Rigorous Stochastic Bounds for the Error in Large Covariance Matrices

Albrecht Böttcher[1] **and David Wenzel**[2]

This note is motivated by recent studies of Ling Huang et al. on distributed PCA and network anomaly detection and contains a rigorous derivation of bounds for the expected value and the variance of the spectral norm of the error in large covariance matrices.

1 Introduction

The covariance matrix $\mathrm{Co}(A)$ associated with a matrix $A \in \mathbf{R}^{m\times n}$ is the $n\times n$ matrix $\mathrm{Co}(A) = (1/m)A^\top A$, where $A^\top$ denotes the transpose of A. Suppose $X \in \mathbf{R}^{m\times n}$ is a random matrix. We are interested in stochastic estimates for

$$\Delta := \mathrm{Co}(A+X) - \mathrm{Co}(A) = \frac{1}{m}\left(A^\top X + X^\top A + X^\top X\right).$$

To be more precise, we want to know bounds for the expected value $E(\|\Delta\|)$ and the variance $\sigma^2(\|\Delta\|)$, where $\|\cdot\|$ is the spectral norm. We assume that $X = Z \circ S := (z_{ij}\sigma_{ij})_{i=1,j=1}^{m,n}$ where $\sigma_{ij} \ge 0$ are given numbers and z_{ij} are i.i.d. random variables such that $E(z_{11}) = 0$, z_{11} is symmetric about the origin, $\sigma^2(z_{11}) = 1$, $E(z_{11}^4) < \infty$.

We learned of this problem from the interesting paper [4] by Ling Huang and coauthors. They consider a network composed of n monitors each of which shows data of size m from a stream of data. The jth column X_j of the $m \times n$ matrix X is constituted by the error (loss of information) in the data of the jth monitor. This error is supposed to depend only on the so-called slack parameter of the monitor and hence it is reasonable to assume that the columns of X are i.i.d. random vectors up to the variance, which may change from column to column. Under this assumption we have $x_{ij} = z_{ij}\sigma_j$ and may therefore write the Hadamard product $X = Z \circ S$ in the form $X = ZD$ with $D = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$.

[1]Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz. Germany, aboettch@mathematik.tu-chemnitz.de

[2]Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz. Germany, david.wenzel@s2000.tu-chemnitz.de

The argument of [4] is as follows. We have

$$E(\|A^\top X\|^2) = E(\|X^\top A A^\top X\|) = E(\lambda_{\max}(X^\top A A^\top X)),$$

where $\lambda_{\max}(\cdot)$ stands for the largest eigenvalue, and since the eigenvalues of symmetric random matrices are tightly concentrated [2], it is justified to put

$$E(\lambda_{\max}(X^\top A A^\top X)) \approx \lambda_{\max}(E(X^\top A A^\top X)).$$

A straightforward computation gives $\|E(X^\top A A^\top X)\| = \|D\|^2 \|A\|_{\mathsf{F}}^2$, where $\|\cdot\|_{\mathsf{F}}$ is the Frobenius norm. Thus,

$$E(\|A^\top X\|^2) \approx \|D\|^2 \|A\|_{\mathsf{F}}^2, \tag{1}$$

Analogously,

$$E(\|X^\top A\|^2) = E(\|A^\top X X^\top A\|) = E(\lambda_{\max}(A^\top X X^\top A)) \approx \lambda_{\max}(E(A^\top X X^\top A)),$$

which yields

$$E(\|X^\top A\|^2) \approx \|D\|_{\mathsf{F}}^2 \|A\|^2. \tag{2}$$

The term $E(\|X^\top X\|^2)$ is considered as negligible. Thus, the bound obtained in [4] is

$$\begin{aligned} E(\|\Delta\|) &\leq \frac{1}{m} E(\|A^\top X\|) + \frac{1}{m} E(\|X^\top A\|) \\ &\leq \frac{1}{m} \sqrt{E(\|A^\top X\|^2)} + \frac{1}{m} \sqrt{E(\|X^\top A\|^2)} \\ &\precsim \frac{1}{m} \left(\|D\| \|A\|_{\mathsf{F}} + \|D\|_{\mathsf{F}} \|A\| \right). \end{aligned} \tag{3}$$

