High Dimensional Statistics

Homework 4

Javad Hezareh (404208723)

Fall 2025 (1404)

1 Theoretical Foundations of Covariance Estimation

(a) Having seen samples $x_1, x_2, \ldots, x_n$ drawn i.i.d. from the underlying distribution of the problem, then for the sample covariance matrix $\hat{\Sigma}$ defined as

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, we can show that $\hat{\Sigma}$ is an unbiased estimator of the true covariance matrix Σ as follows:

$$\begin{aligned}
\mathbb{E}[\hat{\Sigma}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(x_i - \bar{x})(x_i - \bar{x})^T] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[x_i x_i^T] - \mathbb{E}[x_i \bar{x}^T] - \mathbb{E}[\bar{x} x_i^T] + \mathbb{E}[\bar{x}\bar{x}^T]\right).
\end{aligned}$$

Since the samples are i.i.d., we have $\mathbb{E}[x_i] = \mu$ and $\mathbb{E}[x_i x_i^T] = \Sigma + \mu\mu^T$. Also, $\mathbb{E}[\bar{x}] = \mu$. Thus,

$$\begin{aligned}
\mathbb{E}[\hat{\Sigma}] &= \frac{1}{n}\sum_{i=1}^{n}\left(\Sigma + \mu\mu^T - \mu\mu^T - \mu\mu^T + \mu\mu^T\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Sigma \\
&= \Sigma.
\end{aligned}$$

Therefore, $\hat{\Sigma}$ is an unbiased estimator of Σ.

(b) To have the relation provided for estimation error hold we just need to apply the variational definiton of operator norm:

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_2 &= \sup_{v \in S^{d-1}} |v^T(\hat{\Sigma} - \Sigma)v| \\
&= \sup_{v \in S^{d-1}} |v^T\hat{\Sigma}v - v^T\Sigma v| \\
&= \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} v^T x_i x_i^T v - v^T \Sigma v \right| \\
&= \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle x_i, v \rangle^2 - v^T \Sigma v \right|.
\end{aligned}$$

2 Covariance Estimation

(a) Fix $i, j \in \{1, \dots, d\}$, and define

$$Z_k := X_i^{(k)} X_j^{(k)} \qquad \text{and} \qquad Y_k := Z_k - \mathbb{E}[Z_k] = X_i^{(k)} X_j^{(k)} - \Sigma_{i,j}.$$

Then

$$\hat{\Sigma}_{i,j} - \Sigma_{i,j} = \frac{1}{n} \sum_{k=1}^{n} Y_k,$$

where $Y_1, \dots, Y_n$ are i.i.d. mean-zero.

Since $X_i/\sqrt{\Sigma_{i,i}}$ is sub-Gaussian with parameter σ^2, we can write (in Orlicz norm form)

$$\|X_i\|_{\psi_2} \le c\,\sigma\sqrt{\Sigma_{i,i}}, \qquad \|X_j\|_{\psi_2} \le c\,\sigma\sqrt{\Sigma_{j,j}},$$

for some absolute constant $c > 0$. A standard fact is that the product of two sub-Gaussian random variables is sub-exponential and satisfies

$$\|X_i X_j\|_{\psi_1} \le C\,\|X_i\|_{\psi_2}\|X_j\|_{\psi_2} \le C\,\sigma^2\sqrt{\Sigma_{i,i}\Sigma_{j,j}},$$

hence Z_k is sub-exponential. Centering does not change the sub-exponential norm by more than a constant factor, so

$$\|Y_k\|_{\psi_1} \le C'\sigma^2\sqrt{\Sigma_{i,i}\Sigma_{j,j}} =: K_{i,j}.$$

Now apply Bernstein's inequality for i.i.d. mean-zero sub-exponential variables: there exist absolute constants $c_1, c_2 > 0$ such that for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} Y_k\right| > \epsilon\right) \le 2\exp\left(-c_1 n \,\min\left(\frac{\epsilon^2}{K_{i,j}^2}, \frac{\epsilon}{K_{i,j}}\right)\right).$$

