Mathematisches Forschungsinstitut Oberwolfach

Report No. 4/2020

DOI: 10.4171/OWR/2020/4

Statistics meets Machine Learning

Organized by
Fadoua Balabdaoui, Zürich
Lutz Dümbgen, Bern
Klaus-Robert Müller, Berlin
Richard Samworth, Cambridge UK

26 January – 1 February 2020

Abstract. Theory and application go hand in hand in most areas of statistics. In a world flooded with huge amounts of data waiting to be analyzed, classified and transformed into useful outputs, the designing of fast, robust and stable algorithms has never been as important as it is today. On the other hand, irrespective of whether the focus is put on estimation, prediction, classification or other purposes, it is equally crucial to provide clear guarantees that such algorithms have strong theoretical guarantees. Many statisticians, independently of their original research interests, have become increasingly aware of the importance of the numerical needs faced in numerous applications including gene expression profiling, health care, pattern and speech recognition, data security, marketing personalization, natural language processing, to name just a few.

The goal of this workshop is twofold: (a) exchange knowledge on successful algorithmic approaches and discuss some of the existing challenges, and (b) to bring together researchers in statistics and machine learning with the aim of sharing expertise and exploiting possible differences in points of views to obtain a better understanding of some of the common important problems.

Mathematics Subject Classification (2010): 62xx.

Introduction by the Organizers

The Oberwolfach workshop Statistics meets Machine Learning, organized jointly by Fadoua Balabdaoui (ETH Zürich), Lutz Dümbgen (University of Bern), Klaus-Robert Müller (Technical University of Berlin) and Richard Samworth (Cambridge) took place during the period January 26th – February 01st, 2020. The

meeting's main goal was to gather researchers from statistics interested in computational tasks and machine learners to exchange current research methods, problems and techniques between the two areas.

The goals of the workshop were successfully attained. During the 21 main talks, the audience got the chance to learn about some of the latest research in neural networks, including recent applications, advances in their implementation, their fundamental limits, and the asymptotic property of the generalization error in the case of over-parametrized neural networks. On the statistical side, the presentations covered a large spectrum of very relevant topics on both theory and computational issues. This includes nonparametric estimation of spatial covariance operators and related separability tests, recent algorithms for single-index models, new methods in functional analysis with big data and interesting theoretical results related to geodesic convex functions. Tuesday evening was devoted to listening to 6 young researchers (students and post-docs) who offered short lightening talks where they presented their most recent results. The organizers take pride in having 28% female participants at this meeting. They also would like to thank all the participants and the administrative staff of the MFO for their great support and help before and during the workshop.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows".

Workshop: Statistics meets Machine Learning

Table of Contents

Grégoire Montavon and Wojciech Samek
Explaining the decisions of deep neural networks and beyond 235

Robert Tibshirani (joint with Ismael Lemhadri, Feng Ruan)
A feature sparse neural network 238

Francis Bach (joint with Hadrien Hendrikx Sébastien Bubeck, Laurent Massoulié, Yin-Tat Lee)
Distributed Machine Learning over Networks 238

Johannes Schmidt-Hieber
Neural networks in the overparametrized regime 240

Emmanuel Candès (joint with Stephen Bates, Matteo Sesia and Chiara Sabatti)
Causal Inference for Genetics with the Digital Twin Test 241

Helmut Bölcskei (joint with D. Elbrächter, P. Grohs, G. Kutyniok, and D. Perekrestenko)
Fundamental limits of deep neural network learning 241

Natalia Bochkina
Testing validity of statistical inference in approximate models 242

Vladimir Spokoiny
Bayesian inference for nonlinear inverse problems 243

Ryan J. Tibshirani
Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems 246

Piet Groeneboom (joint with Fadoua Balabdaoui)
Profile least squares estimators in the monotone single index model 248

Andrea Montanari (joint with Song Mei, Feng Ruan, Youngtak Sohn, Jun Yan)
The generalization error of overparametrized models: Insights from exact asymptotics 251

Angelika Rohde (joint with Cristina Butucea, Lukas Steinberger)
Estimating functionals under local differential privacy 253

Holger Dette (joint with Pramita Bagchi)
Measuring separability in covariance operators of random surfaces 256

David E. Tyler
The role of geodesic convexity in covariance estimation 258

Po-Ling Loh (joint with Ankit Pensia, Varun Jog)
Extracting robust and accurate features via a robust information bottleneck 259

Laura M. Sangalli
Functional and complex data: new methods merging statistics, scientific computing and engineering 260

Bodhisattva Sen (joint with Nabarun Deb)
Multivariate Rank-based Distribution-free Nonparametric Testing using Optimal Transport 260

Victor M. Panaretos (joint with Anirvan Chakraborty)
Testing for the Rank of a Covariance Kernel by Matrix Completion 261

Yi Yu
Network change point localisation 263

Jennifer G. Dy (joint with Chieh Wu, James Ross)
Learning from Complex Medical Data, Clustering, and Interpretable Kernel Dimensionality Reduction 263

Abstracts

Explaining the decisions of deep neural networks and beyond

Grégoire Montavon and Wojciech Samek
(joint work with Klaus-Robert Müller, Sebastian Lapuschkin, Alexander Binder, Jacob Kauffmann)

Machine learning models have become increasingly complex and this complexity has allowed them to reach high prediction accuracy on challenging datasets. In some cases, improved predictivity has come at the expense of interpretability, in particular, complex models tend to be perceived as black-boxes.

A lack of interpretability is problematic, not only because interpretability is desirable in itself (e.g. to extract useful insights from a model or from the modeled data), but also because common measurements of prediction accuracy can become strongly unreliable when certain assumptions about the training data are not met. Real-world datasets are typically not representative of all possible cases and the truly relevant variables may correlate with other irrelevant variables. In such circumstances, one would need to ensure that the machine learning model does not rely on these irrelevant variables. An assessment based purely on test set accuracy would be oblivious to the exact decision strategy and could overestimate the true prediction performance. This phenomenon has been referred to as the 'Clever Hans' effect [9]. Only an extension of the dataset with specific test cases, or an inspection of the model, e.g. via interpretability techniques [3, 16, 12], is capable of highlighting the improper decision structure.

In this talk, we look at the question of explaining the predictions of *deep neural networks*, a successful machine learning approach that has been used increasingly in real-world applications. A challenge for getting these explanations is the complexity of the decision function, which makes it hard to apply simple explanation methods developed in the context of linear models, e.g. based on first-order Taylor expansions. In particular, DNN decision functions are highly nonlinear and multi-scale, with a gradient that is highly varying or 'shattered' [4]. Also, local searches in the input space easily result in 'adversarial examples' [13] where the prediction no longer corresponds to the observed pattern in the input.

Layer-wise relevance propagation (LRP) [3] is a technique that was proposed to robustly explain the neural network decision in terms of input features. It was shown to work on numerous models in a wide range of applications [14, 5, 15]. LRP departs from the neural network's function representation to consider instead its *graph* structure. Specifically, the LRP algorithm performs an iterative redistribution of the neural network output to the lower layers. Redistribution from each layer to the layer below is achieved by means of propagation rules that satisfy a conservation property analogous to Kirchoff's conservation laws in electrical circuits. The LRP algorithm terminates once the input layer has been reached. The LRP algorithm can be motivated as decomposing a complex problem

(analyzing a highly nonlinear function) into a collection of simpler subproblems (treating each neuron individually).

Furthermore, it was shown that the LRP algorithm can be interpreted as a collection of Taylor expansions performed at each layer and neuron of the neural network [11]. Specifically, the 'relevance' received by a given neuron is approximately the product of the neuron activation and a locally constant term. In turn, the LRP redistribution step can be interpreted as (1) identifying the linear terms of a Taylor expansion of the relevance expressed as a function of activations in the lower layer, and (2) propagating to the lower layer accordingly. A connection can be made between different proposed LRP propagation rules and the choice of reference point at which the Taylor expansion is performed [11, 10]. This Taylor-based view on the LRP algorithm allows in particular to verify that the corresponding reference points are meaningful, for example, that they satisfy domain membership constraints. This interpretation of LRP as a collection of Taylor expansions is referred to as "deep Taylor decomposition" [11].

The LRP algorithm has been successfully applied to various data types and problems, ranging from computer vision and natural language processing tasks such as classification of concepts in images [3], age prediction [8] or categorization of text documents [2], over reinforcement learning tasks such as playing computer games [9], to various medical data analysis tasks, e.g., decoding of fMRI signals [14] or therapy outcome prediction [15]. In these diverse applications, LRP explanations provide additional insights into the decision strategies used by the model, which not only help to better understand the data, including its biases and artifacts [8, 9], but also help to analyze the learning processes and model's decision strategies [9].

In the second part of the talk, two recent advances that broaden the usefulness of explanation methods are discussed. First, Spectral Relevance Analysis (SpRAy) [9], a dataset-wide analysis of individual explanations that summarizes the overall decision structure of the model into a finite and easily interpretable set of prototypical decision strategies. This analysis allows to systematically investigate complex models on large datasets. It has unveiled in commonly used datasets, artifacts, that tend to systematically induce flaws into the decision structure of ML models trained on them. For example, a website logo was found in some images of the class 'truck' of the ImageNet dataset, which the state-of-the-art VGG-16 neural network would then use for its predictions [1].

Another advance brings successful explanation techniques to non-neural network architectures such as kernel-based models. The approach that we term 'neuralization' [6] finds for these non-neural network architectures a functionally equivalent neural network so that state-of-the-art explanation techniques such as LRP can be applied. The approach was successfully applied to various unsupervised models, in particular, kernel one-class SVMs [7] and various k-means clustering models [6], thereby shedding light into what input features make a data point anomalous or member of a given cluster.

Although significant progress has been made to improve the transparency of ML models such as deep neural networks, numerous challenges still need to be addressed both on the methods and theory side. In particular, there is a need for standardized and unbiased evaluation benchmarks for assessing the quality and usefulness of an explanation. Furthermore, an important future work will be to adopt a more holistic view on the problem of explanation, that considers how to make best use of the user's interpretation and feedback capabilities, and that also integrates the end goal of the explanation method, for example, achieving better and more informed decisions, or systematically improving and robustifying a machine learning model.

References

[1] C. J. Anders, T. Marinc, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, *Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed* arXiv:1912.11425 (2019).

[2] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, *"What is Relevant in a Text Document?": An Interpretable Machine Learning Approach*, PLOS ONE, **12**(8):e0181142 (2017).

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, *On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation*, PLOS ONE, **10**(7):e0130140 (2015).

[4] D. Balduzzi, M. Frean, L. Leary, J.P. Lewis, K. Wan-Duo Ma, B. McWilliams, *The Shattered Gradients Problem: If resnets are the answer, then what is the question?* ICML (2017), 342–350.

[5] Y. Ding, Y. Liu, H. Luan, M. Sun, *Visualizing and Understanding Neural Machine Translation*, ACL **1** (2017), 1150–1159.

[6] J. Kauffmann, M. Esders, G. Montavon, W. Samek, K.-R. Müller, *From Clustering to Cluster Explanations via Neural Networks*, arXiv:1906.07633 (2019).

[7] J. Kauffmann, K.-R. Müller, G. Montavon, *Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models*, Pattern Recognition, 107198 (2020).

[8] S. Lapuschkin, A. Binder, K.-R. Müller, W. Samek, *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, IEEE International Conference on Computer Vision Workshops (2019), 1629–1638.

[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, *Unmasking Clever Hans Predictors and Assessing What Machines Really Learn*, Nature Communications **10**:1096 (2019).

[10] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller. *Layer-Wise Relevance Propagation: An Overview*, in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS 11700, (2019).

[11] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, *Explaining NonLinear Classification Decisions with Deep Taylor Decomposition*, Pattern Recognition **65** (2017), 211–222.

[12] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, KDD (2016), 1135–1144.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*, ICLR (2014).

[14] A. Thomas, H. Heekeren, K.-R. Müller, W. Samek, *Analyzing Neuroimaging Data Through Recurrent Deep Learning Models*, Frontiers in Neuroscience, **13**:1321 (2019).

[15] Y. Yang, V. Tresp, M. Wunderle, P. Fasching, *Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks*, ICHI (2018), 152-162.
[16] M. Zeiler, R. Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV **1** (2014), 818–833.

