Mathematisches Forschungsinstitut Oberwolfach

Report No. 29/2018

DOI: 10.4171/OWR/2018/29

Matrix Estimation Meets Statistical Network Analysis: Extracting low-dimensional structures in high dimension

Organised by
Florentina Bunea, Ithaca
Angelika Rohde, Freiburg
Patrick Wolfe, London
Harrison Zhou, New Haven

17 June – 23 June 2018

Abstract. The study of complex relationships among the elements of a large collection of random variables lead to the development of a number of areas in probability and statistics such as probabilistic network analysis or random matrix theory. The aim of the workshop was to address the challenge to develop a coherent mathematical framework within which these areas can be integrated, for a successful analysis of massive and complicated data sets.

Mathematics Subject Classification (2010): 62G99, 60G05.

Introduction by the Organisers

The workshop *Matrix estimation meets statistical network analysis*, organized by Florentina Bunea (Ithaca), Angelika Rohde (Freiburg), Patrick Wolfe (London), and Harrison Zhou (New Haven) was well attended with around 25 participants with broad geographic representation. In summary, the workshop was devoted to the problem of developing a coherent mathematical framework within which the areas of probabilistic network analysis and random matrix theory can be integrated for a successful analysis of massive and complicated data sets. The workshop was a nice blend of researchers with the different backgrounds of matrix estimation and statistical network analysis. Having the two overlapping, but still different, communities together was crucial and we view the intensive interactions in the Oberwolfach setting as instrumental in making a synergy happen. Within lively discussions during and after the talks, and in particular in the evenings, fruitful inspiration for new research topics and directions has developed. Starting on

Monday morning with Boaz Nadler's (Rechovot) first talk and the subsequent representation of Peter Bickel (Berkeley), a remarkably open atmosphere was already generated and remained present for the rest of the week.

Whenever we observe entities and relations between them, either directly or induced from other data as a means of summarizing sparse dependency structure, we must draw inferences from network data. These datasets are growing so rapidly in complexity and dimensionality that our statistical analysis methods struggle to keep pace. For this reason network models have seen a revival in the last decade, one of the goals being to depart from basic random graph models such as those due to Erdős & Rényi (1959), which have limited applicability to genetics, neuroscience, internet networks, astronomy, and other fields where data exhibit a much more complicated dependence structure. Often networks come without labeled nodes, and the goal is to infer these labels (e.g., who belongs to which social group, or which proteins are implicated in what biological processes). At other times we must infer the network structure itself. More generally, as we encounter bigger and more heterogenous networks in the real world, understanding the limiting behavior of networks in the large-sample limit is critical to enabling statistical modeling and inference algorithms with good theoretical properties. An important problem is to develop network models that are highly heterogeneous even in the large-sample limit. By addressing this question, the workshop helped the community to take a large step forward.

Another important statistical question is that of estimating the network. It is immediately clear that the term network is ambiguous and needs further clarification. Whereas in all cases, the pictorial representation is a graph consisting of nodes and edges between some of the nodes, the rule according to which an edge appears in such a graph is crucial in defining a certain type of network. For this reason, another core area of the workshop was the study of sparse matrix models and graphs in high dimension.

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: Matrix Estimation Meets Statistical Network Analysis: Extracting low-dimensional structures in high dimension

Table of Contents

Abstracts

Nonparametric estimation for linear SPDEs

Randolf Altmeyer
(joint work with Markus Reiß)

It is well-known that parameters in the drift of a stochastic *ordinary* differential equation, observed continuously on a time interval $[0,T]$, can generally only be estimated consistently, if either $T \to \infty$, the driving noise becomes small or if a sequence of independent samples is observed. For stochastic *partial* differential equations (SPDEs) this is quite different. For example, consider the stochastic heat equation

$$dX(t,x) = \Delta_\vartheta X(t,x)dt + dW(t,x), \quad t \in [0,T],\ x \in \Omega \subset \mathbb{R}^d, \tag{1}$$

where $\Delta_\vartheta g = div(\vartheta \nabla g)$ is the weighted Laplace operator for an unknown thermal diffusivity ϑ and where W is space-time white noise. In the special case when $\vartheta > 0$ is constant, [1] showed that consistent estimation of ϑ is possible also in finite time $T < \infty$, if the Fourier modes $\langle X(t,\cdot), e_k\rangle$ are observed continuously on $[0,T]$ for $k = 1,\dots,N$ as $N \to \infty$, where the e_k are the eigenfunctions of Δ_ϑ. This approach is not feasible, however, when ϑ is not constant, as the eigenfunctions of Δ_ϑ depend on ϑ in this case and are thus unknown, as well. We therefore introduce in this work a different observation scheme. Let $K_{h,x_0}(x) = h^{-d/2}K(h^{-1}(x-x_0))$, $h > 0$, $x_0 \in \Omega$, for a smooth kernel K with compact support and L^2-norm $\|K\| = 1$. Assume that we can observe the linear functionals

$$X_h(t) = \langle X(t,\cdot), K_{h,x_0}\rangle = X(t,\cdot) * K_h(x_0), \quad t \in [0,T]. \tag{2}$$

These *local measurements* correspond to the intuition that in applications it is generally not possible to observe the solution $X(t,x_0)$ at a point x_0, but only a local average, given by the convolution $X(t,\cdot) * K_h(x_0)$. Our goal is to use these measurements to estimate $\vartheta(x_0)$. X_h satisfies the equation

$$dX_h(t) = \langle X(t,\cdot), \Delta_\vartheta K_{h,x_0}\rangle dt + dW_h(t), \quad t \in [0,T], \tag{3}$$

for a scalar Brownian motion W_h. Even though this means that X_h is not a diffusion process, it can be shown that the MLE for constant ϑ is

$$\hat\vartheta_h^{MLE}(x_0) = \frac{\int_0^T b(\bar X_h(t))dX_h(t)}{\int_0^T b(\bar X_h(t))^2 dt} \tag{4}$$

with $\bar X_h(t) = (X_h(r))_{0\le r\le T}$ and $b(\bar X_h(t)) = \mathbb{E}[\langle X(t,\cdot), \Delta_\vartheta K_{h,x_0}\rangle | \bar X_h(t)]$. Since an explicit computation of the conditional expectation seems impossible, we consider instead $\mathbb{E}[\langle X(t,\cdot), \Delta_\vartheta K_{h,x_0}\rangle | X_h(t)]$, which leads for small h and also for general ϑ to the estimator

$$\hat\vartheta_h^s(x_0) = h^2\|(-\Delta)^{-1/2}K\|^2 \frac{\int_0^T X_h(t)dX_h(t)}{\int_0^T X_h^2(t)dt}. \tag{5}$$

We prove that $\hat{\vartheta}_h^s(x_0) - \vartheta(x_0) = O_{\mathbb{P}}(T^{-1/2}h)$ which shows that observing only the local measurement $(X_h(t))_{0\le t\le T}$ for $T < \infty$ is already sufficient for estimating $\vartheta(x_0)$ consistently, as long as $h \to 0$. A key step in the proof is the following localization property of the semigroup $e^{\Delta_\vartheta t}$:

$$e^{\Delta_\vartheta t} K_{h,x_0} = (e^{\Delta_{\vartheta(h\cdot+x_0)} t/h^2} K)_{h,x_0}. \tag{6}$$

This means that the semigroup, applied to the localized function K_{h,x_0}, corresponds to the semigroup for the localized Laplace operator $\Delta_{\vartheta(h\cdot+x_0)}$ applied to K and with rescaled time t/h^2. In view of this property it follows, in the case of constant ϑ, that

$$\hat{\vartheta}_h^s(x_0) \overset{d}{\sim} \|(-\Delta)^{-1/2} K\|^2 \frac{\int_0^{T/h^2} X_1(t) dX_1(t)}{\int_0^{T/h^2} X_1^2(t) dt}. \tag{7}$$

Formally, this is similar to the MLE for the scalar Ornstein-Uhlenbeck process and therefore demonstrates why the observations contain as $h \to 0$ asymptotically the same information content as for large-time asymptotics, even when T is fixed. While $\hat{\vartheta}_h^s(x_0)$ is rate-optimal, it is not efficient. We therefore consider also other estimators, and provide central limit theorems in some cases, where the bias for non-constant ϑ can be analyzed precisely.

References

[1] M. Huebner and B.L. Rozovskii, *On asymptotic properties of maximum likelihood estimators for parabolic stochastic PDE's*, Probability theory and related fields, **103**, 1995, 143-163.

An issue of large scale networks

Peter Bickel

(joint work with Soumendu Mukherjee, Sharmo Bhattacharyya and Purna Sarkar)

Networks are a complex type of structure presenting itself in many applications. They are usually represented by a graph, with possibly weighted edges plus additional covariates (such as directions). As usual we focus on probability models for an unweighted graph without covariates characterized by an n by n adjacency matrix of 0's and 1's whose elements A_{ij} indicate presence or absence of an edge between i and j and which are independent given unobserved independent latent variables $Z_1, \ldots, Z_n$ (Aldous-Hoover models). Block models have been studied for some time as basic approximations to these both from a computational and inferential point of view. A huge number of fitting methods have been developed for block models, many based on spectral clustering and SDP relaxations. Among fitting methods the mean field method is attractive because it can easily be accommodated to the introduction of covariates, dynamics, etc. Unfortunately, it requires non-convex optimization over spaces of dimension n and if the graph is too large poor behavior of the method can be seen even in situations where theory suggests the true behavior of the empirical optimum should be good. This is an

issue of scale. We have developed and will discuss in this talk divide and conquer methods for fitting such large graphs by using patches (small subgraphs). These methods are generic and not limited to mean field. We show computational and, implicitly, inferential improvement in such situations. The ideas can be applied in principle to other situations such as topic models for documents and overlapping block models. Even in our original application stronger theoretical results and and more extensive simulation is needed. This work is presently on arXiv [1].

References

[1] S. S. Mukherjee, P. Sarkar and P. J. Bickel, *Two provably consistent divide and conquer clustering algorithms for large networks*, arXiv preprint, arXiv:1708.05573, (2017).

A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics

Mike Bing

(joint work with Florentina Bunea and Marten Wegkamp)

Topic models have become popular for the analysis of data that consists in a collection of n independent multinomial observations, with parameters $N_i \in \mathbb{N}$ and $\Pi_i \in [0,1]^p$ for $i = 1, \ldots, n$. The model links all cell probabilities, collected in a $p \times n$ matrix Π, via the assumption that Π can be factorized as the product of two nonnegative matrices $A \in [0,1]^{p \times K}$ and $W \in [0,1]^{K \times n}$. Topic models have been originally developed in text mining, when one browses through n documents, based on a dictionary of p words, and covering K topics. In this terminology, the matrix A is called the word-topic matrix, and is the main target of estimation. It can be viewed as a matrix of conditional probabilities, and it is uniquely defined, under appropriate separability assumptions, discussed in detail in this work. Notably, the unique A is required to satisfy what is commonly known as the anchor word assumption, under which A has an unknown number of rows respectively proportional to the canonical basis vectors in $\mathbb{R}^K$. The indices of such rows are referred to as anchor words. Recent computationally feasible algorithms, with theoretical guarantees, utilize constructively this assumption by linking the estimation of the set of anchor words with that of estimating the K vertices of a simplex. This crucial step in the estimation of A requires K to be known, and cannot be easily extended to the more realistic set-up when K is unknown.

