Contents lists available at ScienceDirect

Statistics and Probability Letters

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

How do statisticians analyse big data—Our story

Jian Qing Shi

School of Mathematics, Statistics & Physics, Newcastle University, UK

ARTICLE INFO

Article history:
Available online 2 March 2018

Keywords:
Big data
Modelling
Pre-processing

ABSTRACT

Analysis of big data is a hot topic, but the first problem encountered by many statisticians is 'where to start'. We would like to share our story with the readers who have less experience in this area, and hopefully it can shed some light on it.

1. Introduction

A big thanks to the editors for editing this special issue to discuss the role of statistics in the era of big data. Instead of 'the role of statistics', a common concern for statisticians, when they are faced with large data sets of many gigabytes or even terabytes, is on 'where to start'. This was exactly the first question we (a team of three statisticians) asked ourselves when we started our joint project *Limbs alive – monitoring of upper limb rehabilitation and recovery after stroke* through gaming with two other teams (one in neuroscience and the other in computing science) in 2012. Given the issues that we faced when first starting out, we would like to share our story with readers who may have less experience on analysing big data, and hopefully to shed some light on the process.

The overall aim of the project was to develop a home-based rehabilitation system using action video games for stroke patients, and the main task of our statistical team was to develop an automatic system to assess upper limb function. Assessing upper limb functions is a difficult task. The current commonly used clinical measure is the CAHAI (Barreca et al., 2005). A patient is asked to complete 9 tasks, e.g. 'open a jar' and 'dial 999'. A therapist gives a score for each task ranging from 1 to 7 based on how well the patient completed the task. The CAHAI is the sum of the 9 scores. Although it is a validated measure, it is certainly expensive (each assessment lasts from 20 to 30 min) and subjective. We designed an assessment game including 38 simple movements (see the details in Serradilla et al., 2014), for example a forward roll. We recorded the 3D position data and 4D directional data of the trajectory for each movement using three wireless controllers, one in each hand and the other attached to the middle of the body; see Fig. 1. The size of the data is quite big, up to several dozen gigabytes. The aim of the system is to use the data to calculate the level of impairment of upper limbs after stroke. This is a typical modelling problem, but many obstacles have to be overcome before we can consider any models.

Our experience shows that, to achieve the target, we need to treat it as a comprehensive scientific project rather than a statistical project. This may be the attitude we should have in analysing any big data problems.

2. Experiment design and data acquisition

In the statistical community, we focus mainly on modelling, although recently many statisticians have been getting involved in data pre-processing as well. For some problems, as usual ideally a statistician should be involved at the earliest stage possible, e.g. in the design of the experiment, the choice of devices or the method of recording data.

E-mail address: jian.shi@newcastle.ac.uk.