The purpose of this note is to point out that the above derivation is based on two critical arguments and that, consequently, (3) may tell a wrong message. We first want to clarify that (1) and (2) are used in (3) in the form

$$E(\|A^\top X\|^2) \precsim \|D\|^2 \|A\|_{\mathsf{F}}^2, \quad E(\|X^\top A\|^2) \precsim \|D\|_{\mathsf{F}}^2 \|A\|^2$$

although actually

$$E(\|A^\top X\|^2) \geq \|D\|^2 \|A\|_{\mathsf{F}}^2, \quad E(\|X^\top A\|^2) \geq \|D\|_{\mathsf{F}}^2 \|A\|^2. \tag{4}$$

Secondly, it turns out that the term $E(\|X^\top X\|^2)$ cannot always be neglected. We show that if $\varepsilon > 0$ is given, then

$$E(\|\Delta\|) \geq \|D\| \left(\|D\| - \frac{3\,\|A\|}{\sqrt{n}} \, \frac{1+\sqrt{y}}{y} \right) \tag{5}$$

and

$$E(\|\Delta\|) \leq (1+\varepsilon) \frac{1+\sqrt{y}}{y} \|D\| \left(\|D\|(1+\sqrt{y}) + \frac{2\,\|A\|}{\sqrt{n}} \right) \tag{6}$$

whenever n and m are sufficiently large and the ratio m/n is close enough to some value $y \in (0, \infty)$. Thus, if $\|A\|/\sqrt{n}$ is small, which happens e.g. for sparse matrices,

for nearly orthogonal matrices, or for rectangular truncations of infinite matrices that induce bounded operators on ℓ^2, then (5) and (6) yield

$$(1-\varepsilon)\|D\|^2 \precsim E(\|\Delta\|) \precsim (1+\varepsilon)\|D\|^2\,\frac{(1+\sqrt{y})^2}{y},$$

which reveals that $\|D\|^2 = \max \sigma_j^2$ is the quantity that governs the limiting behavior of $E(\|\Delta\|)$. In that case $E(\|X^\top X)\|^2)$ is the dominating term.

Now suppose m and n are large and m/n is very close to y. We can then replace y by m/n and put $\varepsilon = 0$ in (6). The resulting approximate upper bound is

$$E(\|\Delta\|) \precsim \frac{\sqrt{n}+\sqrt{m}}{m}\,\|D\|\,\Big(\|D\|(\sqrt{n}+\sqrt{m}) + 2\,\|A\|\Big),$$

which in the case $n \approx m$ simplifies to

$$E(\|\Delta\|) \precsim \frac{2}{\sqrt{m}}\,\|D\|\,\|A\| + \frac{2}{\sqrt{m}}\,\|D\|\,\|A\| + 4\,\|D\|^2. \tag{7}$$

Let us denote by U and B the approximate upper bounds in (3) and (7), respectively. Since $\|A\| \geq \|A\|_{\mathsf{F}}/\sqrt{m}$ and $\|D\| \geq \|D\|_{\mathsf{F}}/\sqrt{m}$, we get

$$B \geq \frac{2}{m}\,(\|D\|\|A\|_{\mathsf{F}} + \|D\|_{\mathsf{F}}\|A\|) + 4\,\|D\|^2 = 2\,U + 4\,\|D\|^2,$$

or equivalently,

$$U \leq \frac{B}{2} - 2\,\|D\|^2.$$

Thus, even for small $\|D\|$, the bound U is less than half of the rigorous bound B. In other words, U is chosen at least twice too small. But an upper bound that is too small may cause a missed detection rate that is too high. This may be an explanation for the high missed detection rate of 4 % observed in [4].

The key result we need to prove (5) and (6) is fortunately already available. It is a beautiful theorem by Yin, Bai, and Krishnaiah that was published in 1988 and sharpens an earlier result by Geman of 1980. The reader will easily observe that all we will do is nothing but some simple estimations that are based on this theorem.

2 Deterministic Matrix Times Random Matrix

Let the $m \times n$ random matrix Z be as in the introduction. A deep result by Yin, Bai, and Krishnaiah [5] says that if $n \to \infty$ and $m/n \to y \in (0,\infty)$, then $n^{-1/2}\|Z\|$ converges to $1+\sqrt{y}$ almost surely. Under the additional hypothesis $E(|z_{11}|^n) \leq n^{\alpha n}$ this is a result of Geman [3]. Moreover, in [5] it is shown that for every $\varepsilon > 0$ there exist $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that if $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$ then

$$n^{-k}E(\|Z\|^{2k}) < (1+\sqrt{y}+\varepsilon)^{2k}$$

for all $k \geq k_n$. Thus, if $n \geq n_0$, $y(1-\delta) \leq m/n \leq y(1+\delta)$, and ℓ is a natural number, then, for $2k \geq \ell$,

$$n^{-\ell/2} E(\|Z\|^\ell) \leq n^{-\ell/2} \left(E(\|Z\|^{2k}) \right)^{\ell/(2k)} = \left(n^{-k} E(\|Z\|^{2k}) \right)^{\ell/(2k)} \leq (1 + \sqrt{y} + \varepsilon)^\ell.$$

