Chapter Abstracts for: Object Oriented Data Analysis

By Marron and Dryden

1 What is OODA?

This chapter motivates OODA in the context of the rapidly changing fields of statistics, data science and data analytics. A wide range of Big Data contexts, including each of the dimension and sample size being large are considered. While Big Data is acknowledged as a major challenge, an important point is that an even greater challenge is Complex Data. Key concepts such as data visualization based on modes of variation are illustrated using the curves as data objects, i.e. functional data analysis, example of the Spanish Mortality data. The analysis links with various societal trends in a number of interesting ways. Another important aspect of OODA is non-Euclidean data objects, as typified by the Bladder-Prostate-Rectum image data set. Various shape representations, which result in data objects naturally lying on curved manifolds, including the skeletal approach are discussed. A Bayesian approach to particularly challenging segmentation problems is seen to be very effective.

2 Breadth of OODA

This chapter highlights the breadth of OODA through a variety of real data examples, which go far beyond the standard data matrix oriented statistical contexts. In the first of these, careful attention is paid to the fact that even in classical functional data analysis, while standard analytic approaches such as Principal Component Analysis are very effective at displaying amplitude, i.e. vertical modes of variation, they can be quite poor at capturing other intuitively important modes of variation, such as phase shifts. The example of the Brain Artery data is used to show that OODA extends beyond the mildly non-Euclidean case of manifold data to the strongly non-Euclidean case of tree and graph (i.e. network) structured data objects. A further example of the breadth of OODA is the case of carefully represented human speech sounds as data objects. Finally images of faces as data objects shows another direction of the reach of OODA through a fun gender classification of face images.

3 Data Object Definition

This chapter treats the foundations of OODA. An important aspect is terminology that is designed to make the main ideas transparent, including the fundamental concept of simultaneously considering both the object space and feature space representations of the data objects. Key ideas are illustrated with a very simple two dimensional toy example that allows explicit visualizations of both spaces. The concept of modes of variation is formally defined in a way that enables extension in a meaningful way to the case of non-Euclidean data objects. Scree plots, which display relationships between modes in terms of proportions of variation explained are introduced. Mathematical notation, used through the rest of the book, is also defined. Finally an overview of the intuitive utility of the object and feature space concepts is given. The importance of careful choice of data objects is demonstrated in the specific context of probability distributions as data objects.

4 Exploratory and Confirmatory Analyses

This chapter investigates statistical analysis approaches, first highlighting the difference between exploratory and confirmatory methods. Many of the aspects of population structure, some of which are not obvious and even surprising, that can be found in an exploratory analysis are demonstrated with the Tilted Parabolas and the Twin Arches toy data sets, and with Lung Cancer and Pan Cancer subsets of The Cancer Genome Atlas. While such methods are very good at revealing important population structure, they also have strong potential to find spurious artifacts of sampling variation, that are not reproducible in independently generated data sets. Hence methods that confirm the actual existence of discovered phenomena are also critical. Pointers are given to detailed development of such methods in later chapters. There is also an overview of major OODA statistical tasks.

5 OODA Preprocessing

This chapter lays out some useful preliminary steps in OODA. The common theme is looking at data. This can be quite challenging in high dimensions, where the common linear regression device of studying the distribution of each predictor would require simultaneous visualization of perhaps tens of thousands of univariate distributions. Effective dealing with that issue is illustrated with the Drug Discovery data set using the idea of Marginal Distribution plots, which considers various types of representative variables. Another important issue is simple linear scaling of variables, where various sensible (and possibly quite divergent) analytic choices are highlighted. Nonlinear scaling of variables is also considered, with a recommended automatic shifted log transformation for data sets with wildly varying amounts of skewness. Registration and alignment issues are also overviewed.

6 Data Visualization

This chapter compares and contrasts the relative strengths and weaknesses of different ways of viewing a Euclidean data matrix. This includes heat-map views where matrix entries are coded as colors, as well as treating each of the sets of matrix rows and columns as bundles of curves, to which Functional Data Analysis methods can be applied. A less standard combined view showing all of the above in a single plot, which is then used to display multiple modes of variation is proposed. Data centering of both rows and columns of the data matrix is explored, and appropriate OODA based terminology is developed to keep these often slippery concepts straight in interdisciplinary discussions. The scores scatterplot matrix view, for understanding relationships between data objects, is also considered in detail. That motivates discussion of a number of insightful alternatives to the standard PCA directions for visual scatterplot exploration of data.

