Statistics conference
in honor of Aad van der Vaart's 60th birthday

June 17-21, 2019 / Leiden, The Netherlands

Contents

Abstracts

About rho-estimators

Lucien Birgé

LPSM - Sorbonne Université - Paris

During the last five years, Yannick Baraud and myself worked on the definition and properties of a robust substitute to the maximum likelihood estimator which deals with the estimation of the joint distribution P of n independant (but non necessarily i.i.d.) random variables $X_1, \ldots, X_n$. The estimator $\widehat{P}$ is based on a model (or a countable family of models to achieve adaptation) which means a family $\mathcal{P}_S = \{P_s,\, s \in S\}$ of probabilities. When $P \in \mathcal{P}_\mathcal{S}$, the performance of $\widehat{P}$ depends on some specific notion of "dimension" of $\mathcal{P}_\mathcal{S}$ (usually related to metric or VC dimensions) while, if $P \notin \mathcal{P}_\mathcal{S}$, this performance is only slightly affected provided that P is not far (with respect to Hellinger distance) from $\mathcal{P}_\mathcal{S}$. This accounts for the robustness properties of the estimator. The method applies to various statistical frameworks, namely density estimation and curve estimation in regression with fixed or random design. I shall describe in my talk some of its properties.

Interested readers could look at the following papers, the last one describing a Bayesian extension of the method.

[Baraud et al., 2017] Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

[Baraud and Birgé, 2016] Baraud, Y. and Birgé, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.*, 126(12):3888–3912.

[Baraud and Birgé, 2018] Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

[Baraud and Birgé, 2017] Baraud, Y. and Birgé, L. (2017). Robust bayes-like estimation: Rho-bayes estimation. Technical report, arXiv:1711.08328v1.

Bayesian Inference for Multivariate Medians and Multivariate Quantiles

Subhashis Ghoshal
North Carolina State University

We consider Bayesian inference on a class of multivariate medians and multivariate quantiles of a joint distribution using a Dirichlet process prior. Since, unlike univarite quantiles, the exact posterior distribution of multivariate median and multivariate quantiles are not obtainable explicitly, we study these distributions asymptotically. We derive a Bernstein-von Mises theorem for the multivariate median with respect to a general ℓ_p-norm, which in particular shows that its posterior concentrates around its true value at the parametric rate and its credible sets have asymptotically correct frequentist coverage. The technique involves approximating the posterior Dirichlet process by a Bayesian bootstrap process and deriving a conditional Donsker theorem. We also obtain analogous results for an affine equivariant version of the multivariate based on an adaptive transformation and re-transformation technique. The results are extended to a joint distribution of multivariate quantiles. The accuracy of the asymptotic result is confirmed by a simulation study. We also use the results to obtain Bayesian credible regions for multivariate medians for Fisher's iris data, which consists of four features measured for each of three plant species.

Fast learning with deep learning architectures for classification

Yongdai Kim
Seoul National University

We derive the fast convergence rates of a deep neural network (DNN) classifier with the rectified linear unit (ReLU) activation function learned using the hinge loss. We consider three cases for a true model: (1) a smooth decision boundary, (2) smooth conditional class probability, and (3) the margin condition (i.e., the probability of inputs near the decision boundary is small). We show that the DNN classifier learned using the hinge loss achieves fast rate convergences for all three cases provided that the architecture (i.e., the number of layers, number of nodes and sparsity). is carefully selected. An important implication is that DNN architectures are very flexible for use in various cases without much modification. In addition, we consider a DNN classifier learned by minimizing the cross-entropy, and give conditions for fast convergence rates. If time is allowed, computational algorithms to achieve a right size of deep architectures for fast convergence rates is discussed.

This is a joint work with my Ph.D. students Ilsang Ohn and Dongha Kim.

Causal organic direct and indirect effects: closer to Baron and Kenny

Judith J. Lok

Department of Mathematics and Statistics, Boston University

Baron and Kenny (1986, over 80,000 Google Scholar citations) proposed estimators of direct and indirect effects: the part of a treatment effect that is mediated by a covariate and the part that is not. Direct and indirect effects are especially important in epidemiology and psychology. Subsequent work on natural direct and indirect effects provides a formal causal interpretation. Natural direct and indirect effects use cross-worlds counterfactuals: outcomes under treatment with the mediator "set" to its value without treatment. Organic direct and indirect effects (Lok 2016) avoid cross-worlds counterfactuals, using "organic" interventions on the mediator while keeping the initial treatment fixed at "treatment". They apply also to settings where the mediator cannot be "set". If there is no treatment-mediator interaction, both natural and organic indirect effects lead to the same estimators as in Baron and Kenny. In this article, I propose organic interventions on the mediator while keeping the initial treatment fixed at "no treatment", leading to an alternative version of organic direct and indirect effects. I will show that the product method, proposed in Baron and Kenny, holds for this new indirect effect even if there is treatment-mediator interaction. Furthermore, I will argue that this alternative organic indirect effect is more relevant for drug development than the traditional natural or organic indirect effect. I will start my presentation with an introduction to direct and indirect effects.

