Testing and Cleaning Covariance Matrices

Big Data Statistics - Final Project

Total of 100 Marks

due Friday 3 November 2017 at 17:00

In this project we consider how to test hypotheses about covariance matrices and "clean" these matrices. Special emphasis is on the situation where the dimensionality p of the data becomes large and classic techniques break down.

Testing Covariance Matrices

Question 1 [5 marks]

As a warm-up, read Section 6.6 in **[A]** and reproduce the calculations of Example 6.12 in R. In this example, Box's M-test is used to study nursing home data from Wisconsin (data found in Example 6.10). If you have slightly different results to the book, briefly explain why.

Question 2 [10 marks]

Box's M-test (aka. Box's χ^2 approximation) is a classic result that is based on a *likelihood ratio test* (LRT). The general philosophy behind a LRT is to maximise the likelihood under the null hypothesis H_0 and also to maximise the likelihood under the alternative hypothesis H_1.

Definition 1. *If the distribution of the random sample $\mathbb{X} = (\mathbb{x}_1, \ldots, \mathbb{x}_n)'$ depends upon a parameter vector θ, and if $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega_1$ are any two hypotheses, then the likelihood ratio statistic for testing H_0 against H_1 is defined as*

$$\lambda_1 = \frac{\mathscr{L}_0^\star}{\mathscr{L}_1^\star}$$

where $\mathscr{L}_i^\star$ is the largest value which the likelihood function takes in the region Ω_i, $i = 1, 2$.

At this point it is good to remember that a multivariate Normal distribution is completely characterised by the parameter vector $\theta = (\mu, \Sigma)$, i.e., only the mean vector and the covariance matrix are needed to know the distribution.

The LRT has the following important asymptotic property as $n \to \infty$ that Box leverages to obtain his χ^2 approximation.

Theorem 1. *If $\Omega_1 \subset \mathbb{R}^q$ and if Ω_0 is an r-dimensional subregion of Ω_1 then (under some technical assumptions) for each $\omega \in \Omega_0$, $-2\log(\lambda_1)$ has an asymptotic χ^2_{q-r} distribution as $n \to \infty$.*

The explanation why Theorem 1 is true starts in Section 10.2 of **[B]** where the LRT is derived, culminating in critical region for λ_1 given by eq. (9). At this point, no assumptions are made about the distribution of the population covariance matrices $\Sigma_1, \ldots, \Sigma_q$ (so we don't know how λ_1 is distributed). Assumptions are made in Section 10.4: covariances are assumed Wishart distributed which occurs when the random samples $\mathbb{x}_1, \ldots, \mathbb{x}_n$ are multivariate Normal. Box's χ^2 asymptotic approximation is obtained in Section 10.5 thanks to a formula for the h-moment of λ_1. As $\mathbb{E}[\lambda_1^h]$ has a specific form (given in terms of ratios of Gamma functions), Theorem 8.5.1 of **[B]** can be applied to get an approximation of $\mathbb{P}(-2\rho\log(\lambda_1) \leq z)$ in terms of the χ^2 distribution.

Now that you understand some of the theory, study the classic "iris" dataset (available in R in the `iris` variable). The populations are *Iris versicolor* (1), *Iris setosa* (2), and *Iris virginica* (3); each sample consists of 50 observations. Use Box's M-test (or otherwise) to:

[5] (a) Test the hypothesis $\Sigma_1 = \Sigma_2$ at the 5% significance level.

[5] (b) Test the hypothesis $\Sigma_1 = \Sigma_2 = \Sigma_3$ at the 5% significance level.

Note: this is Problem 10.1 from **[B]**.

Question 3 [10 marks]

On page 311 in **[A]**, just above Example 6.12, the authors make the comment that *"Box's χ^2 approximation works well if each n_ℓ exceeds 20 and if p and g do not exceed 5"*. Your task is to perform a simulation study (see **[G]**) to show what happens to Box's χ^2 approximation when p exceeds 5 while holding g fixed, e.g., $g = 2$. This means you have to design an experiment to show how badly Box's test performs for large p by choosing appropriate Σ_1 and Σ_2, simulating sample data, etc. Present your results in a clear manner (see **[G]** for presentation tips).