This work takes a different view on anchor word estimation, and on the estimation of A. We propose a new method of estimation in topic models, that is not a variation on the existing simplex finding algorithms, and that estimates K from the observed data. We derive new finite sample minimax lower bounds for the estimation of A, as well as new upper bounds for our proposed estimator. We describe the scenarios where our estimator is minimax adaptive. Our finite sample analysis is valid for any n, N_i, p and K, and both p and K are allowed to increase with n, a situation not handled well by previous analyses. We complement our theoretical results with a detailed simulation study. We illustrate that the new

algorithm is faster and more accurate than the current ones, although we start out with a computational and theoretical disadvantage of not knowing the correct number of topics K, while we provide the competing methods with the correct value in our simulations.

Graphons and graphexes as limits and models for sparse graphs

Christian Borgs

Traditionally, the limit theory of dense graphs is cast in terms of convergence of homomorphism densities, corresponding to subgraph frequencies on one side, and weighted multi-way cuts (or, equivalently, ground state energies of statistical physics models) on the other. By contrast, in this talk I stress the relationship to subsampling and the Aldous Hover Theorem, an approach which gives a tight conceptional connection between graph limits and models of dense, inhomogeneous random graphs (often called exchangeable random graphs in the statistics literature).

One approach to generalize this theory to sparse graphs is via rescaling [1, 2]: if we are concerned with graph limits, we rescale the adjacency matrix by dividing by the edge density before taking the limit, and conversely, if we want to generate a random graph from a given graphon, we multiply the graphon by the target density, leading to what is known as inhomogeneous random graphs with a given target density and their estimation [3, 4, 5].

But the main focus of this talk will be on a different approach to limits and models for sparse graphs, based again on the notion of sampling. However, in contrast to the dense theory, where the natural notion of sampling consists of choosing a fixed number k of vertices at random and then outputting the induces subgraph, for sparse graphs we choose a random number of vertices which grows with the number of vertices in the graph we sample from. Specifically, we will take k to be a Poisson random variable with expectation $t/\sqrt{\rho}$ where t is a parameter and ρ is the edge density, then sample k vertices i.i.d. uniformly at random from the vertex set, and finally output the induces subgraph after stripping it of isolated edges and all labels[1]. Sampling convergence of a sequence of graphs G_n is then defined by requiring that the distribution of random graphs of G_n thus obtained is convergent in distribution for all t [6].

This notion of convergence parallels the notion of left convergence for dense graphs and its relationship to the Aldous Hover Theorem, and provides a dual view of sparse graph limits as processes and random measures, an approach which allows for the generalization of many of the well-known results and techniques for dense graph sequences to sparse graph sequences [7, 6, 8]. In contrast to the rescaled theory of sparse graph convergence, whose natural limit objects are unbounded graphons over a probability space, the natural limit objects in this new theory are graphons over sigma-finite measure spaces, and, more generally, their

[1]With this choice of k, the expected number of edges in the induced subgraph is equal to $t^2/2$ uniformly in the sparsity of the graph we sampled from.

extension to graphexes, which form the completion of the space of sparse graphs under sampling convergence.

References

[1] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. Preprint, arXiv:1401.2906, 2014.

[2] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *Ann. Probab., to appear*, 2018. Preprint available at arXiv:1408.0744.

[3] C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. Preprint, arXiv:1508.06675, 2015.

[4] C. Borgs, J. T. Chayes, and A. Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems 28*, pages 1369–1377, 2015. Extended version with proofs available at arXiv:1506.06162.

[5] C. Borgs, J. T. Chayes, A. Smith, and I. Zadit. On privately estimating graphs. In *59th Annual IEEE Symposium on Foundations of Computer Science*, 2018 (to appear).

[6] C. Borgs, J. T. Chayes, H. Cohn, and V. Veitch. Sampling perspectives on sparse exchangeable graphs. Preprint, arXiv:1708.03237, 2017.

[7] C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research, to appear*, 2018. Preprint available at arXiv:1601.07134.

[8] C. Borgs, J. T. Chayes, H. Cohn, and L.M. Lovász. Identifyability for graphexes and the weak kernel metric. Preprint, in preparation, 2018.

Functional data analysis in the Banach space of continuous functions

Holger Dette

(joint work with K. Kokot and A. Aue)

Most of functional data analysis is based on Hilbert space-based methodology for which there exists by now a fully fledged theory. Since all functions utilized for practical purposes are at least continuous, and often smoother we develop in this talk methodology for functional data in the space of continuous functions.

We concentrate on the spce $C(T)$, the space of continuous functions on the compact interval $T = [0, 1]$ equipped with the sup-norm $\|f\| = \sup_{t\in T} |f(t)|$. If μ_1 and μ_2 are the mean functions corresponding to two samples we are interested in testing hypotheses of the form

$$H_0 \colon \|\mu_1 - \mu_2\| \leq \Delta \qquad \text{and} \qquad H_1 \colon \|\mu_1 - \mu_2\| > \Delta, \tag{1}$$

where $\Delta \geq 0$ denotes a pre-specified constant. The classical case of testing perfect equality, obtained by the choice $\Delta = 0$, is therefore a special case of (1). It turns out that from a mathematical point of view the problem of testing relevant (i.e., $\Delta > 0$) hypotheses is substantially more difficult than the classical problem (i.e., $\Delta = 0$). In particular, it is not possible to work with stationarity under the null hypothesis, making the derivation of a limit distribution of a corresponding test statistic or the construction of a bootstrap procedure substantially more difficult. If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent stationary samples in $C([0, 1])$

with mean functions μ_1 and μ_2 a test for the hypotheses (1) can be based on the statistic

$$\hat{d}_\infty = \|\bar{X}_m - Y_n\|,$$

which is a natural estimate of the unknown distance

$$d_\infty = \|\mu_1 - \mu_2\|.$$

One of our main results establishes the asymptotic distribution of $\hat{d}_\infty$.

Theorem 1. *If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are sampled from independent stationary time series $(X_j : j \in \mathbb{N})$ and $(Y_j : j \in \mathbb{N})$ in $C([0,1])$ with mean functions μ_1 and μ_2, respectively. satisfying the conditions*

(A1) There is a constant K such that $\mathbb{E}[\|X_j\|^{2+\nu}] \leq K$ and $\mathbb{E}[\|Y_j\|^{2+\nu}] \leq K$ for some $\nu > 0$.

(A2) There is a real-valued random variable M with $\mathbb{E}[M^2] < \infty$ such that

$$|X_j(t) - X_j(t')| \leq M|t - t'| \ , \ j = 1, \ldots, m$$
$$|Y_j(t) - Y_j(t')| \leq M|t - t'| \ , \ j = 1, \ldots, n$$

holds almost surely for all $t, t' \in T$.

(A3) $(X_j : j = 1, \ldots, m)$ and $(Y_j : j = 1, \ldots, n)$ are φ-mixing with exponentially decreasing mixing coefficients, that is, there is a constant $a \in [0, 1)$ such that $\varphi(k) \leq a^k$ for any $k \in \mathbb{N}$.

then as $m, n \to \infty$ and $m/(m+n) \to \lambda \in (0,1)$

$$(2) \quad T_{m,n} = \sqrt{n+m}(\hat{d}_\infty - d_\infty) \xrightarrow{\mathcal{D}} T(\mathcal{E}) = \max\Big\{ \sup_{t \in \mathcal{E}^+} Z(t), \sup_{t \in \mathcal{E}^-} -Z(t) \Big\},$$

where the centered Gaussian process Z with covariance structure

$$C(s,t) = \mathrm{Cov}(Z(s), Z(t)) = \frac{1}{\lambda} C_1(s,t) + \frac{1}{1-\lambda} C_2(s,t),$$

$$C_1(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(X_1(s), X_{1+i}(t)) \ , \ C_2(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(Y_1(s), Y_{1+i}(t))$$

and the sets $\mathcal{E}^+$ and $\mathcal{E}^-$ are defined in

$$\mathcal{E}^\pm = \big\{ t \in [0,1] : \mu_1(t) - \mu_2(t) = \pm d_\infty \big\}$$

As the asymptotic distribution in Theorem 1 is not distribution free, we also develop estimates of the extremal sets $\mathcal{E}^+$ and $\mathcal{E}^-$ which are defined by

$$\hat{\mathcal{E}}^+_{m,n} := \Big\{ t \in [0,1] \ \Big| \ \bar{X}_m(t) - \bar{Y}_n(t) \geq \hat{d}_\infty - c\frac{\log(m+n)}{\sqrt{m+n}} \ \Big\}$$

$$\hat{\mathcal{E}}^-_{m,n} := \Big\{ t \in [0,1] \ \Big| \ \bar{X}_m(t) - \bar{Y}_n(t) \leq -\hat{d}_\infty + c\frac{\log(m+n)}{\sqrt{m+n}} \ \Big\}$$

The second main result established consistency of these estimates with respect to the Hausdorff distance d_H.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied, then*

$$d_H(\hat{\mathcal{E}}^{\pm}_{m,n}, \mathcal{E}^{\pm}) \xrightarrow{\mathbb{P}} 0.$$

The results of Theorem 1 and 2 are used to develop a multiplier bootstrap procedure, to generate critical values for the test, which rejects the null hypotheses for large vales the statistic $\hat{d}_\infty$. We prove that this test has asymptotic level α and is consistent.

How well do local algorithms solve semidefinite programs?

Zhou Fan
(joint work with Andrea Montanari)

Semi-definite programming (SDP) relaxations are among the most powerful tools available to the algorithm designer. However, several probabilistic models such as planted clique and planted partition reveal an intriguing dichotomy. Either simple local algorithms succeed in estimating the object of interest, or even sophisticated SDP relaxations fail. The conjectural picture emerging in many problems is that SDP relaxations are no more powerful than local algorithms (supplemented with a small amount of side information to break symmetries), and any information that is genuinely non-local is not exploited even by sophisticated SDP hierarchies. To explore this phenomenon, we ask the question:

Can semidefinite programs be (approximately) solved by local algorithms for a large class of random graph models?

We study this in the context of a classical SDP relaxation of the minimum graph bisection problem, when applied to Erdős-Rényi random graphs and stochastic block models with bounded average degree. We show that for this problem, near-optimal SDP solutions may be constructed using a local algorithms approach.

We consider specifically the *two-groups symmetric stochastic block model* that has attracted considerable attention in recent years as a model for community detection in networks. A random graph G over n vertices is generated by partitioning the vertex set into two subsets $S_+ \cup S_-$ of size $n/2$ uniformly at random. Conditional on this partition, any two vertices i, j are connected by an edge independently with probability a/n if $\{i,j\} \subseteq S_+$ or $\{i,j\} \subseteq S_-$, and with probability b/n otherwise. Given a single realization of G, we are requested to identify the partition.

We focus on the success criterion of weak recovery, i.e., to attribute $\{+,-\}$ labels to the vertices so that at least $(1/2+\varepsilon)n$ vertices are labeled correctly with high probability. It was conjectured in [DKMZ11] that this is possible if and only if $\lambda > 1$, where $\lambda \equiv (a-b)/\sqrt{2(a+b)}$ parametrizes an effective signal-to-noise ratio. This conjecture followed from the heuristic analysis of a local algorithm based on belief propagation, and was subsequently proven in [MNS12, MNS13, Mas14]

through the analysis of spectral algorithms related to the linearization of belief propagation around a non-informative fixed point.

