Séminaire de Probabilités et Statistique

Mardi 5 décembre à 14h00

Salle de conférences

Arshak Minasyan

CREST / ENSAE

Statistically optimal robust mean and covariance estimation for anisotropic gaussians

Assume that $X_1, \dots, X_N$ is an ε-contaminated sample of N independent Gaussian vectors in $\mathbb{R}^d$ with mean μ and covariance Σ. In the strong ε-contamination model we assume that the adversary replaced an ε fraction of vectors in the original Gaussian sample by any other vectors. We show that there is an estimator $\widehat{\mu}$ of the mean satisfying, with probability at least $1-\delta$, a bound of the form

$$\|\widehat{\mu} - \mu\|_2 \leqslant c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{N}} + \varepsilon\sqrt{\|\Sigma\|}\right),$$

where $c > 0$ is an absolute constant, and $\|\Sigma\|$ denotes the operator norm of Σ. In the same contaminated Gaussian setup, we construct an estimator $\widehat{\Sigma}$ of the covariance matrix Σ that satisfies, with probability at least $1-\delta$,

$$\left\|\widehat{\Sigma} - \Sigma\right\| \leqslant c\left(\sqrt{\frac{\|\Sigma\|\,\mathrm{Tr}(\Sigma)}{N}} + \|\Sigma\|\sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon\|\Sigma\|\right).$$

Both results are optimal up to multiplicative constant factors. Despite the recent significant interest in robust statistics, achieving both dimension-free bounds in the canonical Gaussian case remained open. In fact, several previously known results were either dimension-dependent and required Σ to be close to identity, or had a sub-optimal dependence on the contamination level ε. As a part of the analysis, we derive sharp concentration inequalities for central order statistics of Gaussian, folded normal, and chi-squared distributions.
This is a joint work with N. Zhivotovskiy.