A feature sparse neural network

Robert Tibshirani
(joint work with Ismael Lemhadri, Feng Ruan)

We introduce LassoNet, a neural network model with global feature selection [1]. The model uses a residual connection to learn a subset of the most informative input features. Specifically, the model honors a hierarchy restriction that an input neuron only be included if its linear variable is important. This produces a path of feature-sparse models in close analogy with the lasso for linear regression, while effectively capturing complex nonlinear dependencies in the data. Using a single residual block, our iterative algorithm yields an efficient proximal map which accurately selects the most salient features. On systematic experiments, LassoNet achieves competitive performance using a much smaller number of input features. LassoNet can be implemented by adding just a few lines of code to a standard neural network.

We also apply LassoNet and the linear model lasso to convolution features from mammograms to classify cancer images from normal images. We find that the new methods are nearly as accurate as the state-of–the-art using ResNet, and the results are easier to interpret

References

[1] Ismael Lemhadri, Feng Ruan, Robert Tibshirani *LassoNet: Neural networks with Feature Sparsity*, arXiv:1907.12207

Distributed Machine Learning over Networks

Francis Bach
(joint work with Hadrien Hendrikx Sébastien Bubeck, Laurent Massoulié, Yin-Tat Lee)

The success of machine learning models is in part due to their capacity to train on large amounts of data. Distributed systems are the common way to process more data than one computer can store, but they can also be used to increase the pace at which models are trained by splitting the work among many computing nodes. In this work, we study the corresponding problem of minimizing a sum of functions which are respectively accessible by separate nodes in a network. New centralized and decentralized algorithms are analyzed, together with their convergence guarantees in deterministic and stochastic convex settings, leading to optimal algorithms for this particular class of distributed optimization problems.

For generic problems [1], we determine the optimal convergence rates for strongly convex and smooth distributed optimization in two settings: centralized and decentralized communications over a network. For centralized (i.e., *master/slave*) algorithms, we show that distributing Nesterov's accelerated gradient descent is optimal and achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_g}(1 + \Delta\tau)\ln(1/\varepsilon))$, where κ_g is the condition number of the (global) function to optimize, Δ is the diameter of the network, and τ (resp. 1) is the time needed to communicate values between two neighbors (resp. perform local computations). For decentralized algorithms based on gossip, we provide the first optimal algorithm, called the *multi-step dual accelerated* (MSDA) method, that achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_l}(1 + \frac{\tau}{\sqrt{\gamma}})\ln(1/\varepsilon))$, where κ_l is the condition number of the local functions and γ is the (normalized) eigengap of the gossip matrix used for communication between nodes.

Within machine learning, for smooth and strongly convex problems, existing decentralized algorithms [1] are slower than modern accelerated variance-reduced stochastic algorithms when run on a single machine, and are therefore not efficient. Centralized algorithms are fast, but their scaling is limited by global aggregation steps that result in communication bottlenecks. In [2], we propose an efficient **A**ccelerated **D**ecentralized stochastic algorithm for **F**inite **S**ums named ADFS, which uses local stochastic proximal updates and randomized pairwise communications between nodes. On n machines, ADFS learns from nm samples in the same time it takes optimal algorithms to learn from m samples on one machine. This scaling holds until a critical network size is reached, which depends on communication delays, on the number of samples m, and on the network topology. We provide a theoretical analysis based on a novel augmented graph approach combined with a precise evaluation of synchronization times and an extension of the accelerated proximal coordinate gradient algorithm to arbitrary sampling. We illustrate the improvement of ADFS over state-of-the-art decentralized approaches with experiments.

References

[1] K. Scaman, F. Bach, S. Bubeck, Y.-T. Lee, L. Massoulié. *Optimal algorithms for smooth and strongly convex distributed optimization in networks.* Proceedings of the International Conference on Machine Learning (ICML), 2017.

[2] H. Hendrikx, F. Bach, L. Massoulié. *An accelerated decentralized stochastic proximal algorithm for finite Sums.* Advances in Neural Information Processing Systems (NeurIPS), 2019.

[3] H. Hendrikx, L. Massoulié, F. Bach. *Accelerated Decentralized Optimization with Local Updates for Smooth and Strongly Convex Objectives.* Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

Neural networks in the overparametrized regime

Johannes Schmidt-Hieber

Some statistical effects observed in modern machine learning seem to contradict classical statistical theory. The most disturbing empirical finding is that overparametrization tends to generalize well in practice.

The classical statistical theory states that there is a so called bias-variance trade-off. Methods with few parameters are not flexible enough to capture the underlying structure in the data. Increasing the number of parameters in the method reduces the statistical risk, but only up to a certain point after which we enter the overparametrized regime where the method starts to explain the randomness in the data. Overfitting adds a lot of stochastic variability to the estimator leading to poor statistical performance. Overfitting should consequently be avoided in practice.

For fixed sample size, plotting the number of parameters (for instance in a neural network) versus the stochastic risk, classical statistical theory predicts that one should see a U-shaped curve. If done in practice, however, the U-shape explains the behavior for the number of parameters being at most the order of the sample size. For machine learning methods with extreme overparametrization, the statistical risk typically becomes again small. Because of this phenomenon, for neural networks it is now common to train them to have zero training loss. This is somehow the most extreme form of overparametrization as it essentially means interpolation of all points in the dataset. The reconstruction thus also perfectly fits all randomness in the data.

The common explanation why overfitting can work is that the gradient descent type methods employed to fit the parameters in machine learning converge to an interpolant satisfying some minimum norm constraint. This phenomenon is also known as implicit regularization. Based on the earlier work by [3], we show that in a simplified neural network model in which only the parameters in the output layer are learned by stochastic gradient descent and all other parameters are fixed beforehand, the minimum norm is a Sobolev norm, see [2] for a precise statement. The neural network reconstruction converges moreover to the natural cubic spline interpolant in the overparametrized regime. If the data have been generated from the nonparametric regression model, the neural network reconstruction can then be shown to be far away from the regression function, proving that at least in this setup implicit regularization is insufficient to do denoising.

General claims that have been made in the machine learning literature, such as the claim that overfitting works well, should therefore be taken with a grain of salt. Despite being true on standardized datasets that are commonly used to test machine learning methods, these claims can be misleading for different types of data structures, such as for instance extremely noisy data. It is the combination of the statistical model and the method that ultimately determines the statistical properties.

References

[1] J. Schmidt-Hieber, *Nonparametric regression using deep neural networks with ReLU activation function*, Annals of Statistics, *to appear*.

[2] J. Schmidt-Hieber, *Rejoinder to discussions of "Nonparametric regression using deep neural networks with ReLU activation function"*, Annals of Statistics, *to appear*.

[3] T. Strohmer and R. Vershynin *A randomized Kaczmarz algorithm with exponential convergence*, The Journal of Fourier Analysis and Applications **15** (2009), 262–278.

Causal Inference for Genetics with the Digital Twin Test

Emmanuel Candès

(joint work with Stephen Bates, Matteo Sesia and Chiara Sabatti)

This work introduces a method to rigorously draw causal inferences—inferences immune to all possible confounding—from genetic data that include parents and offspring. Causal conclusions are possible with these data because the natural randomness in meiosis can be viewed as a high-dimensional randomized experiment. We make this observation actionable by developing a novel conditional independence test that identifies regions of the genome containing distinct causal variants. The proposed *Digital Twin Test* compares an observed offspring to carefully constructed synthetic offspring from the same parents in order to determine statistical significance, and it can leverage any black-box multivariate model and additional non-trio genetic data in order to increase power. Crucially, our inferences are based only on a well-established mathematical description of the rearrangement of genetic material during meiosis and make no assumptions about the relationship between the genotypes and phenotypes.

Fundamental limits of deep neural network learning

Helmut Bölcskei

(joint work with D. Elbrächter, P. Grohs, G. Kutyniok, and D. Perekrestenko)

Deep neural networks have become state-of-the-art technology for a wide range of practical machine learning tasks such as image classification, handwritten digit recognition, speech recognition, or game intelligence. This talk develops the fundamental limits of learning in deep neural networks by characterizing what is possible if no constraints on the learning algorithm and the amount of training data are imposed. Concretely, we consider information-theoretically optimal approximation through deep neural networks with the guiding theme being a relation between the complexity of the function (class) to be approximated and the complexity of the approximating network in terms of connectivity and memory requirements for storing the network topology and the associated quantized weights. The theory we develop educes remarkable universality properties of deep networks. Specifically, deep networks are optimal approximants for vastly different function classes such as affine systems and Gabor systems. Affine systems are generated by the affine group (scalings and translations) whereas Gabor systems are generated

by the Weyl-Heisenberg group (time-shifts and modulations). This universality is afforded by a concurrent invariance property of deep networks to time-shifts, scalings, and frequency-shifts. In addition, deep networks provide exponential approximation accuracy—i.e., the approximation error decays exponentially in the number of non-zero weights in the network—of vastly different functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, general smooth functions, and even one-dimensional oscillatory textures and fractal functions such as the Weierstrass function, both of which do not have any known methods achieving exponential approximation accuracy. In summary, deep neural networks provide information-theoretically optimal approximation of a very wide range of functions and function classes used in mathematical signal processing. We also show that in the approximation of sufficiently smooth functions finite-width deep networks require strictly smaller connectivity than finite-depth wide networks.

Testing validity of statistical inference in approximate models

Natalia Bochkina

Detecting model misspecification historically has been important in statistics, with the tests known as goodness-of-fit tests. Typical tests that don't involve parameter estimation are Kolmogorov-Smirnov and Anderson-Darling tests. A universal test that takes into the account parameter estimation is the χ^2 test which applies to discrete data, and involves a discretisation of a continuous distribution. This test applies usually to one-dimensional data.

Current demand on statistical modelling for modern data usually involves high dimensional distributions and estimating a large number of parameters. Also, constraints on computing imply that approximate models are often used in challenging high dimensional setting.

So, to address these practical problems, the problem we consider here is wider than the standard goodness of fit tests: we are interested whether the asymptotic Gaussian inference is appropriate here, rather than whether the model is correct. In particular, we want to identify the models that "may be wrong but are useful", i.e. the models that produce appropriate asymptotic inference. Note that the proposed method applies to both classical (frequentist) and Bayesian, as they are asymptotically equivalent for well-specified regular models. So informally, we can formulate the hypothesis we will be testing as

H_0: asymptotic inference is correct vs H_1: asymptotic inference is not correct.

For the inference problems where the asymptotic Gaussian inference is not appropriate, e.g. if n is small or where the model is non-regular (e.g. densities with jump, true parameter is on the boundary of the parameter space), this test should not be used. We will also discuss application of this method to models that are not fully identifiable, e.g. inverse problems.

[2] has proposed a test for model misspecification based on comparing the information and curvature of the model which he named the information matrix test which is based on verifying the second Bartlett identity. His test applies in small dimensions and is based on using the third derivatives of the log likelihood (which do vanish for some models, as explained in [2] as well as a subset of entries of these matrices. For linear models with a parameter of fixed dimension, this was done by [3], with a particular choice of a test statistic specific to the linear model.

We propose a generic test that can apply to any regular model, in the sense of Wald and Cramer-Rao. Some versions of it have been appearing as ad hoc tests, e.g. testing the equality of the mean and the variance for the Poisson distribution. As compared to the information test of [2], we propose an alternative test statistic based on a distance between two Gaussian distributions that does not involve any additional differentiation or an ad hoc selection of elements of the matrices to be compared, it is a general method that applies for any regular model, and it applies to high dimensional problems.

[1] has a nice discussion of geometry of asymptotic statistical inference where the parametric set of probability models $\{P_\theta, \theta \in \Theta\}$ is viewed as a Riemannian manifold, with matrix D being a curvature which defines geodesics on the manifold. For a well-specified model, curvature equals to the Fisher information whereas for a misspecified model this is not always the case. This motivates our name for the test, internal coherence test.

This test is applied to members of exponential family, linear and generalised linear models, inverse problems and Variational Bayes approaches.

References

[1] Robert Kass, *The Geometry of Asymptotic Inference*, Statist. Sci. **4** (1989), p. 188–219.
[2] Halbert White, *Maximum Likelihood Estimation of Misspecified Models*, Econometrica **50** (1982), p. 1–25.
[3] Qian M. Zhou, Peter X.-K. Song and Mary E. Thompson, *Information Ratio Test for Model Misspecification in Quasi-Likelihood Inference*, Journal of the American Statistical Association, **107** (2012), p. 205–213

Bayesian inference for nonlinear inverse problems

Vladimir Spokoiny

1. Introduction

Bayesian inference for inverse problems attracted a lot of attention in the recent literature. We mention only few relevant papers. [Knapik et al., 2011] studied minimax contraction rate for linear inverse problems, [Knapik et al., 2016] discussed adaptive Bayes procedures. [Nickl, 2017] studied the BvM for Schrödinger equation, [Nickl and Söhl, 2019] focused on statistical inverse problems for compound Poisson processes, [Monard et al., 2017] discussed applications to X-Ray

Tomography, [Nickl and Söhl, 2017] studied posterior contraction rates for discretely observed scalar diffusions, [Gugushvili et al., 2018] considered Bayesian inverse problems with partial observations, [Trabs, 2018] discussed a linear inverse problem with an unknown operator, [Lu, 2017] established BvM results for a rather general elliptic inverse problem with an additive noise. Nonlinearity of the model makes the study very involved and the cited results heavily used the recent advances in the theory of partial differential equations, inverse problems, empirical processes. We mention [Nickl, 2017] and [Nickl et al., 2018] as particular illustration of the major difficulties in the study of concentration of the penalized MLE and of posterior concentration.

