Big Data Statistics - Final Project (v1)

Total of 100 Marks

due Monday 29 October 2018 at 17:00

In this project we consider how to test hypotheses about covariance matrices and, since everything "eighties" is trendy again, we will look at some multivariate time series papers from that epoch and make them fresh again. There are interesting connections between these two topics. The aim of the project is to take the modern viewpoint and understand what happens in both situations when the dimensionality p of the observations becomes large.

Testing Covariance Matrices

Question 1 [5 marks]

As a warm-up, read Section 6.6 in **[C]** 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 $\mathbb{X} = (\mathbb{x}_1, \dots, \mathbb{x}_n)'$ 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, $i = 0, 1$.

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 $\Sigma_1, \dots, \Sigma_q$ (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 $\mathbb{x}_1, \dots, \mathbb{x}_n$ 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 **[C]**, 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 **[J]**) 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 **[J]** for presentation tips).

Question 4 [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 $H_0 : \Psi = \Psi_0$ 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_0 is transformed to $H_0 : \Sigma = I$. 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 **[A]**:

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

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

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

where $C_k \sim \chi^2_k$ (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 **[A]**.

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 **[K]**.

Question 5 [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 2. *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

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

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 **[D]**.

?

Question 6 [10 marks]

Continuing the previous question and notation, use the Theorem to construct an algorithm that tests $H_1 : \Sigma \neq I$ 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).

Multivariate Time Series

Let $\mathbb{Z}$ denote the set of integers. A sequence of random vector observations $(\mathbb{X}_t : t = 1, \ldots, T)$ with values in $\mathbb{R}^p$ is called a p-dimensional (vector) time series. We denote the sample mean and sample covariance matrix by

$$\overline{\mathbb{X}} := \frac{1}{T}\sum_{t=1}^{T}\mathbb{X}_t, \qquad \mathbb{S}_0 := \frac{1}{T-1}\sum_{t=1}^{T}(\mathbb{X}_t - \overline{\mathbb{X}}_t)(\mathbb{X}_t - \overline{\mathbb{X}}_t)'.$$

The lag-τ sample cross-covariance (aka. autocovariance) matrix is defined as

$$\mathbb{S}_\tau := \frac{1}{T-1}\sum_{t=\tau+1}^{T}(\mathbb{X}_t - \overline{\mathbb{X}}_t)(\mathbb{X}_{t-\tau} - \overline{\mathbb{X}}_t)'.$$

The lag-τ cross-correlation is given by

$$\rho_\tau = D\mathbb{S}_\tau D$$

where $D = \operatorname{diag}(1/\sqrt{s_{11}}, 1/\sqrt{s_{22}}, \ldots, 1/\sqrt{s_{pp}})$ and the values come from $\mathbb{S}_0 = [s_{ij}]$. Assuming $\mathbb{E}[\mathbb{X}_t] = 0$, some authors (e.g., **[H]**, **[I]**) omit $\overline{\mathbb{X}}_t$ and consider the *symmetrised* lag-τ sample cross-covariance given by

$$\mathbb{C}_\tau := \frac{1}{2T}\sum_{t=1}^{T-\tau}(\mathbb{X}_t\mathbb{X}_{t+\tau}' + \mathbb{X}_{t+\tau}\mathbb{X}_t').$$

Question 7 [12 marks]

Simulation is a helpful way to learn about vector time series. Define the matrices

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.4 \\ -0.3 & 0.6 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 2.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}.$$

Generate 300 observations from the "vector autoregressive" VAR(1) model

$$\mathbb{X}_t = \mathbf{A}\mathbb{X}_{t-1} + \varepsilon_t \tag{2}$$

where $\varepsilon_t \sim N_2(0, \Sigma)$, i.e., they are i.i.d. bivariate normal random variables with mean zero and covariance Σ. Note that when simulating is it customary omit the first 100 or more observations and you can start with $\mathbb{X}_0 = (0, 0)'$.