https://doi.org/10.1016/j.spl.2018.02.043

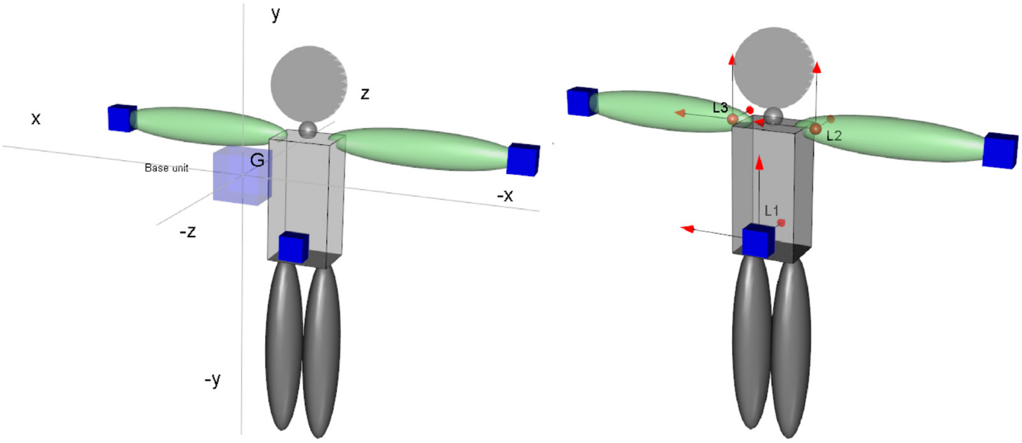

Fig. 1. The coordinate systems before (left) and after (right) transformation: blue squares represent the locations of three controllers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Using the 3D positional data of a trajectory in our project as an example, we used wireless controllers to collect data. Each device recorded a 3-dimensional time series with a frequency of 60 Hz, i.e. at each time point, we had 3 numbers corresponding to x-, y- and z- coordinates. Due to the nature of a home-based system, we can only record the numbers relative to a base unit (G in Fig. 1). To make the position data comparable amongst different patients, we can either require patients to stand on a fixed point relative to the base unit (e.g. exactly 100 cm away with their arms stretched should be orthonormal to the line linking the base unit and the body centre) which is impossible, or we would need to think of a better way to calibrate the data. Thus different coordinate systems are considered (global, moving and somatic, denoted by GCS, MCS and SCS respectively) with the third controller being used even though we were concerned with the movement of the patient's two arms only. The data collected in the third controller is used to do translation and rotation between different coordinate systems and to calibrate the data (Shi et al., 2013). This has been proven to be a key step to success in our project.

Another example is the orientation data. For some movements, for example 'chop & chop', we care about the capability of the patient moving their wrist or the magnitude of the angle that patients can rotate their wrist. The wireless controllers provide orientation information in a 3 × 3 rotation matrix (and thus a 9-dimensional time series for each movement). We first transform them to quaternions, reducing the data to 4-dimensional time series (Shoemake, 1985) and summarising movement rotations to two different metrics: projection angle (the angle formed by projecting a given coordinate-axis in SCS to the plane in GCS) and rotation angle (the angular distance between two given orientations). Those metrics are comparable and can be used in a model.

Without doing this work, data is not comparable, considering modelling is therefore meaningless. Big data usually comes from different resources, recorded in varied circumstances, particularly the free-living data in medical research. Before we start to consider modelling, substantial time and work should be spent on understanding the exact meaning of the numbers, the structure of the data, variation among different data sets, etc. We may need to redesign the experiment and change the way of collecting data if necessary. Those are the keys to success on solving real life problems involving big data.

3. Pre-processing

Pre-processing is a terminology often used in computer science, a programme to process input data to produce an output which can be used in other programmes. Usually the former is much noisier, having a much more complex structure and a larger data size than the latter. More and more statisticians are getting into this step, particularly in statistical applications involving complex data, and have already made significant contributions in segmentation, data calibration, cleaning, smoothing, dimension reduction, etc. Searching on Google, one can find hundreds of references and packages.

In this article, we want to discuss the importance of two issues: finding new 'summary statistics' and data registration. In many disciplines, such as machine learning, finding good 'summary statistics' or features comprises a large proportion or even the majority of the work. For example, the current development of accelerometer data analysis is mainly based on dozens of features calculated from the original signal (Preece et al., 2009), measured by one or several scalar variables (summary statistics) and each of them includes part of the information involved in the original complex data. Conventional statistical methods, usually simple models, are used in modelling using those summary statistics. It has been proven that this is a very efficient way to solve big data problems in many cases, and has also earned a good reputation for machine learning. By contrast, statisticians tend to use more complex and advanced methodologies. This is good, and actually the skill of modelling gives statisticians a big advantage over others on analysing data (big data is no exception). However, our experiences and many lessons show that we should place more attention on developing summary statistics and on conventional methods before we consider any complex models. More discussion will be given in the next section.

