Snapshots of modern mathematics
from Oberwolfach

№ 3/2021

Searching for structure in complex data: A modern statistical quest

Po-Ling Loh[1]

Current research in statistics has taken interesting new directions, as data collected from scientific studies has become increasingly complex. At first glance, the number of experiments conducted by a scientist must be fairly large in order for a statistician to draw correct conclusions based on noisy measurements of a large number of factors. However, statisticians may often uncover simpler structure in the data, enabling accurate statistical inference based on relatively few experiments. In this snapshot, we will introduce the concept of high-dimensional statistical estimation via optimization, and illustrate this principle using an example from medical imaging. We will also present several open questions which are actively being studied by researchers in statistics.

[1] This article is dedicated to Sara van de Geer on the occasion of her birthday, which was celebrated in Oberwolfach in May 2019.

1 The age of modern data

Many modern scientific disciplines are encountering a data revolution. Due to advances in technology and computational power, large volumes of data are now being acquired with unprecedented ease. This provides an unending supply of fascinating new toys for the statistician's playground, many of which exceed those previously encountered in terms of sheer size and complexity. It is thus the statistician's favorite diversion to derive new methods to analyze these datasets, finding ways to interpret the data and draw meaningful and relevant conclusions.

A common characteristic of many contemporary datasets is their *high-dimensional* nature, simultaneously referring to the fact that the number of data points in the acquisition set is rather large, and the property that the number of measurements (features) observed for each point may be larger than the number of points in the dataset. This is partly due to the ability of scientists to collect data on much larger scales than before. A cell biologist can run numerous experiments in parallel, each of which would have taken many months with older technology. An astronomer can collect telescope images of the entire night sky for days on end, and store the data on a computer server for subsequent analysis. However, in addition to the fact that the number of experiments run by the biologist or the number of images collected by the astronomer is several times larger than previously imagined, the complexity of each individual endeavor is also magnitudes larger—the biologist might be acquiring hundreds of measurements in each experiment, whereas the astronomer would be collecting a high-resolution image with a very large number of pixels during a single sweep of the night sky.

The statistician's role is to help make sense of all this data, producing valid conclusions regarding estimation, inference, or prediction. For instance, a statistician might wish to estimate the magnitude of certain physical characteristics in a cell or in the universe; detect changes in the heavens; or predict the effect of an intervention on the biological function of a living organism. At the same time, the statistician should have an eye for quantifying the uncertainty in each of these conclusions, accounting for inherent randomness in the process of data acquisition or fundamental scientific phenomena. Naturally, the amount of uncertainty in the statistician's conclusion reduces as the number of repeated observations collected by the scientist increases—thus averaging away the errors introduced by computing an estimate based on a random subsample—while being inversely affected by the complexity of the data.

Although the number of independent observations may be relatively large in comparison to previous studies, it is often still dwarfed by the marked complexity of modern data. In the biological example, the number of subjects with a certain abnormal condition from which tissue is extracted might be

limited to tens or hundreds, whereas the number of measurements (e.g., genetic) taken on an individual subject is severalfold larger. Similarly, the capacity limit of modern data repositories is reached upon storing days or weeks of telescope data, whereas the number of pixels in a single image of the sky is colossal. It may therefore appear to be a hopeless endeavor to draw accurate conclusions from datasets with such daunting complexity in truly high-dimensional settings.

Fortunately, another natural phenomenon often emerges to succor the statistician: In many cases, although the ambient dimensionality of the problem corresponding to the number of observed features may be discouragingly vast, the fundamental scientific question under study may be answerable based on a small handful of relevant measurements. Continuing our examples, predicting a change in the health of an organism or a shift in a celestial body might be determined easily by focusing on a certain smaller set of biological measurements or the intensity of a small collection of pixels. Although it is unrealistic to believe that a scientist would have the prescience to measure only these features in an experimental study, a trained statistical sleuth might be able to identify the hidden structure and make accurate predictions, despite the seemingly overwhelming complexity of the original dataset.