Convex optimization approaches for this problem are based on the classical SDP relaxation of the minimum-bisection problem. Denoting by $\boldsymbol{A} = \boldsymbol{A}_G$ the adjacency matrix of G, the minimum bisection problem is written as

$$\text{maximize} \quad \langle \boldsymbol{\sigma}, \boldsymbol{A}\boldsymbol{\sigma} \rangle \,, \tag{1}$$

$$\text{subject to} \quad \boldsymbol{\sigma} \in \{+1, -1\}^n \,, \quad \langle \boldsymbol{\sigma}, \mathbf{1} \rangle = 0 \,. \tag{2}$$

The following SDP relaxes the above problem, where $d = (a+b)/2$ is the average degree:

$$\begin{aligned} &\text{maximize} \quad \langle \boldsymbol{A} - (d/n)\mathbf{1}\mathbf{1}^{\mathsf{T}}, \boldsymbol{X} \rangle \,, \\ &\text{subject to} \quad \boldsymbol{X} \succeq 0 \,, \quad \boldsymbol{X}_{ii} = 1 \ \ \forall i \,. \end{aligned} \tag{3}$$

Here, the term $-(d/n)\mathbf{1}\mathbf{1}^{\mathsf{T}}$ is a Lagrangian relaxation of the hard constraint $\langle \boldsymbol{\sigma}, \mathbf{1} \rangle = 0$. This SDP relaxation has a weak recovery threshold λ^{SDP} that appears to be very close to the ideal one $\lambda = 1$. Namely, Guédon and Vershynin [GV16] proved $\lambda^{\mathrm{SDP}} \le C$ for C a universal constant, while [MS16] established $\lambda^{\mathrm{SDP}} = 1 + o_d(1)$ for large average degree d.

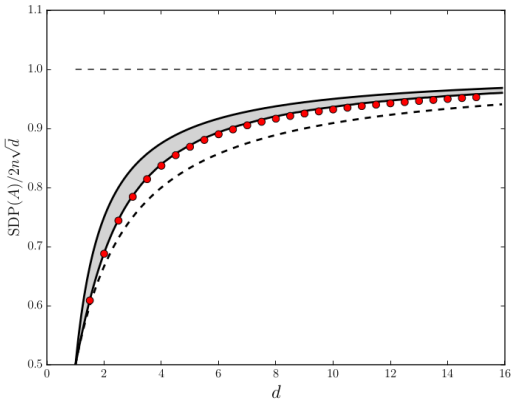

FIGURE 1. Typical value $\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}})$ of the min-bisection SDP for large Erdős-Rényi random graphs with average degree d, normalized by the large degree formula $2n\sqrt{d}$. Circles: Numerical simulations with graphs of size $n = 10^6$. Solid lines: Upper bound $1 - 1/(2d)$ and random-conductance lower bound for the SDP value. Lower dashed line: Explicit lower bound $1 - 1/(d+1)$ for the SDP value.

Denoting by $\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}})$ the value of (3), we establish the following results:

Approximation ratio of local algorithms.: For the Erdős-Rényi random graph $G \sim \mathsf{G}(\mathsf{n}, \mathsf{d}/\mathsf{n})$ with average degree d, we prove that there exists a simple local algorithm that approximates $\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}})$ within a factor $2d^2/(2d^2 + d - 1)$,

asymptotically for large n. In particular, the local algorithm is at most a factor $8/9$ suboptimal, and $1 + O(1/d)$ suboptimal for large degree.

Typical SDP value.: Our proof provides upper and lower bounds on $\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}})$ for $G \sim \mathsf{G}(\mathsf{n}, \mathsf{d}/\mathsf{n})$, showing in particular

$$2\sqrt{d}\left(1 - \frac{1}{d+1}\right) \leq \frac{1}{n}\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}}) \leq 2\sqrt{\mathsf{d}}\left(1 - \frac{1}{2\mathsf{d}}\right).$$

This may be compared with the "Wigner heuristic" for the maximum eigenvalue $2\sqrt{d}$ of the matrix $\boldsymbol{A} - (d/n)\mathbf{1}\mathbf{1}^{\mathsf{T}}$ for dense graphs. While the lower bound is based on the analysis of a local algorithm, the upper bound follows from a dual witness construction using a generalization of the Ihara-Bass identity and a centered variant of the non-backtracking matrix of the graph. Our upper and lower bounds are plotted in Fig. 1 together with the results of numerical simulations.

A local algorithm based on harmonic measures.: The simple local algorithm above aggregates randomness available at each vertex of G uniformly within a neighborhood of that vertex. We analyze a different local algorithm that aggregates information in proportion to the harmonic measure of each vertex, and we characterize the value achieved by this algorithm in terms of the conductance of a random Galton-Watson tree. Numerical data (obtained by evaluating this value and also solving the SDP (3) on large random graphs) suggest that this lower bound is very accurate, cf. Fig. 1.

SDP detection threshold for the stochastic block model.: We then turn to the two-group symmetric block model $G \sim \mathsf{G}(\mathsf{n}, \mathsf{a}/\mathsf{n}, \mathsf{b}/\mathsf{n})$ and prove a local algorithms lower bound for $\mathsf{SDP}(\boldsymbol{A}_{\mathsf{G}})$ in this model, assuming that a small constant fraction of community vertex labels are revealed as a device for breaking symmetries. Our results imply, in particular, that SDP succeeds for weak recovery when $\lambda > \lambda^{\mathrm{SDP}}$, where

$$\lambda^{\mathrm{SDP}} \leq \min\left(2 - \frac{1}{d}, 1 + \frac{C}{d^{1/4}}\right)$$

for C a universal constant (independent of d).

References

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[GV16] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3–4):1025–1049, 2016.

[Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703, 2014.

[MNS12] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. arXiv:1202.1499, 2012.

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. arXiv:1311.4115, 2013.

[MS16] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 814–827, 2016.

Testing for Balance in Social Networks

Derek Feng

(joint work with Randolf Altmeyer, Derek Stafford, Nicholas A. Christakis and Harrison H. Zhou)

Models of social network structure generally build on assumptions about myopic agents, whereby global network features emerge from the dynamic local decision rules of individual agents. For instance, if agents tend to attach to more central or popular actors, scaling emerges in the degree distribution of the graph; if people generally form connections with those who are similar, social networks exhibit homophily; if agents form infrequent but random connections with other agents, the social graph has a small diameter, following the small-world phenomenon.

All of these models, however, are restrictive in that they only apply to positive ties. Much less is theorized or known about the fundamental properties of negative ties. In principle, they need not share the same structural properties as their positive counterparts. Moreover, as most social graphs are *signed* (i.e. have both positive and negative ties), this raises the question of how the presence of the negative ties affects the surrounding positive network structure, and how we should model them concurrently.

One important theory of negative ties advanced by Heider relates to an agent's desire for balance in social relationships. Balance theory postulates that a need for cognitive consistency leads agents to seek to balance the valence in their local social systems. Simply stated, friends should have the same friends and the same enemies. This translates, in graph-theoretic terms, to requiring the product of the signs on a triangle to be positive. Triangles that violate this property are deemed *unbalanced*, and the theory posits that such triangles should be rare compared to their *balanced* counterparts.

As it stands, balance theory has received surprisingly little empirical evaluation. Tests of balance theory require the observation of antagonistic connections between actors, but these ties are often either ignored when the data is gathered, or simply unavailable due to the unwillingness of the actors themselves to divulge such information. Those studies which have been able to observe antagonistic ties have done so in artificial settings – and been very liberal about what constitutes an antagonistic tie – like nominations to adminship on Wikipedia, and user ratings of trustworthiness in an e-commerce website, rather than in face-to-face settings, with few exceptions.

Though the underlying datasets may be vastly different, these studies all resort to exactly the same statistical test to verify balance in their signed networks: for

the test statistic, they use the number of balanced triangles as a measure of the degree of balance in a graph; the null model corresponds to a permutation test on the edge weights of the observed graph. Drawing samples from the null distribution then reduces to shuffling the signs on the graph. The simplicity of this null model belies its principal flaw though – namely, that it treats negative and positive ties as interchangeable. The problem is that, as we shall soon demonstrate, negative ties behave remarkably like random ties drawn from an Erdős-Rényi graph. Meanwhile, research in social network modeling is predominantly focused on showing how *disparate* positive graphs are from independent random graphs. Features like preferential attachment and clustering are fundamental to our understanding of positive ties – features that are clearly absent in negative ties. Thus, by treating positive and negative ties as exchangeable, this null hypothesis creates a test, not for balance, but for differences in the behavior of positive and negative ties.

The main contribution of this work is to provide a new null model that resolves the issues raised above. The crux of the solution is the following key observation: a crucial way in which negative and positive ties differ is through their *embeddedness* level (the number of triangles that tie is a member of) – transitivity and homophily encourage higher levels of embeddedness in positive ties. Our new method, therefore, is to stratify the permutation across embeddedness levels, thereby ensuring that the embeddedness profiles of negative and positive ties remain invariant. This preserves the fundamental differences between the two kinds of ties, creating a more accurate null model of a signed social network without balance. This is supported by both our simulation studies and our theoretical results, where we show that for a reasonable definition of absence of balance in a graph, the true type-I error rate of the old test converges to 1 while the type-I error rate for the new test is consistent with the specified α.

To compare the relative performance of the two tests, we show asymptotic normality of the test statistic under the two null models. Due to the stratified nature of the permutation, this is a nontrivial result, and, to the best of our knowledge, this is the first result showing asymptotic normality of this type of graph statistic under a stratified permutation model. The key insight is that a distribution derived from a permutation test – even a stratified permutation – can be obtained as conditional distribution of independent random variables. This is similar to the dichotomy between the $G(n, p)$ and $G(n, m)$ random graph model. Under certain conditions, the limit and the conditioning operation may be interchanged, enabling us to carry the central limit theorem result in the independent case to the permutation case. This proof technique of Janson has wide applicability, not least in the nascent field of (nonparametric) inference on random graphs.

Our final contribution is that we are the first to collect and analyze a comprehensive dataset capturing both positive and negative ties between individuals in a social network – namely, the networks of 32 villages in rural Honduras. This novel dataset provides a first look into the behavior of interaction between negative and positive human relationships. We find that, unsurprisingly, negative ties behave very differently from positive ties. Applying our new test of balance to the village

networks reveals that balance barely registers as an underlying mechanism dictating the structure of signed networks, which is contrary to the conclusions drawn from the previous literature.

Partial recovery bounds for clustering with the relaxed Kmeans

Christophe Giraud
(joint work with Nicolas Verzelen)

The problem of clustering is that of grouping similar 'objects' in a data set. It encompasses many different instances such as partitioning points in a metric space, or partitioning the nodes of a graph.

