INI programme on 'Statistical scalability': final report

10 January–29 June 2018

ORGANISERS: John Aston (Cambridge), Idris Eckley (Lancaster), Paul Fearnhead (Lancaster), Po-Ling Loh (U. Wisconsin-Madison), Robert Nowak (U. Wisconsin-Madison), Richard Samworth (Cambridge)

SCIENTIFIC ADVISORY COMMITTEE: Peter Bühlmann (ETH Zürich), Tony Cai (U. Pennsylvania), Emmanuel Candès (Stanford), Richard Davis (Columbia), Sara van de Geer (ETH Zürich), Martin Wainwright (UC Berkeley)

ROTHSCHILD PROFESSOR: Peter Bühlmann (ETH Zürich)

SIMONS FELLOWS: Yining Chen (LSE), Holger Dette (Bochum), Edward George (U. Pennsylvania), Claudia Kirch (Magdeburg), Tatyana Krivobokova (Göttingen), Po-Ling Loh (U. Wisconsin-Madison), Jon Wellner (U. Washington), Yi Yu (University of Bristol).

1 Background

We are living in the information age. Modern technology is transforming our ability to collect and store data on unprecedented scales. From the use of Oyster card data to improve London's transport network, to the Square Kilometre Array astrophysics project that has the potential to transform our understanding of the universe, 'Big Data' can inform and enrich many aspects of our lives. Given the prospects of transformational advances to standard practice in a plethora of data-rich industries, government agencies, science and technology, it is unsurprising that Big Data is currently receiving such a high level of media publicity.

Of course, the important role of statistics within Big Data has been clear for some time. However, the current tendency has been to focus purely on algorithmic scalability, such as how to develop versions of existing statistical algorithms that scale better with the amount of data. Such an approach, however, ignores the fact that fundamentally new issues often arise, and highly innovative solutions are required. In particular, the thesis of this programme was that it is only by simultaneous consideration of the methodological, theoretical and computational challenges involved that we can hope to provide robust, scalable methods that are crucial to unlocking the potential of Big Data.

2 Programme scope and outline

The programme was extremely broad, involving 95 programme participants and 249 workshop participants. The following topics are therefore only indicative of general research areas covered during the programme:

Statistical inference after model selection: For many years, it has been a relatively standard practice among applied practitioners to carry out exploratory data analyses in which many different statistical models are fit to a particular dataset, before some model selection algorithm such as Akaike's Information Criterion (Akaike, 1974) is applied to settle on a final model. The difficulty is that uncertainty in parameter estimates is then often reported without reference to the model uncertainty, thereby underestimating the true uncertainty and leading to overconfident conclusions. In Big Data situations where thousands of variables may or may not be needed to adequately capture the data generating mechanism, the perils of naively reusing the same data for model selection and quantifying uncertainty in parameter estimates become even more severe. Of course, one solution is simply to split the sample into two parts, and carry out the model selection and parameter estimation on disjoint sets of observations, but this has always proved unpopular with practitioners due to the reduced 'final' sample size. Very recent developments, e.g. Berk et al. (2013), Zhang and Zhang (2014), Lee et al. (2016), Yu, Bradic and Samworth (2018), Janková and van de Geer (2018) have proposed innovative solutions to this crucial issue in different contexts, and we anticipate significant work in the coming years to develop these ideas to fruition.

Fundamental trade-offs between statistical and computational efficiency: Several other very recent works (e.g. Berthet and Rigollet, 2013; Chandrasekaran and Jordan, 2013; Wang et al., 2016) have sought to quantify trade-offs between statistical and computational efficiency. In one formulation, it is now known that in many common problems in the analysis of Big Data, such as Sparse Principal Components Analysis, and under a widely-believed hypothesis from computational complexity theory, there are regimes in which no randomised polynomial time algorithm can attain the statistically optimal minimax rate of convergence. Such results both create fascinating connections between statistics, theoretical computer science, and information theory, and have fundamental implications for the design of practical algorithms for handling Big Data.

Model misspecification: Data generating mechanisms that underpin Big Data are inevitably enormously complex, and the best statisticians can hope for is that their models represent a useful approximation, and that their methods are robust to departures from these models. It is therefore of great interest to understand how statistical procedures behave when the underlying statistical model is misspecified. One might hope, for instance, that an estimator would converge to the closest element of the statistical model to the true data generating process, in some appropriate sense. Such ideas are relatively familiar in the context of convex models, and certain results (e.g. Dümbgen et al., 2011) are now available in non-convex settings. Nevertheless, given the popularity of non-convex regularisation techniques in modern high-dimensional sparse models (e.g. Loh and Wainwright, 2013), there is current interest in providing greater understanding in these and other related inference problems to inform the development of robust statistical methods.