The main contribution of this paper is a novel approach allowing a unified study of a large class of nonlinear inverse problems. The approach is based on a double relaxation by introducing an auxiliary functional parameter, replacing the structural equation with a penalty, and imposing an additional prior on the auxiliary parameter. This leads to a new model with an extended parameter set but the stochastic term is linear w.r.t. the total parameter set. This fact helps to obtain sharp finite sample bounds for concentration of the penalized maximum likelihood estimator (pMLE) around its population counterpart and for posterior concentration around pMLE. Also we establish a finite sample result about Gaussian approximation of the posterior with an explicit error term in the total variation distance and for the class of centrally symmetric sets around pMLE. All the bounds are given in term of *effective dimension* in place of the total parameter dimension. This helps to compensate the increase of the parameter set and to get the right accuracy of approximation. The approach is "coordinate free" and does not rely on any spectral decomposition and/or any basis representation for the target parameter and penalty term. We focus here on the problem of inverting an known nonlinear smooth operator from noisy discrete data $\boldsymbol{Y}$ following the equation

$$\boldsymbol{Y} \,=\, A(\boldsymbol{f}) + \sigma\boldsymbol{\varepsilon}. \tag{1}$$

A forthcoming paper explains how the proposed approach called "calming" can be extended to many other models including generalized regression, nonparametric diffusion, Bayesian deconvolution, dimension reduction etc.

Now we explain the idea of the method. For the original problem (1), a prior density $\Pi(\boldsymbol{f})$ on the target parameter $\boldsymbol{f}$ yields the posterior

$$\boldsymbol{f} \,\big|\, \boldsymbol{Y} \,\propto\, \exp\bigl\{-\|\boldsymbol{Y} - A(\boldsymbol{f})\|^2/(2\sigma^2)\bigr\}\Pi(\boldsymbol{f}).$$

Now denote by $\boldsymbol{g}$ the image function, $\boldsymbol{g} = A(\boldsymbol{f}) \in \mathcal{Y}^d$ and relax the structural equation $\boldsymbol{g} = A(\boldsymbol{f})$ replacing it with a penalty $\lambda\|\boldsymbol{g} - A(\boldsymbol{f})\|^2/2$. The image function $\boldsymbol{g}$ is modelled using a separate prior. The proposed approach leads to the extended parameter set $(\boldsymbol{f}, \boldsymbol{g})$ which is modelled as

$$(\boldsymbol{f}, \boldsymbol{g}) \,\propto\, \exp\bigl\{-\|\boldsymbol{Y} - \boldsymbol{g}\|^2/(2\sigma^2) - \lambda\|\boldsymbol{g} - A(\boldsymbol{f})\|^2/2\bigr\}\Pi(\boldsymbol{f})\Pi(\boldsymbol{g}).$$

Such a decoupling increases substantially the parameter space. However, by a proper choice of a $\boldsymbol{g}$-prior one can keep the effective dimension of the same order as for the original problem. One can treat the calming approach as a kind of transformation of the original nonlinear problem to a linear one with an extended parameter set and a special prior that includes the structural penalty term. Our theoretical results justify the proposed method and state a number of remarkable features of the total and marginal posteriors.

Theorem 1. *It holds on* Ω_n *with* $\mathbb{P}\big(\Omega_n\big) \geq 1 - 1/n$

$$\sup_{A \in \mathcal{B}(\mathcal{X}^d)} \Big| \mathbb{P}\big(\boldsymbol{f}_G - \tilde{\boldsymbol{f}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\tilde{\mathbb{D}}_G^{-1}\boldsymbol{\gamma} \in A\big)\Big| \leq \mathtt{C}\big\{\delta_{3,n} + 1/n\big\}$$

where $\delta_{3,n}$ *is an explicit error term of order* $n^{-1/2}$.

References

[Gugushvili et al., 2018] Gugushvili, S., van der Vaart, A. W., and Yan, D. (2018). Bayesian inverse problems with partial observations. *Trans. A. Razmadze Math. Inst.*, 172(3, part A):388–403.

[Knapik et al., 2016] Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields*, 164(3-4):771–813.

[Knapik et al., 2011] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.

[Lu, 2017] Lu, Y. (2017). On the bernstein – von mises theorem for high dimensional nonlinear bayesian inverse problems.

[Monard et al., 2017] Monard, F., Nickl, R., and Paternain, G. P. (2017). Efficient Nonparametric Bayesian Inference For X-Ray Transforms. *arXiv e-prints*, page arXiv:1708.06332.

[Nickl, 2017] Nickl, R. (2017). Bernstein - von Mises theorems for statistical inverse problems I: Schrödinger equation. *arXiv e-prints*, page arXiv:1707.01764.

[Nickl and Söhl, 2017] Nickl, R. and Söhl, J. (2017). Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, 45(4):1664–1693.

[Nickl and Söhl, 2019] Nickl, R. and Söhl, J. (2019). Bernstein – von Mises theorems for statistical inverse problems II: compound Poisson processes. *Electronic Journal of Statistics*, 13(2):3513–3571.

[Nickl and Szabó, 2016] Nickl, R. and Szabó, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications*, 126(12):3913 – 3934. In Memoriam: Evarist Gine.

[Nickl et al., 2018] Nickl, R., van de Geer, S. A., and Wang, S. (2018). Convergence rates for Penalised Least Squares Estimators in PDE-constrained regression problems.

[Trabs, 2018] Trabs, M. (2018). Bayesian inverse problems with unknown operators. *Inverse Problems*, 34(8):085001.

Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems

Ryan J. Tibshirani

This talks serves as a postscript of sorts to [6, 7], who developed continuous-time formulations and properties of *trend filtering*, a discrete-time smoothing tool proposed (independently) by [5, 1]. The central object of study is the *falling factorial basis*, as it was called by [6, 7]. Its span turns out to be a space of piecewise polynomials that has a classical place in spline theory, called *discrete splines* [3, 4]. At the time of [6, 7], we were not fully aware of these connections. The current talk reflects ongoing work, which attempts to rectify this by making these connections explicit, reviewing (and making use of) some of the important existing work on discrete splines, and contributing several new perspectives and new results on discrete splines along the way.

To fix ideas, given an integer $k \geq 0$ and input points $x_1 < z \cdots < x_n$, the kth order falling factorial basis h_i^k, $i = 1, \ldots, n$ is defined as

$$
\begin{aligned}
h_j^k(x) &= \frac{1}{(j-1)!} \prod_{\ell=1}^{j-1} (x - x_\ell), \quad j = 1, \ldots, k+1, \\
h_j^k(x) &= \frac{1}{k!} \prod_{\ell=j-k}^{j-1} (x - x_\ell) \cdot 1\{x > x_{j-1}\}, \quad j = k+2, \ldots, n.
\end{aligned} \tag{1}
$$

(This basis is given its name because when the input points are unit-spaced integers, i.e., $x_i = i$, $i = 1, \ldots, n$, the basis functions evaluate to falling factorials.) For $k = 0$ or $k = 1$, this reduces to the well-known truncated power basis; but for $k \geq 2$, these functions have discontinuous derivatives and therefore are *not splines*, but some other "spline-like" piecewise polynomials of degree k.

In [6], and in follow-up work [7], we established several key properties of falling factorial functions, and functions in their span $H_n^k = \text{span}\{h_1^k, \ldots, h_n^k\}$, i.e., of the form

$$
f = \sum_{i=1}^n \alpha_i h_i^k.
$$

The foremost property is that, for any $f \in H_n^k$, the total variation of its kth derivative is *exactly equal to its discrete kth order total variation*, i.e.,

$$
\text{TV}(f^{(k)}) = \sum_{i=1}^n \left| (\Delta^{k+1} f)(x_i) \right| \cdot \frac{x_{i+k+1} - x_i}{k+1}, \tag{2}
$$

where Δ^{k+1} is a discrete derivative operator based on divided differences of order $k+1$. This bridges the gap between the continous-time locally adaptive regression spline variational problem posed by [2] and the discrete-time problem solved by trend filtering. It also allows us to interpolate (extrapolate) the fitted values from trend filtering into a fitted function that is *exactly as smooth* in continuous-time as the fitted values are in discrete-time.

It turns out that there is much more to the story than this. The span H_n^k of the kth degree falling factorial basis functions is in fact quite a special space of piecewise polynomials: a space of kth degree discrete splines. Speaking informally, a discrete spline is a piecewise polynomial that exhibits continuity in all its lower-order discrete derivatives (rather than classical derivatives) at its knot points. Discrete splines have been studied since the early 1970s by applied mathematicians, but have nowhere near the visibility nor popularity of splines. Investigating the connection between falling factorial functions and discrete splines has led us to develop numerous interesting properties underlying these functions, particularly with respect to interpolation and representation, which we describe next.

- We have developed a new perspective on how to construct the falling factorial basis (1) "from scratch". This starts by defining a discrete derivative operator (in continuous-time, i.e., from functions to functions) and its inverse, a discrete integrator; we then show that the falling factorial basis functions are given by kth order discrete integration of appropriate step functions. The importance of this construction is two-fold: first, it reveals an even stronger property of discrete splines than the total variation equality (2): for kth degree discrete splines, their kth discrete derivative matches their kth derivative everywhere, and furthermore, they are the *only* kth degree piecewise polynomials with this property. Second, it suggests an avenue for how to construct multivariate discrete splines (see the last bullet point).
- We have identified a natural dual basis to the falling factorial basis that is based on evaluations of discrete derivatives. As a primary use case, this dual basis allows us to efficiently interpolate within the space of discrete splines, which generalizes Newton's divided difference interpolation formula. More importantly, it turns out that this interpolation formula can be recast in an implicit form, showing that interpolation using discrete splines can be done in *constant-time*. In other words, discrete spline interpolation can be made entirely local and even more efficient than spline interpolation.
- We are working on further representation properties of discrete splines. The fact that their kth discrete derivatives matches their kth derivatives everywhere directly implies that their continuous-time kth degree total variation matches their discrete-time kth degree total variation (2). But preliminary calculations suggests that there are likely other important representational results (i.e., the ability to represent a continuous-time smoothness functional exactly in of simple discrete-time quantities), e.g., for classical Sobolev-type smoothness functions, yet to be worked out. This could allow us to discretize a variety of variational problems or differential equations in a fundamentally new and potentially more efficient way.
- Lastly, an open direction is to construct discrete splines over triangulations. The hope is to follow the construction of univariate discrete splines via iterated discrete integration, and leverage this in appropriate sense.

We note that, in a multivariate setting, extensions of the local interpolation and representation properties of univariate discrete splines would be particularly important (as both of these are especially expensive in multiple dimensions), and would thus mark a particularly big success of the discrete spline view.

References

[1] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. *ℓ_1 trend filtering* SIAM Review **51** (2009) 339–360.

[2] Enno Mammen and Sara van de Geer. *Locally apadtive regression splines* Annals of Statistics **25** (1997) 387–413.

[3] Olvi L. Mangasarian and Larry L. Schumaker. *Discrete splines via mathematical programming* SIAM Journal on Control **9** (1971) 174–183.

[4] Olvi L. Mangasarian and Larry L. Schumaker. *Best summation formulae and discrete splines* SIAM Journal on Numerical Analysis **10** (1973) 448–459.

[5] Gabriel Steidl, Stephan Didas, and Julia Neumann. *Splines in higher order TV regularization* International Journal of Computer Vision **70** (2006) 214–255.

[6] Ryan J. Tibshirani. *Adaptive piecewise polynomial estimation via trend filtering* Annals of Statistics **42** (2014) 285–323.

[7] Yu-Xiang Wang, Alexander Smola, and Ryan J. Tibshirani. *The falling factorial basis and its statistical applications* International Conference on Machine Learning **31** (2014).