Kmeans and a convex relaxation

When these objects can be represented as vectors in a Euclidean space, some of the most standard clustering approaches are based on the minimization of the Kmeans criterion [20]. Observing n objects and writing $X_a \in \mathbf{R}^p$ for the object $a \in \{1, \dots, n\}$, the Kmeans criterion of a partition $G = (G_1, \dots, G_k)$ of $\{1, \dots, n\}$ is defined as

$$\text{Crit}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \left\| X_a - \frac{1}{|G_k|} \sum_{b \in G_k} X_b \right\|^2 , \tag{1}$$

where $\|.\|$ stands for the Euclidean norm. This criterion quantifies the dispersion of each group around its centroid in order to favor homogeneous partitions. A Kmeans procedure then aims at finding a partition $\widehat{G}$ that minimizes, at least locally, the Kmeans criterion. However, solving this problem is NP-hard and it is even hard to approximate [2].

In general, iterative procedures such as Llyod's algorithm [20] and its variants [5] are only shown to converge to a local minimum of the Kmeans criterion. Alternatively, Peng and Wei [25] have suggested to relax the Kmeans criterion to a Semi-Definite Program (SDP) followed by a rounding step. The resulting program is provably solvable in polynomial time. In this talk, we (i) put forward the versatility of Peng and Wei's procedure and some of its variants by handling both vector and general graph clustering problems and (ii) explain its near-optimal performances.

SubGaussian Mixture Models (sGMM) and Stochastic Block Models (SBM)

In the computer-science and statistical literature, the most popular approach to assess the performances of a procedure is the 'model-based' strategy. It assumes there exists a true unknown partition G of the 'objects' and that the data have been randomly generated from a probability distribution rendering this partition. Then, one can assess the performances of a clustering procedure by comparing the partition estimated from the data to G.

For vector clustering, it is classical to assume that the vectors X_a are distributed according to a SubGaussian Mixture Model (sGMM). In a sGMM with partition

G, the random variables X_a are assumed to be independent and for $a \in G_k$, the random variable X_a is assumed to follow a subGaussian distribution centered at $\mu_k \in \mathbb{R}^p$ and with covariance matrix Σ_k. In other words, variables X_a whose indices a belongs to the same group are identically distributed and variables X_a and X_b whose indices belong two different groups have different means.

Node clustering in a network has been widely investigated within the framework of Stochastic Block Models (SBM) [17] and its variants. According to a SBM with partition G, the network edges are sampled independently and the probability of presence of an edge between any two nodes $a \in G_k$ and $b \in G_l$ is equal to some quantity $Q_{kl} \in [0,1]$ only depending on the group. In other words, two nodes a and b belonging to the same group in G share the same probability of being connected to any other node c.

These two random models have attracted a lot of attraction in the last decade. See e.g. [1, 22] for two recent reviews on SBM and [9, 24, 21, 27] for recent contributions on sGMM. A large body of the literature on these two models focuses on pinpointing the right scaling between the model parameters allowing to recover the partition G from the data. For sGMM, this translates into identifying the minimal distance $\min_{k\neq l} \|\mu_k - \mu_l\|$ within the mixtures means, such that, there exists a clustering procedure, if possible running in polynomial time, that recovers G with high probability. Most of the works concentrate on two types of recovery: perfect recovery, where one wants to recover exactly the partition G with high probability and weak-recovery where the estimated partition $\widehat{G}$ is only required to be more accurate than random guessing. The goal is then to identify the precise threshold at which perfect or weak recovery can occur. We refer to [1] for a review of these questions in SBM. Between these two extreme regimes, when the best possible classification is neither perfect nor trivial, the objective is to maximize the proportion of well-classified data. Given two partitions $\widehat{G} = (\widehat{G}_1, \dots, \widehat{G}_K)$ and $G = (G_1, \dots, G_K)$ of $\{1, \dots, n\}$ into K non-void groups, we define the proportion of non-matching points

$$err(\widehat{G}, G) = \min_{\pi \in \mathcal{S}_K} \frac{1}{2n} \sum_{k=1}^{K} \left| G_k \triangle \widehat{G}_{\pi(k)} \right|, \tag{2}$$

where $A \triangle B$ represents the symmetric difference between the two sets A and B and $\mathcal{S}_K$ represents the set of permutations on $\{1, \dots, K\}$. When $\widehat{G}$ is a partition estimating G, we refer to $err(\widehat{G}, G)$ as the misclassification proportion (or error) of the clustering. The problem of minimizing this error has attracted less attention but see [4, 15, 10, 28, 3, 12, 9, 14, 13] for some related contributions.

Among the polynomial-time clustering procedures, Semi-Definite-Programs (SDP) have proved to be versatile and they have been investigated in a large range of clustering problems, including clustering in SBM [11, 15, 26, 19, 18, 13], sGMM [9, 24, 27] or in block covariance models [7, 8]. While not always reaching the exact threshold for weak/perfect clustering in several cases [19, 26], SDP algorithms are versatile enough in order to enjoy some robustness properties [26, 23, 13], which are not met by more specialized algorithms (see [23] for more details). However,

most SDPs require the partition to be balanced or that, at least, the size of each group is known in advance. Besides, all SDPs studied for SBM clustering arise as convex relaxations of min-cut optimization problems [11, 15, 26, 19, 18, 13] and therefore only fall within the framework of assortative SBM where within group probabilities of connection are larger than between group probabilities of connection. In other words, the diagonal entries of Q have to be larger than its off-diagonal entries.

Our Contribution

We provide misclassification error bounds for the relaxed Kmeans of Peng and Wei [25] both in the sGMM and the SBM frameworks. Compared to other SDPs, this convex relaxation of Kmeans has the nice feature to only require the knowledge of the number of groups (which can sometimes be estimated, as in [7]). Hence, there is no need to know the size of the clusters, nor the parameters of the model.

Let us give a glimpse of our results on sGMM, by specifying it to the special case of Gaussian mixture models, with K groups of equal size $m = n/K$ and equal covariance Σ. The general statement of the results for possibly unbalanced groups in sGMM is available in [16]. Write $\Delta = \min_{k\neq l} \|\mu_k - \mu_l\|$ for the minimal Euclidean distance between the means of the components and write $R_\Sigma = |\Sigma|_F^2/|\Sigma|_{op}^2$ for the ratio between the square Frobenius norm of Σ and the the square operator norm of Σ. This ratio can be interpreted as an effective rank of Σ. In the sequel, c stands for a positive numerical constant. Then, with high probability, when applying the relaxed Kmeans, the proportion of misclassified observations decreases exponentially fast with the signal to noise ratio

$$s^2 = \frac{\Delta^2}{|\Sigma|_{op}} \wedge \frac{n\Delta^4}{K|\Sigma|_F^2}, \tag{3}$$

at least, as long as the condition $s^2 \geq cK$, or equivalently

$$\Delta^2 \geq c|\Sigma|_{op}\left(1 \vee \sqrt{\frac{R_\Sigma}{n}}\right)K \tag{4}$$

is met. Since $err(\widehat{G}, G) \leq 1/n$, implies that the partition $\widehat{G}$ is equal to G, this result ensures perfect recovery of the clustering with high-probability when $s^2 \geq c(K \vee \log(n))$, recovering the results of [27]. It also ensures a better than random guess clustering when (4) is met, which improves, in high-dimensional setting, upon state-of-the art results in [24, 21]. We also point out that the right SNR to look at in this setting is s^2 defined by (3) and not $\tilde{s}^2 = \Delta^2/|\Sigma|_{op}$ previously considered in the literature.

On the SBM side, the relaxed Kmeans procedure can be applied to general SBM to cluster nodes presenting similar connectivity profiles. Instead of the previously discussed SDPs that look for a partition with maximal within-group connectivity, this allows us to handle general unknown connection matrices Q and thereby going far beyond the assortative case. Denoting by m the size of the smallest group in G, we prove that, with high probability, the misclassification proportion decreases

exponentially fast with the signal-to-noise ratio

$$s^2 = m \cdot \min_{j \neq k} \frac{\|Q_{j:} - Q_{k:}\|^2}{|Q|_\infty}, \tag{5}$$

at least as long as the condition $s^2 \geq cn/m$ is met. Here, $Q_{j:}$ stands for the j-th column of Q and $|Q|_\infty$ denotes the supremum norm. Note that this result encompasses sparse graph, where the connection probability may scale as a constant divided by n. Our results are (i) the first results about clustering with an SDP in non-assortative cases and (ii) the only known exponential bounds for partial recovery in *general* SBM are those of [3] which handle the sparse setting where the matrix Q scales as $Q = Q_0/n$, with Q_0 fixed and $n \to \infty$. Their results are optimal in the vicinity of the weak recovery threshold. Our results cover a setting with slightly more signal (the missclassification error has to be smaller than e^{-cK}), and the results do not overlap. In particular, our exponential rate (5) is faster by at least a factor K than the exponential rate involved in [3]. Hence both results are more complementary than comparable. When specified to the classical case with all within-group probabilities equal to p and all between-group probabilities equal to q, with $q < p$, and groups with the same size, we recover the results obtained by [13] for a relaxed version of the MLE, but without knowledge of the group sizes.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *ArXiv e-prints*, March 2017.

[2] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The Hardness of Approximation of Euclidean k-means. *arXiv preprint arXiv:1502.03316*, 2015.

[3] Emmanuel Abbe and Colin Sandon. Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Algorithms for Recovery. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 670–688, Washington, DC, USA, 2015. IEEE Computer Society.

[4] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

[5] David Arthur and Sergei Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[6] F. Bunea, C. Giraud, X. Luo, M. Royer, and N. Verzelen. Model Assisted Variable Clustering: Minimax-optimal Recovery and Algorithms. *ArXiv e-prints*, August 2015.

[7] Florentina Bunea, Christophe Giraud, Martin Royer, and Nicolas Verzelen. PECOK: a convex optimization approach to variable clustering. *arXiv preprint arXiv:1606.05100*, 2016.

[8] Q. Berthet, P. Rigollet, and P. Srivastava. Exact recovery in the Ising blockmodel. *Annals of Statistics (to appear)*, page arXiv:1612.03880, 2018.

[9] S. Chrétien, C. Dombry, and A. Faivre. A Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering. *ArXiv e-prints*, June 2016.

[10] Peter Chin, Anup Rao, and Van Vu. Stochastic Block Model and Community Detection in Sparse Graphs: A spectral algorithm with optimal rate of recovery. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 391–423, Paris, France, 03–06 Jul 2015. PMLR.

[11] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57, 2016.
[12] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the binary stochastic block model. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 185–189, July 2016.
[13] Y. Fei and Y. Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *ArXiv e-prints*, 2017.
[14] Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, 18(1):1980–2024, January 2017.
[15] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *arXiv preprint arXiv:1411.4686*, 2014.
[16] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed Kmeans. *arXiv preprint arXiv:1807.07547*, 2018.
[17] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
[18] B. Hajek, Y. Wu, and J. Xu. Semidefinite Programs for Exact Recovery of a Hidden Community. *ArXiv e-prints*, February 2016.
[19] Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.
[20] S. Lloyd. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 1982.
[21] Y. Lu and H. H. Zhou. Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *ArXiv e-prints*, December 2016.
[22] Cristopher Moore. The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. *CoRR*, abs/1702.00467, 2017.
[23] Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841. ACM, 2016.
[24] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.
[25] Jiming Peng and Yu Wei. Approximating K-means-type Clustering via Semidefinite Programming. *SIAM J. on Optimization*, 18(1):186–205, February 2007.
[26] A. Perry and A. S. Wein. A semidefinite program for unbalanced multisection in the stochastic block model. *ArXiv e-prints*, July 2015.
[27] M. Royer. Adaptive Clustering through Semidefinite Programming. *Advances in Neural Information Processing Systems (NIPS)*, 2017.
[28] Se-Young Yun and Alexandre Proutière. Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms. *CoRR*, abs/1412.7335, 2014.