Heterogeneity: Another prototypical feature of Big Data is its heterogeneity. These departures from stylised traditional statistical models of independent and identically distributed observations may take many forms, but include missing data, changepoints, and data combined from multiple sources. Work on these critical issues for the analysis of Big Data is only just beginning (e.g. Städler et al., 2014; Killick et al., 2012; Meinshausen and Bühlmann, 2015; Wang, Yu and Rinaldo, 2017; Wang and Samworth, 2018), and even in examples such as changepoints where the problems have been well studied in simple, univariate contexts, interesting new phenomena

emerge in large-scale data settings (Aston and Kirch, 2018).

New data types: Along with the colossal increases in data volumes, a feature of 21st century data is that they come in widely differing types. In handwriting or speech recognition, for instance, it is often most convenient to model observed data as realisations of random functions; functional data analysis has therefore become an increasingly important area (Ramsay and Silverman, 2005). In other applications such as astrophysics, data may live naturally on a low-dimensional manifold within a higher-dimensional ambient space; the extent to which algorithms can exploit the intrinsic low-dimensional structure is therefore of interest (Genovese et al., 2012). These are important examples of situations where a good understanding of the underlying geometry is crucial for statistical inference. Moreover, recent advances in healthcare technology have sharpened the focus and need for robust methods for processing complex images, which often also possess temporal dependence, such as in the case of fMRI data (e.g. Worsley et al., 2002).

Workshops and other events

The workshops represented high points within the context of the entire programme: a chance for researchers to disseminate their latest ideas to a wider audience and to inspire problems to work on over the subsequent weeks and months.

Workshop 1: Theoretical and algorithmic underpinnings of Big Data (15–19 January). Organisers: Francis Bach, Sara van de Geer, Richard Samworth

This opening workshop served two purposes: first, participants had the chance to describe key recent advances in methodology and algorithms for handling large, complex data structures, along with their theoretical underpinnings. Second, participants were encouraged to use this opportunity to map out what they saw as some of the most important directions to be pursued in the remainder of the programme.

Workshop 2: Statistics of geometric features and new data types (19–23 March). Organisers: John Aston, Richard Davis, Axel Munk

Examples of the new data types facing practitioners, and the geometries they induce, were described above. This was an opportunity for practitioners to learn about the latest developments in these and other challenges, and for more mathematically-oriented scientists to discover the most important new data types that can shape their own research agendas.

Workshop 3: Big Data challenges: heterogeneity, model misspecification and changepoints (16–20 April). Organisers: Peter Bühlmann, Idris Eckley, Po-Ling Loh

As mentioned above, models for Big Data are often rather simplistic, and fail to capture the full complexity of the data generating mechanism. This workshop allowed participants to identify and present progress on some of the outstanding challenges in ensuring robustness of modern statistical methods; a recurring theme of the workshop was how to extend methods in univariate changepoint detection to more complex data types.

This workshop was held in the beautiful setting of Low Wood hotel in the Lake District.

Workshop 4: Future challenges in statistical scalability (25–29 June). Paul Fearnhead, Robert Nowak

The programme concluded with a vision of the future. On the one hand, it summarised how recent advances in the mathematical sciences can be exploited by practitioners suffering from a data deluge; on the other, participants presented outstanding challenges facing the discipline in the coming years.

Open for Business days

In conjunction with the Turing Gateway to Mathematics (and in particular, Lissie Hope, Jane Leeks and Clare Merritt), two Open for Business Days were held during the programme. The first was on 'Big Data and the role of statistical scalability' (28 February), and the second was on 'Statistical scalability for streaming data' (21 June). Here, talks were mainly given by practitioners from the public sector and industry on the manifold challenges they face in extracting useful information from their colossal data sets. Particular topics included discussion of challenges and opportunities in using data to better inform drug development and personalised medicine, and to diagnose faults and security threats on IP networks. These Open for Business days provided excellent opportunities for interdisciplinary engagement and collaboration.

Rothschild lecture

Another significant highlight of the programme was the Rothschild lecture, delivered on 22 June by Peter Bühlmann. Peter spoke on the topic of 'Causality, invariance and robustness', and described both some of the fascinating history of causation in statistics, and also new ways in which heterogeneity can be expoited for causal inference.

Regular seminars

Outside of the workshop weeks and special events, we had a regular schedule of two seminars per week. As well as regular research talks, these times allowed more junior researchers to present their work in a more informal setting, and more established researchers to give tutorial lectures.

