Covariance estimation with missing observation

Karim Lounici Grégoire Pacreau

March 2022

1 Introduction

Let $X, X_1, \dots, X_n$ be i.i.d. zero mean vectors with unknown covariance matrix $\Sigma = \mathbb{E}[X \bigotimes X]$. Our objective is to estimate the unknown covariance matrix Σ when the vectors $X_1, \dots, X_n$ are partially observed, that is, when some of their components are not observed. More precisely, we consider the following framework. Denote by $X_i^{(j)}$ the jth component of the vector X_i. We assume taht each component $X_i^{(j)}$ is observed independently with probability $\delta \in (0, 1]$.

$$Y_i^{(j)} = \delta_{i,j} X_i^{(j)}, \qquad 1 \le i \le n, 1 \le j \le p \tag{1.1}$$

We can think of the $\delta_{i,j}$ as masked variables. If $\delta_{i,j} = 0$, then we cannot observe the jth component of X_i and the default value 0 is assigned to $Y_i^{(j)}$. Our goal is to estimate Σ given the partial observations $Y_1, \dots, Y_n$. This problem has been previously studied in [1], albeit with less sharp bounds than those presented in this paper. Furthermore, this paper applies the techniques of [1] to build robust estimators countering cell-wise contamination of features.

For example of interest in missing values in reinforcement learning: paper by scornet, josse, prost and varoquaux.

2 Tools and definitions

2.1 Sub-exponential random vectors

We recall the definition and some basic properties of sub-exponential random vectors.

Definition 1 *For any $\alpha \ge 1$, the ψ_α-norms of a real-valued random variable V are defined as:*

$$\|V\|_{\psi_\alpha} = \inf\{u > 0, \mathbb{E}\exp(|V|^\alpha / u^\alpha) \le 2\}$$

We say that a random variable V with values in $\mathbb{R}$ is sub-exponential if $\|V\|_{\psi_\alpha} < \infty$ for some $\alpha \ge 1$. If $\alpha = 2$, we say that V is sub-Gaussian.

Here are two well-known properties of sub-exponential random variables:

1. For any real-valued variable V such that $\|V\|_\alpha < \infty$ for some $\alpha > 1$, we have

$$\mathbb{E}[|V|^m] \le 2\frac{m}{\alpha}\Gamma\left(\frac{m}{\alpha}\right)\|V\|_{\psi_\alpha}^m \qquad \forall m \ge 1,$$

   where $\Gamma(\dot{})$ is the Gamma function.

2. If a real-valued random variable V is sub-Gaussian, then V^2 is sub-exponential. Indeed, we have:

$$\|V^2\|_{\psi_1} \le 2\|V\|_{\psi_2}^2$$

Definition 2 *A random vector $X \in \mathbb{R}^p$ is sub-exponential if $\langle X, x\rangle$ are sub-exponential random variables for $x \in \mathbb{R}^p$. The ψ_α-norms of a random vector X are defined as:*

$$\|X\|_{\psi_\alpha} = \sup_{x\in\mathbb{R}^p, |x|_2=1} \|\langle X, x\rangle\|_{\psi_\alpha}, \qquad \alpha \geq 1$$

We recall the Bernstein inequality for sub-exponential real-valued random variables (CITATION NEEDED).

Proposition 1 *Let $Y_1, \dots, Y_n$ be independent centered sub-exponential random variables, and $K = \max_i \|Y_i\|_{\psi_1}$. Then, for every $t \geq 0$, we have with probability at least $1 - e^{-t}$:*

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \leq CK\left(\sqrt{\frac{t}{n}} \wedge \frac{t}{n}\right)$$

where $C > 0$ is an absolute constant.

2.2 Some elements of matrix theory

3 Cell-wise contamination

In 1967, Huber introduced a general framework for the contamination of a dataset known as ϵ-contamination [**?**]. This model supposes that among all rows of the dataset, an unknown but small proportion ϵ are observed with gross error. Here is a formal definition of this form of contamination:

Definition 3 *Let $\mathbb{P}_\Sigma$ and be the true distribution of the data, with Σ the covariance matrix we want to estimate. Let $\mathbb{Q}$ be the distribution of the outliers. Let $X_1, \dots, X_n$ be our finite samples observed. Huber contamination means that there exists $Z_1, \dots, Z_n \sim \mathcal{B}(\epsilon)$, such that the (X_i, Z_i) are iid and, $\forall A \subset \mathbb{R}^p$:*

$$\begin{cases} \mathbb{P}(X_i \in A | Z_i = 0) = \mathbb{P}_\Sigma(A) \\ \mathbb{P}(X_i \in A | Z_i = 1) = \mathbb{Q}(A) \end{cases} \tag{3.1}$$

where $\mathcal{B}(\epsilon)$ is the Bernoulli law of parameter $\epsilon \in [0, 1]$.

This framework is well known in the field of robust statistics, with a plethora of unbiased estimators of both mean and covariance (CITATION NEEDED). However, there exist cases where contamination may appear independently from the row structure of our samples. [2] presents a generalisation of the ϵ-contamination where outliers appear within rows, with specific components of the row being outliers instead of its entire information. We can present this new contamination as follows:

Definition 4 *Using the same notations as before, Cell Wise contamination entails that there exists a set of random variables $(Z_{i,j})_{1\leq i\leq n, 1\leq j\leq p} \sim \mathcal{B}(\epsilon)$, such that the $(X_i^{(j)}, Z_{i,j})$ are iid and, $\forall A \subset \mathbb{R}$:*

$$\begin{cases} \mathbb{P}(X_i^{(j)} \in A | Z_{i,j} = 0) = \mathbb{P}_{\Sigma,j}(A) \\ \mathbb{P}(X_i^{(j)} \in A | Z_{i,j} = 1) = \mathbb{Q}_j(A) \end{cases} \tag{3.2}$$

where $\mathbb{P}_{\Sigma,j}$ and $\mathbb{Q}_j$ are respectively the marginal distribution of $\mathbb{P}_\Sigma$ and $\mathbb{Q}$ at dimension j.