Learning graphical models by covariance queries

Gabor Lugosi

Pompeu Fabra University Barcelona

In this joint work with Jakub Truszowski, Vasiliki Velona, and Piotr Zwiernik, we study the problem of reconstructing the structure behind large Gaussian graphical models. If high-dimensional problems, even storing the sample covariance matrix may be too costly. For such situations we propose efficient randomized algorithms that recover the structure of Gaussian graphical models accessing only a small number of values of the covariance matrix. Our algorithms work in a regime of tree-like graphs or, more generally, for graphs of small treewidth. Our results demonstrate that for large classes of graphs, the structure of the corresponding Gaussian graphical models can be determined much faster than even computing the empirical covariance matrix.

Consistency of Bayes methods for non-linear inverse problems

Richard Nickl

Cambridge University

Bayes methods for inverse problems have become very popular in applied mathematics in the last decade. They provide reconstruction algorithms as well as in-built 'uncertainty quantification' via Bayesian credible sets, and particularly for Gaussian priors can be efficiently implemented by MCMC methodology. For linear inverse problems, they are closely related to classical penalised least squares methods and thus not fundamentally new, but for non-linear and non-convex problems, they give genuinely distinct and computable algorithmic alternatives that cannot be studied by variational analysis or convex optimisation techniques. In this talk we will discuss recent progress in Bayesian Non-Parametric statistics that allows to give rigorous statistical guarantees for posterior mean reconstructions in non-linear non-convex inverse problems arising in some elliptic PDE models and in non-Abelian ('neutron-spin') X-ray tomography.

Bayesian inference for dynamical systems using Gibbs posteriors for families of Gibbs measures

Andrew Nobel

University of North Carolina, Chapel Hill

In this talk I will describe a Bayesian framework for making inferences about dynamical systems. Of interest is a model class consisting of a parametrized family of Gibbs measures on an appropriate shift space. Model classes of this sort include (hidden) Markov chains of arbitrary order, and generalizations of these, which may have long range dependence. Given a prior distribution over the model class and observations from an ergodic process, we obtain a Gibbs posterior distribution on the parameter space using a loss function that relates the state space of the models to the space of the observations. The asymptotic behavior of the Gibbs posterior distribution is characterized by a variational problem in which one seeks to minimize a divergence based rate function over dynamically invariant couplings of the Gibbs measures and the observed process. Moreover, the Gibbs posterior distributions concentrate around the solution set of this variational problem. In the case of properly specified models, the convergence results may be used to establish posterior consistency in general settings. This work establishes tight connections between Gibbs posterior inference and the thermodynamic formalism in dynamical systems.

Joint work with K. McGoff and S. Mukherjee.

How to estimate a density on a spider web ?

Dominique Picard
Université Paris-Diderot, Paris 7

Our purpose is to study the *density estimation problem*, namely, one observes $X_1, \ldots, X_n$ that are i.i.d. random variables defined on a space $\mathcal{M}$ and the problem is to find a good estimation to the common density function.

This problem has a long history in mathematical statistics but here we will consider very general sets $\mathcal{M}$ such as Riemannian manifolds or sets of matrices or graphs or spider webs and prove that with some assumptions, we can build an estimation theory with estimation procedures, regularity sets and upper bounds evaluations quite parallel to what has been neatly done in $\mathbf{R}^d$. In particular we prove that kernel methods can be constructed with minimax and oracle properties.

If we want to roughly summarize the basic assumptions that will be made in this work, let us mention that some of them are concerning the basic dimensional structure of the set (doubling conditions...), whereas others are devoted to construct an environment where regularity spaces and approximation properties hold.

This setting is quite general but at the same time is sufficiently rich in allowing the development of a smooth functional calculus with well localized spectral kernels, Besov regularity spaces, and wavelet type systems. Naturally, the classical setting on $\mathbf{R}^d$ and the one on the sphere are contained in this general framework, but also various other settings are covered. In particular, spaces of matrices, of graphs, of compact Riemannian manifolds, convex subsets of (non-compact) Riemannian manifolds are covered.

