, to the date there exist no polynomial-time algorithm that provably identifies a mixture from incomplete data — except under very strong assumptions. My research aims to develop such algorithm. So far, we have devised a state-of-the-art, optimization-based method with local convergence guarantees [5].

Future work

Theory

- Missing data can make even orthogonal subspaces appear identical [6]. Hence we need to develop new notions of distance between subspaces that can capture notions of missing data.
- There exists no polynomial-time algorithm that will provably identify mixtures of subspaces. I would like to study develop such a method, or determine guarantees for an existing one.
- It is often the case that large datasets are distributed across many data centers, and the data at each center is insufficient on its own to accurately estimate the subspace of the entire dataset. I would like to investigate optimal strategies to share data in order to estimate the underlying subspaces in distributed systems as accurately as possible.
- Multi-dimensional arrays, also known as tensors, can be used to study complex relationships in datasets ranging from chemometrics to neuroscience, graph analysis and many others. Just as a low-rank matrix has a column subspace and a row subspace, a low-rank tensor has a column subspace, a row subspace, an aisle subspace, etc. In [7] I present a simple method to identify these subspaces in closed form. However, if the tensor is only approximately low-rank (e.g., if there is noise), there is no definitive method to identify these subspaces. I believe that my approach in [7] can be extended to noisy tensors to obtain decomposition methods with theoretical, computational and precision advantages.
- Rigidity theory aims to determine whether a set of pairwise distances between a collection of points uniquely determines their position up to rigid transformations (rotations, translations and reflexions). Rigidity has a tight relation with identifying subspaces from missing data — as one only observes a subset of all the distances, and distance matrices are low-rank. Rigidity is a fundamental problem with applications in fields as diverse as chemistry, echolocation and graph theory. However, a characterization of rigid structures remains an open question since the times of Lagrange and Maxwell. I would like to extend the theory that we developed for missing data to expand our knowledge of rigidity.
- I am also interested in extending our results to the case where we can only observe linear transformations of the data, a setting that arises in compression, face recognition and collaborative filtering.

Practice

- Subspaces are such a natural structure, that they arise even in the most unexpected places, for example, in phylogenetics. Phylogenetics study the evolution of species. One of its main goals is to infer the so called tree of life. Fortunately, gene frequencies in quartets (subsets of four species) of a tree lie in low-dimensional subspaces [8]. However, it is now known that many species (e.g., bacteria) hybridize. So, rather than a tree, life has evolved as a network. I am collaborating with biologists and statisticians to determine whether quartets (or other subsets) of phylogenetic networks also lie in low-dimensional subspaces. This will allow to develop methods to infer the now called network of life.
- Creating better drugs to treat diseases is a challenging task. To this end, scientists first identify a target (e.g., a protein) that is related to the disease, and the goal is to develop a drug that will interact with this target (to cure the illness) but not with others (to minimize secondary effects). This process traditionally involves screening millions of drug compounds (e.g. molecules) in search of a few ones that have the desired interactions. This is an expensive endeavor. Some targets behave similarly, so it is reasonable to assume that the interactions have some structure (subspace). If we could find a subspace that explains these interactions, we would be able to predict all interactions only testing a few cleverly chosen ones. I am collaborating with the Morgridge Institute for Research to estimate drug-target interactions with optimal accuracy and create better drugs at lower costs.
• Scientists at the University of Toronto want to understand periodicities in the water velocities of Lake Huron. To this end they record data every 0.25 seconds for months. Unfortunately, due to instrumentation imperfections, data is often lost in non-uniform patterns. In collaboration with these scientists, I plan to use our theory and methods to infer the missing velocities.

• In a way, identifying mixtures of subspaces is a classification problem. One can first cluster points, and then estimate the subspace corresponding to each cluster. In practice, deep neural networks have proven to be a major asset for classification. I would like to explore the use of artificial neural networks to identify mixtures of subspaces from incomplete data.

References


http://danielpimentel.github.io