For a pratical example of such a contamination, one can consider a dataset made where each row corresponds to a series of measures by different sensors. There is no particular reason that an error on one sensor propagates to the others, thus abnormal readings will be restricted to a particular component of the sample.

4 Proofs

4.1 Proof of upper bound

Let $X_1, \dots, X_n$ be i.i.d. zero mean vectors following a $\mathcal{N}(0, \Sigma)$ law, with Σ an unknown positive definite Hermitian matrix of size $p \times p$. Let for $1 \leq i \leq n$ and $1 \leq j \leq p$, δ_{ij} follows an Bernoulli lax $\mathcal{B}(\delta)$, with $\delta \in [0, 1]$, such that δ_{ij} is independent both from $X_i^{(j)}$, that is the jth component of X_i, and of any other Bernoulli random variable. Let finally $Y_i^{(h)} = \delta_{ij} X_i^{(j)}$ the observed random variable with missing values. We want an upper bound to the error between the

4.1.1 Concentration of $Y_i \otimes Y_i$ to its mean

First, let us find an upper bound to $\|n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}[Y \otimes Y]\|_2$.

Let us suppose matrix $\boldsymbol{\delta} = (\delta_i)_{i=1}^n$ fixed and known. By definition of the operator norm, we can express this error in terms of Rayleigh quotients:

$$\begin{aligned} \|n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}[Y \otimes Y]\|_2 &= \max_{\|u\|=1} u \left(n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}[Y \otimes Y] \right) u^\top \\ &= \max_{\|u\|=1} n^{-1} \sum_i u(Y_i \otimes Y_i)u^\top - u\mathbb{E}[Y \otimes Y]\, u^\top \end{aligned} \tag{4.1}$$

Let us first examine the first term. Let $\delta_i = (\delta_{i1}, \dots, \delta ip)$. Let $\odot$ be the Hadamard product of two vectors. We have:

$$\begin{aligned} n^{-1} \sum_i u(Y_i \otimes Y_i)u^\top &= n^{-1} \sum_i u\left((\delta_i \odot X_i) \otimes (\delta_i \odot X_i)\right) u^\top \\ &= n^{-1} \sum_i u(\delta_i \odot X_i)^\top \langle \delta_i \odot X_i, u \rangle \end{aligned}$$

using a basic property of the tensor product: $\forall u, v, w \in \mathbb{R}^p$ vectors, $(u \otimes v)w = u\langle v, w \rangle$. From there, we twice apply the following property of the Hadamard product: $\forall u, v, w \in \mathbb{R}^p, u(v \odot w) = (u \odot v)w$. We have:

$$\begin{aligned} n^{-1} \sum_i u(\delta_i \odot X_i)^\top \langle \delta_i \odot X_i, u \rangle &= n^{-1} \sum_i (u \odot \delta_i) X_i^\top \langle \delta_i \odot X_i, u \rangle \\ &= n^{-1} \sum_i (u \odot \delta_i) X_i^\top \langle X_i, u \odot \delta_i \rangle \\ &= n^{-1} \sum_i \langle u \odot \delta_i, X_i \rangle^2 \end{aligned}$$

Furthermore, by rearanging the second term, we get:

$$\|n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}[Y \otimes Y]\|_2 = n^{-1} \sum_{i=1}^n \langle X_i, u \odot \delta_i \rangle - \langle \Sigma(u \odot \delta_i), u \odot \delta_i \rangle$$

4.1.2 Bias of $Y_i \otimes Y_i$

Let us now find an upper bound to $\|n^{-1}\mathbb{E}(Y_i \otimes Y_i) - \Sigma\|$:

$$\begin{aligned} \|n^{-1}\mathbb{E}(Y_i \otimes Y_i) - \Sigma\| &= \|\mathbb{E}(\delta_i \odot X_i) \otimes (\delta_i \odot X_i) - \Sigma\| \\ &= \|\mathbb{E}(\delta_i \otimes \delta_i) \odot (X_i \otimes X_i) - \Sigma\| \end{aligned}$$

By the tower property,

$$\begin{aligned}\mathbb{E}(\delta_i \otimes \delta_i) \odot (X_i \otimes X_i) &= \mathbb{E}\left(\mathbb{E}_\delta(\delta_i \otimes \delta_i) \odot (X_i \otimes X_i)\right) \\ &= \mathbb{E}\left((\delta_i \otimes \delta_i) \odot \Sigma\right)\end{aligned}$$

and thus:

$$\begin{aligned}\|n^{-1}\mathbb{E}(Y_i \otimes Y_i) - \Sigma\| &= \|\mathbb{E}\left((\delta_i \otimes \delta_i) - \mathbf{1}\right) \odot \Sigma\| \\ &\leq \|\mathbb{E}(\delta_i \otimes \delta_i) - \mathbf{1}\|\|\Sigma\|\end{aligned}$$

with $\mathbf{1}$ being the square matrix filled with 1. By mutual independence of the Bernoulli variables, we know that $\mathbb{E}\delta_i \otimes \delta_i$ is the matrix with diagonals equal to δ and non diagonal terms equal to δ^2.

References

[1] K. Lounici. High-dimensional covariance matrix estimation with missing observations. 20(3):1029–1058.

[2] J. Raymaekers and P. J. Rousseeuw. Handling cellwise outliers by sparse regression and robust covariance.