Applications

(1) Covariance estimation: assume that $Y_1, \dots, Y_n \in \mathbb{R}^d$ are iid random vectors such that $\|Y_j\| \le K$ a.s., $\mathbb{E} Y_j = 0$ and $\mathbb{E}(Y_j Y_j^T) = \Sigma$.

How good is the estimator $\hat{\Sigma}_n = \frac{1}{n} \sum_{j=1}^n Y_j Y_j^T$, the empirical covariance matrix?

Recall matrix Bernstein's inequality:

$$P\left(\left\| \frac{1}{n} \sum_{j=1}^n X_j \right\| \ge \max\left[2\sigma \sqrt{\frac{t}{n}}, \frac{4}{3} L \frac{t}{n} \right] \right) \le 2d\, e^{-t}.$$

In our case, $X_j = Y_j Y_j^T - \Sigma$, so that $\|X_j\| \le 2K^2$, and

$$\sigma^2 = \|\mathbb{E} X_1^2\| = \|\mathbb{E}(Y_1 Y_1^T - \Sigma)^2\| = \|\underbrace{\mathbb{E}\|Y_1\|_2^2 Y_1 Y_1^T - \Sigma^2}_{\succeq 0}\|$$
$$\le \|\mathbb{E}\|Y_1\|_2^2 Y_1 Y_1^T\|$$

(i) Trivial bound: $\sigma^2 \le K^2 \|\Sigma\|$, whence

$$\|\hat{\Sigma}_n - \Sigma\| = O_p\left(K \sqrt{\frac{\|\Sigma\| \log(2d)}{n}} + \frac{K^2 \log(2d)}{n} \right).$$

$$\|\hat{\Sigma}_n - \Sigma\| \lesssim \varepsilon \|\Sigma\| \Rightarrow n \gtrsim \max\left(\frac{K^2 \log(2d)}{\varepsilon^2 \|\Sigma\|}, \frac{K^2 \log(2d)}{\varepsilon \|\Sigma\|} \right).$$
$$\simeq \frac{K^2 \log(2d)}{\|\Sigma\| \varepsilon^2}.$$

(ii) The bound $\sigma^2 \le K^2 \|\Sigma\|$ can be too crude.

Assume in addition that $\sup_{v \ne 0} \frac{\mathbb{E}\langle Y, v\rangle^4}{(\mathbb{E}\langle Y, v\rangle^2)^2} \le \tau$ (*)

("bounded kurtosis"). Then

$$\sigma^2 \le \tau \cdot \mathrm{tr}(\Sigma) \|\Sigma\|$$

Proof: $\left\| \mathbb{E}\, \|Y\|_2^2 YY^T \right\| = \sup_{\|v\|_2 = 1} \mathbb{E}\, \|Y\|_2^2 \langle Y, v\rangle^2$

$\leq \left(\mathbb{E}\, \|Y\|_2^4\right)^{1/2} \cdot \sup_{\|v\|_2=1} \mathbb{E}^{1/2} \langle Y, v\rangle^4$. Moreover,

$\mathbb{E}^{1/2} \langle Y, v\rangle^4 \leq \sqrt{\tau}\, \mathbb{E} \langle Y, v\rangle^2 \leq \sqrt{\tau}\, \|\Sigma\|$, and

$$\mathbb{E}^{1/2} \|Y\|_2^4 = \left(\sum_1^d \mathbb{E}\, Y_j^4 + \sum_{j \neq k} \mathbb{E}\, Y_j^2 Y_k^2 \right)^{1/2} \leq \left(\underbrace{\sum_1^d \mathbb{E}\, Y_j^4 + \sum_{j \neq k} \left(\mathbb{E}\, Y_j^4 \mathbb{E}\, Y_k^4\right)^{1/2}}_{\left[\sum_1^d \left(\mathbb{E}\, Y_j^4\right)^{1/2}\right]^2} \right)^{1/2}$$

$\leq \sqrt{\tau} \sum_1^d \mathbb{E}\, Y_j^2 = \sqrt{\tau}\, \mathrm{tr}\, \Sigma$,

and the inequality follows.