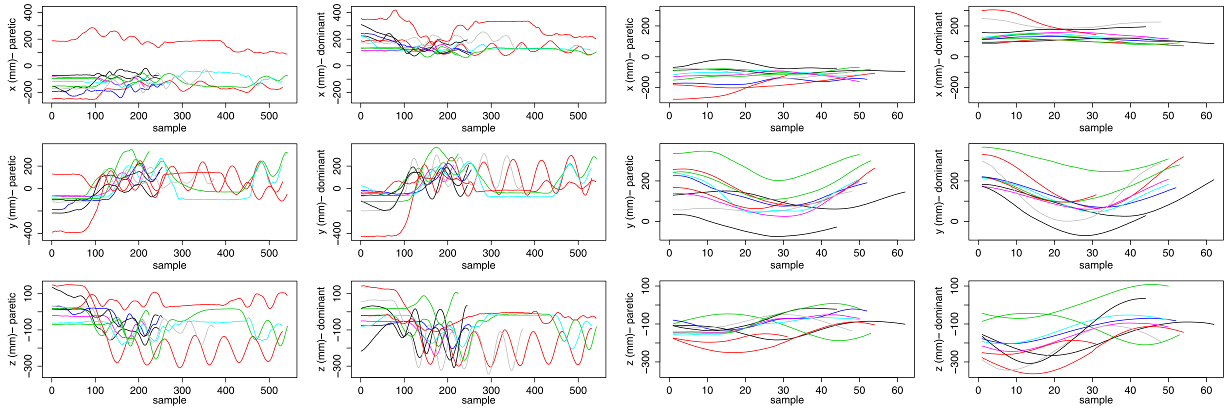

Fig. 2. Raw (left) and segmented data (right).

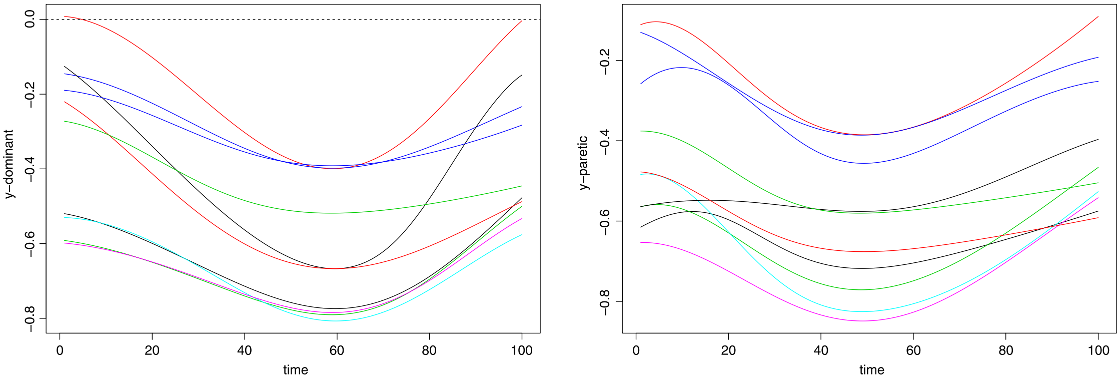

Fig. 3. Data after registration.

In our projects, several dozens of summary statistics have been calculated, linked to features of speed, fluency, accuracy and the synchrony between two hands when patients do each movement, providing information on assessing their upper limbs function. Some summary statistics are easy to define and calculate from raw data (a sample is shown in Fig. 2), such as speed, range of movement (to measure accuracy), but the others need a lot of effort by collaborating statistical analysis with knowledge from experts of the field. For example, we proposed to use the maximum cross-correlations between lags of −5 to 5 to measure synchrony between two hands after lengthy discussion and numerical comparison. This measure is suitable for both mirrored and in-phase movements, and has been proved to be an important variable in the final model.

Data registration is a key step in pre-procession when we need to use curve or shaped data in modelling, but is still less familiar in statistics outside the community of functional data analysis. Fig. 2 shows the segmented data (one replication of the movement of a forward roll) from several people. The huge variation among the curves blocks the possibility of using them directly in modelling. Simply using a landmark registration (Ramsay and Silverman, 2005) makes those curves comparable (Fig. 3) and it is therefore possible to use them in a functional regression model. Registration has now been used in many different areas (see e.g. Raket et al., 2016) particularly those involving big spatial or temporal data. It is still a hot and challenging topic. The difficulties of developing registration methods are attributed to finding transformations making the curves comparable and keeping individual characteristics simultaneously. Some recent development can be found in Marron et al. (2015).

4. Modelling

This is certainly the territory of statisticians, although we have lost our authority in many areas particularly in big data analysis. In the Centre of Doctoral Training of Cloud Computing for Big Data in Newcastle University, UK, we provide industrial projects to students every year. Each project usually involves very big data, and is conducted by a group of mixed students having a computer science or statistics/mathematics background. Usually, students with a computer science background perform very well at the beginning. They can define processing pipelines to process data coming from different resources and having different formats using rapidly developed technologies in cloud computing and big data, such as visual studio, Microsoft Azure, Hadoop, etc. However, after the first phase, almost all the problems the companies are really concerned with are converted to problems of modelling, including how to use statistical language to describe the problems, which model should be used, and how to interpret the outcomes. These examples show that modelling is the key in analysing big data. The reality is that mathematicians/statisticians are lagging a bit behind our contenders in machine learning. This is mainly caused by the different culture in the two disciplines. Researchers in machine learning focus more on problem solving. They pay much more attention on pre-processing and on feature extraction and development (obtaining summary statistics), and usually start modelling from simple and conventional models, and this allows them to address many problems efficiently and quickly.