Frequentist and Bayesian Nonparametrics for Boundary Models

Markus Reiß
Humboldt University Berlin

We consider a nonparametric boundary or frontier recovery problem where the observation support (of a regression or point process model) lies above a boundary function g. While nonparametric estimation of g follows mainly known approaches from standard nonparametrics [1], the estimation of linear functionals like $\int_0^1 g(x)dx$ [2] or even nonlinear functionals like $\int_0^1 |g(x)|^p dx$ [3] can be achieved in an unbiased way, which is non-asymptotically of minimal variance and attains (unexpected) optimal minimax rates over Hölder and monotone function classes. The key idea is a bias correction for the nonparametric MLE. Adopting a nonparametric Bayesian method the posterior contraction rates in [1] can be obtained as well, even though the model is not regular [4]. We then ask whether the Bayesian approach for the estimation of $\int_0^1 g(x)dx$ is more intrinsic in the sense that it performs an automatic bias correction around the frequentist MLE and thus provides credible intervals with asymptotic correct coverage and minimal length. This is achieved to some extent by a Bernstein-von Mises Theorem for monotone g under a compound Poisson prior, but not always [5]. The talk will give an overview of these results and discuss in particular the major structural questions.

[1] Jirak, M., Meister, A. and Reiß, M. *Adaptive estimation in nonparametric regression with one-sided errors.* Annals of Statistics 42(5), 1970-2002, 2014.
[2] Reiß, M. and Selk, L. *Efficient estimation of functionals in nonparametric boundary models.* Bernoulli 23(2), 1022-1055, 2017.
[3] Reiß, M. and Wahl, M. *Functional estimation and hypothesis testing in nonparametric boundary models.* Bernoulli, to appear, 2018.
[4] Reiß, M. and Schmidt-Hieber J. *Posterior contraction rates for support boundary recovery.* Preprint, arXiv:1703.08358, 2017.
[5] Reiß, M. and Schmidt-Hieber, J. *Nonparametric Bayesian analysis of the compound Poisson prior for support boundary recovery.* Annals of Statistics, to appear, 2019.

On assumption free tests and confidence intervals for causal effects estimated with machine learning

James M. Robins
Harvard University

For many causal parameters ψ of interest, doubly robust machine learning estimators $\widehat{\psi}_1$ are the state-of-the-art, incorporating the benefits of the low prediction error of machine learning (ML) algorithms; the decreased bias of doubly robust estimators; and.the and the bias reduction of sample splitting with cross fitting. Nonetheless, even if unmeasured confounding is absent, when the vector of potential confounders is high dimensional, the associated $(1-\alpha)$ Wald confidence intervals $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\text{se}}[\widehat{\psi}_1]$

may still undercover even in large samples, because the bias of the estimator may be of the same or even larger order than its standard error of order $n^{-1/2}$.

In this talk, we introduce novel tests that (i) can have the power to detect whether the bias of $\widehat{\psi}_1$ is of the same or larger order than its standard error of order $n^{-1/2}$,(ii) can provide a lower confidence limit on the degree of undercoverage of the interval $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\text{se}}[\widehat{\psi}_1]$ and (iii) strikingly, require essentially no assumptions whatsoever. We also introduce an estimator $\widehat{\psi}_2 = \widehat{\psi}_1$-$\widehat{\mathbb{IF}}_{22}$ with bias generally less, often much less, than that of $\widehat{\psi}_1$, yet whose standard error is not much greater than $\widehat{\psi}_1 s$ .The tests, as well as the estimator $\widehat{\psi}_2$, are based on a U-statistic $\widehat{\mathbb{IF}}_{22}$ that is the second-order influence function for the parameter that encodes the estimable part of the bias of $\widehat{\psi}_1$. When the covariance matrix of the potential confounders is known, $\widehat{\mathbb{IF}}_{22}$ is an unbiased estimator of its parameter. When the covariance matrix is unknown, we propose several novel estimators of $\widehat{\mathbb{IF}}_{22}$ that perform almost as well as the known covariance case in simulation experiments.

Our impressive claims need to be tempered in an important way. No test, including ours, of the null hypothesis that the ratio of the bias to its standard error can be consistent [without making additional assumptions (e.g. smoothness or sparsity) that may be incorrect].

This is joint work with Lin Liu and Rajarshi Mukherjee.

Efficient two-sample functional estimation and the super-oracle phenomenon

Richard Samworth

Cambridge University

We consider the estimation of two-sample integral functionals, of the type that occur naturally, for example, when the object of interest is a divergence between unknown probability densities. Our first main result is that, in wide generality, a weighted nearest neighbour estimator is efficient, in the sense of achieving the local asymptotic minimax lower bound. Moreover, we also prove a corresponding central limit theorem, which facilitates the construction of asymptotically valid confidence intervals for the functional, having asymptotically minimal width. One interesting consequence of our results is the discovery that, for certain functionals, the worst-case performance of our estimator may improve on that of the natural 'oracle' estimator, which is given access to the values of the unknown densities at the observations.

