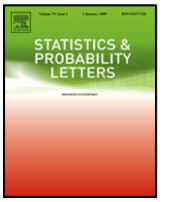
ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro



Journeys in big data statistics

Ian L. Dryden *, David J. Hodge

School of Mathematical Sciences, University of Nottingham, UK

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Big data
Object-oriented data
Transport
Networks

ABSTRACT

The realm of big data is a very wide and varied one. We discuss old, new, small and big data, with some of the important challenges including dealing with highly-structured and object-oriented data. In many applications the objective is to discern patterns and learn from large datasets of historical data. We shall discuss such issues in some transportation network applications in non-academic settings, which are naturally applicable to other situations. Vital aspects include dealing with logistics, coding and choosing appropriate statistical methodology, and we provide a summary and checklist for wider implementation.

1. A new natural resource

We will be the first to admit that it is difficult to keep up. How can you expect someone who is trained in dealing with datasets of $n = 30$ observations with $p = 3$ variables to suddenly cope with a 100 K-fold increase of $n = 3\,000\,000$ observations and $p = 300\,000$ for example, or even worse? Everything has to change. Summarizing a dataset becomes a major computational challenge and p-values take on a ludicrous role where everything is significant. Yet dealing with a wide range of sizes of datasets has become vital for the modern statistician.

Virginia Rometty, chairman, president and chief executive officer of IBM said the following at Northwestern University's 157th commencement ceremony in 2015:

> *What steam was to the 18th century, electricity to the 19th and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource.*

The need to make sense of the huge rich seams of data being produced underlines the great importance of Statistics, Mathematical and Computational Sciences in today's society. But what is 'new' about data? Data has been used for centuries, for example data collected on the first bloom of cherry blossoms in Kyoto, Japan starting in 800AD and now highlighting climate change (Aono, 2017); Gauss' meridian arc measurements in 1799 used to define the metre (Stigler, 1981); and Florence Nightingale's 1859 mortality data and graphical rose diagram presentation on causes of death in the Crimean War leading to modern nursing practice (Nightingale, 1859). All of these old, small datasets are at the core of important issues for mankind, so it is not the data or its importance but the size, structure and ubiquity of data that is new.

Many of the challenges in the new world of Statistics in the Age of Big Data are of a different nature from traditional scenarios. Statisticians are used to dealing with bias and uncertainty, but how can this be handled when datasets are so large and collected in the wild without traditional sampling protocols? What do you do with all the data is an important question. The last 20 years has seen an explosion of statistical methodology to handle large p, often with sparsity assumptions (Hastie

* Corresponding author.
E-mail address: ian.l.dryden@gmail.com (I.L. Dryden).

https://doi.org/10.1016/j.spl.2018.02.013
0167-7152/

et al., 2015). Large n used to be the realm of careful asymptotic theory or thought experiments, but in reality one often does encounter large n now in practice.

Two possible routes to practical inference are conditioning and sampling. Conditioning on a small window of values of a subset of covariates will very quickly reduce the size of data available as the number of covariates increases, due to the curse of dimensionality. Such small subsets of the dataset can be used to estimate predictive distributions conditional on the values of the covariates, leading to useful predictions. We give some further detail below in a case study from the transport industry. Sampling sensibly on the other hand is a more difficult task. Although it is straightforward to sample at random of course, given the inherent biases in most big data one needs to carry out sampling to counteract the bias in the data collection.

A further aspect of the avalanche of new data being available is that it is often highly-structured. For example, large quantities of medical images are routinely collected each day in hospitals around the world, each containing highly-complicated structured information. The emerging area of Object Oriented Data Analysis (Marron and Alonso, 2014) provides a new way of thinking of statistical analysis for such data. Examples of object data include functions, images, shapes, manifolds, dynamical systems, and trees. The main aims of multivariate analysis extend more generally to object data, e.g. defining a distance between objects, estimation of a mean, summarizing variability, reducing dimension to important components, specifying distributions of objects, carrying out hypothesis tests, prediction, classification and clustering.

From Marron and Alonso (2014), in any study an important consideration is to decide what are the atoms (most basic parts) of the data. A key question is 'what should be the data objects?', and the answer will then lead to appropriate methodology for statistical analysis. The subject is fast developing following initial definitions in Wang and Marron (2007), and a recent summary with discussion is given by Marron and Alonso (2014) with applications to Spanish human mortality functional data, shapes, trees and medical images. One of the key aspects of object data analysis is that registration of the objects must be considered as part of the analysis. In addition the identifiability of models, choice of regularization and whether to marginalize or optimize as part of the inference are important aspects of object data analysis, as they are in statistical shape analysis (Dryden and Mardia, 2016).

It is obvious that the realm of big data is a very wide and varied one. In some realms the difficulties lie with truly astronomical quantities of data which are not even feasibly stored for future retrieval, for which online algorithm development is a key area of research; whereas in other realms the challenge is in discerning patterns and learning from large datasets of historical data. We shall discuss the latter, in generality, below for what can loosely be thought of as transportation network applications in non-academic settings. Many of the approaches and recommendations discussed below are naturally applicable to other applications, such as a general practice of data retention, while others related to origin–destination filtering are clearly more specific to transportation problems.

2. Case study: transportation big data

The classification of problems into different areas of interest can be greatly beneficial in allowing techniques of particular relevance to all problems in a particular area to be discussed as one. The contemporary challenge we shall now discuss surrounds the use of statistics in real-world infrastructure problems that can arise for public or mass transportation, such as train travel, bus travel, or similar networked transportation methods. Collaborations between universities and businesses up and down the country already exist, and will continue to grow in the coming years for trying to share best practices and perform statistical analysis on datasets harvested by businesses about their customers, to either improve customer experience or to improve business efficiency. We concern ourselves here with the challenges one will meet in embedding good practice and developing useful models for exploitation of data in businesses where perhaps even the initial data handling task has so far seemed daunting.

Studying transportation systems as networked queues has been one of the most natural approaches, borne out of the queueing theory literature of previous decades. Courtesy of advances in computing, larger and larger network problems are now attempted to be 'solved' or at least approximately solved. Much of the focus in recent years lies with proposing online algorithms for live traffic management. With big data, opportunities arise to try and optimize these local dynamic decision problems: of re-routing a vehicle; skipping stops (if permitted); or allocating platforms, all in light of a wealth of additional statistical information. Approaches to dynamic resource allocation laid out in Glazebrook et al. (2014) would often benefit from a serious statistical analysis to first properly understand the dynamics of a network-based model, so that when formulating the problem in a queueing framework an appropriate level of confidence can be placed on the stochastic quantities. In particular, if you were to consider traffic management decisions on a railway surrounding the choice of platforms or use of signals outside a busy station, an effective algorithm for allocating the resource that is the station platform at a particular time can only function with a well-calibrated cost function which accounts for knock-on effects of such a decision. Possessing years of historical data during which a wealth of such decisions have been made and their consequences mapped, leads us very naturally to first want to perform some robust statistical analyses.

For statistical analyses, the natural starting point to a statistician is gaining access to the appropriate historical data. There are already two large data types of interest, customer-centric journey counting or vehicle journeys. In the world of buses it is estimated that over 5 billion passenger journeys occur each year in the UK Department of Transport, UK (2016), for example, and the alternative approach lies with vehicle logging data. For our discussion we shall concern ourselves more with logged vehicle movements, such as the approximately three million individual train movements which are logged on a

Please cite this article in press as: Dryden I.L., Hodge D.J., Journeys in big data statistics. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.013.