2 Solving a minimization problem

How might one endeavor to perform such a challenging task? An approach which has recently become popular among statisticians is to use optimization. The idea is to minimize the quantity

$$\text{Loss}(\beta) + \lambda \cdot \text{Penalty}(\beta), \tag{1}$$

where β is a variable that represents the quantity one wishes to estimate. In a classical statistical setting, one would simply take the approach of minimizing the quantity $\text{Loss}(\beta)$, which computes the amount of error ("loss") incurred by a certain choice of β. The $\text{Penalty}(\beta)$ function is included in formula (1) for the express purpose of discovering unknown structure when the dataset is highly complex, so that searching for the value of β that minimizes $\text{Loss}(\beta)$ alone would lead to a highly inaccurate result based on the relatively small number of experiments. The function $\text{Penalty}(\beta)$ thus "penalizes" assignments of β that are overly complex. Finally, the quantity λ is a positive number that determines the relative importance of minimizing the loss and penalty criteria, so that values of λ which are close to 0 place little importance on minimizing complexity, whereas larger values of λ place a higher emphasis on simplicity in relation to accuracy.

As a concrete example, suppose the goal is to predict whether a tissue sample has been excised from a healthy or diseased individual. In this case, β might

correspond to a particular method for combining the measurements from a tissue sample into a statistician's predictive model, for instance, $\beta = (\beta_1, \beta_2, \beta_3)$, while Loss($\beta$) is the fraction of incorrectly classified samples when the prediction is computed as a function of the weighted sum

$$\beta_1 \times (\text{feature 1}) + \beta_2 \times (\text{feature 2}) + \beta_3 \times (\text{feature 3}). \tag{2}$$

Due to the unwieldy complexity of the data, the scientist might wish to identify a prediction method which makes a decision based on only a small subset of features. The Penalty(β) function would therefore be chosen to produce a larger value whenever β corresponds to a prediction method that involves a large number of features, perhaps by returning a number between 0 and 3 which simply counts the number of nonzero components of β. The solution to the minimization problem (1) would therefore be a value of β which is simultaneously accurate (making Loss(β) small) and simple (making Penalty(β) small).

The primary interest of the statistician is not to devise a method for minimizing the expression (1), per se—although when the problem in question is highly complex, the set of possible assignments for β is likewise very large, and clever algorithms are needed for finding the optimal value of β in a matter of seconds rather than days. Instead, the role of the statistician is to determine the "best" ways to quantify the loss and penalty of an assignment of β so that the optimal value corresponds most closely to the desired scientific outcome. Furthermore, the accuracy of the outcome depends on the amount of data provided by the scientist, since the expression (1) is of course computed using a randomly sampled dataset. Thus, a statistician strives to determine how many replicated experiments must be performed to achieve a certain desirable accuracy; characterize the best choices of the functions Loss(β) and Penalty(β); and assess the proper value of λ that will result in the best possible scientific conclusion for a given dataset. Additionally, the statistician's sensitivities are attuned to sources of uncertainty in the data acquisition process, and he or she may draw a variety of conclusions by modeling different forms of uncertainty in the statistical model, all of which may be valuable to the scientist.

3 A diversion on medical imaging

Having described the general framework under which complex problems and their underlying structure are studied, we now explain a particular scientific example taken from the field of radiology. Routine imaging procedures such as X-ray, magnetic resonance imaging (MRI), or computed tomography (CT) translate measurements from electromagnetic fields in the scanning device into a 2D or 3D image that is then examined by the radiologist. The precise method for

converting these digital signals into images is based on mathematical optimization procedures such as the one described above, where the goal is to "estimate" the physical structure of a patient's internal functions based on inherently noisy observations. Quantifying the number of replicated measurements necessary to attain a desired level of accuracy in the reconstructed image is then critical for minimizing time and cost. Furthermore, as in pediatric or abdominal imaging, it may be impossible for a patient to remain immobile for long periods of time, thus necessitating scans of shorter duration which do not compromise accuracy.

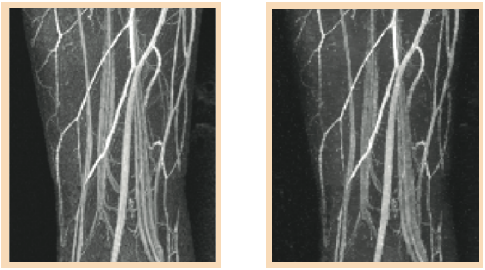

Figure 1: Angiograms showing the location of blood vessels in human tissue. Relatively few pixels in the image are relevant to a physician who wishes to assess the configuration of the patient's blood vessels. Modern statistical methods lead to the relatively clear reconstruction in the right image, based on collecting data for 10% of the time previously required for clinical scans.