Consequently, if $n \to \infty$ and $m/n \to y \in (0, \infty)$ then

$$\limsup\, n^{-\ell/2}\, E(\|Z\|^\ell) \leq (1 + \sqrt{y})^\ell \tag{8}$$

for $\ell = 1, 2, 3, \ldots$.

Lemma 2.1 *Let $X = Z \circ S$. Then*

$$\|E(A^\top X X^\top A)\| = \left\| \left(\operatorname{diag} \left(\sqrt{\sum_{j=1}^{n} \sigma_{ij}^2} \right)_{i=1}^{m} \right) A \right\|^2, \tag{9}$$

$$\|E(X^\top A A^\top X)\| = \max_{1 \leq j \leq n} \left| \sum_{i=1}^{m} (AA^\top)_{ii}\, \sigma_{ij}^2 \right|. \tag{10}$$

Proof. We have $E[(XX^\top)_{ij}] = E[\sum_{\ell=1}^n x_{i\ell} x_{j\ell}]$ and this is zero for $i \neq j$ (due to the symmetry of z_{11} about the origin) and $r_i^2 := \sum_{\ell=1}^n \sigma_{i\ell}^2$ for $i = j$. Thus, $\|E(A^\top X X^\top A)\|$ equals

$$\|A^\top \operatorname{diag}(r_1^2, \ldots, r_m^2)\, A\| = \|\operatorname{diag}(r_1, \ldots, r_m)\, A\|^2,$$

which is (9). Taking again into account symmetry about the origin, we obtain that $E(X^\top A A^\top X)$ is equal to

$$\operatorname{diag} \left(\sum_{i=1}^{m} (AA^\top)_{ii}\, \sigma_{i1}^2, \ldots, \sum_{i=1}^{m} (AA^\top)_{ii}\, \sigma_{in}^2 \right),$$

which implies (10). □

Proposition 2.2 *Let $X = Z \circ S$ and put*

$$\xi = E(\|A^\top X\|^2) = E(\|X^\top A\|^2).$$

Then the maximum of the right-hand sides of (9) and (10) is a lower bound for ξ for all m and n. Given $\varepsilon > 0$, there exist $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that if $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$, then

$$\xi \leq (1+\varepsilon) \|A\|^2\, \|S\|^2 (1 + \sqrt{y})^2 n. \tag{11}$$

Proof. First of all it is clear that $\|A^\top X\|^2 = \|X^\top A\|^2$. From the triangle inequality we infer that

$$\begin{aligned} E(\|A^\top X\|^2) &= E(\|X^\top A A^\top X\|) = \int_\Omega \|X^\top(\omega) A A^\top X(\omega)\| dP(\omega) \\ &\geq \left\| \int_\Omega X^\top(\omega) A A^\top X(\omega) dP(\omega) \right\| = \|E(X^\top A A^\top X)\| \end{aligned}$$

and, analogously,

$$E(\|X^\top A\|^2) \geq \|E(A^\top X X^\top A)\|.$$

Lemma 2.1 therefore shows that ξ is greater than or equal to the maximum of the right-hand sides of (9) and (10).

From (8) we deduce that $E(\|Z\|^2) \leq (1+\varepsilon)(1+\sqrt{y})^2 n$ if n_0 is large enough, $\delta > 0$ is sufficiently small, $n \geq n_0$, and $y(1-\delta) \leq m/n \leq y(1+\delta)$. An inequality by M. Marcus says that $\|Z \circ S\| \leq \|Z\|\,\|S\|$ (see, for example, [1, Problem I.6.13]). Consequently,

$$\xi = E(\|A^\top (Z \circ S)\|^2) \leq \|A\|^2\,\|S\|^2\,E(\|Z\|^2), \tag{12}$$

which implies (11). □

Things look nicer in the case of interest in [4], that is, in the case where S is constant along the columns and thus $X = ZD$ with $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$.