In particular, for $\epsilon \le K_{i,j}$ we get a purely quadratic tail:

$$\mathbb{P}\left(\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| > \epsilon\right) \le 2\exp\left(-c_1 n \frac{\epsilon^2}{K_{i,j}^2}\right).$$

Therefore, by absorbing constants (and the dependence on $K_{i,j}$) into $C_1, C_2 > 0$, we can write the required form:

$$\mathbb{P}\left(\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| > \epsilon\right) \le C_1 e^{-C_2 n \epsilon^2}.$$

(b) Using the bound from part (a) and a union bound over all (i, j):

$$\begin{aligned}\mathbb{P}\left(\max_{i,j}\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| > \epsilon\right) &\le \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbb{P}\left(\left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| > \epsilon\right) \\ &\le d^2\, C_1 e^{-C_2 n \epsilon^2}.\end{aligned}$$

We want the RHS to be at most $1/n$. It is enough to choose ϵ such that

$$d^2 C_1 e^{-C_2 n \epsilon^2} \le \frac{1}{n}.$$

Taking logs,

$$\log d^2 + \log C_1 - C_2 n\epsilon^2 \le -\log n \quad \Longrightarrow \quad C_2 n\epsilon^2 \ge 2\log d + \log C_1 + \log n.$$

Hence one valid choice is

$$\epsilon \;\ge\; \sqrt{\frac{2\log d + \log C_1 + \log n}{C_2 n}} \;=\; O\left(\sqrt{\frac{\log d + \log n}{n}}\right),$$

and with this choice we obtain

$$\max_{i,j} \left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| = O\left(\sqrt{\frac{\log d + \log n}{n}}\right) \quad \text{with probability at least } 1 - \frac{1}{n}.$$

3 Covariance Estimation under Missing Data

(a) We compute $\mathbb{E}[Y_{ki}Y_{kj}]$. Since $Y_{ki}Y_{kj} = \delta_{ki}\delta_{kj}X_{ki}X_{kj}$ and δ's are independent of X_k,

$$\mathbb{E}[Y_{ki}Y_{kj}] = \mathbb{E}[\delta_{ki}\delta_{kj}]\,\mathbb{E}[X_{ki}X_{kj}].$$

If $i \neq j$, then δ_{ki} and δ_{kj} are independent Bernoulli(π), so

$$\mathbb{E}[\delta_{ki}\delta_{kj}] = \mathbb{E}[\delta_{ki}]\,\mathbb{E}[\delta_{kj}] = \pi^2, \qquad \mathbb{E}[X_{ki}X_{kj}] = \sigma_{ij},$$

hence $\mathbb{E}[Y_{ki}Y_{kj}] = \pi^2\sigma_{ij}$.

If $i = j$, then $\delta_{ki}^2 = \delta_{ki}$ and X_{ki}^2 has mean σ_{ii}, so

$$\mathbb{E}[Y_{ki}Y_{ki}] = \mathbb{E}[\delta_{ki}X_{ki}^2] = \mathbb{E}[\delta_{ki}]\,\mathbb{E}[X_{ki}^2] = \pi\sigma_{ii}.$$

Therefore,

$$\mathbb{E}[Y_{ki}Y_{kj}] = \begin{cases} \pi^2\sigma_{ij}, & i \neq j, \\ \pi\sigma_{ii}, & i = j. \end{cases}$$

(b) For $i \neq j$,

$$\mathbb{E}[\hat{\sigma}_{ij}] = \frac{1}{n\pi^2}\sum_{k=1}^{n}\mathbb{E}[Y_{ki}Y_{kj}] = \frac{1}{n\pi^2}\sum_{k=1}^{n}\pi^2\sigma_{ij} = \sigma_{ij},$$

so $\hat{\sigma}_{ij}$ is unbiased off-diagonal.