The complexity of the image reconstruction problem is naturally determined by the dimensions of the image; for 3D images, the number of pixels can be fairly large even for small physical regions. Nonetheless, medical images possess inherent structure that may be leveraged though careful statistical analysis. For instance, in the angiogram image in Figure 1, the intensities of only a small number of pixels corresponding to locations of blood vessels are relevant for reconstructing the image. Although this is not the case for the brain image in Figure 2, it is nonetheless possible to express the image in terms of a small number of standard "building block" images known as wavelets, corresponding to the underlying simplified structure. This brings us back to the optimization problem (1), where β represents the unknown image, Loss(β) is the error of the MRI signal measurements assuming the image is β, and Penalty(β) records the complexity of the image (which is large when the number of non-blank pixels or wavelets used to represent the image is large). More precisely, a single point in the dataset corresponds to a (measurement, readout)

pair, where the features are computed based on the acquisition frequencies set by the imaging technician within the imaging machine. During a scanning session, data will be collected by repeatedly scanning the object under different acquisition frequencies, corresponding to different (measurement, readout) pairs. The function Loss(β) is then obtained by aggregating the differences between predicted readout signals and observed readout signals over all data acquired during the scan:

$$\begin{aligned}\text{Loss}(\beta) = (\text{prediction } 1 - \text{observation } 1)^2 \\ + (\text{prediction } 2 - \text{observation } 2)^2 + \cdots,\end{aligned}$$

where the predicted readout signals are a function of β computed using the same weighted sum expression given in Equation (2) above.

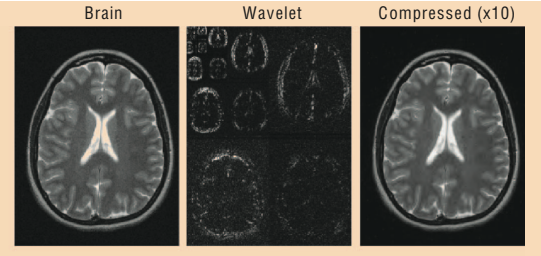

Figure 2: Images of the human brain. The leftmost image shows a brain image from a standard MRI scan. The center image shows some of the simpler wavelet images that can be combined to represent the full brain image. The right image has been reconstructed using 10% of the data, based on the statistical methods discussed in the text.

Reconstructed images based on this procedure are shown in the rightmost panels of Figures 1 and 2. In both cases, the number of measured signals used to create the images is ten times smaller than the number of measurements required by methods that do not take advantage of underlying structure. In practical terms, this means a *tenfold reduction* in the amount of time required for a clinical scan, which can have substantial consequences on the quality of healthcare for a patient! The method we have described is known as "compressed sensing," since the method operates by compressing the complexity of the problem into one that becomes tractable for accurate reconstruction.

4 Current fascinations

The examples and applications described above only provide a glimpse of the many interesting data-analytic problems encountered by contemporary statisticians. We now mention, in broad strokes, an additional assortment of problems that are active topics of discussion among researchers in statistics.

4.1 The problem of inference

Confidence intervals are a commonly used statistical tool for quantifying the uncertainty of estimates with respect to noise in generating or acquiring data. The intervals are often centered around an estimate, with wider intervals indicating more uncertainty. However, the methods for constructing confidence intervals can be fairly complicated in complex data settings, since one incurs additional uncertainty while locating the underlying structure that enables accurate estimation. More concretely, if one wished to quantify the uncertainty in reconstructing the pixel intensities in an angiogram, the level of uncertainty would be larger than for a method which focused on accurately estimating the intensities of a small number of pixels *known a priori* to contain blood vessels. Although these ideas are very natural and intuitive, it is an ongoing challenge for statisticians to quantify exactly how much the level of uncertainty inflates in different scientific settings and for different types of statistical procedures [8, 3, 29].

4.2 Network data

In some datasets, the measurements collected from one experiment may have relationships that can be extracted via a careful statistical analysis of the data. For example, based on surveying a population of individuals for their opinions and interests, one may be able to infer the pattern of friendships between pairs of people, thus constructing the "social network" of the population. Statistical tools have been developed in recent years to reconstruct such networks from indirect observations as in the example previously described [7, 20, 22, 38]. If the network structure were known, an important task might be to infer the clusters of communities in the network based on the relative density of connections between groups of individuals [4, 5], or use the network structure to guide subsequent predictions [28]. As a final example, some statistical work has been conducted on inferring properties of a dynamic process on a dataset collected over a network, to try to locate the source of an epidemic spread based on partial information [21].