Also generate 300 observations from the "vector moving average" VMA(1) model

$$\mathbb{X}_t = \varepsilon_t + \mathbf{A}\varepsilon_{t-1}. \tag{3}$$

[1] (a) Plot the time series $\mathbb{X}_t$ for the VAR(1) model given by (2)

[1] (b) Obtain the first five lags of sample cross-correlations of $\mathbb{X}_t$ for the VAR(1) model, i.e., $\rho_1, \ldots, \rho_5$.

[1] (c) Plot the time series $\mathbb{X}_t$ for the MA(1) model given by (3).

[1] (d) Obtain the first two lags of sample cross-correlations of $\mathbb{X}_t$ for the MA(1) model.

[5] (e) Implement the test from **[F]** and reproduce the simulation experiment given in Section 5. This means you need to generate Table 1 from **[F]**. the number in table might be "count" instead of percentage

[3] (f) The file `q-fdebt.txt` contains the U.S. quarterly federal debts held by (i) foreign and international investors, (ii) federal reserve banks, and (iii) the public. The data are from the Federal Reserve Bank of St. Louis, from 1970 to 2012 for 171 observations, and not seasonally adjusted. The debts are in billions of dollars. Take the log transformation and the first difference for each time series. Let $(\mathbb{X}_t)$ be the differenced log series.

Test $H_0 : \rho_1 = \ldots = \rho_{10} = 0$ vs $H_a : \rho_\tau \neq 0$ for some $\tau \in \{1, \ldots, 10\}$ using the test from **[F]**. Draw the conclusion using the 5% significance level.

Question 8 [13 marks]

More generally, a p-dimensional time series $\mathbb{X}_t$ follows a VAR model of order ℓ, VAR(ℓ), if

$$\mathbb{X}_t = \mathbf{a}_0 + \sum_{i=1}^{\ell} \mathbf{A}_i \mathbb{X}_{t-i} + \varepsilon_t \tag{4}$$

where $\mathbf{a}_0$ is a p-dimensional constant vector and $\mathbf{A}_i$ are $p \times p$ (non-zero) matrices for $i > 0$, and i.i.d. $\varepsilon_t \sim N_p(0, \Sigma)$ for all t with $p \times p$ covariance matrix Σ.

One day you might want to "build a model" using the VAR(ℓ) framework. One of the first things you need to do is to determine the optimal order ℓ. Tiao and Box (1981) suggest using sequential likelihood ratio tests; see Section 4 in **[G]**. Their approach is to compare a VAR(ℓ) model with a VAR($\ell - 1$) model and amounts to considering the hypothesis testing problem

$$H_0 : \mathbf{A}_\ell = 0 \qquad \text{vs.} \qquad H_1 : \mathbf{A}_\ell \neq 0.$$

We can do this by determining model parameters using a least-squares approach. We rewrite (4) as

$$\mathbb{X}'_t = X'_t \mathbb{A} + \varepsilon'_t$$

where $X_t = (1, \mathbb{X}'_{t-1}, \dots, \mathbb{X}'_{t-\ell})'$ is a $(p\ell + 1)$-dimensional vector and $\mathbb{A} = [\mathbf{a}_0, \mathbf{A}_1, \dots, \mathbf{A}_\ell]$ is a $p \times 1 + \ell \times (p \times p) = p \times (p\ell + 1)$ matrix. With observations at times $t = \ell + 1, \dots, T$, we write the data as

$$\mathbf{X} = X\mathbb{A} + E \tag{5}$$

where $\mathbf{X}$ is a $(T - \ell) \times p$ matrix with the ith row being $\mathbb{X}'_{\ell+i}$, X is a $(T - \ell) \times (p\ell + 1)$ design matrix with the ith row being $X'_{\ell+i}$, and E is a $(T - \ell) \times p$ matrix with the ith row being $\varepsilon'_{\ell+i}$. The matrix $\mathbb{A}$ contains the coefficient parameters of the VAR(ℓ) model and let $\Sigma_{\epsilon,\ell}$ be the corresponding innovation covariance matrix. Under a normality assumption, the likelihood ratio for the testing problem is

$$\lambda_1 = \left(\frac{|\hat{\Sigma}_{\epsilon,\ell}|}{|\hat{\Sigma}_{\epsilon,\ell-1}|} \right)^{(T-\ell)/2}.$$