Profile least squares estimators in the monotone single index model

Piet Groeneboom

(joint work with Fadoua Balabdaoui)

The monotone single index model tries to predict a response from the linear combination of a finite number of parameters and a function linking this linear combination to the response via a monotone *link function* ψ_0 which is unknown. So, more formally, we have the model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) + \epsilon,$$

where Y is a one-dimensional random variable, $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional random vector with distribution function G, ψ_0 is monotone and ϵ is a one-dimensional random variable such that $\mathbb{E}[\epsilon|\boldsymbol{X}] = 0$ G-almost surely. For identifiability, the regression parameter $\boldsymbol{\alpha}_0$ is a vector of norm $\|\boldsymbol{\alpha}_0\| = 1$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, so $\boldsymbol{\alpha}_0 \in \mathcal{S}_{d-1}$, the boundary of the unit sphere in $\mathbb{R}^d$.

The ordinary profile least squares estimate of $\boldsymbol{\alpha}_0$ is an M-estimate in two senses: for fixed $\boldsymbol{\alpha}$ the least squares criterion

$$\psi \mapsto n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \psi(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\}^2 \tag{1}$$

is minimized for all monotone functions ψ (either decreasing or increasing) which gives an $\boldsymbol{\alpha}$ dependent function $\hat{\psi}_{n,\boldsymbol{\alpha}}$, and the function

$$\boldsymbol{\alpha} \mapsto n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\}^2 \tag{2}$$

is then minimized over $\boldsymbol{\alpha}$. This gives a profile least squares estimator $\hat{\boldsymbol{\alpha}}_n$ of $\boldsymbol{\alpha}_0$, which we will call the LSE.

Although this estimate of $\boldsymbol{\alpha}_0$ has been known now for a very long time (more than 30 years probably), it is not known whether it is $\sqrt{n}$ convergent (under appropriate regularity conditions), let alone that we know its asymptotic distribution. Also, simulation studies are rather inconclusive. Other simulation studies, presented in [1], are also inconclusive. In that paper, it was also proved that an ordinary least squares estimator (which ignores that the link function could be non-linear) is $\sqrt{n}$-convergent and asymptotically normal under elliptic symmetry of the distribution of the covariate $\boldsymbol{X}$. Another linear least squares estimator of this type, where the restriction on $\boldsymbol{\alpha}$ is $\boldsymbol{\alpha}^T \boldsymbol{S}_n \boldsymbol{\alpha} = 1$, where S_n is the usual estimate of the covariance matrix of the covariates, and where a renormalization at the end is not needed (as it is in the just mentioned linear least squares estimator) was studied in [2] and there shown to have similar behavior. If this suggests that the profile LSE should also be $\sqrt{n}$-consistent, the extended simulation study in [2] shows that it is possible to find other estimates which exhibit a better performance in these circumstances.

An alternative way to estimate the regression vector is to minimize the criterion

$$\boldsymbol{\alpha} \mapsto \left\| n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\} \boldsymbol{X}_i \right\|^2 \tag{3}$$

over $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$, where $\| \cdot \|$ is the Euclidean norm. Note that this is the sum of d squares. We prove that this minimization procedure leads to a $\sqrt{n}$ consistent and asymptotically normal estimator, which is a more precise and informative result compared to what we know now about the LSE.. Using the well-known properties of isotonic estimators, it is easily seen that the function (3) is piecewise constant as a function of $\boldsymbol{\alpha}$, with finitely many values, so the minimum exists and is equal to the infimum over $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$. Notice that this estimator does not use any tuning parameters, just like the LSE.

In [2], a similar Simple Score Estimator (SSE) $\hat{\boldsymbol{\alpha}}_n$ was defined as a point $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$ where all components of the function

$$\boldsymbol{\alpha} \mapsto n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\} \boldsymbol{X}_i$$

cross zero. If the criterion function were continuous in $\boldsymbol{\alpha}$, this estimator would have been the same as the least squares estimator, minimizing (3), with a minimum equal to zero, but in the present case we cannot assume this because of the discontinuities of the criterion function.

The definition of an estimator as a crossing of the d-dimensional vector $\mathbf{0}$ makes it necessary to prove the existence of such an estimator, which we found to be a rather non-trivial task. Defining our estimator directly as the minimizer of (3), so as a least squares estimator, relieves us from the duty to prove its existence. Since our estimator is asymptotically equivalent to the SSE, we refer to it here under the same name.

A fundamental function in our treatment is the function $\psi_{\boldsymbol{\alpha}}$, defined as follows.

Definition 1. Let $\mathcal{S}_{d-1}$ denote again the boundary of the unit ball in $\mathbb{R}^d$. Then, for each $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$, the function $\psi_{\boldsymbol{\alpha}} : \mathbb{R} \to \mathbb{R}$ is defined as the nondecreasing function which minimizes

$$\psi \mapsto \mathbb{E}\{Y - \psi(\boldsymbol{\alpha}^T \boldsymbol{X})\}^2$$

over all nondecreasing functions $\psi : \mathbb{R} \to \mathbb{R}$. The existence and uniqueness of the function $\psi_{\boldsymbol{\alpha}}$ follows for example from the results in [5].

The importance of the function $\psi_{\boldsymbol{\alpha}}$ arises from the fact that we can differentiate this function w.r.t. $\boldsymbol{\alpha}$, in contrast with the least squares estimate $\hat{\psi}_{n,\boldsymbol{\alpha}}$, and that $\psi_{\boldsymbol{\alpha}}$ represents the least squares estimate of ψ_0 in the underlying model for fixed $\boldsymbol{\alpha}$, if we use $\boldsymbol{\alpha}^T \boldsymbol{x}$ as the argument of the monotone link function.

It is also possible to introduce a tuning parameter (bandwidth) h and use a kernel estimate $\tilde{\psi}'_{n,h,\boldsymbol{\alpha}}(u) = \frac{1}{h} \int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n,\boldsymbol{\alpha}}(x)$ for $\frac{d}{du}\psi_{\boldsymbol{\alpha}}(u)\big|_{u=\boldsymbol{\alpha}^T \boldsymbol{X}}$, where we minimize

$$\boldsymbol{\alpha} \mapsto \left\| n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\} \boldsymbol{X}_i \, \tilde{\psi}'_{n,h,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\|^2 \tag{4}$$

instead of (3). All estimators are computed by an augmented Lagrange method, embedded in the Hooke-Jeeves pattern search method, which avoids reparametrization. Details of the method are given and the implementation in `R`, using `Rcpp`, can be found in [3].

We note that this method is, for several reasons, rather different from the heuristic Lagrange method, suggested in Section 4.2 of [2]. The method in Section 4.2 of [2] was still based on the "crossing of zero" definition instead of the least squares definition of the estimators above and in fact tried to eliminate the Lagrange parameter. The result of that procedure could not ascertain that the solution $\hat{\boldsymbol{\alpha}}_n$ had norm 1, and a renormalization at the end was needed to enforce this constraint, which has a somewhat unpredictable influence on the convergence of the algorithm. The augmented Lagrange method, on the other hand, has *two* penalty terms, a linear and a quadratic one, and does not eliminate the two Lagrange parameters. In this case we may assume that the solution, provided by the method, has indeed norm 1 in the number of decimals, set by the procedure.

We finally give simulation results for these different methods, where we make a comparison with the results of the Effective Dimension Reduction (EDR) method, proposed in [4] and implemented in the `R` package `edr`. The results show that the profile least squares estimators have a much better performance for the present simulation models, which were also used in [1].

References

[1] Fadoua Balabdaoui, Cécile Durot, and Hanna Jankowski. *Least squares estimation in the monotone single index model* Bernoulli **25** (2019) 3276–3310

[2] Fadoua Balabdaoui, Piet Groeneboom, and Kim Hendrickx. *Score estimation in the monotone single-index model* Scand. J. Stat. **46** (2019) 517–544

[3] Piet Groeneboom. *Algorithms for computing estimates in the single index model* https://github.com/pietg/single_index (2018)

[4] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. *Direct estimation of the index coefficient in a single-index model* Ann. Statist. **29** (2001) 595–623

[5] Dieter Landers and L. Rogge. *Isotonic approximation in L_s* J. Approx. Theory **31** (1981) 199–223

The generalization error of overparametrized models: Insights from exact asymptotics

Andrea Montanari

(joint work with Song Mei, Feng Ruan, Youngtak Sohn, Jun Yan)

Modern neural networks are often so complex that they can perfectly fit the data, achieving vanishing error on the training set. Empirical work [5] demonstrated that this is often the case even if the actual labels in the training set are replaced by purely random labels, and if training error is measured using any one of several loss functions: zero-one loss, hinge loss, or square loss. Despite their huge complexity, these models generalize well to unseen data. Finally, the generalization error degrades gracefully with labels noise.

As shown empirically in [1], these phenomena are not specific to neural networks, and instead arise in a variety of statistical models. When the number of parameters increases (for fixed sample size), the test error first decreases, driven by a decrease in approximation error, and then increases, driven by an increase in generalization error. This familiar U-shaped curve is normally taken as indication that the number of parameter should not be 'too large' not to overfit the data. However, it turns out that –as the number of parameters increases further– the test error decreases again and is minimal when the number of parameters is much larger than the number of samples. An illustration of this behavior is given in Figure 1 (from [3]), which is further discussed below.

These observations pose several challenges to statistical theory. Let us outline two of them: (i) In the overparametrized regime, the training error vanishes and the test error remains different from zero: as a consequence, it is difficult to understand the behavior of these models in terms of uniform convergence; (ii) The fact that the error vanishes (even for square loss) implies that good generalization performances are achieved with vanishing or nearly vanishing explicit regularization. This is particularly surprising in the presence of noise.

In order to understand these phenomena, in two recent papers [3, 4] we study the following random features function class

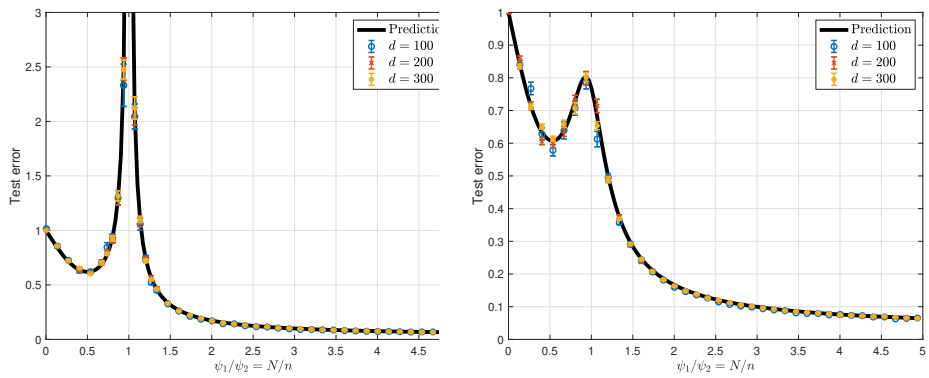

FIGURE 1. Random features ridge regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated via $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle$ (zero noise) with $\|\boldsymbol{\beta}_1\|_2^2 = 1$, and $\psi_2 = n/d = 3$. Left frame: regularization $\lambda = 10^{-8}$ (we didn't set $\lambda = 0$ exactly for numerical stability). Right frame: $\lambda = 10^{-3}$. The continuous black line is our theoretical prediction, and the colored symbols are numerical results for several dimensions d. Symbols are averages over 20 instances and the error bars report the standard error of the means over these 20 instances.

$$\mathcal{F}_{\mathsf{RF}}^N(\boldsymbol{W}) \equiv \left\{ \hat{f}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^N a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \leq N \right\}.$$

The function class is parametrized by $\boldsymbol{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$, and the weights $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N]$ are fixed and random with $\boldsymbol{w}_i \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$ (where $\mathbb{S}^{d-1}(r)$ is the sphere in $\mathbb{R}^d$, with radius r). This model can be regarded as a two-layers neural network with random first-layer weights or as a randomized approximation of the reproducing kernel Hilbert space, with kernel

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}\{\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_2 \rangle)\}$$

(where expectation is taken with respect to $\boldsymbol{w} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$).