Question 4 [10 marks]

The Bartlett statistic, see **[B]** page 413 Eq. (10), is for $g = 2$ given by

$$V_1 = \frac{|\mathbb{A}_1|^{N_1/2}|\mathbb{A}_2|^{N_2/2}}{|\mathbb{A}_1 + \mathbb{A}_2|^{N/2}}$$

where $N_g := n_g - 1$ and $N := N_1 + N_2$. Setting $\mathbb{S}_g = \mathbb{A}_g/N$, multiplying through the numerator and denominator by $|\mathbb{S}_2^{-1}|$ and using the fact that $|AB| = |A||B|$ for matrices A and B, we can instead consider

$$V_1^* = \frac{|\mathbb{S}_1\mathbb{S}_2^{-1}|^{N_1/2}}{|c_1\mathbb{S}_1\mathbb{S}_2^{-1} + c_2|^{N/2}}$$

where $c_g = N_g/N$. Notice we are in the Fisher regime $\mathbb{S}_1\mathbb{S}_2^{-1}$. A recent result, Theorem 4.1 in **[F]**, shows that

Theorem 2. *Assume $N_1 \to \infty$, $N_2 \to \infty$, and $p \to \infty$ such that $y_{N_1} = p/N_1 \to y_1 \in (0, 1)$ and $y_{N_2} = p/N_2 \to y_2 \in (0, 1)$. Then*

$$-\frac{2}{N}\log V_1^* - p\, F_{y_{N_1}, y_{N_2}}(f) \to N(\mu_2, \sigma_2^2),$$

where

$$F_{a,b}(f) := \frac{a + b - ab}{ab}\log\left(\frac{a+b}{a+b-ab}\right) + \frac{a(1-b)}{b(a+b)}\log(1-b) + \frac{b(1-a)}{a(a+b)}\log(1-a),$$

and μ_2 and σ_2 can be determined.

Read the paper to determine the appropriate μ_2 and σ_2 and use these constants to develop and algorithm in R to test the hypothesis $H_1 : \Sigma_1 = \Sigma_2$ for large p. Perform a simulation study to compare its performance (type I error and power) to Box's M-test for varying p.

Note that:

- Instead of calculating `log(det(A))` for some matrix `A`, it might be advisable in R to use the equivalent `determinant(A, logarithm=True)` as it is more numerically accurate for matrices with small determinant.

- If your observations are real-valued, then μ_1 and σ_2^2 are given in the paper by (4.8) and (4.9).

Question 5 [10 marks]

We are now going to look at the problem of testing that a covariance matrix is equal to a given matrix. If observations $\mathbb{y}_1, \ldots, \mathbb{y}_n$ are multivariate Normal $N_p(\nu, \Psi)$, we wish to test the hypothesis $H_1 : \Psi = \Psi_0$ where Ψ_0 is a given positive definite matrix. Let Q be the matrix such that

$$Q\Psi_0 Q' = I,$$

then set $\mu := Q\nu$ and $\Sigma := Q\Psi Q'$. If we define $\mathbb{x}_i := Q\mathbb{y}_i$ it follows that $\mathbb{x}_1, \ldots, \mathbb{x}_n$ are observations from $N_p(\mu, \Sigma)$ and the hypothesis H_1 is transformed to $H_1 : \Sigma = I$. Using the LRT approach, we can find the test statistic

$$\lambda_1 = \left(\frac{e}{n}\right)^{\frac{1}{2}pn} |\mathbb{A}|^{\frac{1}{2}n} e^{-\frac{1}{2}\operatorname{tr}\mathbb{A}},$$

where

$$\mathbb{A} := \sum_{k=1}^{n} (\mathbb{x}_k - \bar{\mathbb{x}})(\mathbb{x}_k - \bar{\mathbb{x}})'.$$

Unfortunately λ_1 is a biased statistic. The following unbiased estimator was proposed in **[E]**:

$$\lambda_1^* := e^{\frac{1}{2}pN} \left(|\mathbb{S}| e^{-\operatorname{tr}\mathbb{S}}\right)^{\frac{1}{2}N}$$

where $N := n - 1$ and $\mathbb{S} := \mathbb{A}/n$. The distribution of λ_1^* has the following χ^2 approximation

$$\mathbb{P}(-2\rho \log \lambda_1^* \le z) = \mathbb{P}(C_f \le z) + \frac{\gamma_2}{\rho^2 (n-1)^2} \left(\mathbb{P}(C_{f+4} \le z) - \mathbb{P}(C_f \le z)\right) + O(n^{-3}). \qquad (1)$$

where $C_k \sim \chi_k^2$ (i.e., χ^2 distributed with k degrees of freedom), $f := \frac{1}{2}p(p+1)$, $\rho := 1 - (2p^2 + 3p - 1)/[6(n-1)(p+1)]$, and $\gamma_2 := p(2p^4 + 6p^3 + p^2 - 12p - 13)/[288(p+1)]$. All the details can be found in **[B]** Section 10.8.1, **[B]** around Eq. (19) on p. 441, and **[E]**.