Statistical estimation for age-structured models in a large population limit

Marc Hoffmann
(joint work with A. Boumezoued and P. Jeunesse)

Motivated by improving mortality tables from human demography databases, we investigate statistical inference of a stochastic age-evolving density of a population alimented by time inhomogeneous mortality and fertility. Asymptotics are taken as the size of the population grows within a limited time horizon: the observation gets closer to the solution of a PDE (a inhomogeneous version of the Von Foerster Mc Kendrick equation) and the difficulty lies in controlling simultaneously the stochastic approximation to the limiting PDE in a suitable sense together with an appropriate parametrisation of the anisotropic solution. In this setting, we prove new concentration inequalities that enable us to implement the Goldenshluger-Lepski algorithm and derive oracle inequalities. Minimax adaptation under local Hölder smoothness constraints are also investigated.

Optimal hypothesis testing for stochastic block models with growing degrees

Zongming Ma
(joint work with Debapratim Banerjee)

We consider optimal hypothesis testing for distinguishing a stochastic block model from an Erdős–Rényi random graph. Let A denote an $n \times n$ adjacency matrix of a random undirected graph, we are interested in testing

$$H_0 : A \sim \mathcal{G}_1\left(n, \frac{p_n + q_n}{2}\right) \quad \text{vs.} \quad H_1 : A \sim \mathcal{G}_2\left(n, p_n, q_n\right).$$

Here the null model is Erdős–Rényi random graph with edge probability $(p_n+q_n)/2$ and the alternative the balanced stochastic block model with two blocks where the within block connection probability is p_n and the between block connection probability is q_n. For simplicity assume that $p_n > q_n$, though all the results generalize naturally to the case of $p_n < q_n$. We are interested in the asymptotic regime where $n \to \infty$, $p_n, q_n \to 0$, $n(p_n + q_n) \to \infty$ while

$$t = \sqrt{\frac{n(p_n - q_n)^2}{2(p_n + q_n)}}$$

remains a constant. We derive central limit theorems for a collection of linear spectral statistics under both the null and local alternatives. In addition, we show that linear spectral statistics based on Chebyshev polynomials can be used to approximate signed cycles of growing lengths which in turn determine the likelihood ratio

test asymptotically when the graph size and the average degree grow to infinity together. For example, let $\hat{p}_{n,\mathrm{av}} = \frac{1}{n(n-1)} \sum_{i \neq j} A_{ij}$,

$$A_{\mathrm{res}} = \left(\frac{A_{ij} - \hat{p}_{n,\mathrm{av}} \mathbf{1}_{i \neq j}}{\sqrt{n \hat{p}_{n,\mathrm{av}} (1 - \hat{p}_{n,\mathrm{av}})}} \right),$$

and $P_m(x) = 2T_m(x/2)$ for $m = 0, 1, 2, \ldots$, where T_m is the Chebyshev polynomial of degree m. We showed that if as $n \to \infty$, $np_n^2 \to \infty$ and $t \in (0, 1)$ is known, then for any $k_n \to \infty$ such that $k_n = o(\min(\sqrt{\log n}, \log(np_n^2)))$, a test that rejects for large values of

$$L_a = \sum_{i=3}^{k_n} \frac{t^i}{2i} \mathrm{Tr}(P_i(A_{\mathrm{res}}))$$

achieves the same asymptotic power as the likelihood ratio test which is optimal by the Neyman–Pearson lemma. Therefore, one achieves sharp asymptotic optimal power of the testing problem within polynomial time complexity provided that the average degree grows sufficiently fast.

References

[1] D. Banerjee & Z. Ma. *Optimal hypothesis testing for stochastic block models with growing degrees*, arXiv preprint arXiv:1705.05305 (2017).

Nonparametric inference for continuous-time event counting and link-based dynamic network models

Enno Mammen

(joint work with Alexander Kreiß and Wolfgang Polonik)

In this talk we consider a model for a time series of random networks. We denote by $V_n = \{1, \ldots, n\}$ the set of nodes (actors, agents) of the network. We will discuss models for two types of data:

- Counting of interaction events:

  $$N_{n,ij}(t) = \#\{\text{interaction events between } i \text{ and } j \text{ before or at } t\}$$

  for a pair of actors (i, j). Examples for this model are data of e mails or of phone calls between persons $i, j \in V_n$.
- Dynamic networks:

  $$Z_{n,ij}(t) = \begin{cases} 1, & \text{if there is a link between } i \text{ and } j \\ 0, & \text{if there is no link} \end{cases}$$

  for a pair of actors (i, j). Here an example is a network of friendships between persons $i, j \in V_n$.

In this note, we only discuss models for interaction events. The results for these models can be easily also applied to dynamic networks. All results are formulated for undirected interactions, i.e., we assume that $N_{n,ij} = N_{n,ji}$ for all pairs (i, j). This assumption is made for simplicity. All results can be formulated for the directed case as well.

Our basic assumption is that for all (i, j), the processes $N_{n,ij}$ are one-dimensional counting processes with respect to an increasing, right continuous filtration $\mathcal{F}_t$, $t \in [0, T]$. The σ-field $\mathcal{F}_t$ contains all information available up to the time point t. The intensities of the counting processes $N_{n,ij}$ are modelled by

$$\lambda_{n,ij}(\theta, t) := \exp(\theta^T(t)Y_{n,ij}(t)), \quad \text{if } C_{n,ij}(t) = 1,$$

for $i, j = 1, ..., n$,

- where $\theta(t)$ is an unknown function,
- where $Y_{n,ij}(t)$ are $\mathcal{F}_t$-predictable co-variates,
- and where the functions $C_{n,ij}(t)$ are $\mathcal{F}_t$-predictable indicator functions only taking values in $\{0, 1\}$.

This model is very flexible:

- We do not assume that θ is a fixed parameter. We allow that it is a parameter that develops in time.
- The variables $Y_{n,ij}(t)$ can contain information about the global and local history of the network and external information.
- Inference will be based on the truncated process $N^C_{n,ij}$ with intensity function

  $$\lambda_{n,ij}(\theta, t) := C_{n,ij}(t) \exp(\theta^T(t)Y_{n,ij}(t))$$

  for $i, j = 1, ..., n$. This means that we do not model the dynamics of $N_{n,ij}$ in case that $C_{n,ij}(t) = 0$. The presence of the function $C_{n,ij}(t)$ enhances the modelling flexibility significantly. For instance, we can model them as being equal to zero for a certain subset of edges (i, j) not possessing a certain property at time t. One example is, to set $C_{n,ij}(t)$ equal to zero, if there was no event between i and j for a certain period. In this case our model is only fitted to "active" pairs.

Our approach is an adaptation from methods of classical survival analysis to network analysis:

- An edge (i, j) compares to a patient i.
- An interaction on the edge (i, j) at time point t compares to the death of a patient i at t.

Our network/interaction model differs from classical survival analysis in the following points:

- – Survival analysis: the intensity/hazard function depends on the history of a patient i.
  – Interaction analysis: the intensity/hazard function depends on the history of the whole network.

- – Survival analysis: the hazard $\lambda_i(\theta,t)=\lambda_0(t)C_{n,i}(t)\exp(\theta Y_{n,i}(t))$ contains a baseline hazard $\lambda_0(t)$
  – Interaction analysis: choice of $Y_{n,ij1}(t)\equiv 1$ and of time-dependent parameters $\theta(t)$ makes introduction of baseline hazard/intensity superfluous.
- – Survival analysis: $C_{n,i}(t)$ is determined by the data: it is equal to zero if patient i is dead at t or if he has left the study at t.
  – Interaction analysis: $C_{n,ij}(t)$ is defined by the statistician: e.g. it is defined to be equal to zero if $\lambda_{n,ij}(\theta,t)$ is very small.

There is an active literature on dynamic networks. This includes discrete time models based on Markov processes, on dynamic latent space models, on dynamic exponential random graph models, on dynamic block models or on dynamic multigroup membership models. Furthermore, time-continuous models have been proposed based on link-based or actor-based continuous-time Markov processes. Our model is most closely related to Butts (2008), and Perry and Wolfe (2013), where also models based on counting processes have been proposed.

For the estimation of θ_0 we use a local likelihood criterion at points t_0:

$$\ell_T(\theta;t_0)=\sum_{(i,j)}\int_0^T\frac{1}{h}K\left(\frac{s-t_0}{h}\right)\log\lambda_{i,j}(\theta,s)dN^C_{i,j}(s)$$
$$-\sum_{(i,j)}\int_0^T\frac{1}{h}K\left(\frac{s-t_0}{h}\right)\lambda_{i,j}(\theta,s)ds.$$

Here, inference is based on the truncated process $N^C_{n,ij}$. The local MLE is defined as

$$\hat\theta(t_0)=\arg\max_{\theta\in\Theta}\ell_T(\theta,t_0),$$

where Θ is the assumed range of the parameter functions. For the estimator $\hat\theta(t_0)$ we have developed asymptotic theory at a fixed point t_0. Our main result is that with probability tending to one, the derivative of the local log-likelihood function $\ell_T(\theta,t_0)$ has a root $\hat\theta_n(t_0)$, satisfying

$$\sqrt{l_nh}\left(\hat\theta_n(t_0)-\theta_0(t_0)-bias+o_P(h^2)\right)\to N(0,\int K(u)^2du\cdot\Sigma^{-1}),$$

in distribution, as $n\to\infty$, where

$$bias=bias_1+bias_2.$$

Here, we have put

$$\begin{aligned} g(\theta,t) &:= \mathbb{E}(e^{\theta^TY_{n,12}(t)}-\theta^TY_{n,12}(t)e^{\theta_0(t)^TY_{n,12}(t)}|C_{n,12}(t)=1)\\ \Sigma &:= \partial_{\theta^2}g(\theta_0(t_0),t_0)\\ l_n &:= \frac{n(n-1)}{2}\mathbb{P}(C_{n,12}(t_0)=1)\end{aligned}$$

The bias terms are defined as follows:

$$\begin{aligned} bias_1 &:= \frac{1}{2}h^2 \int K(u)u^2 du \cdot \Sigma^{-1}\partial_\theta \partial_{t^2} g(\theta_0(t_0), t_0), \\ bias_2 &:= \frac{h}{l_n} 2 \sum_{i,j=1}^{n} \int_0^T \frac{1}{h} K\left(\frac{s-t_0}{h}\right)(1 - C_{n,ij}(t_0))C_{n,ij}(s) \\ &\qquad Y_{n,ij}(s)Y_{n,ij}(s)^T \exp(\theta_0^T(s)Y_{n,ij}(s))\theta_0'(t_0)\frac{t_0 - s}{h}\, \mathrm{d}s. \end{aligned}$$

For more details on the theoretic result, see Kreiß et al. (2017).