We assume that coefficients $\boldsymbol{a}$ are trained using iid data samples $\{(y_i, \boldsymbol{x}_i)\}_{i \leq n}$, whereby $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and we consider two models for the labels y_i:

$$\text{Continuous response [3]:} \quad y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i \,, \quad \varepsilon_i \sim \mathsf{N}(0, \tau^2) \,,$$

$$\text{Binary labels [4]:} \quad y_i \in \{+1, -1\} \,, \quad \mathbb{P}(y_i = +1 | \boldsymbol{x}_i) = g_*(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle) \,.$$

where $f_* : \mathbb{S}^{d-1}(\sqrt{d}) \to \mathbb{R}$, and $g_* : \mathbb{R} \to \mathbb{R}$ satisfy certain technical conditions. In the first setting, we fit the parameters $\boldsymbol{a}$ using ridge regression. In the second, we consider max margin classification. In both cases we consider the proportional asymptotics $N, n, d \to \infty$ with $N/d \to \psi_1$, $n/d \to \psi_2$, and obtain the following results: (i) For ridge regression, we characterize the asymptotics of the test and

training error, under an isotropicity assumption on the nonlinear component of f_*; (ii) For max-margin classification we obtain the same type of results, for a Gaussian proxy of the original model that we refer to as the 'noisy linear activation model.' We conjecture that the latter has the same asymptotics as the original random features model.

Figure 1 compares our analytical predictions for the test error of random features ridge regression with numerical simulations. The agreement is excellent already at $d = 100$, and the analytical curves provide a clear picture of the decrease of test error in the overparametrized regime. We obtain several new insights:

(1) Only the linear component of function f_* is learnt in this regime, while the nonlinear component is treated as noise. This is consistent with the findings of [2].
(2) For fixed sample size-to-dimension ratio ψ_2, the minimum test error is achieved at large overparametrization $\psi_1 \to \infty$.
(3) The optimum value of the regularization parameter (at large overparametrization), depends on the signal-to-noise ratio (SNR). At low SNR the optimum regularization vanishes, while at high SNR it is nonzero.
(4) The latter phenomenon is related to a 'built-in' regularization mechanism. In high dimension the nonlinear component of the activation functions effectively acts as a ridge regularizer.

References

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
[2] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv:1904.12191*, 2019.
[3] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355*, 2019.
[4] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*, 2019.
[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

Estimating functionals under local differential privacy

Angelika Rohde
(joint work with Cristina Butucea, Lukas Steinberger)

Consider n individuals who possess data $X_1, \ldots, X_n$, assumed to be iid from some probability distribution $\mathbb{P} \in \mathcal{P}$. The statistician does not get to see the original data $X_1, \ldots, X_n$, but only a *privatized* version of observations Z. The conditional distribution of Z given $X = (X_1, \ldots, X_n)'$ is denoted by Q and referred to as a

channel distribution or a privatization scheme, i.e. $Pr(Z \in A|X = x) = Q(A|x)$. For $\alpha \in (0, \infty)$, the channel Q is said to provide α-differential privacy if

$$\sup_{A} \sup_{x,x':d_0(x,x')=1} \frac{Pr(Z \in A|X = x)}{Pr(Z \in A|X = x')} \quad \leq \quad e^{\alpha}, \tag{1}$$

where the first supremum runs over all measurable sets and $d_0(x, x')$ denotes the number of distinct entries of x and x'. An α-DP channel Q even provides local privacy, if individual i can generate its privatized data Z_i on its local machine, without ever giving away its original data X_i. We consider two particular classes $\mathcal{Q}_\alpha^{(SI)}$ and $\mathcal{Q}_\alpha^{(NI)}$ of locally α-DP channels, namely sequentially interactive (SI) and non-interactive (NI) channels. A channel $Q : \mathcal{B}(\mathcal{Z}^n) \times \mathcal{X}^n \to [0, 1]$ is called sequentially-interactive, if there exist Markov kernels Q_i, $i = 1, \ldots, n$, such that

$$Q(dz|x) = Q_n(dz_n|x_n, z_1, \ldots, z_{n-1}) \cdots Q_2(dz_2|x_2, z_1)Q_1(dz_1|x_1).$$

The corresponding priviazation scheme is illustrated in the following diagram

$$\begin{array}{ccccccc}
X_1 & & X_2, Z_1 & & & & X_n, Z_1, \ldots, Z_{n-1} \\
Q_1 \downarrow & \searrow & Q_2 \downarrow & \searrow & \cdots\cdots & \searrow & Q_n \downarrow \\
Z_1 & & Z_1, Z_2 & & & & Z_1, \ldots, Z_n.
\end{array}$$

A channel is said to be non-interactive, if there exist Markov kernels Q_i such that

$$Q(dz|x) = \bigotimes_{i=1}^{n} Q_i(dz_i|x_i).$$

Of course, $\mathcal{Q}_\alpha^{(NI)} \subset \mathcal{Q}_\alpha^{(SI)}$. We study the problem of estimating a functional $\theta(\mathbb{P})$ of an unknown probability distribution $\mathbb{P} \in \mathcal{P}$ in which the original iid sample $X_1, \ldots, X_n$ is kept private from the statistician via an α-local differential privacy constraint. Let

$$\mathcal{M}_{n,\alpha}^{(SI)} = \inf_{Q \in \mathcal{Q}_\alpha^{(SI)}} \inf_{\hat{\theta}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{Q\mathbb{P}^n} |\hat{\theta}_n - \theta(\mathbb{P})|$$
$$\mathcal{M}_{n,\alpha}^{(NI)} = \inf_{Q \in \mathcal{Q}_\alpha^{(NI)}} \inf_{\hat{\theta}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{Q\mathbb{P}^n} |\hat{\theta}_n - \theta(\mathbb{P})|$$

and

$$\mathcal{M}_n = \inf_{\hat{\theta}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}^n} |\hat{\theta}_n - \theta(\mathbb{P})|$$

denote the α-SIDP, the α-NIDP and the conventional minimax risk, respectively. We write ω_{TV} and ω_H for the modulus of continuity of the functional θ over $\mathcal{P}$ with respect to total variation and Hellinger distance. Our first result complements the theory developed by [2], who established the characterization

$$\mathcal{M}_n \asymp \omega_H \left(n^{-1/2} \right)$$

under the same conditions on θ and $\mathcal{P}$ as in the following theorem.

Theorem 1 (Rohde and Steinberger, 2019 [3]). *Suppose that $\mathcal{P}$ is convex and dominated, and that $\theta : \mathcal{P} \to \mathbb{R}$ is linear and bounded. Then, for any sequence $\alpha_n \in (0, 1]$,*

$$\left.\begin{array}{l}\mathcal{M}_{n,\alpha_n}^{(SI)}\\ \mathcal{M}_{n,\alpha_n}^{(NI)}\end{array}\right\} \asymp \omega_{TV}\left((n(e^{\alpha_n}-1)^2)^{-1/2}\right), \quad \text{as } n \to \infty.$$

Somewhat surprisingly, the difficulty of estimating a linear functional based on privatized data is characterized by ω_{TV}, whereas, it is characterized by the Hellinger modulus of continuity ω_H if the original data $X_1, \ldots, X_n$ are available. This may, and typically will, lead to different rates of convergence in private and non-private problems. Note that even in cases where we do or can not compute the moduli ω_{TV} and ω_H explicitly, we always have the a priori information that

$$\omega_H(\varepsilon) \leq \omega_{TV}(\varepsilon) \leq \omega_H(\sqrt{2\varepsilon}).$$

This means that the private rate of estimation is never faster than the non-private rate and is never slower than the square root of the non-private rate. Next, we point out that for estimation of linear functionals, sequentially interactive procedures do not improve in terms of rate over non-interactive ones.

We also provide a general construction of α-locally differentially private estimation procedures that is minimax rate optimal if $\mathcal{P}$ is convex and dominated and θ is linear and bounded. The construction relies on a bounded functional parameter $\ell : \mathcal{X} \to \mathbb{R}$. Each individual generates Z_i independently and binary distributed on $\{-z_0, z_0\}$, with

$$Pr(Z_i = z_0 | X_i = x_i) = \frac{1}{2}\left(1 + \frac{\ell(x_i)}{z_0}\right)$$

and $z_0 = \|\ell\|_\infty \frac{e^\alpha+1}{e^\alpha-1}$. The next theorem states that there exists always a non-interactive channel and an estimator which is then given by an affine transformation of the sample mean $\bar{Z}_n = n^{-1}\sum_{i=1}^n Z_i$, which attains the optimal minimax rate of Theorem 1.

Theorem 2 (Rohde and Steinberger, 2019 [3]). *Fix $\alpha \in (0, 1]$ and $n \in \mathbb{N}$. Suppose that $\mathcal{P}$ is convex and dominated, and that $\theta : \mathcal{P} \to \mathbb{R}$ is linear and bounded. Then there exists a bounded function $\ell^* : \mathcal{X} \to \mathbb{R}$ and a number $b \in \mathbb{R}$, such that*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{[Q_1^{(\alpha,\ell^*)}\mathbb{P}]^n}\left[\left|\Pi[\bar{Z}_n + b] - \theta(\mathbb{P})\right|\right] \lesssim \omega_{TV}\left((n(e^\alpha-1)^2)^{-1/2}\right),$$

where

$$\Pi(x) = \begin{cases}\sup_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}), & \sup_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}) < x,\\ x, & \inf_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}) \leq x \leq \sup_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}),\\ \inf_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}), & x < \inf_{\mathbb{P}\in\mathcal{P}}\theta(\mathbb{P}).\end{cases}$$

The same rate of $\mathcal{M}_{n,\alpha}^{(SI)}$ and $\mathcal{M}_{n,\alpha}^{(NI)}$ in case of linear functionals raises the question whether these two minimax risks always coincide. Turning to a quadratic functional, however, we find that using a sequentially interactive privatization

scheme considerably improves over the non-interactive one in terms of minimax-optimal rates of convergence. Let $\mathcal{H}_{\beta,L}$ denote the Hölder ball on $[0,1]$ of radius L to the exponent β.

Theorem 3 (Butucea, Rohde and Steinberger, 2020)**.** *Let $X_1,\dots,X_n$ be iid with density $p \in \mathcal{H}_{\beta,L}([0,1])$, $\beta > 0$. For the functional $\theta(p) = \int_0^1 p^2(x)\,dx$, the α-SIDP minimax risk of is of order*

$$\mathcal{M}_{n,\alpha}^{(SI)} \asymp n^{-\frac{4\beta}{4\beta+2}} + n^{-1/2}, \qquad \text{'elbow' at } \beta = \frac{1}{2}$$

and the α-NIDP minimax risk is of order

$$\mathcal{M}_{n,\alpha}^{(NI)} \asymp n^{-\frac{4\beta}{4\beta+3}} + n^{-1/2}, \qquad \text{'elbow' at } \beta = \frac{3}{4}.$$

Recall from [1] that the non-privatized minimax risk of estimating $\theta(p) = \int_0^1 p^2(x)\,dx$ is of order

$$\mathcal{M}_n \asymp n^{-\frac{4\beta}{4\beta+1}} + n^{-1/2}, \qquad \text{'elbow' at } \beta = \frac{1}{4}.$$

References

[1] Bickel, P. J. and Ritov, Y., *Estimating integrated squared density derivatives: sharp best order of convergence estimates*, Sankhya A **50** (1988), 381–393.

[2] D. Donoho and R. Liu, *Geometrizing rates of convergence, II*, Ann. Statist. **19** (1991), 633–667.

[3] A. Rohde and L. Steinberger (2019). *Geometrizing rates of convergence under local differential privacy constraints*, Ann. Statist., to appear.

Measuring separability in covariance operators of random surfaces

Holger Dette

(joint work with Pramita Bagchi)

A common approach to obtain reasonable estimates of the covariance operator from surface data is the structural assumptions of separability, which has become very popular, for example in the analysis of geostatistical space-time models. Roughly speaking, this assumption allows to write the covariance

$$c(s,t,s',t') = \mathbb{E}[X(s,t)X(s',t')]$$

of a (real valued) space-time process $\{X(s,t)\}_{(s,t)\in S\times T}$ as a product of the space and time covariance function, that is

$$c(s,t,s',t') = c_1(s,s')c_2(t,t').$$

It was pointed out by many authors that the assumption of separability - although rarely satisfied in real applications - yields a substantial simplification of the estimation problem and thus reduces computational costs in the estimation of the covariance in high dimensional problems.

