ISSN 2186-7437

NII Shonan Meeting Report

No. 2018-3

Analysing Large Collections of Time Series

Anne Driemel
Rob J Hyndman
Galit Shmueli

February 12–15, 2018

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Analysing Large Collections of Time Series

Organizers:
Anne Driemel (TU Eindhoven, The Netherlands)
Rob J Hyndman (Monash University, Australia)
Galit Shmueli (National Tsing Hua University, Taiwan)

February 12–15, 2018

Due to technological advances in sensor technology, there is a tremendous increase in the availability of massive data streams—many of these are time series in nature. For examples, sensors could measure various parameters of a manufacturing environment, vital parameters of a medical patient, or fitness parameters of a healthy person, or movement sensors could be installed in a fixed environment, traffic sensors measure the number of people or vehicles in a network. Other examples where massive time series data are generated include web applications tracking user clicks, machine log data generated in an IT-infrastructure, and point-of-sales data in a large store. The aim of capturing this data could be monitoring production quality, monitoring the state of health of a patient, detecting intruders climbing over fences, predicting if a user is likely to click on an advertisement, forecasting the number of passengers on a particular train route, and so on. These advances in sensor and cloud technologies have led to the Internet of Things (IoT) phenomenon, where huge sets of time series are collected in a distributed way and either the data or some aspect of them is transferred to a centralized cloud.

As a result of the deluge of information, new paradigms are needed for working with time series data. Instead of working at the level of individual observations, we can consider each time series as a single data point in a space of time series. Data analysis tasks such as forecasting, clustering, density estimation and outlier detection, have been largely developed for Euclidean (feature) spaces, and cannot easily be applied in these spaces of time series. We need new algorithmic methods in order to handle the infinite- dimensional geometry of the space of time series.

Indeed, the study of forecasting methods has a long history, and has been studied in various scientific communities, including statistics, econometrics, control engineering and computer science. Many of the widely-used techniques (such as exponential smoothing and the Kalman filter) were developed in the 1960s. From an algorithmic perspective, these methods are elegant and efficient, which make them very appealing when computational power is scarce. Since then, much progress has been made with respect to both theoretical and computational aspects of forecasting. However, the focus has been limited to forecasting individual time series, or a small number of time series. New methods are required in order to develop algorithms and models designed for forecasting millions of related series.

Once we take the perspective of studying a space of time series, we can consider potential time series that have not yet been observed (e.g., the data that will be observed after we install a new sensor). We may wish to forecast these unobserved time series, but the existing paradigms provide no way of doing so. Visualization of large collections of time series is also challenging, and impossible using classical time series graphics. Similarly, identifying outliers in a space of time series, or defining the median of a large collection of time series are difficult tasks and existing tools are very limited or non-existent.

The workshop brought together 23 researchers across statistics, econometrics and computer science, and in particular from the areas of Computational Geometry and Forecasting, as well as Topological Data Analysis (TDA), Functional Data Analysis (FDA), and Machine Learning. The program of the workshop included survey talks on selected topics and ample time for work in groups on focus topics that resulted from discussions among participants. Almost all of the workgroups were interdisciplinary—leading to new research collaborations and showing that there is a huge potential in the cross-fertilization of these research communities under the lens of time series data.

Overview of Survey Talks

Feature-based time-series analysis

Ben D Fulcher, Monash University, Australia

I will give an introduction to feature-based approaches to time-series analysis. I will summarize the range of feature-based representations for time series that have been developed to aid interpretable insights into time-series structure. Particular emphasis will be given to emerging research that facilitates wide comparison of feature-based representations that allow us to understand the properties of a time-series dataset that make it suited to a particular feature-based representation or analysis algorithm. I argue that the future of time-series analysis is likely to embrace approaches that exploit machine learning methods to partially automate human learning to aid understanding of the complex dynamical patterns in the time series we measure from the world.

