TOPICS IN STATISTICS

by
Yannis G. Yatracos
Yau Mathematical Sciences Center
Tsinghua University in Beijing
e-mail: yatracos@tsinghua.edu.cn

Lecture Times: Tuesday and Wednesday 3:20 pm.- 4:55 pm. (First from Feb. 21 to March 18, then I think there is a break. We continue after the break for another 8 weeks).

You can use the following information to log in zoom.

Zoom Account: kaixu1996@gmail.com
Passcode: YMSCymsc106

Meeting ID: 849 963 1368
Passcode: YMSC

Prerequisites

1. A course in Mathematical Statistics (unbiasedness, UMVUE, sufficiency, completeness, consistency, admissibility).

2. A course in Probability with Calculus, including probability inequalities and convergence concepts (in probability, in law and almost surely, Borel-Cantelli Lemma).

3. A course in Regression Analysis.

4. Calculus/Analysis and notions of Metric spaces, Matrix Algebra.

A student who took already these courses will follow the course/ideas more easily.

The Content

Mainly my research results over the years: in Statistical Theory with model assumptions, and more recent results with Algorithmic models à la Breiman (2001), i.e. the data $\mathbf{X}$, is obtained from a Black-Box, with input either data, $\mathbf{Y}$, or a parameter θ. The output of the Black-Box is approximated by a Learning machine $f(\mathbf{Y}, \theta)$; θ may denote also the parameter used in the approximation.

Topics we will discuss

1. MLE's Bias Pathology, model updated MLE and MME and Wallace's Minimum Message Length method.
2. Bootstrap Pathologies (natural from part 1, since Efron and Tibshirani in their Bootstrap book present ir as extension of the ML "plug-in" method.
3. Artificially augmented samples, shrinkage and MSE reduction, Pitman's Efficiency.

4. Tukey's Poly-efficiency (for k-models instead of one).
5. Minimum Distance Estimation of a density with convergence rates via Kolmogorov's entropy, respectively, of the space of densities or a regression type function, i.e. not necessarily a mean. Upper and lower convergence rates will be presented.

6. Estimating a parameter θ without model assumptions (in Algorithmic models), obtaining upper error rates in probability. The approach is based on an extension of Wolfowitz's Minimum Distance Estimation without models, using Rubin's Matching idea and Data Generating Machines (Sampler, Black Box.)
7. Approximate Bayesian Computations (ABC) using a sufficient statistic: the Fiducial (F)-ABC. Tools from part 6 are used.
8. How to choose among two or more Learning Machines? Using new graphical tools to detect

identifiability almost surely.
9. Classification: Projection Pursuit Cluster Detection without model assumptions, with Maximum Variance Components Split (MVCS) and other methods. Separation of Cryptocurrencies.

Some topics are controversial. The presentations is not necessarily in this order. Parts 1-5 are using parametric models, parts 6-9 are Algorithmic models. Part 5 needs to be done before parts 6 and 7.

References

Breiman, Leo (2001), Statistical Modeling: The Two Culture, *Statistical Science*, **16**, 199-231)
Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159-183. Correction (1974) **30**, 728.
Wolfowitz, J. (1957) The Minimum Distance Method *The Annals of Mathematical Statistics*, **28**, **1**, 75-88.

THE GOAL: Present some Mathematical Foundations of Data Science and new results. Why these results are needed? Deep Learning has its limitations and in its current form cannot be used to solve all problems.

• In Le Cun *et al.* (2015, p. 442), the last section "The future of deep learning", it is mentioned that "major progress in artificial intelligence will come about through systems that combine learning *with complex reasoning*, replacing the simple reasoning used so far." It is also mentioned that unsupervised learning will become more important in the future, "... we discover the structure of the world by observing it, not by being told the name of every object."

• The complex reasoning will not involve only computational extensions of simple reasoning.

• For new results, we will revisit classical approaches where model assumptions are used. The methods will be altered and new methods

are proposed to be used when the data, X, is obtained from a Black-Box with inputs Y and θ, and X is approximated by a Learning Machine $f(Y, \theta)$, and the learned model is $f(Y, \hat{\theta})$; $\hat{\theta}$ is an estimate of θ.