Corollary 2.3 *If $X = ZD$ and ξ is as in Proposition 2.2, then*

$$\max\left(\|D\|_{\mathsf{F}}^2 \|A\|^2, \|D\|^2 \|A\|_{\mathsf{F}}^2\right) \leq \xi \tag{13}$$

for all m and n, and for each $\varepsilon > 0$ there exist $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that

$$\xi \leq (1+\varepsilon)\|A\|^2\,\|D\|^2(1+\sqrt{y})^2 n \tag{14}$$

whenever $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$.

Proof. In the case at hand, $\sigma_{ij} = \sigma_j$ and hence the right-hand sides of (9) and (10) become

$$\left\| \left(\operatorname{diag}\left(\sqrt{\sum_{j=1}^{n} \sigma_j^2} \right)_{i=1}^{m} \right) A \right\|^2 = \left\| \|D\|_{\mathsf{F}}\, I_{m\times m}\, A \right\|^2 = \|D\|_{\mathsf{F}}^2\,\|A\|^2,$$

$$\max_{1\leq j\leq n} \left| \sum_{i=1}^{m} (AA^\top)_{ii}\,\sigma_j^2 \right| = \max_{1\leq j\leq n} \sigma_j^2 \left| \operatorname{tr}(AA^\top) \right| = \|D\|^2\,\|A\|_{\mathsf{F}}^2.$$

This in conjunction with Proposition 2.2 proves (13). We have

$$S = (\,\sigma_1\ \ldots\ \sigma_n\,) \otimes (\,1\ \ldots\ 1\,)^\top$$

and thus $\|S\| = \|D\|_{\mathsf{F}}\sqrt{m}$. Estimate (11) therefore yields

$$\xi \leq (1+\varepsilon)\|A\|^2\,\|D\|_{\mathsf{F}}^2(1+\sqrt{y})^2 nm,$$

which is weaker than (14). However, the analogue of (12) is

$$\xi = E(\|A^\top (ZD)\|^2) \leq \|A\|^2\,\|D\|^2\,E(\|Z\|^2),$$

and this gives exactly (14). □

Note that (13) proves (4).

3 The Error in Covariance Matrices

Here is our main result.

Theorem 3.1 *Let $X = Z \circ S$. Given $\varepsilon > 0$, there exist $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that if $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$, then*

$$E(\|\Delta\|) \leq (1+\varepsilon)\frac{1+\sqrt{y}}{y}\|S\|\left(\|S\|(1+\sqrt{y}) + \frac{2\,\|A\|}{\sqrt{n}}\right), \tag{15}$$

$$\sigma^2(\|\Delta\|) \leq 3\,(1+\varepsilon)\frac{(1+\sqrt{y})^2}{y^2}\|S\|^2\left(\|S\|^2(1+\sqrt{y})^2 + \frac{2\,\|A\|^2}{n}\right), \tag{16}$$

Proof. Fix $\varepsilon > 0$ and define $\gamma > 0$ by $1+\gamma = \sqrt[4]{1+\varepsilon}$. We have

$$\begin{aligned} E(\|\Delta\|) &\leq \frac{1}{m}\,E(\|A^\top X\|) + \frac{1}{m}\,E(\|X^\top A\|) + \frac{1}{m}\,E(\|X^\top X\|) \\ &= \frac{2}{m}\,E(\|A^\top X\|) + \frac{1}{m}\,E(\|X\|^2) \\ &\leq \frac{2}{m}\,\sqrt{\xi} + \frac{\|S\|^2}{m}\,E(\|Z\|^2). \end{aligned}$$

and

$$\begin{aligned} E(\|\Delta\|^2) &= \frac{1}{m^2}\,E\left(\|A^\top X + X^\top A + X^\top X\|^2\right) \\ &\leq \frac{3}{m^2}\,E\left(\|A^\top X\|^2 + \|X^\top A\|^2 + \|X^\top X\|^2\right) \\ &= \frac{6}{m^2}\,\xi + \frac{3}{m^2}\,E(\|X\|^4) \leq \frac{6}{m^2}\,\xi + \frac{3}{m^2}\,\|S\|^4\,E(\|Z\|^4). \end{aligned}$$

By Proposition 2.2 and (8) there are $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that

$$\frac{1+\gamma}{1-\delta} < \sqrt{1+\varepsilon} \tag{17}$$

and such that if $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$, then

$$\begin{aligned} &\xi \leq (1+\gamma)^2\|A\|^2\|S\|^2(1+\sqrt{y})^2 n, \\ &E(\|Z\|^2) \leq (1+\gamma)(1+\sqrt{y})^2 n, \\ &E(\|Z\|^4) \leq (1+\gamma)^2(1+\sqrt{y})^4 n^2. \end{aligned}$$