Challenges in Forecasting High-Dimensional Time Series

Julie Novak, Netflix, USA

This talk provides an overview of the challenges faced when forecasting high-dimensional time series data and the methods used to address them. We motivate the topic by describing issues that arise when forecasting IBMs quarterly revenue for all their divisions and markets. It is often the case that such high-dimensional time series are naturally structured in a hierarchical manner. As a result, the forecast reconciliation problem becomes a critical one, where the goal is to make sure that forecasts produced independently at each node of the hierarchy are aggregate consistent while remaining as accurate as possible. We review the state of the art methodology that practitioners currently use and highlight recent advances in the field. Finally we discuss open questions and future research directions in this area.

Core-sets for learning streaming signals in real-time

Dan Feldman, University of Haifa, Israel

A coreset (or, core-set) for a given problem is a compressed representation of its input, in the sense that a solution for the problem with the (small) coreset as input would yield a provable (1+epsilon) factor approximation to the problem with the original (large) input. Using traditional techniques, a coreset usually implies provable linear time algorithms for the corresponding optimization problem, which can be computed in parallel on the cloud/GPU, via one pass over Big data, and using only logarithmic space (i.e, in the streaming model). In this talk I will survey main coresets techniques, with applications for real-time signal processing such as localization of nano-drones, GPS data, and new coresets for deep learning.

Functional data analysis, with a view on current time series methods

Alexander Aue, University of California, Davis, USA

In this talk, I will trace the broader developments within the field of functional data analysis that have taken place during the past two or so decades, with attention focused on the case of dependent functional observations. I will discuss by way of examples the most important tools of statistical inference, such as dimension reduction techniques, for independent data, explain what issues arise under dependence and how these may be resolved. These general considerations will then be utilized to give an overview of more specialized prediction algorithms and estimation strategies for functional time series. The talk will conclude with some speculation about future research directions.

Topological Data Analysis In a Nutshell

Bei Wang, University of Utah, USA

Topological Data Analysis (TDA) is an emerging area in exploratory data analysis and data visualization that has had a growing interests and notable successes with an expanding research community. Topological techniques which capture the shape of data have the potential to extract salient features and to provide robust descriptions of large and complex (i.e., high throughput, high-dimensional, incomplete and noisy) data. In this talk, I will survey some of the classic topological techniques, with a focus on their applications in data analysis and data visualization. I will also briefly touch on the new opportunities connecting TDA with time series analysis.

Trajectory Segmentation and Clustering

Kevin Buchin, TU Eindhoven, the Netherlands and
Maike Buchin, TU Dortmund, Germany

Nowadays more and more movement data is being collected, of people, animals, and vehicles. Analysing such data requires efficient algorithms. We first give a brief overview of work in this field, and then focus on algorithms for two analysis tasks: segmentation and clustering. Segmentation asks to split and possibly group trajectories such that they have similar movement characteristics. We present geometric and model-based approaches to segmentation, and show how these can be used to classify subtrajectories based on their characteristics. Clustering asks to group similar trajectories or subtrajectories. We present algorithmic results for clustering based on geometric similarity measures

Workgroup Reports

Visualizing nested and crossed aggregation structures for time series

George Athanasopoulos, Rob Hyndman, Hanlin Shang, Galit Shmueli

Collections of time series are often disaggregated using different grouping factors, which may be nested. For example, we may have time series of population by sex and by country, state and region; the geographical factor is nested, while sex is not. In many examples, there may be several of these grouping factors, and some of them have several levels of nesting. It is important to be able to visualize the structure of the groupings and nesting to better understand the possibilities of analyzing data at different aggregation levels. We explored several visualization possibilities, including existing solutions (e.g. UpSet) and developing new ones. We concluded that the structure is most easily visualized using a set of tree maps. An interactive display could be built where elements of the tree maps are clickable, and a linked plot shows the time series associated with each selected element, and their interactions, and possibly also their children.

Forecasting trajectories

Anne Driemel, Kevin Buchin, Mahsa Ashouri, Alexander Aue, Julie Novak, Anastasios Panagiotelis

Forecasting models such as ARMA have been successfully applied to vector data and functional data. The starting question of the discussion was whether we can extend this to trajectories of moving objects such as seagulls. The trajectory would be observed either over the period of several days, or several trips starting from a fixed location, or trips between two fixed locations. Repeated observations of the same trip, ordered in time, could be used to learn the distribution in a suitable functional space. One of the challenges lies in mapping the trajectories to the functional space, normalizing the different lengths and local time-deviations. The main question, however, is whether there is any persistence in the behavior from trip to trip that would justify using a forecasting model. We looked at a small subset of trips of seagull trajectories of the same bird over several days and did not find any evidence of this persistence. There might be a different data set that shows such persistence.