Synthesis and analysis for structured regression

Sara van de Geer

Seminar for Statistics, ETH Zürich

Suppose we have noisy observations $Y \in \mathbb{R}^n$ of a signal $f^0 \in \mathbb{R}^n$:

$$Y = \underbrace{f^0}_{\text{signal}} + \underbrace{\epsilon}_{\text{noise}}.$$

The synthesis problem is to minimize the residual sum of squares with ℓ_1-regularization

$$\text{minimize over } b \in \mathbb{R}^p : \ \|Y - Xb\|_2^2/n + 2\lambda\|b\|_1, \tag{1}$$

with $X \in \mathbb{R}^{n\times p}$ a given design matrix and $\lambda > 0$ a tuning parameter. A minimizer $\hat{\beta}$ of problem (1) is well-known under the name Lasso. It is called a synthesis problem as an estimator of the signal f^0 is synthesized using the coefficients $\hat{\beta}$ by putting $\hat{f} = X\hat{\beta}$.

The analysis problem is

$$\text{minimize over } f \in \mathbb{R}^n : \ \|Y - f\|_2^2/n + 2\lambda\|Df\|_1 \tag{2}$$

where $D \in \mathbb{R}^{m\times n}$ is a given matrix (for example the incidence matrix of a graph). The analysis problem directly focusses on estimating f^0. A minimizer $\hat{f}$ of problem (2) is structure-induced by the penalty $\|D\hat{f}\|_1$.

Since the Lasso is well-studied, one way to examine the analysis problem is to reformulate it as a synthesis problem (if possible). We however propose to study the analysis problem directly. It turns out that in a sense oracle inequalities for the analysis problem are easier to derive than for the synthesis problem.

As detailed example we present the total variation penalty

$$\|Df\|_1 := \sum_{i=2}^{n} |f_i - f_{i-1}|,$$

where we re-derive and refine some known oracle bounds. Further examples concern total variation penalties on graphs and higher order total variation penalties.

Targeted Machine Learning for Causal Inference

Mark van der Laan

UC Berkeley (Biostatistics Division, and Statistics Department)

We review targeted minimum loss estimation (TMLE), which provides a general template for the construction of asymptotically efficient plug-in estimators of a target estimand for infinite dimensional models. TMLE involves maximizing a parametric likelihood along a so-called least favorable parametric model through an initial estimator (e.g., ensemble super-learner) of the relevant functional of the data distribution. The asymptotic normality and efficiency of the TMLE relies on the asymptotic negligibility of a second-order term. This typically requires the initial estimator to converge at a rate faster than $n^{-1/4}$. We propose a new estimator of functional parameters of the data distribution, the Highly Adaptive LASSO (HAL), that converges at a sufficient rate regardless of the dimensionality of the data/model, under almost no additional regularity. This allows us to propose a general TMLE that is asymptotically efficient in great generality. We demonstrate the practical performance of HAL and its corresponding TMLE for the average causal effect for dimensions up till 10. We also present a nonparametric bootstrap method for inference taking into account the higher order contributions of the HAL-TMLE. Finally, we demonstrate the construction of super-efficient TMLE, and discuss the pros and cons.

A look at the connection between weak and strong posterior consistency

Stephen Walker

University of Texas

It is well known that a sequence of posterior distributions can be weakly consistent; i.e. accumulate in neighborhoods of the true density with respect to the Levy Prokhorov metric, yet fail to do so with respect to a stronger metric such as the L_1. A counter example has been provided by A. Barron. This talk will take a deeper look at this phenomenon and will be connecting the two metrics via a smoothing operation on densities. An automatic correction to achieve strong consistency given weak consistency is provided: to a predictive sample add a random uniform random variable of decreasing size. This procedure can be understood mathematically and also intuitively in light of Barron's counterexample. (joint work with Minwoo Chae)

Statistical and Computational Guarantees of EM with Random Initialization

Harry Zhou

Yale University

This talk considers parameter estimation in the two-component symmetric Gaussian mixtures in d dimensions with n independent samples. We show that, even in the absence of any separation between components, with high probability, the EM algorithm converges to an estimate in at most $O(\sqrt{n}\log n)$ iterations, which is within $O((d/n)^{1/4}(\log n)^{3/4})$ in Euclidean distance to the true parameter, provided that $n = \Omega(d\log^2 d)$. This is within a logarithmic factor to the minimax optimal rate of $(d/n)^{1/4}$. The proof relies on establishing (a) a non-linear contraction behavior of the population EM mapping (b) concentration of the EM trajectory near the population version, to prove that random initialization works. This is in contrast to previous analysis in Daskalakis, Tzamos, and Zampetakis (2017) that requires sample splitting and restart the EM iteration after normalization, and Balakrishnan, Wainwright, and Yu (2017) that requires strong conditions on both the separation of the components and the quality of the initialization. Furthermore, we obtain the asymptotic efficient estimation when the signal is stronger than the minimax rate.