Lemma $\left\| \mathbb{E}\, \|Y\|_2^2 YY^T \right\| \geq \mathrm{tr}(\Sigma) \|\Sigma\|$

Proof: A version of the FKG (Fortuin-Kasteleyn-Ginibre) inequality states that, given $f, g : \mathbb{R}^d \to \mathbb{R}$ that are non-decreasing in each variable, we have that $\mathbb{E}[f(Y) g(Y)] \geq \mathbb{E} f(Y) \cdot \mathbb{E} g(Y)$

The claim of the lemma can be shown as follows: let v be a unit vector. Then $\langle \mathbb{E}\, \|Y\|_2^2 YY^T v, v\rangle = \mathbb{E}\, \|Y\|_2^2 \langle Y, v\rangle^2$. Let $\{v, u_2, \ldots, u_{d-1}\}$ be an orthonormal basis. Then $Y = (Y_1, \ldots, Y_d)^T$ in this basis. Take $f(Y) = Y_1^2$ and $g(Y) = \|Y\|_2^2$, then the claim follows by taking sup over v.

Remark (*) is satisfied by (a) elliptically symmetric distributions

$Y \overset{d}{=} g \cdot B U$, $U \sim \mathrm{Unif}(\text{sphere})$, g, U are independent

In particular, Gaussian and Multivariate t-distribution with $\nu > 4$ d.f.

(b) vectors with independent coordinates that possess exponential moments

(in this case,

$$\|\hat{\Sigma}_n - \Sigma\| = O_p\left(\sqrt{\tau}\,\|\Sigma\|\sqrt{\frac{r(\Sigma)\log(2d)}{n} + \frac{\kappa^2\log(2d)}{n}}\right).$$

$$\|\hat{\Sigma}_n - \Sigma\| \lesssim \varepsilon\|\Sigma\| \Rightarrow n \gtrsim \max\left(\frac{\tau\, r(\Sigma)\log(2d)}{\varepsilon^2}, \frac{\kappa^2\log(2d)}{\varepsilon\|\Sigma\|}\right)$$ - better than before!

- $\log(2d)$ factor is necessary: e.g., consider $Y = \sqrt{n}\, e_j$ w.p. $\frac{1}{n}$, $j = 1..d$. Then $\mathbb{E}\, YY^T = I_d$ but $\|\hat{\Sigma}_n - I_d\| \geq 1$ unless each e_j is sampled at least once $\Rightarrow n \geq c\, d\log d$ ["coupon collector" bound].

- What if $\mu = \mathbb{E}\, Y$ is also unknown?

  Then $\hat{\Sigma}_n = \frac{1}{n-1}\sum_{j=1}^n (Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T$ where $\bar{Y}_n = \frac{1}{n}\sum_1^n Y_j$.

  Note that $\hat{\Sigma}_n = \underbrace{\frac{1}{n-1}\sum_{j=1}^n (Y_j - \mu)(Y_j - \mu)^T}_{\text{previous case}} + \underbrace{\frac{n}{n-1}(\mu - \bar{Y}_n)(\mu - \bar{Y}_n)^T}_{\text{term of smaller order (Exercise)}}$

  Alternative approach is based on noticing that

  $\hat{\Sigma}_n = \frac{1}{\binom{n}{2}}\sum_{i<j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$ - a matrix-valued U-statistic.

  Matrix Bernstein inequality for such U-statistic holds with n replaced by $\lfloor \frac{n}{2} \rfloor$.

- Important application: Principal Component Analysis

  Let Π_j be the orthogonal projection onto the first j eigenvectors of Σ, and $\delta_j := \frac{1}{2}(\lambda_j - \lambda_{j+1}) > 0$. Then

  $\|\Pi_j(\Sigma) - \Pi_j(\hat{\Sigma}_n)\| \leq \frac{\|\Sigma - \hat{\Sigma}_n\|}{\delta_j}$ [this is a version of Davis-Kahan theorem for the spectral norm].

- Many other extensions of covariance estimation are known:

(i) "Masked Covariance" (Chen, Gittens, Tropp '18 — [PDF] The Masked Sample Covariance Estimator: An Analysis via Matrix Concentration Inequalities, RY CHEN, A GITTENS, JA TROPP - arXiv preprint arXiv:1109.1637, 2011 - Citeseer)

estimate $(\Sigma \odot M) = \begin{pmatrix} \Sigma_{11} \cdot M_{11} & \dots & \Sigma_{1d} M_{1d} \\ \vdots & \ddots & \\ \Sigma_{d1} \cdot M_{d1} & \dots & \Sigma_{dd} M_{dd} \end{pmatrix}$

Matrix M quantifies the "importance" of entries, or contains known prior information (e.g. about zeros of Σ).

(ii) Estimation with missing entries (Lounici '14 — High-dimensional covariance matrix estimation with missing observations, K Lounici - arXiv preprint arXiv:1201.2577, 2012 - arxiv.org)

each coordinate of Y_j is observed with probability $(1-\delta)$ and is unknown with probability δ.

