Epilogue

The question is not what you look at, but what you see.

Henry David Thoreau

Distances & Measurements

The concept of distance is basic to the human mind. We often do not qualitatively compare the objects of our thoughts, but prefer to explicitly express our ideas about how similar or dissimilar we judge them to be. Even with regard to psychological states, we like to distinguish different degrees of involvement, excitement, attachment, etc. The urge to classify and order the phenomena in the world seems to be a basic human need. Aristotle, the great ancient classifier, constructed extensive systems of thought in which phenomena were categorized and assigned to disjoint classes with distinct properties. Although the application of the “Aristotelean knife” (Robert Pirsig) has led to many unnatural and problematic classifications in hindsight, it was nevertheless extremely fruitful in that it imposed order on the world, enabling a large and still ongoing scholarly activity.

The next important step in the scientific enterprise was the shift from a purely mental exercise to actual experimentation. Instead of reasoning about possible causes and relations in the natural world, researchers were actively asking questions and trying to construct theories that were consistent with the facts obtained. Abstracting from the individual researcher and his situation, science was founded on the notion of universality: Patterns observed under a given experimental situation should be reproducible in a different location and time, even by a different scientist. The basic tool that allows for such an objective approach is the notion of a measurement. Thereby, the objects of our inquiry are related in a prescribed way to standardized
models, allowing the scientist to extract universal information. This information is conveniently represented in the strict language of mathematics, which is universal by its underlying logical foundation.

The different levels of measurement have been defined by Stanley Smith Stevens in an influential article (Stevens, 1946). Nominal measurements correspond to Aristotle’s legacy of classification: Objects and their properties are assigned and distinguished by labels. Mathematically, this is the domain of set theory. On the ordinal level, objects are ordered, corresponding to a totally (or linearly) ordered set. Next, interval measurements allow for quantitative statements. Properties measured are mapped to numbers, and the notion of distance surfaces. An example is the Celsius scale for temperature, where one degree Celsius is defined as one hundredth of the difference in temperature between water at the melting and the boiling point, respectively. Remarkably, measurements on an interval scale are relative, i.e., only distances are well-defined, and there does not exist a designated origin\(^7\). Mathematically, such measurements correspond to affine spaces. Finally, ratio measurements are expressed on a scale that possesses a non-arbitrary zero value.

A great deal of early science was involved with the search for the most elementary properties by which we can compare the objects in our world. This has led to the development of systems of measurement, sets of units specifying anything which can be measured. The international system of units (SI) identifies seven distinct kinds of physical quantities that can be measured: length, mass, time, electric current, temperature, luminous intensity and amount of substance. Notwithstanding its great success in the commercial and scientific domain, it can be argued whether this is a complete or natural list.

**Looking back**

Here we were concerned with more abstract quantities. The objects considered in this thesis are complex systems that can not be easily reduced to one or more fundamental properties: the respiratory system, the brain, dynamical systems. Even if it were possible to project the state of such a complex entity to a single number, the great loss in information incurred does not make this an attractive proposal. Therefore, instead of extracting a single property from such a complex system, we have considered ways in which we can compare systems quantitatively with each other. Although this did also result in a single numerical quantity, namely, a distance between each pair of systems under consideration, the totality of all such distances contains a much greater amount of information. This simple fact was the starting point for the

\(^7\) Although it is commonly said that water freezes at 0 degree Celsius, the Celsius scale was not intended to be used for such absolute statements. Rather, the correct way would be to say that “water freezes at a temperature difference of 0 degrees from the temperature where water freezes.”
methods developed and applied in the rest of this thesis. It is not obvious, and special methods were needed to extract this information from the measured distances.

The central questions considered in this thesis were the following:

- How can we define a useful distance for complex systems?
- What kind of information is obtained from such a distance and how can we analyze it?
- What does this tell us about the original systems?

These questions are increasingly difficult to answer. It is not too difficult to define interesting “distances” between complex systems, although a few pitfalls need to be avoided. In particular, in order to allow for sensible comparisons between more than two distinct systems, a “distance” measure needs to be a true distance (in the mathematical sense), i.e., it needs to exhibit metric properties that allow for a consistent and natural interpretation in such a multivariate setting. This seriously restricts the class of possible “distance” measures, and involves an important principle: Being a true distance allows for a natural representation of complex systems as points in an abstract functional space, which is a very powerful way to visualize and analyze differences and commonalities between complex systems. For general “distance” measures such a representation is usually not possible. Indeed, it is well known that bivariate measures (such as “distances”) can, and generally do, lead to spurious or even false results when applied in a multivariate setting (Kus et al., 2004). This problem is completely avoided by using a true distance. Of course there is a price to pay for this convenience: It might not be easy to find a suitable, true distance for the systems we want to study. And even if we obtain such a measure, it is not clear that it then also captures the relevant information about a system. Fortunately, the class of optimal transportation distances, are general enough to be both applicable in most settings, and in such a way that they capture interesting information.

The geometrical and statistical analysis of distances is also a rather well-developed topic, so we mostly did connect results scattered in the literature and closed a few gaps. However, what is actually measured in such an interval-scale approach is a completely different matter. The first two questions were addressed in a phenomenological setting: it is not necessary to know exactly what causes differences in complex systems, if one is primarily interested in the existence of such differences. For example, in the application to the respiratory system, we were interested in distinguishing healthy breathing from breathing with a diseased lung, which is a simple supervised classification task — albeit one of considerable interest. Since such classification was possible, we might now ask why this is the case. Then the question of how to reverse-engineer the information obtained from abstract distances becomes important. This road is mostly unexplored so far.
The future

The examples in this thesis demonstrate that the combination of optimal transportation distances, reconstruction of these distances by multidimensional scaling, and canonical discriminant analysis of the resulting coordinates is a powerful and versatile approach to the classification and study of complex systems. This thesis is finished, but many paths remain still to be explored. Let me mention a few here that have not been discussed in the earlier chapters.

- On the practical side: The calculation of the Wasserstein distances is still too complex (i.e., slow) to handle large datasets (with more than a few hundred to thousand sample points per subject). Bootstrapping smaller subsets helps a long way in reducing the computational complexity, but algorithmic improvements would be preferable. A number of interesting approximation algorithms have been developed in recent years, and implementing these as actually usable software would be desirable.

- How can classification based on nonmetric multidimensional scaling be cross-validated? Since nonmetric reconstructions are usually obtained by an iterative procedure, this is not as simple as it sounds. Optimally matching the resulting point configurations (de Leeuw and Meulman, 1986) might be one possibility to proceed.

- To avoid the quadratic dependence on sample size when computing all pairwise distances between $N$ samples, is it possible to reconstruct Euclidean configurations locally, i.e., by only using the distances of the closest $k \ll N$ points?