4.3 Deep learning

The last few years have seen a flurry of attention focused on a machine learning tool known as deep learning. This comprises a class of algorithms that have been extremely useful for a plethora of tasks, including automatic object detection in photographs, voice recognition, and image editing. On the other hand, although the success of deep learning is widely recognized, many opportunities remain for further investigation. Deep learning methods have often exceeded the performance that might be expected by theoreticians; thus, it remains important to understand the circumstances under which deep learning can be relied upon to succeed, and also the limitations of deep learning in comparison to other preexisting methods [10, 16, 17, 23]. In addition, since much flexibility is given in choosing a deep learning method, statisticians wish to understand which models lead to more robust decisions with less overall uncertainty. For an interesting introduction to deep learning, see Snapshot 15/2019 Deep Learning and Inverse Problems by Arridge *et al.*

4.4 Further topics

The aforementioned sampling of topics is by no means comprehensive, and we now mention several additional areas that are currently being explored by statisticians in the international community. These include causality [9], optimal transport [2, 25, 30, 34], privacy [6, 14], dimension reduction [12, 24, 31, 35, 37], online learning [32, 36], differential equations [1, 11, 15, 27, 33], and regression [13, 18, 19, 26]. Each of these areas is very interesting in their own right, and all share the common theme of uncovering an appropriate type of structure in complex data. The interested reader is encouraged to explore the writings of these authors to gain additional exposure to modern research in statistics.

Image credits

Figure 1 and Figure 2 M. Lustig, D. L. Donoho, J. Santos, and J. Pauly, *Compressed sensing MRI*, IEEE Signal Processing Magazine, March 2008.

References

[1] R. Altmeyer and M. Reiß, *Nonparametric estimation for linear SPDEs from local measurements*, arXiv preprint arXiv:1903.06984 (2019).

[2] F. Bachoc, A. Suvorikova, D. Ginsbourger, J.-M. Loubes, and V. Spokoiny, *Gaussian processes with multidimensional distribution inputs via optimal*

transport and Hilbertian embedding, Electronic Journal of Statistics **14** (2020), no. 2, 2742–2772.

[3] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, *Predictive inference with the jackknife+*, Annals of Statistics **49** (2021), no. 1, 486–507.

[4] M. Behr, M. Ansari, A. Munk, and C. Holmes, *Testing for dependence on tree structures*, Proceedings of the National Academy of Sciences **117** (2020), no. 18, 9787–9792.

[5] X. Bing, F. Bunea, Y. Ning, and M. Wegkamp, *Adaptive estimation in structured factor models with applications to overlapping clustering*, Annals of Statistics **48** (2020), no. 4, 2055–2081.

[6] C. Butucea, A. Dubois, M. Kroll, and A. Saumard, *Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids*, Bernoulli **26** (2020), no. 3, 1727–1764.

[7] A. Caballe, N. Bochkina, and C. Mayer, *Joint estimation of sparse networks with application to paired gene expression data*, arXiv preprint arXiv:1608.05533 (2016).

[8] A. Carpentier, O. Klopp, M. Löffler, and R. Nickl, *Adaptive confidence sets for matrix completion*, Bernoulli **24** (2018), no. 4A, 2429–2460.

[9] D. Cevid, P. Bühlmann, and N. Meinshausen, *Spectral deconfounding via perturbed sparse linear models*, Journal of Machine Learning Research **21** (2020), 1–41.

[10] L. Chizat and F. Bach, *A note on lazy training in supervised differentiable programming*, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), 2019.

[11] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79** (2017), no. 3, 651–676.

[12] G. Darnell, S. Georgiev, S. Mukherjee, and B. E. Engelhardt, *Adaptive randomized dimension reduction on massive data*, The Journal of Machine Learning Research **18** (2017), no. 1, 5134–5163.

[13] H. Dette, A. Pepelyshev, and A. Zhigljavsky, *Optimal designs in regression with correlated errors*, Annals of Statistics **44** (2016), no. 1, 113.

[14] J. Duchi and R. Rogers, *Lower bounds for locally private estimation via communication complexity*, Conference on Learning Theory, PMLR, 2019, pp. 1161–1191.

[15] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, *Log-concave sampling: Metropolis-Hastings algorithms are fast!*, Conference On Learning Theory, 2018, pp. 793–797.

[16] K. Eckle and J. Schmidt-Hieber, *A comparison of deep networks with ReLU activation function and linear spline-type methods*, Neural Networks **110** (2019), 232–242.