7 Distance Based Methods

This chapter considers many aspects of distance based methods of statistical analysis. Several common metrics, i.e. distances, are compared on the basis of relative properties of their Fréchet means. An overview of Multi-Dimensional Scaling shows how this method gives data object representations similar to PCA for the Euclidean metric, and extends the scores visualization aspect of PCA to the case of many other metrics. Several important examples of distances are discussed, such as the Wasserstein, i.e. Earth Mover’s, i.e. Mallows, Metric in the context of Functional Data Analysis, and the Procrustes and Generalized Procrustes approaches to landmark based shape analysis. Choice of metric is seen to have a major impact in the case of covariance matrices as data objects via visualization of geodesic paths.

8 Manifold Data Analysis

This chapter considers the important and challenging OODA case of data objects that naturally lie on a curved manifold. The value of analyzing data in this way is first illustrated using the perhaps most straightforward example of data objects on the unit circle, known as directional data. A brief introduction to aspects of manifold geometry needed for OODA is given. A series of improved analogs of PCA is developed for data objects lying on manifolds. The important case of landmark based shape space is illustrated in more detail, and illustrated using real data examples. Versions of the Central Limit Theorem for probability distributions on manifolds are briefly explored. A fundamental concept stemming from these developments is Backwards PCA, leading to principal nested spheres and principal nested shape spaces. More discussion of covariance matrices as data objects, from the manifold data space viewpoint, is given.

9 FDA Curve Registration

This chapter considers curve alignment, i.e. registration, in Functional Data Analysis. The main goal is to extend the fundamental concept of modes of variation into both amplitude and phase modes. The need for this is demonstrated with a toy phase shift example where conventional PCA modes are clearly very non-intuitive. Much more insightful decomposition is accomplished using the Fisher-Rao approach that improves upon most proposed methods in this area through directly tackling the problem using the notion of warp equivalence classes, and development of the warp invariant Fisher-Rao metric. Careful handling of the quotient space structure, i.e. getting the mathematics really right, results in a fully automatic method, which does not need the heavy manual tuning required by other approaches. The especially strong peak alignment properties of this metric are demonstrated with a toy example. Deeper analysis comes from combining the Fisher Rao approach with a Principal Nested Spheres Analysis, as developed in Chapter 8.

10 Graph Structured Data Objects

This chapter is an overview of graph, i.e. network, structured data objects. In contrast to the manifold data objects in Chapters 8 and 9, which can be called “mildly non-Euclidean” because manifolds admit approximating tangent planes, the present case is “strongly non-Euclidean”, because even that local approximation is no longer present. A number of approaches to such data are illustrated using the Brain Artery Tree data. These include a simple combinatorial approach which ignores all geometric properties, an approach that leverages the large amount known about the complicated mathematics of phylogenetic trees, the Dyck Path approach which provides a curve representation enabling the use of Functional Data Analysis methods, and a Topological Data Analysis approach. A Graph Laplacian approach to networks as data objects is also explored. The main ideas and utility of that approach are demonstrated using an example of corpus linguistics, i.e. natural language processing.

11 Classification - Supervised Learning

This chapter considers approaches to the classification, i.e. discrimination problem, also called supervised learning. That is a generalization of automated disease diagnosis, where the goal is to use measurements made on healthy and sick people to develop an automatic rule for classifying additional people. Classical methods are explored, with a novel nonparametric derivation of Linear Discriminant Analysis being given. Kernel methods, which lie at the heart of much of Machine Learning are also discussed and compared. Those motivate the optimization based Support Vector Machine, whose visualization in high dimensions reveals the concept of data piling. The loss of generalizability entailed by data piling is addressed by the development of Distance Weighted Discrimination. Its strong data separation properties make it a useful approach to data heterogeneity such as batch adjustment challenges. Alternative classification approaches are also considered.

12 Clustering - Unsupervised Learning

This chapter is about clustering, i.e. unsupervised learning. Like classification, discussed in Chapter 11, clustering methods focus on sub-groups of a data set. But instead of using known group labels as in the latter, the goal of the former is to discover reasonable groups. Well known methods, such as k-means and hierarchical clustering are contrasted and compared. For the latter, various distance methods and linkage functions are considered. Both two dimensional and high dimensional examples show strong differences in behavior of linkage functions across settings. Use of data visualizations to find clusters is also considered. An overview of hybrid methods, such as fuzzy clustering, and integration of clustering with other statistical tasks is also given.