Since $m \geq ny(1-\delta)$, it follows that

$$E(\|\Delta\|) \leq 2\frac{1+\gamma}{1-\delta}\|A\|\,\|S\|\frac{1+\sqrt{y}}{y}\,\frac{1}{\sqrt{n}} + \frac{1+\gamma}{1-\delta}\|S\|^2\frac{(1+\sqrt{y})^2}{y},$$

which gives (15) because of (17). Analogously,

$$E(\|\Delta\|^2) \leq 6\,\frac{(1+\gamma)^2}{(1-\delta)^2}\|A\|^2\,\|S\|^2\frac{(1+\sqrt{y})^2}{y^2}\,\frac{1}{n} + 3\,\frac{(1+\gamma)^2}{(1-\delta)^2}\|S\|^4\frac{(1+\sqrt{y})^4}{y^2},$$

which, by (17), implies (16). □

In the case where S is constant along the columns we have the following.

Theorem 3.2 *Let $X = ZD$ with $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. For each $\varepsilon > 0$ there exist $n_0 = n_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that if $n \geq n_0$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$, then*

$$\|D\| \left(\|D\| - \frac{3\,\|A\|}{\sqrt{n}} \, \frac{1+\sqrt{y}}{y} \right) \leq E(\|\Delta\|), \tag{18}$$

$$E(\|\Delta\|) \leq (1+\varepsilon) \frac{1+\sqrt{y}}{y} \|D\| \left(\|D\|(1+\sqrt{y}) + \frac{2\,\|A\|}{\sqrt{n}} \right), \tag{19}$$

$$\sigma^2(\|\Delta\|) \leq 3\,(1+\varepsilon) \frac{(1+\sqrt{y})^2}{y^2} \|D\|^2 \left(\|D\|^2(1+\sqrt{y})^2 + \frac{2\,\|A\|^2}{n} \right). \tag{20}$$

Proof. The upper bounds (19) and (20) do not result from Theorem 3.1, because, as already observed in the proof of Corollary 2.3, $\|S\| = \|D\|_{\mathsf{F}} \sqrt{m}$. However, replacing in the proof of Theorem 3.1 the reference to Proposition 2.2 by reference to Corollary 2.3, we obtain precisely the bounds (19) and (20). To get (18) note first that

$$m\,E(\|\Delta\|) \geq E(\|X\|^2) - 2\,E(\|A^\top X\|).$$

Corollary 2.3 implies that there are absolute constants n_1 and $\delta > 0$ such that

$$2\,E(\|A^\top X\|) \leq 2\,\sqrt{\xi} \leq 2.5\,\|A\|\|D\|(1+\sqrt{y})\sqrt{n}$$

for $n \geq n_1$ and $y(1-\delta) \leq m/n \leq y(1+\delta)$. We also assume that $2.5/(1-\delta) \leq 3$. We have

$$E(\|X\|^2) = E(\|ZD\|^2) = E(\|D^\top Z^\top Z D\|) \geq \|E(D^\top Z^\top Z D)\|$$

(recall the proof of Proposition 2.2 for the last inequality). Because

$$E(D^\top Z^\top Z D) = \operatorname{diag}(m\sigma_1^2, \ldots, m\sigma_n^2),$$

we see that

$$\|E(D^\top Z^\top Z D)\| = m \max_{1 \leq j \leq n} \sigma_j^2 = m\,\|D\|^2.$$

In summary,

$$E(\|\Delta\|) \geq \|D\|^2 - \frac{2.5}{m}\,\|A\|\|D\|(1+\sqrt{y})\sqrt{n}.$$

Taking into account that $m \geq ny(1-\delta)$, we arrive at (18). □

References

[1] R. Bhatia: *Matrix Analysis.* Springer-Verlag, New York 1997.

[2] N. Alon, M. Krivelevich, and V. H. Vu: On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.* **131** (2002), 259–267.

[3] S. Geman: A limit theorem for the norm of random matrices. *Ann. Probab.* **8** (1980), 252–261.

[4] L. Huang, X. L. Nguyen, M. Garofalakis, M. Jordan, A. Joseph, and N. Taft: Distributed PCA and network anomaly detection. To appear.

[5] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah: On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** (1988), 509–521.