In the talk the methods were illustrated by a data set on bike sharing in Washington D.C., where the vertices i are the bike stations and where an interaction on the edge (i, j) is a person renting a bike at station i and returning it at station j or vice versa. As covariates we use the numbers of tours between i and j in the recent past, the numbers of stations to which tours were undertaken from or to i or j and the numbers of joint destinations among others. For details see again Kreiß et al. (2017).

References

[1] C.T. Butts, *A relational event framework for social action*, Sociol. Methodol., **38** (2008), 155–200.

[2] A. Kreiß, E. Mammen, and W. Polonik, *Nonparametric inference for continuous-time event counting and link-based dynamic network models*, eprint arXiv:1705.03830 (2017).

[3] P.O. Perry and P.J. Wolfe, *Point process modelling for directed interaction networks*, J. Royal Statist. Soc. Ser. B, **75** (2013), 821–849.

Strong Gaussian approximation of the mixture Rasch model

Alexander Meister

(joint work with F. Liese and J. Kappus)

We consider the famous Rasch model, which is applied to psychometric surveys when n individuals under test answer m questions. The score is given by a realization of a random binary matrix. Its (j, k)th entry indicates whether or not the answer of the jth person to the kth question is correct. In the mixture Rasch model one assumes that the individuals are chosen randomly from a huge population. We prove that the mixture Rasch model is asymptotically equivalent to a Gaussian observation scheme in Le Cam's sense as n tends to infinity and m is allowed to increase slowly in n. For that purpose we show a general result on strong Gaussian approximation of the sum of independent high-dimensional binary random vectors. As a first application we construct an asymptotic confidence region for the difficulty parameters of the questions. Moreover we discuss nonparametric estimation of the ability density. This talk is based on a joint work with F. Liese (1944-2018) and J. Kappus (Univ. Rostock).

References

[1] F. Liese, A. Meister and J. Kappus, *Strong Gaussian approximation of the mixture Rasch model*, Bernoulli, to appear.

Learning binary latent variable models: A tensor eigenpair approach

Boaz Nadler

(joint work with Ariel Jaffe, Roi Weiss, Yuval Kluger and Shai Carmi)

Latent variable models with hidden binary units appear in various applications. A common model is of the form

$$\mathbf{x} = W^T \mathbf{h} + \sigma \xi$$

where $\mathbf{x} \in \mathbb{R}^m$ is the observed vector, $\mathbf{h} \in \{0,1\}^d$ is the latent unobserved vector with $d \leq m$, $W \in \mathbb{R}^{d \times m}$ is an unknown mixing matrix of rank d, $\xi \in \mathbb{R}^m$ is a zero mean noise vector with i.i.d. entries of unit variance, and σ is the noise level. Given n i.i.d. samples $\mathbf{x}_i$ of this model the problem is to estimate the number of latent variables d and the unknown matrix W. Without making simplifying assumptions, estimating W, in particular in the presence of noise, is a challenging computational problem. In this work we propose a novel spectral approach to this problem, based on the eigenvectors of both the second order moment matrix and third order moment tensor of the observed data. We prove that under mild non-degeneracy conditions, our method consistently estimates the model parameters at the optimal parametric rate. Our tensor-based method generalizes previous orthogonal tensor decomposition approaches, where the hidden units were assumed to be either statistically independent or mutually exclusive. The complexity of our approach is polynomial in the observed dimension m and number of samples n, but exponential in the number of hidden variables d. The reason for the latter is that our approach computes all $O(2^d)$ eigenvectors of a suitable tensor, and then extracts from them only the d relevant ones. An interesting open question is whether there exist statistically consistent procedures to estimate W under general identifiability conditions which have polynomial run-time in d. We illustrate the consistency of our method on simulated data and demonstrate its usefulness in learning a common model for population mixtures in genetics.

References

[1] A. Jaffe, R. Weiss, S. Carmi, Y. Kluger and B. Nadler, *Learning Binary Latent Variable Models: A Tensor Eigenpair Approach*, International Conference on Machine Learning (ICML) **35** (2018).

[2] A. Jaffe, R. Weiss and B. Nadler, *Newton correction methods for computing real eigenpairs of symmetric tensors*, SIAM Journal on Matrix Analysis and Applications (2018), to appear.

Choice of network motif in network analyses

Sofia Olhede
(joint work with Patrick Wolfe)

Designing the basic building blocks of network analysis is complicated by the fact that many networks are not labelled–e.g. when comparing two networks we cannot use statistics that depend on the choice of labelling of nodes [1, 4, 3]. The theory of exchangeable graphs is naturally matched to this lack of labelling [3]. For this reason, when analysing networks, often statistics are therefore used that intrinsically do not depend on the node labelling. Counts of isomorphic copies of shapes present in a graph, or motifs, are a concrete example of this approach. Subsequent analysis are often implemented that is permutation invariant [2].

There are many outstanding questions remaining before counts of motifs can be used routinely to make better supported inferences of networks. The answers to these questions depend on the modelling assumptions that we are ready to make of the underlying network we are seeking to analyse. Understanding has traditionally been developed for Erdos–Renyi graphs. But many real-world graphs have more complex characteristics than Erdos–Renyi graphs, requiring us to develop more general methods.

Some properties of counts of small graphs are already well–known under the assumption of exchangeability [2]. This however, does not answer what small graphs we should be counting. Ideally they should either be related to the inference problem we are seeking to address or how we think the data might have been generated.

We then seek to answer, what makes certain counts more important? Our argument is that those which contribute more or are more numerous, are more important. Starting from simple assumptions we show that eventually if we consider shapes that can be mapped out by "walking" each edge, and having a limited budget of steps, only cycles, trees, and uni–cyclic graphs are of any great significance [5].

We explore this realisation, and its practical utility, by redefining a summary that has been normalized to reflect departures from Erdos–Renyi graphs. We further explore the variation in the graph by using sub sampling methods. This produces a tool able to characterise the tree like properties versus the cycles in the studied networks. A number of well-studied examples demonstrate the practical usage of counts of subgraphs, when renormalized.

References

[1] P. J. Bickel, A. Chen, *A nonparametric view of network models and Newman– Girvan and other modularities*, Proceedings of the National Academy of Sciences of the USA **106** (2009), 21068–21073.

[2] P. J. Bickel, A. Chen, E. Levina, *The method of moments and degree distributions for network models*, The Annals of Statistics **39** (2011), 2280–2301.

[3] P. Diaconis, S. Janson, *Graph limits and exchangeable random graphs*, Rend. Mat. Appl. **28** (2008), 33–61.

[4] L. Lovasz, *Large Networks and Graph Limits*, American Mathematical Society, Providence Rhode Island, (2013).
[5] P.-A. G. Maugis, S. C. Olhede, P. J. Wolfe, *Topology reveals universal features for network comparison*, arXiv:1705.05677.

Estimation and clustering in the Dynamic Stochastic Block Model

Marianna Pensky
(joint work with Teng Zhang)

In the talk, we consider a dynamic network defined as an undirected graph with n nodes with connection probabilities changing in time. Assume that we observe the values $B_{i,j,l} \in \{0,1\}$ of a tensor $B \in \{0,1\}^{n\times n\times L}$ at times t_l where $0 < t_1 < \cdots < t_L = T$. For simplicity, we assume that time instants are equispaced and the time interval is scaled to one, i.e. $t_l = l/L$. Here $B_{i,j,l} = 1$ if a connection between nodes i and j is observed at time t_l and $B_{i,j,l} = 0$ otherwise. We set $B_{i,i,l} = 0$ and $B_{i,j,l} = B_{j,i,l}$ for any $i, j = 1, \cdots n$ and $l = 1, \cdots, L$, and assume that $B_{i,j,l}$ are independent Bernoulli random variables with $\Lambda_{i,j,l} = P(B_{i,j,l} = 1)$ and $\Lambda_{i,i,l} = 0$. In the talk, we examine a Dynamic Stochastic Block Model (DSBM) which can be viewed as a natural extension of the Stochastic Block Model. In a DSBM, all n nodes are grouped into m classes $\Omega_1, \cdots, \Omega_m$, and probability of a connection $\Lambda_{i,j,l}$ is entirely determined by the groups to which the nodes i and j belong at the moment t_l. In particular, if $i \in \Omega_k$ and $j \in \Omega_{k'}$, then $\Lambda_{i,j,l} = G_{k,k',l}$. Here, G is the connectivity tensor at time t_l with $G_{k,k',l} = G_{k',k,l}$. We assume that the connection probabilities as functions of time have low complexity in the sense that they allow a sparse representation in some standard orthonormal basis H. If those functions are smooth, then one can choose H to be the Fourier transfrom while if they can be functions with jumps, a kind of wavelet transfrom will satisfy the requirement. The objective is estimation of the tensor of connection probabilities Λ and clustering of the nodes into m clusters.

In the first part of the talk, in order to construct an estimator of Λ, we derive a penalized least squares estimator $\widehat{\Lambda}$ and show that $\widehat{\Lambda}$ satisfies an oracle inequality and also attains minimax lower bounds for the risk. The estimators constructed in the paper are adaptive to the unknown number of blocks and to the sparsity of the connection probabilities as functions of time. The technique relies on the vectorization of the model and leads to much simpler mathematical arguments than the ones used previously in the stationary set up. In addition, all results are non-asymptotic and allow a variety of extensions.

In the second part of the talk we consider clustering of the DSBM under the assumptions that the connection probabilities, as functions of time, are smooth and that at most s nodes can switch their class memberships between two consecutive time points. We estimate the edge probability tensor by a kernel-type procedure and extract the group memberships of the nodes by spectral clustering. The procedure is computationally viable, adaptive to the unknown smoothness of the functional connection probabilities, to the rate s of membership switching and to

the unknown number of clusters. In addition, it is accompanied by non-asymptotic guarantees for the precision of estimation and clustering.

References

[1] Pensky, M. Dynamic network models and graphon estimation. ArXiv1607.00673
[2] Pensky, M., Zhang, T. Spectral clustering in the dynamic stochastic block model. ArXiv1705.01204

On the reconstruction error of PCA

Markus Reiss
(joint work with Martin Wahl)

We identify principal component analysis (PCA) as an empirical risk minimisation (ERM) problem with respect to the reconstruction error. By applying the usual machinery of ERM techniques in combination with the Davis-Kahan inequality for matrices we are able to bound the excess risk in the reconstruction error of PCA by the minimum of a global rate (of order $n^{-1/2}$ in the sample size n) and a local rate (of order n^{-1}, but depending on a spectral gap condition). These results extend or complement previous result by Blanchard, Bousquet and Zwald [1].

In Reiß and Wahl [2] it is then argued that these upper bounds are still too crude because they do not catch the zero reconstruction error in the isotropic case of the identity as covariance matrix. To tighten the bounds, spectral projector calculus and local eigenvalue concentration results have to be developed and applied. The new bounds give for standard examples from functional data analysis and learning an excess risk smaller than the oracle risk and thus good oracle inequalities. A CLT in the parametric case exemplifies the inhomogeneity of the error with respect to the eigenvalue spacings.