Perform a simulation study to understand the performance (type I error and power) of (1) for $n = 500$ and $p = 5, 10, 50, 100, 300$; see **[H]**.

Question 6 [10 marks]

Continuing the previous question (and its notation), notice that

$$-\frac{2}{N} \log \lambda_1^* = \operatorname{tr}\mathbb{S} - \log|\mathbb{S}| - p.$$

Setting $T_1 := \operatorname{tr}\mathbb{S} - \log|\mathbb{S}| - p$, prove the following theorem.

Theorem 3. *Assume that $n \to \infty$, $p \to \infty$, and $p/n \to y \in (0, 1)$. Then*

$$T_1 - p\, d_1(y_N) \to N(\mu, \sigma_1^2)$$

where $N := n - 1$, $Y_N := p/N$ and

$$d_1(y) := 1 + \frac{1-y}{y} \log(1-y),$$
$$\mu_1 := -\frac{1}{2} \log(1-y),$$
$$\sigma_1^2 := -2\log(1-y) - 2y.$$

Hint: Apply Theorem in Lecture 6 on page 7 with $\frac{1}{p}T_1 := F^{\mathbb{S}}(f)$ with $f(x) = x - \log x - 1$. Also see **[F]**.

Question 7 [10 marks]

Continuing the previous question and notation, use the Theorem to construct an algorithm that tests $H_1 : \Sigma = I$ and perform a simulation study to understand its performance (type I error and power) for $p = 5, 10, 50, 100, 300$. Comment on how it performs compared to (1).

Question 8 [10 marks]

The recent paper **[D]** provides some improvements on **[F]** in that it allows the Fisher matrix $\mathbb{F} = \mathbb{S}_1\mathbb{S}_2^{-1}$ to be of the form $\mathbb{F} = S_1\mathbb{T}^*S_2\mathbb{T}$ where $\mathbb{T}$ is a deterministic matrix. It uses a new CLT for random matrices of the form $\mathbb{S}^{-1}\mathbb{T}$ where $\mathbb{T}$ is nonnegative definite and deterministic Hermitian matrix.

Discuss the difficulties of reproducing Figure 2 in **[D]** by numerical methods (4.7), (4.8) and (4.9). For example, consider the case of $H = \delta_1$ in (4.8) and (4.9) and write out the expressions for the approximations of $\mathbb{E}X_f$ and $\mathbf{Cov}(X_{f_i}, X_{f_j})$ in this case. Also describe what is required to generate the $m_0(z)$ for every z.

Cleaning Covariance Matrices

Question 9 [5 marks]

We are now going to look at the recently proposed "Rotationally Invariant Estimator" (RIE) technique to clean covariance matrices. The aim of cleaning is to construct the best estimate of what the true population covariance matrix could be. This is acheived by combining theoretical knowledge about how the eigenvalues of a covariance should behave and tweaking the sample covariance accordingly.

The recent paper **[C]** gives a long survey of results in this area with the RIE given in Section 6 and implementation in Section 8. After reading through these two sections, implement the RIE approach given by Algorithm 1 of **[C]** (see page 76).

Question 10 [15 marks]

Reproduce the first row (i.e., only IW-regularization case) of Table 1, Table 2, and Table 3. That is: check the consistency over 100 samples, check the consistency with respect to dimension, check the consistency with respect to the dimension ratio.

Conclusion

Question 11 [5 marks]

Explain why, in the era of "Big Data", it is crucial to correctly estimate or understand the (population) structure of covariance matrices. Give an example, explain what could go wrong if a classic approach was naively applied, and draw conclusions from your results in the previous questions.

Hint: See Section 1.1 **[C]** and Section 1 **[D]** for inspiration.

References

[A] Johnson, Wichern (2007). Applied Multivariate Statistical Analysis. Pearson Prentice Hall.

[B] Anderson (2003). An introduction to Multivariate Statistical Analysis. Wiley.

[C] Bun, Bouchard, Potters (2017). Cleaning large correlation matrices: tools from Random Matrix Theory. Physics Reports 666, 1–109.

[D] Zheng, Bai, Yao (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. Bernouilli 23(2), 1130–1178.

[E] Sugiura, Nagao (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrices. Annals of Mathematical Statistics Vol. 39, No. 5, 1686–1692.

[F] Bai, Jiang, Yao, Zheng (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. Annals of Statistics Vol 37, No. 6B, 3822–3840.

[G] `http://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf`

[H] `https://stats.stackexchange.com/a/40874`