Reference

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436-444.

Questions we will study in order to complete our trip/tour.

1) Do you prefer more or less randomness in a Statistical Experiment?

2) Does Data from a Model evolve, providing additional Information? How do you use the additional information? In Deep Learning each layer provides additional information. How is

it used?

3) Which estimation criteria to use?

You prefer estimate T of θ over estimate S if

$$E(T-\theta)^2 \leq E(S-\theta)^2 \qquad (1)$$

or use the Pitman's Closeness criterion,

$$P[|S-\theta| > |T-\theta|] > 1/2, \text{ or ...} \qquad (2)$$

Do you think (1) and (2) hold both for T?

How you would feel if this is not the case?

4) The efficiency of an estimate T_n is usually studied under one model. Data rarely follows the assumed model. Is there an alternative?

5) Do you see a sample as i.i.d. r.vs $X_1, \ldots, X_n$ or as vector $(X_1, \ldots, X_n)$?

6) In Analysis of Variance we have two terms: the between-groups variations and the within-groups variation. It is used to detect clusters when MacQueen (1967, p. 288) who introduced it wrote "The point of view taken in this application is not to find some unique, definitive

grouping, but rather to simply aid the investigator in obtaining qualitative and quantitative understanding of large amounts of N-dimensional data by providing him with reasonably good similarity groups."

Is there an alternative Analysis?

7) Is it easier to identify with statistical methods clusters in R^{10} or in R^{20}?

8) Did you think of estimating a random function of a parameter with risk the Mean Squared Error?

9) Should the estimation error of $\theta \in \Theta$ depend on the "massiveness" of Θ?

10) How can you decide between two learning machines $f_1(Y, \theta)$ and $f_2(Y, \theta)$ wghich one to use to represent the output of a Black-Box?

The material is not going to be presented in the same order.

Reference

MacQueen, J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281-297, Berkeley, Calif., 1967. University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512992.

Preview of some of 1)-10) with Data to get a feeling.

A) SEPARATING CRYPTOS FROM OTHER ASSETS (non-supervised)

• Data=log(Returns of prices) for: cryptos, stocks, bonds, commodities, real estate indices, commodities.

• Collected, summarised, presented using Factor Analysis.

Think of Regression of Y on undetermined $\mathbf{X}$'s you define according to a criterion, usually via Variance decomposition in components.

CRYPTOS: THE MOVIE! A FIRST LOOK(IN 01B)

The procedure:

• First, we "estimate" the 24D dataset for the interval [03.01.2014; 22.04.2016].

• At every moment t, we project the 24D dataset on the 2 main factors extracted via Factor Analysis. The idea was to show that over time the cryptos tend to form a separate cluster.

• Second, we extend the interval on a daily basis, "estimate" the 24D dataset from the beginning till the current time and project on the 2D space defined by the tail and memory factor, explaining, respectively, 76% and 6% of Total Variance. Third, the moment factor, explaining 6% of total variance, correlated with skewness and kyrtosis of log-returns. Memory factor: relates to decay of statistical dependence over time, measured via autocovariance function or change in variance of consecutive sums.

• The last observation is 31.11.2020.
• At every moment we also include the past.
• The number of assets vary by time, as we don't have all the cryptos for the entire period. At the end, we have 906 assets: 234 cryptocurrencies, 635 stocks, 13 exchange rates, 17 commodities, 5 bonds, and 2 real estate indexes.
• Green dots are the Cryptos. Part of the Green dots is surrounded by a curve, which is a 95% confidence region for the cryptos to be a potential cluster, based on Kernel Density Estimation.
• We plot the confidence region only for cryptos in this video, we also tried for other assets, but it looked messy.
• USDT is closer to CHF if we project on the 3D space.

QUESTIONS AND ANSWERS

• Why BCH appears in the beginning in Black, in the other assets writing BCH, and at the end

is in the middle of cryptos? At the beginning it was close to the other classical assets.

• Why BCH was circled in the beginning? It was in separate cluster.