To develop measures for the deviation from separability we consider a random element X in the Hilbert space H with $\mathbb{E}\|X\|^4 < \infty$, which is the tensor product

$H_1 \otimes H_2$ of two real separable Hilbert spaces H_1 and H_2. A typical example is the space of all real-valued square integrable functions $H = L^2(S \times T)$ defined on $S \times T$, where $H_1 = L^2(S)$ and $H_2 = L^2(T)$. The covariance operator of X is an element of $S_2(H)$, the space of all Hilbert Schmidt operators, and defined by

$$C := \mathbb{E}\left[(X - \mathbb{E}X) \otimes_o (X - \mathbb{E}X)\right],$$

where for $f, g \in H$ the operator $f \otimes_o g : H \to H$ is defined by

$$(f \otimes_o g)h = \langle h, g \rangle f \text{ for all } h \in H \ .$$

We also assume $\|C\|_2 \neq 0$, which essentially means the random variable X is non-degenerate. The covariance operator is called separable if

$$C = C_1 \widetilde{\otimes} C_2 \text{ for some } C_1 \in S_2(H_1) \text{ and } C_2 \in S_2(H_2).$$

To measure separability we consider the bounded linear operator $T_1 : S_2(H) \times S_2(H_1) \mapsto S_2(H_2)$ defined by

$$T_1(A \widetilde{\otimes} B, C_1) = \langle A, C_1 \rangle_{S_2(H_1)} B$$

for all $C_1 \in S_2(H_1)$. Similarly, let $T_2 : S_2(H) \times S_2(H_2) \to S_2(H_1)$ be the bounded linear operator defined by

$$T_2(A \widetilde{\otimes} B, C_2) = \langle B, C_2 \rangle_{S_2(H_2)} A$$

for all $C_2 \in S_2(H_2)$. It can be shown that the operators T_1 and T_2 are well-defined, bi-linear and continuous. Interestingly, we can define a nonnegative measure, which vanishes under the assumption of separability and depends only on the covariance operator C. To be precise, let Ψ be any fixed element of $S_2(H_2)$ such that $T_2(C, \Psi) \neq 0$, and define

$$D(C) = \|C\|_2^2 - \frac{\|T_1(C, T_2(C, \Psi))\|_2^2}{\|T_2(C, \Psi)\|_2^2}. \tag{1}$$

It can be shown that $D(C) \geq 0$ and equality holds, if and only if C is separable.

Formally we estimate the distance $D(C)$ by plugging in the estimator

$$\widehat{C}_N := \frac{1}{N} \sum_{i=1}^{N} \left[(X_i - \overline{X}) \otimes_o (X_i - \overline{X})\right].$$

for C based on a sample $X_1, X_2, \ldots, X_N$. The resulting statistic is given by

$$\widehat{D}_N = \|\widehat{C}_N\|_2^2 - \frac{\|T_1(\widehat{C}_N, T_2(\widehat{C}_N, \Psi))\|_2^2}{\|T_2(\widehat{C}_N, \Psi)\|_2^2}.$$

As this representation only involves norms we do not have to store the complete estimate of the covariance kernel. The following results provide the asymptotic properties of the statistic $\widehat{D}_N$

Theorem 1. *If* $\mathbb{E}\|X\|_2^4 < \infty$ *and the assumptions of separability holds we have*

$$N\widehat{D}_N \xrightarrow{d} \left\| \mathcal{G} - \frac{T_2(\mathcal{G},\Psi)\widetilde{\otimes}T_1(C,T_2(C,\Psi))}{\|T_2(C,\Psi)\|_2^2} \right\|_2^2 - \frac{\|T_1(\mathcal{G},T_2(C,\Psi)) - T_1(C,T_2(\mathcal{G},\Psi)\|_2^2}{\|T_2(C,\Psi)\|_2^2}$$

$$= \left\| \mathcal{G} - \frac{T_2(\mathcal{G},\Psi)\widetilde{\otimes}C_2}{\langle C_2,\Psi\rangle_{S_2(H_2)}} \right\|_2^2 - \left\| \frac{T_1(\mathcal{G},C_1)}{\|C_1\|_2} - \frac{\langle C_1,T_2(\mathcal{G},\Psi)\rangle_{S_2(H_1)}C_2}{\langle C_2,\Psi\rangle_{S_2(H_2)}\|C_1\|_2} \right\|_2^2,$$

where $\mathcal{G}$ *is a centered Gaussian process with covariance operator*

$$\Gamma := \lim_{N\to\infty} Var(\sqrt{N}\widehat{C}_N) \;=\; Var(X_1 \otimes_o X_1)$$

Theorem 2. *If* $\mathbb{E}\|X\|_2^4 < \infty$*, then the statistic*

$$\sqrt{N}\left(\widehat{D}_N - D(C)\right)$$

converges in distribution to a centered normal distribution with variance $\nu^2 := 4\langle\Gamma(A-B),(A-B)\rangle_{\mathrm{HS}}$*, where* $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ *is the inner product on* $S_2(H)$*,*

$$A = C - \frac{T_2(C,\Psi)\widetilde{\otimes}T_1(C,T_2(C,\Psi))}{\|T_2(C,\Psi)\|_2^2},$$

$$B = \frac{1}{\|T_2(C,\Psi)\|_2^2}\left[T_2(C,T_1(C,T_2(C,\Psi)))\widetilde{\otimes}\Psi - \frac{\|T_1(C,T_2(C,\Psi))\|_2^2}{\|T_2(C,\Psi)\|_2^2}T_2(C,\Psi)\widetilde{\otimes}\Psi\right],$$

and the centering term $D(C)$ *is defined in* (1)*.*

References

[1] P. Bagchi, H. Dette (2020), *A test for separability in covariance operators of random surfaces*, Annals of Statistics, to appear.

The role of geodesic convexity in covariance estimation

David E. Tyler

It is well known that the negative log-likelihood for the covariance matrix Σ based on a random sample $X_1,\ldots,X_n$ from a multivariate normal distribution, i.e. $L(\Sigma) = tr\{\Sigma^{-1}S_n\} + \log\det\Sigma$ with S_n being the sample covariance matrix, is convex in the precision matrix Σ^{-1}. When penalizing $L(\Sigma)$ it is then natural to add a penalty term which is convex in Σ^{-1}, as is the case for the graphical lasso. Perhaps a lesser know property of $L(\Sigma)$ is that it is also geodesically convex (or g-convex) in Σ, or equivalently in Σ^{-1}. Consequently, penalized covariance matrices based on g-convex penalties are also fairly tractable.

In this talk, a number of results based on g-convex penalties are discussed. For example, a class of non-smooth g-convex penalty functions are introduced for which the eigenvalues of the corresponding penalized covariance matrices has distinct groups of equal eigenvalues. This penalization method can be viewed as lassoing eigenvalues. A particularly promising member of this class of non-smooth g-convex penalties arises from an application of the Marčenko-Pasteur law.

G-convexity also has applications to penalized M-estimators of covariance matrices. The M-estimators of covariance can be defined as the value of Σ which minimizes the loss function $L_\rho(\Sigma) = n^{-1}\sum_{i=1}^n \rho(X_i'\Sigma^{-1}X_i) + \log\det\Sigma$ for a given ρ-function. Under the standard condition that $\rho(s)$ is convex in $\log(s)$, the function $L_\rho(\Sigma)$, although not convex in Σ^{-1}, is known to be g-convex, as recently shown within the area of signal processing. Consequently, the minimization problem arising from adding a g-convex penalty to $L_\rho(\Sigma)$, or when restricting Σ to a g-convex set, has a unique solution. A simple reweighting algorithm for computing the resulting penalized M-estimator of scatter is shown to always converges to the the minimum regardless of the initial value. Finally, it is noted that the popular spatial sign covariance matrix, defined as $V = n^{-1}\sum_{i=1}^n \theta_i\theta_i'$ where $\theta_i = X_i/\|X_i\|$ is the spatial sign of X_i, can be viewed as a limiting version of a penalized M-estimator of scatter.

Extracting robust and accurate features via a robust information bottleneck

Po-Ling Loh
(joint work with Ankit Pensia, Varun Jog)

We propose a novel strategy for extracting features in supervised learning that can be used to construct a classifier which is more robust to small perturbations in the input space. Our method builds upon the idea of the information bottleneck by introducing an additional penalty term that encourages the Fisher information of the extracted features to be small, when parametrized by the inputs. By tuning the regularization parameter, we can explicitly trade off the opposing desiderata of robustness and accuracy when constructing a classifier. We derive the optimal solution to the robust information bottleneck when the inputs and outputs are jointly Gaussian, proving that the optimally robust features are also jointly Gaussian in that setting. Furthermore, we propose a method for optimizing a variational bound on the robust information bottleneck objective in general settings using stochastic gradient descent, which may be implemented efficiently in neural networks. Our experimental results for synthetic and real data sets show that the proposed feature extraction method indeed produces classifiers with increased robustness to perturbations.

References

[1] A. Pensia, V. Jog, and P. Loh, *Extracting robust and accurate features via a robust information bottleneck*, arXiv preprint (2019).

Functional and complex data: new methods merging statistics, scientific computing and engineering

Laura M. Sangalli

Recent years have seen an explosive growth in the recording of increasingly complex and high-dimensional data. Classical statistical methods are often unfit to handle such data, whose analysis calls for the definition of new methods merging ideas and approaches from statistics, applied mathematics and engineering. This work in particular focuses on functional data defined over complex multidimensional domains, including curved bi-dimensional domains. I will present an innovative class of methods, based on regularizing terms involving partial differential equations. The proposed methods make use of advanced numerical techniques such as finite element analysis and isogeometric analysis. An illustration to the analysis of neuroimaging data is provided. In this applied field, the proposed methods offer important advantages with respect to the best state of the art techniques, allowing to correctly take into account to complex anatomy of the brain.

Multivariate Rank-based Distribution-free Nonparametric Testing using Optimal Transport

Bodhisattva Sen
(joint work with Nabarun Deb)

In the talk, we propose a general framework for distribution-free nonparametric testing in multi-dimensions, based on a notion of multivariate ranks defined using the theory of optimal transport (see e.g., Villani (2003)). Unlike other existing proposals in the literature, these multivariate ranks share a number of useful properties with the usual one-dimensional ranks; most importantly, these ranks are distribution-free (i.e., its joint distribution does not depend on the underlying data generating process). This crucial observation allows us to design nonparametric tests that are exactly distribution-free under the null hypothesis. We demonstrate the applicability of this approach by constructing exact distribution-free tests for two classical nonparametric problems: (I) testing for mutual independence between random vectors, and (II) testing for the equality of multivariate distributions. In particular, we propose (multivariate) rank versions of distance covariance ([3]) and energy statistic ([4]) for testing scenarios (I) and (II) respectively. In both these problems we derive the asymptotic null distribution of the proposed test statistics. We further show that our tests are consistent against all fixed alternatives. We also study the asymptotic (Pitman) efficiency of these multivariate rank-based tests and show that these are the only computationally-feasible tests that have non-zero Pitman efficiency among asymptotically distribution-free procedures.

Moreover, the proposed tests are tuning-free, computationally feasible and are well-defined under minimal assumptions on the underlying distributions (e.g., they do not need any moment assumptions). We also demonstrate the efficacy of these procedures via extensive simulations. In the process of analyzing the theoretical

properties of our procedures, we end up proving some new results in the theory of measure transportation and in the limit theory of permutation statistics using Stein's method for exchangeable pairs, which may be of independent interest.

These multivariate rank maps are optimal transport maps that transport the (empirical) data distribution to a (discretization of a) reference measure (e.g., uniform distribution on the unit hypercube). Although strong consistency properties of these empirical rank maps are known in the literature (see e.g., [2] and [1]), the rate of convergence of these estimators is still unknown (and is an open problem).

References

[1] Deb, Nabarun and Sen, Bodhisattva (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. arxiv e-prints, art. *arXiv preprint arXiv:1909.08733*.

[2] del Barrio, E., J. A. Cuesta-Albertos, M. Hallin, and C. Matrán (2018). Center-outward distribution functions, quantiles, ranks, and signs in rd. arxiv e-prints, art. *arXiv preprint arXiv:1806.01238*.

[3] Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist. 35*(6), 2769–2794.

[4] Székely, G. J. and M. L. Rizzo (2013). Energy statistics: a class of statistics based on distances. *J. Statist. Plann. Inference 143*(8), 1249–1272.