For $i = j$,

$$\mathbb{E}[\hat{\sigma}_{ii}] = \frac{1}{n\pi^2}\sum_{k=1}^{n}\mathbb{E}[Y_{ki}^2] = \frac{1}{n\pi^2}\sum_{k=1}^{n}\pi\sigma_{ii} = \frac{1}{\pi}\sigma_{ii},$$

so the diagonal entries are biased (inflated by a factor $1/\pi$). Intuitively, Y_{ki}^2 is observed only when $\delta_{ki} = 1$, which happens with probability π, so dividing by π^2 over-corrects on the diagonal.

A natural correction is to use different normalizations:

$$\tilde{\sigma}_{ij} := \begin{cases} \frac{1}{n\pi^2}\sum_{k=1}^{n} Y_{ki}Y_{kj}, & i \neq j, \\ \frac{1}{n\pi}\sum_{k=1}^{n} Y_{ki}^2, & i = j, \end{cases}$$

which makes both cases unbiased.

(c) Assume $\|X_k\|_\infty \leq M$ almost surely. Fix $i \neq j$. Define

$$W_k := \frac{Y_{ki}Y_{kj}}{\pi^2}.$$

Then

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n} W_k, \qquad \mathbb{E}[W_k] = \frac{1}{\pi^2}\mathbb{E}[Y_{ki}Y_{kj}] = \sigma_{ij} \quad (\text{since } i \neq j).$$

Moreover, since $|\delta_{ki}\delta_{kj}| \leq 1$ and $|X_{ki}X_{kj}| \leq M^2$,

$$|W_k| = \left|\frac{\delta_{ki}\delta_{kj}X_{ki}X_{kj}}{\pi^2}\right| \leq \frac{M^2}{\pi^2} \quad \Rightarrow \quad W_k \in \left[-\frac{M^2}{\pi^2}, \frac{M^2}{\pi^2}\right] \text{ a.s.}$$

Applying Hoeffding's inequality to the bounded i.i.d. variables W_k gives, for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} W_k - \sigma_{ij}\right| > t\right) \le 2\exp\left(-\frac{2nt^2}{\left(\frac{2M^2}{\pi^2}\right)^2}\right) = 2\exp\left(-\frac{n\pi^4 t^2}{2M^4}\right).$$

Hence the claim holds with some absolute constant $c > 0$ (e.g. $c = 1/2$):

$$\mathbb{P}\big(|\hat{\sigma}_{ij} - \sigma_{ij}| > t\big) \le 2\exp\left(-c\,\frac{n\pi^4 t^2}{M^4}\right), \qquad i \ne j.$$

(d) Missing entries effectively reduce the amount of information used to estimate Σ. In particular, an off-diagonal product $X_{ki}X_{kj}$ is only observed when *both* coordinates are present, which occurs with probability $\mathbb{P}(\delta_{ki}\delta_{kj} = 1) = \pi^2$. Thus, the raw empirical average of $Y_{ki}Y_{kj}$ is shrunk by a factor π^2 in expectation.

It is therefore necessary to correct by π^2 in the estimator to remove this systematic shrinkage and avoid bias toward 0 for off-diagonal entries. On the diagonal, the observation probability is only π (since it depends on a single coordinate), which is why a π^2 correction produces bias and one should instead correct the diagonal using π.

4 Sparse Covariance Estimation via Hard Thresholding

(a) Fix (i, j), and define $Z_k := X_{ki}X_{kj}$ so that $\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^n Z_k$ and $\mathbb{E}[Z_k] = \sigma_{ij}$. Since X_{ki} and X_{kj} are sub-Gaussian with $\|X_{ki}\|_{\psi_2} \le \kappa$ and $\|X_{kj}\|_{\psi_2} \le \kappa$, their product is sub-exponential and

$$\|Z_k\|_{\psi_1} = \|X_{ki}X_{kj}\|_{\psi_1} \le C\, \|X_{ki}\|_{\psi_2}\|X_{kj}\|_{\psi_2} \le C\kappa^2,$$

for some absolute constant $C > 0$. Let $Y_k := Z_k - \mathbb{E}[Z_k]$ so that $\mathbb{E}[Y_k] = 0$ and $\|Y_k\|_{\psi_1} \le C'\kappa^2$.