Social activities

The programme participants very much enjoyed the regular tea and cakes in the Isaac Newton Institute, as well as regular dinners, punting outings and trips to watch the May Bumps, to name just a few!

Scientific outcomes

Many participants commented on the oustanding environment and research facilities provided by the Isaac Newton Institute for the purposes of collaboration and interaction. The layout of the building really helps people to meet and start chatting, and the release from regular university duties, the support provided by the INI staff and the proximity of both the excellent Betty and Gordon Moore library and (of course!) free coffee are all conducive to research ideas and productivity. Progress was made on problems of machine learning debugging, scale calibration in robust regression, multivariate ranks, changepoint detection, estimation of stationary covariance matrices, partial least squares for classification, the need to automatically identify appropriate

structure in Big Data, bringing ideas from category theory and theoretical computer science into scalable statistical computation, and shape-constrained estimation, among many others. Several talks were mentioned as particular highlights, including those by Tim Cannings (U. Southern California) on 'Classification with imperfect training labels', Adam Sykulski (Lancaster) on 'Spatiotemporal modelling and parameter estimation of anisotropic particle trajectories' and Guy Bresler (MIT) on 'Reducibility and computational lower bounds for problems with planted sparse structure'. The fact that the talks are streamed is an invaluable resource.

Several international programme participants gave seminars at other UK universities, including York, Lancaster, LSE, Glasgow and Bristol. It was also beneficial that the other concurrent INI programme, on 'Uncertainty quantification for complex systems: theory and methodologies', was on a loosely cognate area, and there was plenty of fruitful interaction between the programmes.

It was extremely rewarding to see that the programme participants felt they benefited so greatly from the Statistical Scalability programme, and we thank the Isaac Newton Institute staff for all their hard work on our behalf.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **19**, 716–723.

Aston, J. A. D. and Kirch, C. (2014) High dimensional efficiency with applications to change point tests. *Electron. J. Statist.*, **12**, 1901–1947.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *Ann. Statist.*, **41**, 802–837.

Berthet, Q. and Rigollet P. (2013) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. W&CP*, **30**, 1046–1066.

Chandrasekaran, V. and Jordan, M. I. (2013) Computational and statistical tradeoffs via convex relaxation. *Proc. Nat. Acad. Sci.*, **110**, E1181–E1190.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011) Approximation by log-concave distributions, with applications to regression. *Ann. Statist.*, **39**, 702–730.

Genovese, C., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012) Minimax manifold estimation. *J. Mach. Learn. Res.*, **13**, 1263–1291.

Jamieson, J., Malloy, M., Bubeck, S. and Nowak, R. (2014) lil UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits. *COLT 2014*.

Janková, J. and van de Geer, S. (2018) Inference in high-dimensional graphical models. `https://www.arxiv.org/abs/1801.08512`.

Killick, R., Fearnhead, P. and Eckley, I. (2012) Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, **107**, 1590–1598.

Loh, P. and Wainwright, M. J. (2013) Regularized M-estimators with non convexity: Statistical and algorithmic theory for local optima. *NIPS 2013*.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. (2016) Exact post-selection inference with the Lasso. *Ann. Statist.*, **44**, 907–927.

Meinshausen, N. and Bühlmann, P. (2015) Maximin effects in inhomogeneous large-scale data. *Ann. Statist.*, **43**, 1801–1830.

Ramsay, J. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer-Verlag, New York.

Städler, N., Stekhoven, D.J. and Bühlmann, P. (2014) Pattern alternating maximization algorithm for missing data in large p, small n problems. *J. Mach. Learn. Res.* **15**, 1903–1928.

Wang, T., Berthet, Q. and Samworth, R. J. (2016) Statistical and computational trade-offs in the estimation of sparse principal components. *Ann. Statist.*, **44**, 1896–1930.

Wang, T. and Samworth, R. J. (2018) High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, **80**, 57–83.

Wang, D., Yu, Y. and Rinaldo, A. (2017) Optimal covariance change point detection in high dimension. `https://www.arxiv.org/abs/1712.09912`.

Worsley, K. J., Liao, C., Aston, J. A. D., Petre, V., Duncan, G. H. and Evans, A. C. (2002) A general statistical analysis for fMRI data. *Neuroimage*, **15**, 1–15.

Yu, Y., Bradic, J. and Samworth, R. J. (2018) Confidence intervals for high-dimensional Cox models. `https://arxiv.org/abs/1803.01150`.

Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc., Ser. B*, **76**, 217–42.