Predicting Movement Patterns

Anne Driemel, Kevin Buchin, Alexander Aue

In trajectory analysis local time warping plays an important role to align similar trajectories and to detect repeated movement patterns. The starting question of the discussion was how local time warping can be integrated with methods from functional time series analysis. This was discussed by examples of data and analysis problems in sports analysis and ecology. The three participants of the discussion group plan to continue investigating this interesting problem in more detail in the future.

Protecting Privacy in Time Series Analysis

Maike Buchin, Jie Gao, Michael Horton, Maarten Löffler, Nalini Ravishanker, Galit Shmueli, Frank Staals

The goal of this work is exploring and mapping the terrain of issues associated with privacy protection in the context of time series analysis. While privacy protection is widely researched in the context of cross-sectional data, we identified a gap in the context of time series. Time series that measure potentially private information have now become extremely widely available. Examples include smart meter readings, fitness band health data, and driving sensors data. Scoping the map requires identifying the following major components: analysis goals of interest, determining what is sensitive about the data, and the different approaches. What is sensitive information? We distinguish between sensitive "periods" within an individual time series, cases when the entire time series is sensitive, and the case when the time series becomes sensitive when integrated/correlated with public data In order to decide what is sensitive information and what is the utility of the analysis, we must determine (1) who has the original (input) data (2) who does the computing and (3) who receives the output of the analysis.

Key distinct purposes of analyzing time series data include forecasting aggregate time series and subset aggregations (e.g. electricity load management), describing a "typical" individual time series, comparing an individual time series to a benchmark (e.g. for calibration, anomaly detection, classification), identifying correlated sets of time series (e.g. for detecting contagion or gaming); and forecasting individual time series using auxiliary/ancillary data. We identified several different potential approaches that are either borrowed from non-time-series privacy protection or that are time-series specific. These include statistical models, data replacement/imputation with artificial values, data swapping across series, adding fake series, hiding sensitive periods, time scale swapping/reversal/shifting, and series normalization.

New topological features for time series

Bei Wang, Kate Smith-Miles, Ben Fulcher, Nalini Ravishanker, Kevin Verbeek, Dilini Talagala, Thiyanga Talagala

We are interested in knowing whether topological data analysis (TDA) are useful for discriminating features of time series, in particular for nonlinear, nonstationary time series. In order to test this, we start with three types of datasets. First, we experiment with 3 different types of time series derived from the Lorenz attractor where the best existing features achieve roughly a 80% classification accuracy. We are interested in knowing whether TDA features can achieve a higher classification accuracy alone (usefulness); or how different are TDA features from existing features (uniqueness). Second, we will experiment with 1000 empirical datasets to address the usefulness and uniqueness questions. We are interested in knowing whether there is a sweet spot, that is, is there a class of time series with certain properties for which TDA features perform the best, i.e. in terms of classification and regression. Last, we are interested in classifying labeled datasets from earthquakes and explosions. In particular, we would like

to study the use of TDA features in classification task in comparison to using higher order spectral (HOS) methods. For instance, can TDA by itself, or TDA in addition to HOS methods lead to lower/zero misclassification rates.

We want to explore the following research questions:

1. How do we map a time series dataset to metric space for computing persistent homology?
2. How do we map points in the persistent diagrams to informative TDA features?
3. How do we characterize information loss or sufficiency of using TDA features?
4. Are there sweet spots (i.e., broad classes of time series and datasets) for TDA type methods? Are there classes for which TDA helps capture additional information over and beyond well accepted methods in such areas.
5. Scalable computation for TDA features for time series datasets?
6. What are the metrics for success?