A conjecture about the decision-theoretic optimality of PCA is formulated:
Conjecture. *For any given dimension p consider an i.i.d. sample $X_1, \dots, X_n \sim N(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{p\times p}$ is a covariance matrix with known eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, but unknown eigenvectors. The PCA spectral projector $\widehat{P}_{\leq d}$ onto the first $d < p$ principal components (eigenvectors of the empirical covariance matrix) is non-asymptotic minimax and admissible with respect to reconstruction error $R(\hat{P}_d, \Sigma) = \mathbb{E}[trace((I - \hat{P}_d)\Sigma)]$ among all orthogonal projectors $\hat{P}_d$ on subspaces of dimension d.*

It is discussed how this conjecture follows from the Bayesian conjecture that the PCA spectral projector is the Bayes-optimal estimator when we assume that the eigenspaces of Σ are generated by the Haar measure on the orthogonal group $O(\mathbb{R}^p)$. It is then proved that the Bayes-optimal estimator is diagonal in the eigenbasis of the empirical covariance matrix. It seems, however, non-trivial to prove or disprove whether it always projects onto the subspace of the d largest empirical eigenvalues.

References

[1] G. Blanchard, O. Bousquet, L: Zwald *Statistical properties of kernel principal component analysis*. Mach. Learn. **66** (2007), 259–294.

[2] M. Reiß, M. Wahl *Non-asymptotic upper bounds for the reconstruction error of PCA* , arXiv:1609.03779, Preprint 2018.

Geometrizing rates of convergence under local differential privacy

Lukas Steinberger

(joint work with Angelika Rohde)

In [1], 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 $X_1, \ldots, X_n \sim \mathbb{P}$ is kept private even from the statistician via an α-local differential privacy constraint. That is, the statistician only gets to see 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

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

where the first supremum runs over all measurable sets and $d_0(x, x') := |\{i : x_i \neq x'_i\}|$ denotes the number of distinct entries of x and x'. This definition is due to [4]. In this work, we consider only the local paradigm of differential privacy. A channel Q is said to provide α-local differential privacy if it satisfies (1) and it is such that every individual i can produce a private version Z_i of its original data X_i 'on its local machine' without having to know any of the confidential data X_j, $j \neq i$.

Suppose now that we want to estimate a real parameter $\theta(\mathbb{P})$ based on the privatized observation vector Z, whose unconditional distribution is equal to $Q\mathbb{P}^{\otimes n}(dz) := \int Q(dz|x)\, \mathbb{P}^{\otimes n}(dx)$, where $\mathbb{P}^{\otimes n}$ is the n-fold product measure of $\mathbb{P}$. The Q-privatized minimax risk of estimation under a loss function $l : \mathbb{R} \to \mathbb{R}$ is therefore given by

$$\mathcal{M}_n(Q, \mathcal{P}, \theta) := \inf_{\hat{\theta}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{Q\mathbb{P}^{\otimes n}} \left[l(|\hat{\theta}_n - \theta(\mathbb{P})|) \right], \tag{2}$$

where the infimum runs over all estimators $\hat{\theta}_n$ taking Z as input data. Note that if the channel Q is given by $Q(A|x) = Pr(Z \in A|X = x) = \mathbb{1}_A(x)$, then there is no privatization at all and the Q-privatized minimax risk reduces to the conventional minimax risk of estimating $\theta(\mathbb{P})$. If we want to guarantee (local) α-differential privacy, then we may choose any channel Q that satisfies (1) and we will try to make (2) as small as possible. This leads us to the α-private minimax risk

$$\mathcal{M}_{n,\alpha}(\mathcal{P}, \theta) := \inf_{Q \in \mathcal{Q}_\alpha} \mathcal{M}_n(Q, \mathcal{P}, \theta),$$

where $\mathcal{Q}_\alpha$ is some set of α-local differentially private channels. It is this additional infimum over $\mathcal{Q}_\alpha$ that makes the theory of private minimax estimation deviate fundamentally from the conventional minimax estimation approach. A sequence of channels $Q^{(n)} \in \mathcal{Q}_\alpha$, for which $\mathcal{M}_n(Q^{(n)}, \mathcal{P}, \theta)$ is of the order of $\mathcal{M}_{n,\alpha}(\mathcal{P}, \theta)$, is referred to as a minimax rate optimal channel and may depend on the specific estimation problem under consideration, i.e., on θ and $\mathcal{P}$. We write $\mathcal{M}_{n,\infty}(\mathcal{P}, \theta)$ for the classical (non-private) minimax risk.

Our contribution is to characterize the rate at which $\mathcal{M}_{n,\alpha}(\mathcal{P}, \theta)$ converges to zero as $n \to \infty$, in high generality, and to provide concrete minimax rate optimal α-locally differentially private channel distributions. To this end, we utilize the modulus of continuity of the functional $\theta : \mathcal{P} \to \mathbb{R}$ with respect to the total variation distance $d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1)$, that is,

$$\omega_{TV}(\varepsilon) := \sup\{|\theta(\mathbb{P}_0) - \theta(\mathbb{P}_1)| : d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1) \le \varepsilon, \mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}\},$$

and we show that under some regularity conditions, and for any fixed $\alpha \in (0, \infty)$,

$$\mathcal{M}_{n,\alpha}(\mathcal{P}, \theta) \quad \asymp \quad l\left(\omega_{TV}\left(n^{-1/2}\right)\right). \tag{3}$$

Here, $a_n \asymp b_n$ means that there exist constants $0 < c_0 < c_1 < \infty$ and $n_0 \in \mathbb{N}$, not depending on n, so that $c_0 b_n \le a_n \le c_1 b_n$, for all $n \ge n_0$.

It is important to compare (3) to the analogous result for the non-private minimax risk. This was established in the seminal paper by [2], who, under regularity conditions similar to those imposed here, showed that

$$\mathcal{M}_{n,\infty}(\mathcal{P}, \theta) \quad \asymp \quad l\left(\omega_H\left(n^{-1/2}\right)\right), \tag{4}$$

where $\omega_H(\varepsilon) = \sup\{|\theta(\mathbb{P}_0) - \theta(\mathbb{P}_1)| : d_{\mathrm{H}}(\mathbb{P}_0, \mathbb{P}_1) \le \varepsilon, \mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}\}$ and d_{H} is the Hellinger distance. Comparing (4) to (3), we notice that the Hellinger modulus ω_H of θ is replaced by the total variation modulus ω_{TV}. 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

$$\omega_H(\varepsilon) \le \omega_{TV}(\varepsilon) \le \omega_H(\sqrt{2\varepsilon}),$$

because $d_{\mathrm{TV}} \le d_{\mathrm{H}} \le \sqrt{2 d_{\mathrm{TV}}}$. 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.

That differential privacy leads to slower minimax rates of convergence was already observed by [3], for specific estimation problems. Here, we develop a unifying general theory to quantify the privatized minimax rates of convergence in a large class of different estimation problems, including (even irregular) parametric and non-parametric cases. This is also the first step towards a fundamental theory of adaptive estimation under differential privacy that will be pursued elsewhere.

We also exhibit a general construction scheme for minimax rate optimal α-locally differentially private channels that applies in many classical estimation problems. Suppose that for some $s \ge 0$, $t > 0$, there is an estimator of the form

$\frac{1}{n}\sum_{i=1}^n \ell_h(X_i)$ in the direct (non-private) estimation problem, that has a bias which decays at least as fast as h^t, as the tuning parameter $h \to 0$, and such that $\|\ell_h\|_\infty \lesssim h^{-s}$. If $\omega_{TV}(\varepsilon) \asymp \varepsilon^{\frac{t}{s+t}}$ as $\varepsilon \to 0$, and the regularity conditions on θ and $\mathcal{P}$ are satisfied, then generating Z_i independently and binary distributed on $\{-z_0, z_0\}$, with

$$Pr(Z_i = z_0 | X_i = x_i) = \frac{1}{2}\left(1 + \frac{\ell_{h_n}(x_i)}{z_0}\right), \qquad h_n = \left(\frac{e^\alpha + 1}{\sqrt{n}(e^\alpha - 1)}\right)^{\frac{t}{s+t}}$$

and $z_0 = \|\ell_{h_n}\|_\infty \frac{e^\alpha+1}{e^\alpha-1}$, yields an α-locally differentially private channel that attains the minimax rate in (3). We also treat the anisotropic multivariate case, where h_n may be a vector of tuning parameters. The conditions on ℓ_h are satisfied in many classical moment or density estimation problems. We point out that there are cases where the estimator $\frac{1}{n}\sum_{i=1}^n \ell_h(X_i)$ in the direct problem has the properties required above, even though a minimax optimal estimator in that problem is not of linear form.

Finally, we illustrate the general theory by a number of examples. Our theory allows to quantify the price to be paid for local differential privacy in a large class of estimation problems.

References

[1] Rohde, A. and Steinberger, L. Geometrizing rates of convergence under differential privacy constraints. *arXiv:1805.01422*, 2018.

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

[3] Duchi, J. C., Jordan, M. I. and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.*, 113(521): 182–201, 2018.

[4] Dwork, C., McSherry, F., Nissim, K. and Smith, A. Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography*, (S. Halevi and T. Rabin, eds.). Lecture Notes in Computer Science: 265–284, 2006.

Does data interpolation contradict statistical optimality?

Alexandre B. Tsybakov

(joint work with Mikhail Belkin and Alexander Rakhlin)

We show that the classical Nadaraya-Watson estimator with an appropriately chosen kernel interpolates the data, yet achieves optimal rates of convergence for the problems of nonparametric regression and prediction with square loss. This curious observation goes against the usual intuition that a good statistical procedure should forego the exact fit to data in favor of a more smooth representation. The family of estimators we consider do exhibit a bias-variance trade-off with a tuning parameter, yet this "regularization" co-exists in harmony with data interpolation.

Let (X, Y) be a random pair on $\mathbb{R}^d \times \mathbb{R}$ with distribution P_{XY}, and let $f(x) = \mathbb{E}[Y|X = x]$ be the regression function. A goal of nonparametric estimation is to construct an estimate f_n of f, given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn independently from P_{XY}. A classical approach to this problem is kernel smoothing.

In particular, the Nadaraya-Watson estimator [Nadaraya(1964), Watson(1964)] is defined as

$$f_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}, \tag{1}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function and $h > 0$ is a bandwidth and we assume that the denominator does not vanish. Appropriate choices of K and h lead to optimal rates of estimation, under various assumptions, cf. [Tsybakov(2009)] and references therein.

We consider singular kernels that approach infinity when their argument tends to zero. It has been observed, at least since [Shepard(1968)], that the resulting function in (1) interpolates the data. We will focus on the particular kernel

$$K(u) = \|u\|^{-a} \mathbf{I}\{\|u\| \le 1\}, \tag{2}$$

for some $a > 0$. Here, $\|\cdot\|$ denotes the Euclidean norm and $\mathbf{I}\{\cdot\}$ stands for the indicator function. Our results can be extended to other related singular kernels, for example, to

$$K(u) = \|u\|^{-a} [1 - \|u\|]_+^2 \tag{3}$$

where $[c]_+ = \max\{c, 0\}$, and

$$K(u) = \|u\|^{-a} \cos^2(\pi \|u\| /2) \mathbf{I}\{\|u\| \le 1\}, \tag{4}$$

considered in [Lancaster and Salkauskas(1981), Katkovnik(1985)]. Also, $\|\cdot\|$ can be any norm on $\mathbb{R}^d$, not necessarily the Euclidean norm.

To state the results, we use the following definition.

