Prof. Mohamed NDAOUD

Chair of Excellence in Data Science (CY Initiative)
IDS Department, ESSEC Business School

Thursday January 13, 2022

12:30 pm (CET)

"On potential benefits of overfitting in high dimensions"

ESSEC La Défense (CNIT) - Amphi 201
and to participate via Zoom, <u>please click here</u>
Password/Code : WGRisk

We consider the supervised clustering problem under the two-component anisotropic Gaussian mixture model in high dimensions and in the non-asymptotic setting. We characterize precisely the risk of l_2 regularized supervised least squares classifiers. We deduce the fact that the interpolating solution may outperform regularized classifiers, under mild assumptions on the covariance structure of the noise. Our analysis also shows that overfitting can be robust to corruption in the covariance of the noise when the signal is aligned with the ``clean" part of the covariance, for the properly defined notion of alignment. To the best of our knowledge, this peculiar phenomenon has not yet been investigated in the rapidly growing literature related to interpolation. We conclude that overfitting is not only benign but can also be robust.
This is joint work with Stanislav Minsker and Yiqiu Shen.

CREAR ESSEC Research Center on Risk

Institut des ACTUAIRES

Labex MME-DII
Modèles Mathématiques et Économiques de la Dynamique, de l'Incertitude et des Interactions

SFdS

Banque Finance Assurance

Prof. Mohamed NDAOUD

Chair of Excellence in Data Science (CY Initiative)
IDS Department, ESSEC Business School

Dr. Mohamed Ndaoud, Assistant Professor in the **I**nformation System, **D**ecision Sciences and **S**tatistics (IDS) Department at ESSEC Business School in Paris, holds the Junior Chair of Excellence in Data Science, supported by the CY Initiative in the framework of the PIA program. After receiving his Ph.D. degree in Statistics from CREST-ENSAE Paris in 2019, Mohamed spent two years as an instructor in the Mathematics Department at University of Southern California. His research interests include variable selection, community detection and robust statistics in high dimensional settings.