Measuring the discriminatory power and usefulness of a time series feature

Ben Fulcher, Bei Wang, Kevin Verbeek, Nalini Ravishanker, Dilini Talagala, Thiyanga Talagala, Rob Hyndman, Julie Novak, Kate Smith-Miles, Anastasios Panagiotelis, George Athanasopoulos

We are interested to develop a new methodology for evaluating any new candidate feature of a time series: to establish its robustness, its uniqueness and discrimination power, and its usefulness. Of course, this is with reference to an existing set of features (around 7000 features from Fulcher), and across an existing set of instances (for example, the more than 10000 time series from Fulcher, the 100000 time series from Kang, Hyndman and Smith-Miles, or a specific set of time series associated with an application domain). Tests will be developed to provide statistically rigorous evidence of these properties of new candidate features:

1. Robustness - how well does the feature respond to small perturbations in the time series?
2. Discrimination - if we regress on the new feature against the existing features, can we conclude that the new feature explains any variation not captured by the existing features?
   a. If the new feature is discriminating, we still need to test for usefulness.
   b. If the new feature is not discriminating on these instances, can we find subsets of the instances, or can we evolve new instances, where the new feature proves to be powerful to explain variation in the data. If we can find no instances where the new feature is discriminating, then it is probably not useful (although we may wish to keep it and replace some existing features due to advantages such as computational efficiency or interpretability). If the new feature is discriminating on some subset of instances, we still need to check for usefulness.
3. Usefulness - this depends on the application: regression, classification, clustering, forecasting, anomaly detection, etc. Each of these purposes/goals has a performance metrics (e.g. classification accuracy), which can be used to perform a simple test: does the performance metric increase (statistically

significantly) if we augment the feature set with the new feature, compared to the existing feature set?

We intend to write a methodological paper that establishes these statistical tests within the above framework, and then demonstrates the methodology on a case study taken from the other project: new topological features for time series.

Data reduction for time series

Maike Buchin, Maarten Löffler, Frank Staals

We considered the following question. Let $f(t)$ be an unknown function of time, and suppose we want to collect information about f. We can do this by measuring f at any time. The challenge is that we have limited battery / limited budget / want to minimise the number of measurements while ensuring that we have a sufficiently close approximation of f.

Trajectory Clustering

Dan Feldman, Jie Gao, Michael Horton

Consider trajectories passively collected by cell towers or WiFi access points in the following manner: an agent with ID in the vicinity of a cell tower P at time t is recorded by the cell tower as a tuple (ID, t, P). Data collected by a cell tower is locally stored at the cell tower. We would like to compute compact data structures at the cell towers such that one can answer queries efficiently about the cost of proposed bus schedules—defined by the total cost of moving the trajectories to be aligned with the nearest bus schedule (a k-means type of cost). We expect to use coreset and min-hash to address this problem.

Coresets for the Frechet distance

Dan Feldman, Anne Driemel, Kevin Buchin

Coresets provide a powerful theoretical framework for developing approximately correct techniques to compute sophisticated statistics over streaming data. Examples of problems that can be solved with this technique include k-means clustering, eigendecomposition and regression. One of the open problems is whether the framework can be extended to clustering under the Frechet distance. We discussed some possible solutions in the restricted setting where queries are line segments. We hope to be able to generalize these solutions to more complex queries.

List of Participants

- Mahsa Ashouri, National Tsing Hua University, Taiwan
- George Athanasopoulos, Monash University, Australia
- Alexander Aue, University of California, Davis, USA
- Maike Buchin, TU Dortmund, Germany
- Kevin Buchin, TU Eindhoven, the Netherlands
- Dan Feldman, University of Haifa, Israel
- Ben Fulcher, Monash University, Australia
- Jie Gao, Stony Brook University, USA
- Michael Horton, University of Sydney, Australia
- Maarten Löffler, Utrecht University, the Netherlands
- Julie Novak, Netflix, USA
- Anastasios Panagiotelis, Monash University, Australia
- Nalini Ravishankar, University of Connecticut, USA
- Hanlin Shang, Australian National University, Australia
- Kate Smith-Miles, University of Melbourne, Australia
- Frank Staals, Utrecht University, the Netherlands
- Dilini Talagala, Monash University, Australia
- Thiyanga Talagala, Monash University, Australia
- Kevin Verbeek, Eindhoven Technical University, the Netherlands
- Bei Wang, University of Utah, USA
- Qiwei Yao, London School of Economics, UK