Statistica Sinica **34** (2024), 1-19
doi:https://doi.org/10.5705/ss.202021.0388

ROBUST ESTIMATION OF COVARIANCE MATRICES: ADVERSARIAL CONTAMINATION AND BEYOND

Stanislav Minsker* and Lang Wang

University of Southern California

Abstract: We consider the problem of estimating the covariance structure of a random vector $Y \in \mathbb{R}^d$ from an independent and identically distributed (i.i.d.) sample $Y_1, \dots, Y_n$. We are interested in the situation in which d is large relative to n, but the covariance matrix Σ of interest has (exactly or approximately) low rank. We assume that the given sample is either (a) ε-adversarially corrupted, meaning that an ε-fraction of the observations can be replaced by arbitrary vectors, or (b) i.i.d., but the underlying distribution is heavy-tailed, meaning that the norm of Y possesses only finite fourth moments. We propose estimators that are adaptive to the potential low-rank structure of the covariance matrix and to the proportion of contaminated data, and that admit tight deviation guarantees, despite rather weak underlying assumptions. Finally, we show that the proposed construction leads to numerically efficient algorithms that require minimal tuning from the user, and demonstrate the performance of such methods under various models of contamination.

Key words and phrases: Adversarial contamination, covariance estimation, heavy-tailed distribution, low-rank recovery, U-statistics.

1. Introduction

We focus on the problem of covariance estimation under various types of contamination, emphasizing practical methods that admit an efficient implementation. Assume that we are given independent copies $Y_1, \dots, Y_n$ of a random vector $Y \in \mathbb{R}^d$ that follows an unknown distribution $\mathcal{D}$ over $\mathbb{R}^d$, with mean $\mu := \mathbb{E}[X]$ and covariance matrix $\Sigma := \mathbb{E}[(Y-\mu)(Y-\mu)^T]$. The observations $Y_1, \dots, Y_n$ are assumed to be either *ε-adversarially corrupted*, meaning that an "adversary" could replace a fraction $\varepsilon < 0.5$ of observations with arbitrary (possibly random) vectors, or that the underlying distribution $\mathcal{D}$ is heavy-tailed, meaning that the Euclidean norm $\|Y\|_2$ is assumed to possess only four finite moments. Our goal is to construct an estimator of the covariance matrix Σ that performs well in the present framework.

As attested by, among others, Tukey (1960) and Huber (1964), robust estimation has a long history. During the past two decades, a growing number of applications has created high demand for practical tools for recovering high-

*Corresponding author.