• Why some names appear and what they are? USDT, BCH, CHF, ETH, XRR(?). We chose only top ten cryptos, based on market capitalization.

SEPARATING THE ASSETS

• There is no complete separation of clusters with K-Means or Support Vector Machines.

• Complete separation with Variance Components Split (VCS) method, seen in this course.

Figures in 01A

TIME FOR THE MOVIE AGAIN! 01B

• K-Means and its disadvantage. From Francis Bach notes on unsupervised clustering.

Figure 2 in 02

B) THE BOOTSTRAP

$\mathbf{X}$ denotes sample $X_1, \ldots, X_n$. Include it in a box and draw B samples $X_1^*, \ldots, X_n^*$, of size n with replacement. Construct an estimate T_B of a parameter θ. Then $T = E(T_B|\mathbf{X})$ is an estimate, since $\mathbf{X}$ is sufficient statistic. Then, T_B and T have the same mean, so they have the same bias. It holds:

$$Var(T_B) = Var(E(T_B|\mathbf{X})) + EVar(T_B|\mathbf{X})$$
$$= Var(T) + EVar(T_B|\mathbf{X})$$

and the last term I called "Cushion Error" depends on θ, so can be infinite. Therefore, $E(T-\theta)^2 < E(T_B - \theta)^2$.

Why somebody would use T_B?

Simulations provide $\hat{E}(T-\theta)^2 - \hat{E}(T_B - \theta)^2$. If they are mostly negative, below the line at 0, the Bootstrap fails. See the results in Bootstrap simulations.

Figure 03

C) MINIMIZING THE MSE OR USE PITMAN'S CLOSENESS CRITERION?

Estimate T_n is often preferred to S_n if $E(T_n - \theta)^2 < E(S_n - \theta)^2$.

Often, when S_n is unbiased for θ, there is a biased estimate, usually shrinkage estimate, $T_n = c_n S_n, 0 < c_n < 1$, with

$$E(T_n - \theta)^2 < E(S_n - \theta)^2.$$

Do you expect most often $|S_n - \theta| > |T_n - \theta|$, i.e. $P[|S_n - \theta| > |T_n - \theta|] > 1/2$?

We will see this is not often the case, as simulations for normal samples with $\theta = \sigma^2$ and unknown mean show.

The graph shows $P[|T_n - \theta| > |S_n - \theta|])$ is much larger than 1/2 for small samples, and it is still larger than 1/2 for all n and not only for the normal model; S_n is the unbiased estimate of σ^2 and T_n the shrinkage estimate.

Figure 04

D) FOUNDATIONS OF ALGORITHMIC DATA SCIENCE

Identifiability of parameters cannot be confirmed so far with intractable or unavailable data models. The samples are obtained using a Black-Box or Quantile function with input $\theta \in \Theta$. Machine Learners want to evaluate identifiability of parameters used in Learning Machines but still the underlying models are unknown. *EDI* is introduced to confirm identifiability taking advantage of the available samplers.

Consider a normal mixture:

$$pN(\mu_1, 1) + (1-p)N(\mu_2, 1),$$

$p \in [0,1], (\mu_1, \mu_2) \in [0,2]^2$. The model, with $\theta = (p, \mu_1, \mu_2)$, is not identifiable, since $\theta = (.25, .4, 1.2)$ and $\theta^* = (.75, 1.2, .4)$ provide data from the same model and $\theta \neq \theta^*$.

EDI will confirm whether θ is identifiable, comparing $EDI(\theta, \theta; n)$ with $EDI(\theta, \theta_i^*; n), 1 \leq$

$i \leq M$. If there is $\theta^*_{i_0} \neq \theta$ such that both $EDI(\theta, \theta)$ and $EDI(\theta, \theta^*_{i_0})$, for large n, *a)* are not near 0 and *b)* take similar value, then θ is non-identifiable because of $\theta^*_{i_0}$.

EDI in action. Figure 05

We continue with

MLE's Bias Pathology, Model Updated MLE and MME, and Wallace's Minimum Message Length (MML) method

unless there is some other strong preference.