[5] Villani, C. (2003). *Topics in optimal transportation*, Volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

Testing for the Rank of a Covariance Kernel by Matrix Completion

Victor M. Panaretos

(joint work with Anirvan Chakraborty)

How can we discern whether a continuous time stochastic process $\{X(t) : t \in [0,1]\}$ is finite dimensional, and if so, what its precise dimension is? And how can we do so at a given level of confidence? This question is central to a great deal of methods for functional data, which require low-dimensional representations whether by functional PCA or other methods. The determination is to be made on the basis of i.i.d. replications of the process $\{X_1, \ldots, X_n\}$ (assumed continuous and centered), with the twist that these are measured with measurement error contamination on a grid of finite size,

$$W_{ip} = X_i(t_p) + \varepsilon_{ip}, \qquad \begin{cases} i = 1, \ldots, n \\ p = 1, \ldots, L \end{cases},$$

where $0 < t_1 < \ldots < t_L < 1$ is a fixed grid (assumed regular without loss of generality) and the array ε_{ip} is i.i.d. with zero mean, variance $\sigma_p^2 > 0$, and independent of $\{X_i\}_{i=1}^n$. This measurement scheme obfuscates the underlying dimension: the $L \times L$ covariance matrix $K_{\boldsymbol{W},L}$ of $\boldsymbol{W}_i = (W_{i,1}, ..., W_{i,L})^\top$ is the superposition of that of $\boldsymbol{X}_i = (X_i(t_1), ..., X_i(t_L))^\top$and that of $\boldsymbol{\varepsilon} = (\varepsilon_{i,1}, ..., \varepsilon_{i,L})^\top$,

$$K_{\boldsymbol{W}_i,L} = K_{\boldsymbol{X}_i,L} + K_{\boldsymbol{W}_i,L}$$

and so its rank is always equal to $L \wedge n$. This leads to the following conundrum: if one were to smooth in order to annihilate the noise, then the resulting rank would depend on the choice of smoothing parameter. Without smoothing, the ridge added by the noise variance confounds the rank. For this reason, it was remarked by Hall & Vial [2] that the problem of inferring the rank of the underlying covariance kernel $k(u,v) = \mathbb{E}[X(u)X(v)]$ may not be amendable to approaches based on formal hypothesis testing.

Observing that the smoothing step might be entirely circumvented, we construct a formal testing approach based on matrix completion. The key observation is that only the diagonal is affected by the presence of noise, and that since $K_{\boldsymbol{X}_i,L}$ is a discrete version of an otherwise continuous continuous covariance $k(u,v)$, we might be able to determine the rank relying solely on its off-diagonal entries, where $K_{\boldsymbol{X}_i,L}$ coincides with $K_{\boldsymbol{W}_i,L}$, provided L is sufficiently large.

We show that when the true rank r_{true} of $k(u,v)$ is finite, there exists a critical $1 \leq L_* < \infty$ such that for all $L > L_*$, the polynomial

$$\Theta \mapsto \|P_L \circ (K_{W,L} - \Theta)\|_F^2$$

is strictly positive when Θ ranges over matrices of rank $< r_{\text{true}}$, whereas it has a unique root at $\Theta = K_{X,L}$ when Θ ranges over matrices of rank $\leq r_{\text{true}}$.

We use this identifiability result to build a matrix-completion test statistic that measures the best possible least square fit of the off-diagonal elements of the empirical version of $K_{\boldsymbol{W}_i,L}$, $\hat{K}_{\boldsymbol{W}_i,L} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_i\boldsymbol{W}_i^\top$, optimised over covariances of given finite rank:

$$T_j = \min_{\Theta^{L\times L}:rank(\Theta)\leq j} ||P_L \circ (\widehat{K}_{W,L} - \Theta)||_F^2.$$

For a given grid of supercritical but fixed size, we determine the statistic's asymptotic distribution as the number of replications n diverges, under the null hypothesis $\{H_{0,j} : \text{rank} = j\}$. We then use it to construct an appropriate bootstrap calibration scheme. This is combined with a stepwise testing procedure controlling the family-wise error rate corresponding to the collection of hypotheses $\{H_{0,j} : j = 1, ..., L-1\}$ formalising the question at hand. Under minimal regularity assumptions we prove that the procedure is consistent and that its bootstrap implementation is valid. The procedure rests on minimal regularity assumptions, involves no tuning parameters or pre-smoothing, and is indifferent to the homoskedasticity or lack of it in the measurement errors.

References

[1] A. Chakraborty and V.M. Panaretos *What is the dimension of a stochastic process? Testing for the ran of a covariance operator*, arXiv:1901.02333 (2019).

[2] P. Hall, and C. Vial *Assessing the finite dimensionality of functional data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(4), pp.689-705 (2006), 689–705.

Network change point localisation

Yi Yu

In this talk I will start with a parametric network offline change point detection problem, showing consistent and optimal change point estimators and a phase transition phenomenon [1]. I will then move on to a nonparametric dynamic network problem allowing for both across-time and within network dependence, accompanied with its application on a zebrafish neuronal activity data set [2].

References

[1] D. Wang, Y. Yu, A. Rinaldo *Optimal Change Point Detection and Localization in Sparse Dynamic Networks*, arXiv:1809.09602 (2018)

[2] O. H. M. Padilla, Y. Yu, C. E. Priebe *Change point localization in dependent dynamic nonparametric random dot product graphs*, arXiv:1911.07494 (2019)

Learning from Complex Medical Data, Clustering, and Interpretable Kernel Dimensionality Reduction

Jennifer G. Dy

(joint work with Chieh Wu, James Ross)

Machine learning as a field has become more and more important due to the ubiquity of data collection in various disciplines. Coupled with this data collection is the hope that new discoveries or knowledge can be learned. In many applications, data is often complex, high-dimensional and multi-faceted, where multiple possible interpretations are inherent in the data. In the talk, I highlight these challenges through my experience in a collaborative research working on discovering disease subtypes and then provide examples of how these challenges led to innovations in machine learning and to new discovery.

Chronic Obstructive Pulmonary Disease (COPD) is a lung disease characterized by airflow limitation usually associated with chronic inflammatory responses to noxious particles, such as cigarette smoke. COPD is a heterogeneous disease. The COPDGene Study[1], which involves 21 clinical sites throughout the US, collected high-resolution computed tomography images, physiological features, demographic features, and genetics data from 10,000 patients. Our goal is to discover disease subtypes that leads to better stratification of the patients so as to provide better prognosis and personalized therapies.

Subtyping from a machine learning point of view is a clustering problem. Clustering is the process of grouping instances together based on some notion of similarity (typically in the form of a metric or a probability model). The first challenge in working on real data is finding the right model. Most standard clustering methods do not take the structure of the problem into account and treat all the features/variables in the same way; however, in our COPD sub-typing problem, we have variables such as age and smoking that are causative agents of variables that indicate lung function and disease severity. The type of grouping we are interested

in discovering relates to how different groups of individuals respond to exposure. The manner in which lung health changes as a function of age and smoke exposure can be used to identify meaningful subgroups. Some people are genetically resistant to the effects of smoke exposure and have preserved lung health even after years of smoking. On the other hand, others are highly sensitive to smoke and experience rapid health decline given similar levels of exposure. This led us to the notion of *disease trajectories*. We model the unknown disease trajectory by a flexible distribution over functions with Gaussian processes (GPs). We define a cluster component as belonging to one of k possible GPs. In our problem, we one have a few (two) time points per patient. We introduce a variational Dirichlet process mixture of Gaussian processes that can also learn from must-link and cannot-link constraints [2]. Our model is able to learn the number of clusters (trajectories) automatically for a mixture of GPs, learn the trajectories, and learn the cluster membership (which trajectory) a patient belongs to. We utilize the must-link constraint to allow us to guide the few time points belonging to the same patient to be in the same trajectory.

The second challenge is high-dimensionality. Not all features are important. Only subsets of the features are useful for describing each cluster. We allow both instances and features to belong to more than one cluster. We utilize GPs to represent trajectories and dual beta process priors for instance and feature assignment to the latent clusters (subtypes) [3].

The third challenge is that data is often multi-faceted, where multiple possible interpretations are inherent in the data. Given a face image data set. One can cluster it based on person's identity; another reasonable clustering is based on pose. Data can be grouped in multiple ways and different subspaces reveal different possible groupings. However, typical clustering algorithms output a single clustering solution. The solution found by the algorithm may not be what the domain scientist is interested in. We introduced a new clustering paradigm: *Find multiple alternative clustering views (perspectives) from data, where data points belonging to the same cluster in one view can belong to different clusters in another view* [4]. There are two modes of discovering multiple alternative clustering views: simultaneously [5] or iteratively [6].

In the talk, I focused on our recent work for learning alternative clustering, the Kernel Dimension Alternative Clustering (KDAC) via an Iterative Spectral Method (ISM) [7]. In alternative clustering, the goal is to find solutions that are of high cluster quality and non-redundant to the existing (previously found) clustering. Moreover, we noticed that typically, the different alternative clusterings reside in different subspaces. Thus, in our formulation, we also simultaneously learn the subspace in which the clustering reside. To enable capturing arbitrarily-shaped clusters, we employ the spectral clustering [8] objective to define cluster quality. We would like the clustering solutions we discover to be non-redundant with each other. There are several possible criteria for measuring non-redundancy: correlation or mutual information. Correlation can capture only linear dependencies. Mutual information can capture non-linear dependencies, but requires estimating

the joint probability distribution. We suggest the Hilbert-Schmidt Independence Criterion (HSIC) [9]. HSIC measures dependence by mapping variables into a reproducing kernel Hilbert space such that correlations measured in that space correspond to high-order joint moments between the original distributions. This approach is able to estimate dependence between variables without explicitly estimating the joint distribution of the random variables. In addition, as shown in [6], HSIC between the data and the latent clustering is mathematically equivalent to the spectral clustering objective.

Let $X \in \mathcal{R}^{n\times d}$ be a dataset with n samples and d features, along with an existing clustering $Y \in \mathcal{R}^{n\times k}$, where k is the number of clusters. If x_i belongs to cluster j, then $Y_{i,j} = 1$; otherwise, $Y_{i,j} = 0$. We wish to discover an alternative clustering $U \in \mathcal{R}^{n\times k}$ on some lower dimensional subspace of dimension $q \ll d$. Let $W \in \mathcal{R}^{d\times q}$ be a projection matrix such that $XW \in \mathcal{R}^{n\times q}$. We seek the optimal projection W and clustering U that maximizes the statistical dependence between XW with U, yielding a high clustering quality, while minimizing the dependence between XW and Y, ensuring the novelty of the new clustering.

Using HSIC as a dependence measure, the objective of KDAC becomes

$$\begin{aligned} &\text{Maximize:} && \mathrm{HSIC}(XW, U) - \lambda \mathrm{HSIC}(XW, Y) \\ &\text{subject to:} && W^T W = I, U^T U = I. \end{aligned} \tag{1}$$

where $\mathrm{HSIC}(X, Y) \equiv \frac{1}{(n-1)^2}\mathrm{Tr}(K_X H K_Y H)$. Here, the variables K_X and K_Y are Gram matrices, and the H matrix is a centering matrix where $H = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ with $\mathbf{1}$ the n-sized vector of all ones. The elements of K_X and K_Y are calculated by kernel functions $k_X(x_i, x_j)$ and $k_Y(y_i, y_j)$. The kernel functions for Y and U used in KDAC are $K_Y = YY^T$ and $K_U = UU^T$, and the kernel function for XW is the Gaussian $k_{XW}(x_i, x_j) = \exp(-\mathrm{Tr}[(x_i - x_j)^T WW^T(x_i - x_j)]/(2\sigma^2))$. Due to the equivalence of HSIC and spectral clustering, the practice of normalizing the kernel K_{XW} is adopted from spectral clustering by [6]. That is, for K_{XW} the unnormalized Gram matrix, the normalized matrix is defined as $D^{-1/2}K_{XW}D^{-1/2}$ where $D = \mathrm{diag}(\mathbf{1}_n^T K_{XW})$ is a diagonal matrix whose elements are the column-sums of K_{XW}.

We optimize Equation (1) using alternating optimization. Holding U and W constant, D is computed as $D = \mathrm{diag}(\mathbf{1}_n^T K_{XW})$. Holding W and D constant and solving for U reduces to spectral clustering [8].

While holding U and D constant to solve for W, (1) reduces to:

$$\begin{aligned} &\text{Maximize:} && F(W) = \textstyle\sum_{i,j} \gamma_{i,j} e^{-\frac{\mathrm{Tr}[W^T A_{i,j} W]}{2\sigma^2}} \\ &\text{subject to:} && W^T W = I \end{aligned} \tag{2}$$

where $\gamma_{i,j}$ are the elements of matrix $\gamma = D^{-1/2}H(UU^T - \lambda YY^T)HD^{-1/2}$, and $A_{i,j} = (x_i - x_j)(x_i - x_j)^T$. This objective, along with a Stiefel Manifold constraint, $W^T W = I$, pose a challenging optimization problem as neither is convex.