- What about covariance estimation from "heavy-tailed" measurements, e.g. for $X \in \mathbb{R}^d$ such that only $\mathbb{E}\|X\|_2^4 < \infty$? In this case, sample covariance does not admit tight deviation bounds any longer, and different estimators based on delicate truncation are required. See the papers below for example:

(i) Robust covariance estimation under $L_4 - L_2$ norm equivalence, S Mendelson, N Zhivotovskiy - arXiv preprint arXiv:1809.10462, 2018 - arxiv.org

(ii) Robust Modifications of U-statistics and Applications to Covariance Estimation Problems, S Minsker, X Wei - arXiv preprint arXiv:1801.05565, 2018 - arxiv.org

(iii) Estimation of the covariance structure of **heavy-tailed** distributions, S **Minsker**, X **Wei** - arXiv preprint arXiv:1708.00502, 2017 - arxiv.org

---

(2) Stochastic Block Model: assume that $G = (V, E)$ is a random undirected graph with n vertices, and

$$Y_{i,j} = \begin{cases} 1, & V_i \text{ and } V_j \text{ are connected by an edge} \\ 0, & \text{else} \end{cases} \quad \text{are independent.}$$

[In other words, Y is the <u>adjacency matrix</u> [$Y_{i,i} = 0$ by convention].

Assume that $V = V^{(1)} \cup V^{(2)}$ - two "communities", and that

$$p_{i,j} = P(Y_{i,j} = 1) = \begin{cases} p, & i,j \in V^{(1)} \text{ or } i,j \in V^{(2)} \\ q < p, & \text{else} \end{cases}$$

<u>Goal</u>: recover $V^{(1)}$ and $V^{(2)}$.

Let M_* be the matrix of probabilities, $(M_*)_{i,j} = p_{i,j}$, $i,j = 1, \dots n$

Then $\operatorname{rank}(M_*) = 2$ since $M_* = Z^T \begin{pmatrix} p_{1,1} & p_{1,2} \\ p_{2,1} & p_{2,2} \end{pmatrix} Z$,

$Z \in \mathbb{R}^{n \times 2}$, $Z_{i,j} = \mathbb{I}\{V_i \in V^{(j)}\}$, and

(1) $M_* = \frac{p+q}{2} \mathbb{1}_n \mathbb{1}_n^T + \frac{p-q}{2} \eta \eta^T$ where $\eta \in \{+1, -1\}^n$ is s.t.

$\eta_j = 1 \Rightarrow V_j \in V^{(1)}$

$\eta_j = -1 \Rightarrow V_j \in V^{(2)}$

<u>If</u> moreover $\sum_j \eta_j = 0$ [equal communities]

$\Rightarrow$ (1) is the eigenvalue decomposition of M_*, and $\frac{\eta}{\sqrt{n}}$ is the eigenvector corresponding to $\lambda_2 = n(p-q)$.

Note that $\mathbb{E} Y = M_* - \operatorname{diag}(M_*)$, and

$$\|Y - M_*\| \le \|Y - M_* - \operatorname{diag}(M_*)\| + \underbrace{\|\operatorname{diag}(M_*)\|}_{\le 1}$$

$$Y - \mathbb{E} Y = \sum_{i<j} A_{ij}, \text{ with } A_{ij} = (Y_{i,j} - p_{i,j})(e_i e_j^T + e_j e_i^T)$$

$$\|A_{ij}\| \le 1 \quad , \quad \mathbb{E} A_{ij}^2 = p_{ij}(1-p_{ij})\left(e_i e_i^T + e_j e_j^T\right)$$

$$\Rightarrow \left\| \sum_{i<j} \mathbb{E} A_{ij}^2 \right\| \le \frac{1}{4} n$$

$$\Rightarrow \|Y - \mathbb{E}Y\| \le \max\left(\sqrt{n(t+\log(2n)}, \frac{4}{3}(t+\log(2n)\right)$$

$$\Rightarrow \|Y - M^*\| \le \underbrace{1 + \sqrt{n(t+\log(2n)} + \frac{4}{3}(t+\log(2n))}_{\delta_n} \quad \text{w.p} \ge 1 - e^{-t}.$$

$\Rightarrow$ as long as $\min\left(q, \frac{p-q}{2}\right) \cdot n > \delta_n$ $\Rightarrow$ the eigenvectors are close with high probability.

- For recent advances on the topic, see this paper for instance:

Improved clustering algorithms for the bipartite stochastic block model
M Ndaoud, S Sigalla, AB Tsybakov - arXiv preprint arXiv:1911.07987, 2019 - arxiv.org