Statisticians have advantages in having a good understanding on the theory in modelling, on the structure of data, on advanced methodology for complex data, but sometimes do less work on other parts of a project as described in the previous sections and are over ambitious on some complex models. If we can borrow some experience from the machine learning community and combine our advantages, we can achieve more together.

In our project, to develop models used in the automatic assessment system, we first try the models using more than two hundred scalar variables including all patient specific information and summary statistics calculated from the movement data as discussed earlier. We tried different regression models, linear and nonlinear mixed-effect models all based only on those scalar variables before we consider functional variables. The final modelling system is based on a nonlinear mixed-effects scalar-on-function model, where the use of functional covariates does improve results, and the inclusion of nonlinear mixed-effects gives a further improvement for acute patients (within 6 month after stroke); see the details in Cheng et al. (2017). The development of the system is successful (Serradilla et al. (2014) won the best paper award in IEEE international conference on serious games and applications for health), and this is an example showing the power of using advanced statistical techniques. However, it is impossible to develop the final model without doing a substantial amount of work (more than eighty percent) on designing the experiment, understanding the data, data cleaning and pre-procession and developing new features (summary statistics).

5. Conclusion

As a conclusion, we think modelling is the key in big data analysis. Advanced and complex models can help to address difficult problems and improve performance. Statisticians have advantages to be a strong competitor. However, to achieve the desired goal, statisticians need to be better involved in other steps of the project including experiment design, data acquisition, cleaning and pre-processing (see other papers in this special issue, e.g. Smirnova et al., 2017). We need to pay more attention on how to extract information from big noisy data using summary statistics and to use simple models before we consider any complex and advanced methodology.

Acknowledgements

I would like to thank my colleagues Prof. Eyre, Drs Cheng and Serradilla for many stimulating discussions on the topic, thank Shan and the reviewer for improving the presentation, and thank the editor, Laura Sangalli, for inviting me.

References

Barreca, S., Stratford, P., Lambert, C., Masters, L., Streiner, D., 2005. Test-retest reliability, validity, and sensitivity of the Chedoke arm and hand activity inventory: a new measure of upper-limb function for survivors of stroke. Arch. Phys. Med. Rehabil. 86, 1616–1622.

Cheng, Y., Shi, J.Q., Eyre, J., 2017. Nonlinear Mixed-effects Scalar-on-function Models and Variable Selection for Kinematic Upper Limb Movement Data. arXiv:1605.06779.

Marron, J.S., Ramsay, J.O., Sangalli, L.M., Srivastava, A., et al., 2015. Functional data analysis of amplitude and phase variation. Statist. Sci. 30, 468–484.

Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D., 2009. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Trans. Biomed. Eng. 56, 871–879.

Raket, L.L., Grimme, B., Schoner, G., Igel, C., Markussen, B., 2016. Separating timing, movement conditions and individual differences in the analysis of human movement. PLoS Comput. Biol. 12.

Ramsay, J.O., Silverman, B.O., 2005. Functional Data Analysis, Second ed. Springer.

Serradilla, J., Shi, J.Q., Cheng, Y., Morgan, G., Lambden, C., Eyre, J.A., 2014. Automatic Assessment of Upper Limb Function During Play of the Action Video Game, Circus Challenge: Validity and Sensitivity to Change.SEGAH 2014.

Shi, J.Q, Cheng, Y., Serradilla, J., Morgan, G., Lambden, C., Ford, G.A., Price, C., Rodgers, H., Cassidy, T., Rochester, L., Eyre, J.A., et al., 2013. Evaluating functional ability of upper limbs after stroke using video game data. In: Imamura, K., et al. (Eds.), BHI 2013. In: LNAI, vol. 8211, Springer, pp. 181–192.

Shoemake, K., 1985. Animating rotation with quaternion curves. In: Proceedings of SIGGRAPH 1985, vol. 19. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, pp. 245–254.

Smirnova, E., Ivanescu, A., Bai, J., Crainiceanu, C.M., 2018. A practical guide to big data. Statist. Probab. Lett. 136, 25–29. Special Issue on "The role of Statistics in the era of Big Data".