Applying Bernstein's inequality for i.i.d. mean-zero sub-exponential random variables, there exist absolute constants $c_1, c_2 > 0$ such that for all $t > 0$,

$$\mathbb{P}\Big(\big|\hat{\sigma}_{ij} - \sigma_{ij}\big| > t\Big) = \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^n Y_k\Big| > t\Big) \le 2\exp\Big(-c_1 n\, \min\Big(\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\Big)\Big).$$

Take $t = \lambda$ with

$$\lambda = C_1\kappa^2\sqrt{\frac{\log d}{n}}$$

for a large enough universal constant $C_1 > 0$. For this choice, the quadratic term dominates (for large d and the above scaling), and we obtain

$$\mathbb{P}\Big(\big|\hat{\sigma}_{ij} - \sigma_{ij}\big| > \lambda\Big) \le 2\exp(-c_2\log d) = 2d^{-c_2}.$$

Now apply a union bound over all d^2 pairs (i, j):

$$\mathbb{P}\Big(\|\hat{\Sigma} - \Sigma\|_{\max} > \lambda\Big) \le \sum_{i,j}\mathbb{P}\Big(\big|\hat{\sigma}_{ij} - \sigma_{ij}\big| > \lambda\Big) \le d^2 \cdot 2d^{-c_2}.$$

Choosing C_1 (hence c_2) so that $c_2 \ge 4$ gives

$$\mathbb{P}\Big(\|\hat{\Sigma} - \Sigma\|_{\max} \le \lambda\Big) \ge 1 - 2d^{-2}.$$

(b) Condition on the event

$$\mathcal{E} := \{\|\hat{\Sigma} - \Sigma\|_{\max} \le \lambda\},$$

we would have $\forall i, j : |\hat{\sigma}_{ij} - \sigma_{ij}| \le \lambda$. We bound each entry of $\hat{\Sigma}_\lambda - \Sigma$. Fix (i, j) and consider two cases.

Case 1: $|\sigma_{ij}| \ge 2\lambda$. On $\mathcal{E}$, we have

$$|\hat{\sigma}_{ij}| \ge |\sigma_{ij}| - |\hat{\sigma}_{ij} - \sigma_{ij}| \ge 2\lambda - \lambda = \lambda,$$

hence $(\hat{\Sigma}_\lambda)_{ij} = \hat{\sigma}_{ij}$. Therefore

$$|(\hat{\Sigma}_\lambda)_{ij} - \sigma_{ij}| = |\hat{\sigma}_{ij} - \sigma_{ij}| \le \lambda.$$

Case 2: $|\sigma_{ij}| < 2\lambda$. If $|\hat{\sigma}_{ij}| \ge \lambda$, then $(\hat{\Sigma}_\lambda)_{ij} = \hat{\sigma}_{ij}$ and again

$$|(\hat{\Sigma}_\lambda)_{ij} - \sigma_{ij}| = |\hat{\sigma}_{ij} - \sigma_{ij}| \le \lambda.$$

If $|\hat{\sigma}_{ij}| < \lambda$, then $(\hat{\Sigma}_\lambda)_{ij} = 0$ and thus

$$|(\hat{\Sigma}_\lambda)_{ij} - \sigma_{ij}| = |\sigma_{ij}| < 2\lambda.$$

Combining both cases, we obtain the uniform entrywise bound on $\mathcal{E}$:

$$|(\hat{\Sigma}_\lambda)_{ij} - \sigma_{ij}| \le 2 \min\big(|\sigma_{ij}|, \lambda\big), \qquad \forall i, j.$$