13 High Dimensional Inference

This chapter is about confirmation and validation of discoveries made by the exploratory methods proposed in earlier chapters, with a focus on OODA contexts. The gravity of the issue is demonstrated with an example showing that spurious sampling artifacts can easily appear to be important population structure in high dimensional situations. Because reasonable probability models are often not available, the focus is on nonparametric approaches. The DiProPerm hypothesis testing method is seen to be a useful application of permutation methods for exploring distributional differences in two class situations. It is seen that DiProPerm is inappropriate in clustering situations (where class labels are not given) which motivates the development of the SigClust approach to understanding when clusters are “really there”, i.e. statistically significant in an important sense.

14 High Dimensional Asymptotics

This chapter considers aspects of classical mathematical statistics that are especially relevant to OODA, in particular high dimensional asymptotic analysis. One important domain is random matrix theory, where the limit is taken as both the sample size (standard) and the dimension of the data vectors (non-standard) tend to infinity. This leads to many useful ways of thinking about high dimensional data and its analysis, including the insightful Marchenko-Pastur distribution of the PCA eigenvalues. A different, but equally insightful, domain is High Dimension Low Sample Size asymptotics, based on limits as the dimension goes to infinity for fixed sample size. Perhaps surprisingly, many insightful lessons are available from this approach. These include the idea of Geometric Representation where modulo rotation Gaussian data tends to have a simple non-random rigid structure, which allows quantitative elucidation of surprising observations made in both earlier and later chapters.. Also included are some unexpected but quite informative consistency properties of PCA. The High Dimension Medium Sample Size domain provides a connection between the above two domains.

15 Smoothing and SiZer

This chapter gives an overview of statistical smoothing, including the contexts of density estimation and nonparametric regression, also known as scatterplot smoothing. It is seen that the popular histogram for understanding one dimensional distributions can be seriously hampered by the “bin edge effect”, which is elucidated using a kernel density estimator, suggesting the latter should be the natural default view (which is used throughout this book). The main ideas are illustrated by the Hidalgo Stamps and Bralower Fossils data sets. A brief discussion of data based smoothing parameter selection is included. The utility of the SiZer (SIgnificance of ZERo Crossings) methodology for meaningful statistical inference is demonstrated with several real data sets, including the British Family Incomes data.

16 Robust Methods

This chapter gives an overview of robust statistical methods in the context of OODA. First there is a high level discussion of the issue that dominated research in robustness: whether outliers are bad data that should be completely eliminated from the analysis, or should be viewed as containing useful information but should have a reduced influence on the analysis. An important case study is the Cornea Curvature data, which because the dimension of the Zernike basis representation was higher than the sample size, motivated the invention of Spherical PCA and Elliptical PCA. The former has since become a common benchmark in high dimensional robust PCA studies. Another important example is the Genome Wide Association data, which had clear outliers which were not at all addressed by spherical PCA, because of the Geometric Representation ideas presented in Chapter 14. That motivated the invention of Visual L1 PCA to give a robust version of PCA in such very high dimensional contexts.

17 PCA Details and Variants

This chapter delves deeply into the linear algebra of PCA and related methods. Several types of mean centering are carefully investigated as projections in various linear spaces including row, column and full matrix spaces. Many aspects of PCA are revealed through a straightforward singular value representation. A more classical Gaussian likelihood derivation is also considered, along with computational issues including lower rank exact representation of high dimensional data. Singular value representation ideas also give insights into two block, i.e. multi-view, analysis methods such as Partial Least Squares and Canonical Correlation Analysis and how the modes of variation generated by each relate to the other. Brief introduction is given to the more sophisticated two block analysis method of (Angle-based) Joint and Individual Variation Explained, which gives modes of variation some of which are joint between blocks, and others that are individual describing variation in one block that is not present in the other. The latter method relies on the relatively recent linear algebraic method of Principal Angle Analysis.

18 OODA Context and Related Areas

This chapter wraps up the book with some peripheral OODA issues, including acknowledgement of preceding ideas, history and terminology. A number of connections between OODA and Object Oriented Programming are established. Relationships between OODA and the parallel research areas of Symbolic Data Analysis and Compositional Data Analysis are also explored. The former simultaneously addresses the analysis of multiple types of data objects, illustrated by an example. In the latter context a large number of methods for analyzing probability vectors as data objects have been developed. Other related research areas are also discussed.