[17] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *Linearized two-layers neural networks in high dimension*, arXiv preprint arXiv:1904.12191 (2019).

[18] K. Gregory, E. Mammen, and M. Wahl, *Optimal estimation of sparse high-dimensional additive models*, arXiv preprint arXiv:1603.07632 (2016).

[19] L. Györfi and H. Walk, *On the asymptotic normality of an estimate of a regression functional.*, Journal of Machine Learning Research **16** (2015), 1863–1877.

[20] J. Kelner, F. Koehler, R. Meka, and A. Moitra, *Learning some popular Gaussian graphical models without condition number bounds*, arXiv preprint arXiv:1905.01282 (2019).

[21] J. Khim and P. Loh, *Confidence sets for the source of a diffusion in regular trees*, IEEE Transactions on Network Science and Engineering **4** (2017), no. 1, 27–40.

[22] O. Klopp, A. B. Tsybakov, and N. Verzelen, *Oracle inequalities for network models and sparse graphon estimation*, The Annals of Statistics **45** (2017), no. 1, 316–354.

[23] M. Mardani, Q. Sun, D. Donoho, V. Papyan, H. Monajemi, S. Vasanawala, and J. Pauly, *Neural proximal gradient descent for compressive imaging*, Advances in Neural Information Processing Systems, 2018, pp. 9573–9583.

[24] C. McWhirter, D. G. Mixon, and S. Villar, *SqueezeFit: Label-aware dimensionality reduction by semidefinite programming*, IEEE Transactions on Information Theory **66** (2019), no. 6, 3878–3892.

[25] F. Memoli, Z. Smith, and Z. Wan, *The Wasserstein transform*, International Conference on Machine Learning, PMLR, 2019, pp. 4496–4504.

[26] N. Müecke, *Reducing training time by efficient localized kernel regression*, The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 2603–2610.

[27] R. Nickl, S. van de Geer, and S. Wang, *Convergence rates for penalised least squares estimators in PDE-constrained regression problems*, arXiv preprint arXiv:1809.08818 (2018).

[28] F. Ortelli and S. van de Geer, *On the total variation regularized estimator over a class of tree graphs*, Electronic Journal of Statistics **12** (2018), no. 2, 4517–4570.

[29] T. Patschkowski and A. Rohde, *Locally adaptive confidence bands*, The Annals of Statistics **47** (2019), no. 1, 349–381.

[30] P. Rigollet and J. Weed, *Uncoupled isotonic regression via minimum Wasserstein deconvolution*, Information and Inference: A Journal of the IMA **8** (2019), no. 4, 691–717.

[31] K. Schnass, *On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD*, Applied and Computational Harmonic Analysis **37** (2014), no. 3, 464–491.

[32] J. Shin, A. Ramdas, and A. Rinaldo, *On the bias, risk and consistency of sample means in multi-armed bandits*, arXiv preprint arXiv:1902.00746 (2019).

[33] J. Söhl and M. Trabs, *Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift*, ESAIM: Probability and Statistics **20** (2016), 432–462.

[34] C. Tameling, M. Sommerfeld, and A. Munk, *Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications*, Annals of Applied Probability (2019).

[35] Y. S. Tan and R. Vershynin, *Polynomial time and sample complexity for non-Gaussian component analysis: Spectral methods*, Proceedings of the 31st Conference On Learning Theory (S. Bubeck, V. Perchet, and P. Rigollet, eds.), Proceedings of Machine Learning Research, vol. 75, PMLR, 2018, pp. 498–534.

[36] E. Tanczos, R. Nowak, and B. Mankoff, *A KL-LUCB algorithm for large-scale crowdsourcing*, Advances in Neural Information Processing Systems, 2017, pp. 5894–5903.

[37] T. M. Tang and G. I. Allen, *Integrated principal components analysis*, arXiv preprint arXiv:1810.00832 (2018).

[38] Y. Zhang, E. Levina, and J. Zhu, *Estimating network edge probabilities by neighbourhood smoothing*, Biometrika **104** (2017), no. 4, 771–783.

Po-Ling Loh *is a professor of statistics at the University of Cambridge*

Mathematical subjects
Probability Theory and Statistics

DOI
10.14760/SNAP-2021-003-EN

Connections to other fields
Computer Science, Engineering and Technology

Snapshots of modern mathematics from Oberwolfach provide exciting insights into current mathematical research. They are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the interested public worldwide. All snapshots are published in cooperation with the IMAGINARY platform and can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

ISSN 2626-1995