Squaring and summing over i, j yields

$$\|\hat{\Sigma}_\lambda - \Sigma\|_F^2 = \sum_{i,j} \big((\hat{\Sigma}_\lambda)_{ij} - \sigma_{ij}\big)^2 \le \sum_{i,j} 4 \min\big(\sigma_{ij}^2, \lambda^2\big) = 4 \sum_{i,j} \min\big(\sigma_{ij}^2, \lambda^2\big).$$

(c) On the event $\mathcal{E}$ from part (b),

$$\frac{1}{d}\|\hat{\Sigma}_\lambda - \Sigma\|_F^2 \le \frac{4}{d} \sum_{i,j} \min(\sigma_{ij}^2, \lambda^2).$$

Use the inequality (given in the hint) that for $0 \le q < 2$,

$$\min(a^2, b^2) \le |a|^q\, b^{2-q}.$$

Applying this with $a = \sigma_{ij}$ and $b = \lambda$ gives

$$\min(\sigma_{ij}^2, \lambda^2) \le |\sigma_{ij}|^q\, \lambda^{2-q}.$$

Therefore,

$$\sum_{i,j} \min(\sigma_{ij}^2, \lambda^2) \le \lambda^{2-q} \sum_{j=1}^{d} \sum_{i=1}^{d} |\sigma_{ij}|^q \le \lambda^{2-q} \sum_{j=1}^{d} c_0 = d c_0\, \lambda^{2-q},$$

where we used $\Sigma \in \mathcal{U}_q(c_0)$, that is $\max_j \sum_i |\sigma_{ij}|^q \le c_0$. Plugging back,

$$\frac{1}{d}\|\hat{\Sigma}_\lambda - \Sigma\|_F^2 \le 4 c_0\, \lambda^{2-q}.$$

Using the choice from part (a), $\lambda = C_1 \kappa^2 \sqrt{\frac{\log d}{n}}$, we obtain

$$\lambda^{2-q} = (C_1 \kappa^2)^{2-q} \left(\frac{\log d}{n}\right)^{1-\frac{q}{2}}.$$

Hence, with probability at least $1 - 2d^{-2}$,

$$\frac{1}{d}\|\hat{\Sigma}_\lambda - \Sigma\|_F^2 \le C'\, c_0 \left(\frac{\log d}{n}\right)^{1-\frac{q}{2}},$$

where $C' > 0$ is a constant absorbing $(C_1 \kappa^2)^{2-q}$ and the factor 4.

5 Covariance Estimation under Toeplitz Structure

(a) Fix $h \in \{0, \dots, d-1\}$. By linearity of expectation,

$$\mathbb{E}[\tilde{\sigma}_h] = \frac{1}{d-h} \sum_{k=1}^{d-h} \mathbb{E}[\hat{\sigma}_{k,k+h}].$$

For any (i,j), $\hat{\sigma}_{ij}$ is the (i,j) entry of $\hat{\Sigma}$, and since $\mathbb{E}[\hat{\Sigma}] = \Sigma$ (because $\hat{\Sigma}$ is the sample mean of $X_k X_k^T$),

$$\mathbb{E}[\hat{\sigma}_{ij}] = \Sigma_{ij}.$$

Therefore,

$$\mathbb{E}[\hat{\sigma}_{k,k+h}] = \Sigma_{k,k+h} = \sigma_{|k-(k+h)|} = \sigma_h,$$

and hence

$$\mathbb{E}[\tilde{\sigma}_h] = \frac{1}{d-h} \sum_{k=1}^{d-h} \sigma_h = \sigma_h.$$

Consequently, for any i, j,

$$\mathbb{E}[\tilde{\Sigma}_{ij}] = \mathbb{E}[\tilde{\sigma}_{|i-j|}] = \sigma_{|i-j|} = \Sigma_{ij},$$

hence we have $\mathbb{E}[\tilde{\Sigma}] = \Sigma$.