Definition 1. *For $L > 0$ and $\beta \in (0, 2]$, the (β, L)-Hölder class, denoted by $\Sigma(\beta, L)$, is defined as follows:*

- *If $\beta \in (0, 1]$, the class $\Sigma(\beta, L)$ consists of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying*

$$\forall x, y \in \mathbb{R}^d, \ |f(x) - f(y)| \le L \|x - y\|^\beta . \tag{5}$$

- *If $\beta \in (1, 2]$, the class $\Sigma(\beta, L)$ consists of continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying*

$$\forall x, y \in \mathbb{R}^d, \ |f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \le L \|x - y\|^\beta \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

We assume the following.

(A1) For any $x \in \mathbb{R}^d$, the expectation $\mathbb{E}[Y|X = x] = f(x)$ exists and $\mathbb{E}[\xi^2|X = x] \le \sigma_\xi^2 < \infty$, where $\xi = Y - \mathbb{E}[Y|X] = Y - f(X)$.

(A2) The marginal density $p(\cdot)$ of X exists and satisfies $0 < p_{\min} \le p(x) \le p_{\max}$ for all x on its support.

The Nadaraya-Watson estimator for a singular kernel K is defined as

$$f_n(x) = \begin{cases} Y_i & \text{if } x = X_i \text{ for some } i = 1, \dots, n, \\ 0 & \text{if } \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) = 0, \\ \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} & \text{otherwise.} \end{cases} \tag{7}$$

We show that this estimator has the properties stated in the next two theorems.

Theorem 1. *Assume that $f \in \Sigma(\beta, L_f)$ for $\beta \in (0, 1]$, $L_f > 0$. Let Assumptions $(A1)$ and $(A2)$ be satisfied, and $0 < a < d/2$. Then for any fixed $x_0 \in \mathbb{R}^d$ in the support of p the estimator (7) with kernel (2) and bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\mathbb{E}[(f_n(x_0) - f(x_0))^2] \le C n^{-\frac{2\beta}{2\beta+d}}$$

where $C > 0$ is a constant that does not depend on n.

Theorem 2. *Assume that $f \in \Sigma(\beta, L_f)$ for $\beta \in (1, 2]$, $L_f > 0$. Let Assumptions $(A1)$ and $(A2)$ be satisfied, and $0 < a < d/2$. Assume in addition that, for all x, y in the support of p, we have $|p(x) - p(y)| \le L_p \|x - y\|^{\beta-1}$, $L_p > 0$. Then for any fixed $x_0 \in \mathbb{R}^d$ such that the Euclidean ball of radius h centered at x_0 is contained in the support of p, the estimator (7) with kernel (2) and bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\mathbb{E}[(f_n(x_0) - f(x_0))^2] \le C n^{-\frac{2\beta}{2\beta+d}}$$

where $C > 0$ is a constant that does not depend on n.

In particular, the pointwise mean squared error (MSE) bound of Theorem 1 immediately implies that the integrated MSE with respect to the marginal distribution of X satisfies

$$\mathbb{E} \int_{\mathbb{R}^d} (f_n(x) - f(x))^2 p(x) dx \le C n^{-\frac{2\beta}{2\beta+d}},$$

assuming that f is bounded on the support of the marginal density p.

References

[Katkovnik(1985)] V Katkovnik. *Nonparametric identification and smoothing of data (Local approximation methods)*. Nauka, Moscow, 1985.

[Lancaster and Salkauskas(1981)] Peter Lancaster and Kes Salkauskas. Surfaces generated by moving least squares methods. *Mathematics of computation*, 37(155):141–158, 1981.

[Nadaraya(1964)] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[Rakhlin et al.(2017)] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[Shepard(1968)] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.

[Tsybakov(2009)] Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009.

[Watson(1964)] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Gaussian network reconstruction using prior information

Aad van der Vaart

(joint work with Gino Kpogbezan, Stéphanie van der Pas, Botond Szabó, Mark van der Wiel and Wessel van Wieringen)

We introduced a Bayesian model based on the horseshoe prior to fit a coupled collection of high-dimensional regression models, allowing for a soft incorporation of prior information on the regression parameters. A main application is to recovery of the partial correlation network based on the observation of a sample of Gaussian vectors, where the coordinates of the vectors correspond to, say, observations of gene expression. This problem can be reduced to a set of regression models that explain the collected observations on a given gene by a linear regression on the other genes. The analysis then should borrow information between the different regression models. We used the horseshoe prior to model sparsity of the regressions and couple the regression models by setting joint hyperparameters. Furthermore, since the number of observations is typical small, we allow for a soft encoding of prior information on the network, by modelling the hyperparameters in groups. If the prior information happens to be wrong, then the data can estimate as equal and not much is lost, but it turns out that correct prior information can make the results much more accurate.

Implementation of the model is through a variational Bayesian approximation to the posterior distribution.

Our talk had a practical part, explaining the preceding modelling procedure in some detail, and showing it in action on simulated and real datasets, and a theoretical part. In the latter part we study the frequentist properties of credible intervals obtained from the horseshoe prior, which underly the model selection procedure.

Sharp $\sin\Theta$ theorems under a relative rank condition

Martin Wahl

(joint work with Moritz Jirak)

The study of general perturbation bounds has a long tradition in matrix analysis, functional analysis, and operator theory. A basic problem is to estimate how a small perturbation effects the eigenvalues and eigenvectors of a self-adjoint compact operator. Classical perturbation bounds include the Weyl inequality and the Davis-Kahan $\sin\Theta$ theorem. Due to its importance in many areas of pure and applied mathematics, generalizations and refinements have been intensively

investigated in the literature, see e.g. Bhatia [1]. A powerful machinery to derive perturbation bounds is given by the holomorphic functional calculus for linear operators, see e.g. the monograph by Kato [5].

In many situations, the perturbation is random and the unperturbed operator has certain structural properties. In such scenarios, classical perturbation results, such as Weyl and Davis-Kahan, are often far from optimal, see e.g. Jirak and Wahl [3, 4] and O'Rourke, Vu, and Wang [6]. A main example is given by the empirical covariance operator, a central object in high-dimensional probability due to its importance for statistics and machine learning. Another example is given by random perturbations of low-rank matrices and the matrix recovery problem.

The objective of relative perturbation bounds is to improve upon absolute bounds by exploiting certain structure of the perturbation. For instance, there are relative versions of the Weyl inequality and the Davis-Kahan $\sin\Theta$ theorem which benefit from considering relative errors and relative spectral gaps. Relative bounds are a well-studied object in other branches of mathematics, see e.g. the review paper by Ipsen [2]. On the other hand, this appears to be a rarely studied topic in probability theory and statistics. Only recently, it has been shown in problems related to empirical covariance operators that relative techniques may lead to substantial improvements.

In this talk, we derive $\sin\Theta$ theorems, taylored for relative perturbations. We show that a sharp bound can be achieved under a relative rank condition. As a main example, we apply our results to empirical covariance operators. Besides, we demonstrate that our general result also applies to other structured random perturbations. The proof is based on a novel contraction phenomenon, contrasting previous spectral perturbation approaches.

References

[1] R. Bhatia, *Matrix analysis*, Springer-Verlag, New York (1997).

[2] I. C. F. Ipsen, *An overview of relative* $\sin\Theta$ *theorems for invariant subspaces of complex matrices*, J. Comput. Appl. Math. **123** (2000), 131–153.

[3] M. Jirak and M. Wahl, *Relative perturbation bounds with applications to empirical covariance operators*, preprint, available at https://arxiv.org/pdf/1802.02869.

[4] M. Jirak and M. Wahl, *Perturbation bounds for eigenspaces under a relative rank condition*, preprint, available at https://arxiv.org/pdf/1803.03868.

[5] T. Kato, *Perturbation theory for linear operators*, Springer-Verlag, Berlin, reprint of the 1980 edition (1995).

[6] S. O'Rourke, V. Vu and K. Wang, *Random perturbation of low rank matrices: improving classical bounds*, Linear Algebra Appl. **540** (2018), 26–59.

Detecting relevant changes in the mean of non-stationary processes - a mass excess approach

Weichi Wu

This paper considers the problem of testing if a sequence of means $(\mu_t)_{t=1,\dots,n}$ of a non-stationary time series $(X_t)_{t=1,\dots,n}$ is stable in the sense that the difference of the means μ_1 and μ_t between the initial time $t = 1$ and any other time is smaller than a given threshold, that is $|\mu_1 - \mu_t| \leq c$ for all $t = 1, \dots, n$. A test for hypotheses of this type is developed using a bias corrected monotone rearranged local linear estimator and asymptotic normality of the corresponding test statistic is established. As the asymptotic variance depends on the location of the roots of the equation $|\mu_1 - \mu_t| = c$ a new bootstrap procedure is proposed to obtain critical values and its consistency is established. As a consequence we are able to quantitatively describe relevant deviations of a non-stationary sequence from its initial value. The results are illustrated by means of a simulation study and by analyzing data examples.

Reporter: Lukas Steinberger

Participants

Randolf Altmeyer
Fachbereich Mathematik
Humboldt Universität Berlin
Unter den Linden 6
10099 Berlin
GERMANY

Prof. Dr. Peter J. Bickel
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Mike Bing
Department of Statistical Science
Cornell University
1188 Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

Prof. Dr. Christian Borgs
Microsoft Research
1 Memorial Drive
Cambridge, MA 02142
UNITED STATES

Prof. Dr. Florentina Bunea
Department of Statistical Science
Cornell University
Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Dr. Zhou Fan
Department of Statistics
Stanford University
390 Serra Mall
Stanford, CA 94305-4065
UNITED STATES

Derek Feng
Department of Statistics
Yale University
P.O. Box 208290
New Haven, CT 06520-8290
UNITED STATES

Prof. Dr. Christophe Giraud
Centre de Mathématiques Appliquées
UMR 7641 - CNRS
École Polytechnique
91128 Palaiseau Cedex
FRANCE

Prof. Dr. Marc Hoffmann
CEREMADE
Université Paris-Dauphine, PSL
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

Dr. Zongming Ma
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

Prof. Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Alexander Meister
Fachbereich Mathematik
Universität Rostock
18051 Rostock
GERMANY

Prof. Dr. Boaz Nadler
Department of Computer Science
and Applied Mathematics
The Weizmann Institute of Science
P.O.Box 26
Rehovot 76100
ISRAEL

Dr. Sofia Olhede
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Marianna Pensky
Department of Mathematics
University of Central Florida
Orlando, FL 32816-1364
UNITED STATES

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

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

Dr. Lukas Steinberger
Fakultät für Mathematik
Albert-Ludwigs-Universität Freiburg
Abeilung für Mathematische Stochastik
Ernst-Zermelo-Strasse 1
79104 Freiburg i. Br.
GERMANY

Prof. Dr. Alexandre B. Tsybakov
Laboratoire de Statistique
CREST - UMR 9194
5, Avenue Henry le Châtelier
91764 Palaiseau Cedex
FRANCE

Prof. Dr. Aad W. van der Vaart
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
2300 RA Leiden
NETHERLANDS

Dr. Martin Wahl
Institut für Mathematik
Humboldt Universität zu Berlin
Unter den Linden 6
10099 Berlin
GERMANY

Dr. Weichi Wu
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Prof. Dr. Huibin Zhou
Department of Statistics
Yale University
P.O. Box 208290
New Haven CT 06520-8290
UNITED STATES