We introduce ISM to solve (2). ISM attempts to find such a W in the following iterative fashion. Let W_0 be an initial matrix. Given W_k at iteration k, the matrix

W_{k+1} is computed as:

$$W_{k+1} = \mathrm{eig}_{\max}(\Phi(W_k)), \quad k = 0, 1, 2, \ldots,$$

where the operator $\mathrm{eig}_{\max}(\Phi(W_k))$ returns a matrix whose columns are the q eigenvectors corresponding to the q largest eigenvalues of $\Phi(W_k)$.

$$\Phi(W) = \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} \exp(-\frac{\mathrm{Tr}[W^T A_{i,j} W]}{2\sigma^2}) A_{i,j}, \tag{3}$$

We iteratively update W with W_k until convergence.

We provide ISM a natural initialization, W_0, constructed through a second order Taylor approximation of the objective. This resulted in high quality results without random restarts in search of a better initialization. Furthermore, we provide theoretical guarantees on its fixed point. In particular, we establish conditions under which the fixed point of ISM satisfies both the 1st and 2nd order necessary conditions for local optimality. Empirical experiments show that ISM outperforms gradient ascent on a Stiefel manifold and dimension growth in clustering quality measures along with significantly lower computational cost.

Finally, we end our talk showing how we can utilize ISM as an algorithm for solving more generally Interpretable Kernel Dimensionality Reduction (IKDR) problems [10]. Kernel dimensionality reduction (KDR) algorithms find a low dimensional representation of the original data by optimizing kernel dependency measures (such as HSIC) that are capable of capturing nonlinear relationships. The standard strategy is to first map the data into a high dimensional feature space using kernels prior to a projection onto a low dimensional space. While KDR methods (e.g., kernel principal component analysis [11]) can be easily solved by keeping the most dominant eigenvectors of the kernel matrix, the new features are no longer easy to interpret. To make KDR interpretable, IKDR projects the original input onto a subspace *before* the kernel feature mapping; therefore, the projection matrix can indicate how the original features linearly combine to form the new features. Unfortunately, the IKDR objective requires a non-convex manifold optimization that is difficult to solve and can no longer be solved by eigendecomposition.

IKDR for a variety of machine learning (ML) paradigms – supervised, semi-supervised, unsupervised, alternative clustering, can be expressed as the following optimization problem:

$$\max_W \mathrm{Tr}(\Gamma K_{XW}) \quad \text{s.t. } W^T W = I, \tag{4}$$

where Γ is a symmetric matrix commonly derived from K_Y. Refer to [10] for the explicit form of Γ for the various ML paradigms. $K_{XW} \in \mathbb{R}^{n \times n}$ is a kernel matrix with each entry defined as $K_{XW_{ij}} = k(W^T x_i, W^T x_j)$ where $k : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is a kernel function. Let Y be the one-hot encoding of the labels with its corresponding kernel matrix denoted as K_Y.

We show that an efficient iterative spectral (eigendecomposition) method (ISM) can be utilized to solve the general IKDR optimization problem (4). Previously, in [7] ISM only provides theoretical guarantees for the Gaussian kernel. In [10], we generalize the theoretical guarantees of ISM to an entire family of kernels

and propose the necessary criteria for a kernel to be a member of the family. In identifying this family, we prove that each kernel within the family has a surrogate Φ matrix and the optimal projection is formed by its most dominant eigenvectors. We further show that conic combinations of kernels from the ISM family belong to the ISM family and their respective Φ matrix is the conic combination of the corresponding Φ matrices. With this extension, we establish how a wide range of IKDR applications across different learning paradigms can be solved by ISM.

References

[1] E.A. Regan, J.E. Hokanson, J.R. Murphy, B. Make, D.A. Lynch, T.H. Beaty, D. Curran-Everett, E.K. Silverman, J.D. Crapo, *Genetic epidemiology of COPD (COPDGene) study design*, COPD **7** (2010), 32–43.

[2] J.C. Ross and J.G. Dy, *Nonparametric mixture of Gaussian processes with constraints*, Proceedings of the 30th International Conference on Machine Learning (ICML), JMLR Workshop and Conference Proceedings **28** (2013), 1346–1354.

[3] J.C. Ross, P.J. Castaldi, M.H. Cho, J.G. Dy, *Dual beta process priors for latent cluster discovery in chronic obstructive pulmonary disease*, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014), 155–162.

[4] Y. Cui, X.Z. Fern, J.G. Dy, *Non-redundant multi-view clustering via orthogonalization*, Proceedings of the 7th IEEE International Conference on Data Mining (2007), 133–142.

[5] D. Niu, J.G. Dy, M.I. Jordan, *Multiple non-redundant spectral clustering views*, Proceedings of the 27th International Conference on Machine Learning (2010), 831–838.

[6] D. Niu, J.G. Dy, M.I. Jordan, *Iterative discovery of multiple alternative clustering views*, IEEE Transactions on Pattern Analysis and Machine Intelligence **36(7)** (2014), 1340–1353.

[7] C. Wu, S. Ioannidis, M. Sznaier, X. Li, D.R. Kaeli, J.G. Dy, *Iterative spectral method for alternative clustering*, International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research **84** (2018), 115–123.

[8] A.Y. Ng, M.I. Jordan, Y. Weiss, *On Spectral Clustering: Analysis and an algorithm*, Advances in Neural Information Processing Systems **14** (2001), 849–856.

[9] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, *Measuring statistical dependence with Hilbert-Schmidt norms*, International conference on algorithmic learning theory (2005), 63–77.

[10] C. Wu, J. Miller, Y. Chang, M. Sznaier, J.G. Dy, *Solving interpretable kernel dimension reduction*, Conference on Neural Information Processing Systems (2019).

[11] B. Schölkopf, A. Smola, K-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural computation **10(5)** (1998), 1299–1319.

Reporters: Leonard Henckel, Solt Kovács

Participants

Prof. Dr. Francis Bach
INRIA - SIERRA project-team
Departement d'Informatique
Ecole Normale Superieure
Voie DQ 12
2, rue Simone Iff
75012 Paris Cedex
FRANCE

Prof. Dr. Fadoua Balabdaoui
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Natalia Bochkina
School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
King's Buildings
Peter Guthrie Tait Road
Edinburgh EH9 3FD
UNITED KINGDOM

Prof. Dr. Helmut Bölcskei
Mathematical Information Sciences
ETH Zürich
Room: ETF E 122
Sternwartstrasse 7
8092 Zürich
SWITZERLAND

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Emmanuel J. Candès
Department of Statistics
Stanford University
Sequoia Hall
Stanford CA 94305-4065
UNITED STATES

Yuansi Chen
Seminar für Statistik
ETH Zürich (HG G 19)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Laëtitia Comminges
CEREMADE
Université Paris-Dauphine
Bureau: C 612
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
Gebäude IB 2/65
44780 Bochum
GERMANY

Lutz Dümbgen
Institut für Mathematische Statistik
und Versicherungslehre
Universität Bern
Alpeneggstrasse 22
3012 Bern
SWITZERLAND

Prof. Dr. Cécile Durot
Département de Mathématiques et Informatique
Université Paris Nanterre
E 26, Bâtiment G
200 Avenue de la République
92001 Nanterre Cedex
FRANCE

Dr. Jennifer Dy
Department of Electrical and Computer Engineering
Northeastern University
409 Dana Building
360 Huntington Avenue
Boston, MA 02115-5000
UNITED STATES

Prof. Dr. Piet Groeneboom
Delft Institute of Applied Mathematics
Delft University of Technology
Van Mourik Broekmanweg 6
2628 XE Delft
NETHERLANDS

Prof. Dr. László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Stoczek u. 2
1521 Budapest
HUNGARY

Leonard Henckel
Seminar für Statistik
ETH Zürich (HG G 11)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Jana Janková
Statistical Laboratory, DPMMS
Center for Mathematical Sciences
University of Cambridge
Pavilon D2.08
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Solt Kovács
Seminar für Statistik
ETH Zürich (HG G 18)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Yulia Kulagina
Seminar für Statistik
ETH Zürich (HG G 18)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Elizaveta Levina
Department of Statistics
Michigan Institute for Data Science
University of Michigan
323 West Hall, Office 459
1085 S. University Avenue
Ann Arbor MI 48109-1107
UNITED STATES

Matthias Löffler
Departement Mathematik
ETH-Zentrum (HG G 10.3)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Po-Ling Loh
Department of Statistics
Wisconsin Institute for Discovery
University of Wisconsin, Madison
1300 University Avenue
Madison,WI 53715
UNITED STATES

Dr. Zongming Ma
Department of Statistics
The Wharton School
University of Pennsylvania
468 Jon M. Huntman Hall
3730 Walnut Street
Philadelphia PA 19104
UNITED STATES

Prof. Dr. Marloes Maathuis
Seminar für Statistik
ETH Zürich (HG G 15.1)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Nicolai Meinshausen
Seminar für Statistik
ETH Zürich (HG G 24.2)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Andrea Montanari
Department of Electrical Engineering
and Department of Statistics
Stanford University
Stanford CA 94305-4065
UNITED STATES

Dr. Grégoire Montavon
School IV: Electrical Engineering and
Computer Science
Technical University of Berlin
Sekr. MAR 4-1
Marchstrasse 23
10587 Berlin
GERMANY

Prof. Dr. Klaus-Robert Müller
Computer Science Department
Technical University of Berlin
Sekr. MAR 4-1
Marchstrasse 23
10587 Berlin
GERMANY

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Robert D. Nowak
Department of Electrical and
Computer Engineering
University of Wisconsin-Madison
Engineering Hall # 3627
1550 Engineering Drive
Madison WI 53706
UNITED STATES

Dr. Guillaume Obozinski
E P F L
INN 216 (Bâtiment INN)
Station 14
1015 Lausanne
SWITZERLAND

Prof. Dr. Victor M. Panaretos
Institut de Mathématiques
E P F L
MA B1 503
Station 8
1015 Lausanne
SWITZERLAND

Jonas Peters
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Prof. Dr. Angelika Rohde
Fakultät für Mathematik
Albert-Ludwigs-Universität Freiburg
LST für Stochastik
Ernst-Zermelo-Strasse 1
79104 Freiburg i. Br.
GERMANY

Dr. Wojciech Samek
Fraunhofer Institute for
Telecommunications
Heinrich Hertz Institute, HHI
Department of Video Coding and
Analytics
Einsteinufer 37
10587 Berlin
GERMANY

Prof. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Laura M. Sangalli
Dipartimento di Matematica
Politecnico di Milano
Piazza L. da Vinci, 32
20133 Milano
ITALY

**Prof. Dr. Johannes
Schmidt-Hieber**
Department of Applied Mathematics
University of Twente
P.O.Box 217
7500 AE Enschede
NETHERLANDS

Dr. Bodhisattva Sen
Department of Statistics
Columbia University
1255 Amsterdam Avenue
New York, NY 10027
UNITED STATES

Prof. Dr. Vladimir G. Spokoiny
Weierstrass Institute for Applied
Analysis
and Stochastics (WIAS)
Mohrenstrasse 39
10117 Berlin
GERMANY

Dr. Armeen Taeb
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Robert Tibshirani
Department of Statistics
Stanford University
Sequoia Hall, Room 106
Stanford, CA 94305
UNITED STATES

Prof. Dr. Ryan Tibshirani
Departments of Statistics and
Machine Learning
Carnegie Mellon University
229B Baker Hall
Pittsburgh, PA 15213
UNITED STATES

Prof. Dr. Alexandre B. Tsybakov
CREST - ENSAE
5, Avenue Henry Le Châtelier
91120 Palaiseau Cedex
FRANCE

Prof. Dr. David E. Tyler
Department of Statistics
RUTGERS, The State University of New
Jersey
567 Hill Center, Busch Campus
110 Frelinghuysen Road
Piscataway, NJ 08854
UNITED STATES

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich (HG G 24.1)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Tengyao Wang
Department of Statistical Science
University College London
Room 133
1-19 Torrington Place
London WC1E 7HB
UNITED KINGDOM

Dr. Yuting Wei
Department of Statistics and Data Science
Carnegie Mellon University
Baker Hall 229J
Pittsburgh, PA 15213
UNITED STATES

Prof. Dr. Jon A. Wellner
Statistics Department
University of Washington
P.O. Box 35 43 22
Seattle, WA 98195
UNITED STATES

Prof. Dr. Fan Yang
Department of Computer Science
ETH Zürich (CAB G 68)
Universitätsstrasse 6
8092 Zürich
SWITZERLAND

Dr. Yi Yu
Department of Statistics
University of Warwick
Room 4.11
Coventry CV4 7AL
UNITED KINGDOM