(b) Let $\|A\|_\infty$ denote the induced ℓ_∞ operator norm:

$$\|A\|_\infty := \max_{1 \le i \le d} \sum_{j=1}^{d} |A_{ij}|.$$

Assume $A \in \mathbb{R}^{d \times d}$ is symmetric. Let λ be an eigenvalue of A with eigenvector $v \neq 0$, so $Av = \lambda v$. Choose an index $i^\star$ such that $|v_{i^\star}| = \|v\|_\infty$.

Looking at the $i^\star$-th coordinate of $Av = \lambda v$,

$$\lambda v_{i^\star} = (Av)_{i^\star} = \sum_{j=1}^{d} A_{i^\star j} v_j.$$

Taking absolute values and using $|v_j| \le \|v\|_\infty = |v_{i^\star}|$,

$$|\lambda| \, |v_{i^\star}| \le \sum_{j=1}^{d} |A_{i^\star j}| \, |v_j| \le \sum_{j=1}^{d} |A_{i^\star j}| \, |v_{i^\star}| = \left(\sum_{j=1}^{d} |A_{i^\star j}| \right) |v_{i^\star}|.$$

Since $|v_{i^\star}| > 0$, we can cancel it to get

$$|\lambda| \le \sum_{j=1}^{d} |A_{i^\star j}| \le \max_{1 \le i \le d} \sum_{j=1}^{d} |A_{ij}| = \|A\|_\infty.$$

For symmetric A, the operator norm equals the largest absolute eigenvalue:

$$\|A\|_2 = \max_{\lambda \in \mathrm{spec}(A)} |\lambda|.$$

Combining with the eigenvalue bound above yields

$$\|A\|_2 \le \|A\|_\infty = \max_{1 \le i \le d} \sum_{j=1}^{d} |A_{ij}|.$$

(c) Let $E := \tilde{\Sigma} - \Sigma$. Then E is symmetric Toeplitz with entries

$$E_{ij} = \tilde{\sigma}_{|i-j|} - \sigma_{|i-j|}.$$

Using part (b),

$$\|\tilde{\Sigma} - \Sigma\|_2 = \|E\|_2 \le \|E\|_\infty = \max_{1\le i\le d} \sum_{j=1}^{d} |E_{ij}|.$$

For a fixed row i, the value $|i-j|$ ranges from 0 to $d-1$. For each lag $h \ge 1$, there are at most two indices j such that $|i-j| = h$ (one on each side), and for $h = 0$ there is exactly one. Therefore, for every i,

$$\sum_{j=1}^{d} |E_{ij}| \le |\tilde{\sigma}_0 - \sigma_0| + 2\sum_{h=1}^{d-1} |\tilde{\sigma}_h - \sigma_h|.$$

Taking the maximum over i gives the same bound, hence

$$\|\tilde{\Sigma} - \Sigma\|_2 \le |\tilde{\sigma}_0 - \sigma_0| + 2\sum_{h=1}^{d-1} |\tilde{\sigma}_h - \sigma_h| \le 2\sum_{h=0}^{d-1} |\tilde{\sigma}_h - \sigma_h|.$$

Explanation: The Toeplitz structure reduces the effective number of parameters from d^2 to only d lags $\{\sigma_h\}_{h=0}^{d-1}$. The estimator $\tilde{\sigma}_h$ averages all sample covariance entries along the h-th diagonal, which decreases variance compared to estimating each Σ_{ij} separately. Moreover, the operator norm error is controlled by the sum of lag errors rather than a maximum over d^2 entries. If the correlations σ_h decay sufficiently fast with h (so large lags are small and/or can be truncated), then only a moderate number of lags contribute meaningfully to the bound. This is why $\tilde{\Sigma}$ can be consistent in operator norm even when $d \gg n$: one is effectively estimating a low-dimensional structured object by pooling many repeated entries per lag.