EXTREME COMPRESSIVE SAMPLING FOR COVARIANCE ESTIMATION

MARTIN AZIZYAN, AKSHAY KRISHNAMURTHY, AND ARTI SINGH

[HTTPS://ARXIV.ORG/ABS/1506.00898](https://arxiv.org/abs/1506.00898)

PROBLEM SETUP

COVARIANCE ESTIMATION FROM COMPRESSIVE MEASUREMENTS

- Vectors $x_1, x_2, \dots x_n \in \mathbb{R}^d$
- Sample covariance $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$
- Measurements: $(A_t, A_t^T x_t)_{t=1}^n, \quad A_t \in \mathbb{R}^{d\times m}$ equivalently $(\Phi_t, \Phi_t x_t)_{t=1}^n, \quad \Phi_t \in \mathbb{R}^{d\times d}$
  - A_t orthonormal basis for an *m*-dimensional subspace of R^d; Φ_t m-dimensional orthogonal projection matrix, both drawn uniformly at random
- Distribution-free setting: no assumption on how x_i are generated
  - Goal: estimate the sample covariance matrix from the measurements
- Distributional setting: $x_i \sim \mathcal{N}(0, \Sigma)$
  - Goal: estimate the population covariance matrix Σ from the measurements
- Goal: bounds on the sample complexity $n = f(m, d, \epsilon, \delta)$ to achieve error ϵ w. p. $\geq 1 - \delta$

COVARIANCE ESTIMATE

- Intuitive estimator: $\hat{\Sigma}_1 = \frac{d^2}{nm^2} \sum_{t=1}^{n} \Phi_t x_t (\Phi_t x_t)^T$
- Debiased estimator: $\hat{\Sigma} = \frac{m((d+2)(d-1)\hat{\Sigma}_1 - (d-m)\mathrm{tr}(\hat{\Sigma}_1)\mathbf{I}_d)}{d(dm+d-2)}$
- Reason: Proposition 1:

$$\mathbb{E}\{\hat{\Sigma}_1\} = \frac{d(dm+d-2)\Sigma + d(d-m)\mathrm{tr}(\Sigma)\mathbb{I}_d}{m(d+2)(d-1)}$$

- The paper derives upper and lower bounds on the sample complexity of the debiased estimator

DISTRIBUTION FREE RESULTS

UPPER BOUNDS

- Theorem 2: Let $d \geq 2$, $\delta \in (0,1)$, $\delta \geq 4d^2 e^{-n/12}$. Then, there exist k_1 and k_2 such that, with probability $\geq 1 - \delta$

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq k_1 \|X\|_\infty^2 \sqrt{\frac{d^2 \log^2\left(\frac{nd}{\delta}\right)}{nm^2}} + k_2 \|X\|_\infty^2 \frac{d^2 \log^2\left(\frac{nd}{\delta}\right)}{nm^2}$$

- Theorem 3: Let

$$S_1 = \left\| \frac{1}{n} \sum_{t=1}^{n} \|x_t\|_2^2 x_t x_t^T \right\|_2 \qquad S_2 = \frac{1}{n} \sum_{t=1}^{n} \|x_t\|_2^4$$

Then under the same conditions as Theorem 2,

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq k_1 \left(\sqrt{\frac{d}{m} S_1} + \sqrt{\frac{d}{m^2} S_2} \right) \sqrt{\frac{\log(d/\delta)}{n}} + k_2 \frac{d\|X\|_{2,\infty}^2}{nm} \log(d/\delta)$$

DISTRIBUTIONAL SETTING

UPPER BOUND

- Corollary 4: when x_i are Gaussian distributed:

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq k_1 \|\Sigma\|_\infty \left(\sqrt{\frac{d^2 \log^6(nd/\delta)}{nm^2}} + \sqrt{\frac{\log(d/\delta)}{n}} \right) + k_2 \|\Sigma\|_\infty \left(\frac{d^2 \log^3(nd/\delta)}{nm^2} \right)$$

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq k_3 \|\Sigma\|_2 \left(\sqrt{\frac{d^3 \log^2(nd/\delta)}{nm^2}} + \frac{d^3 \log^2(nd/\delta)}{nm^2} + \sqrt{\frac{\log(2d/\delta)}{n}} \right)$$

- Corollary 5: $\mathrm{rank}(\Sigma) \leq k$

$$\|\hat{\Sigma}_k - \Sigma\|_2 \leq \kappa \|\Sigma\|_2 \left(\sqrt{\frac{dk}{nm} + \frac{dk^2}{nm^2}} + \frac{dk}{nm} \right) \log^2 \left(\frac{nd}{\delta} \right)$$

  - Interpretation: if $n = \Theta(d)$, can set $m = O(k \log^2(d)/\epsilon^2)$ to get

$$\|\hat{\Sigma} - \Sigma\|_2 \leq \kappa_1 \epsilon$$

REMARKS

- When $n \gg \frac{d^2}{m^2}$, ignoring logarithmic factors,

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq \tilde{O}\left(\sqrt{\frac{d^2}{nm^2}}\right)$$

$$\|\hat{\Sigma} - \Sigma\|_2 \leq \tilde{O}\left(\sqrt{\frac{d^3}{nm^2}}\right)$$

- To est. the population cov. matrix in the fully observed setting:

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq \tilde{O}\left(\sqrt{\frac{1}{n}}\right)$$

$$\|\hat{\Sigma} - \Sigma\|_2 \leq \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

- Sample size shrinks from n to nm^2/d^2 due to compressed meas.

- The above does not assume any structure on the covariance matrix

SIMULATION RESULTS

PERFORMANCE OF PROPOSED ESTIMATOR