The likelihood ratio test of H_0 is equivalent to rejecting H_0 for large values of

$$-2\log(\lambda_1) = -(T - \ell)\log\left(\frac{|\hat{\Sigma}_{\epsilon,\ell}|}{|\hat{\Sigma}_{\epsilon,\ell-1}|} \right).$$

A commonly used statistic is Bartlett's approximation given by

$$M(\ell) = -(T - \ell - 1.5 - p\ell)\log\left(\frac{|\hat{\Sigma}_{\epsilon,\ell}|}{|\hat{\Sigma}_{\epsilon,\ell-1}|} \right),$$

which follows asymptotically (as $n \to \infty$ and p fixed) a χ^2 distribution with p^2 degrees of freedom. The following methodology is suggested for selecting the order ℓ:

1. Select a positive integer P, which is the maximum VAR order that we would like to consider.
2. Setup the regression framework (5) for the VAR(P) model. That is, there are $T - P$ observations (i.e., rows) in the $\mathbf{X}$ matrix.
3. For $\ell = 0, \dots, P$ compute the least-squares estimate of the AR coefficient matrix $\mathbb{A}$. For $\ell = 0$, we have $\mathbb{A} = \mathbf{a}_0$. Then compute the ML estimate for Σ_ϵ, ℓ given by

$$\hat{\Sigma}_{\epsilon,\ell} := (1/T - P)\mathbf{R}'_\ell \mathbf{R}_\ell$$

   where $\mathbf{R}_\ell = \mathbb{X} - X\mathbb{A}$ is the residual matrix of the fitted VAR(ℓ) model.
4. For $\ell = 1, \dots, P$, compute test statistic $M(\ell)$ and its p-value, which is based on the asymptotic $\chi^2_{k^2}$ distribution. ? p^2
5. Examine the test statistics sequentially starting with $\ell = 1$. If all the p-values of the $M(\ell)$ test statistics are greater than the specified type I error for $\ell > m$, then a VAR(m) model is specified. This is so because the test rejects the null hypothesis $\mathbf{A}_\ell = 0$, but fails to reject $\mathbf{A}_\ell = 0$ for $\ell > m$.

$p = 2$

Consider a bivariate time series $\mathbb{X}_t = (\mathbb{x}_t^{\text{TB}}, \mathbb{x}_t^{\text{CPI}})$ where $\mathbb{x}_t^{\text{TB}}$ is the change in monthly US treasury bills with maturity 3 months and $\mathbb{x}_t^{\text{CPI}}$ is the inflation rate, in percentage, of the U.S. monthly consumer price index (CPI). This data from the Federal Reserve Bank of St. Louis. The CPI rate is 100 times the difference of the log CPI index. The sample period is from January 1947 to December 2012. The data are in the file `m-cpib3m.txt`.

[1] (a) Plot the time series $\mathbb{X}_t$.

[6] (b) Select a VAR order for $\mathbb{X}_t$ using the methodology (described above).

[6] (c) Drawing on your results obtained in this project and the theory discussed in class, explain and demonstrate (e.g., simulation study) what might happen with this methodology if the dimensionality p of the time series becomes large.

Question 9 [20 marks]

The recent paper **[H]** is concerned with extensions of the classical Marchenko-Pastur to the time series case. Reproduce their simulation study which is found in Section 5 and Figure 1.

MA(1) problem

References

[A] 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.

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

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

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

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

[F] Li, McLeod (1981). Distribution of the Residual Autocorrelations in Multivariate ARMA Time Series Models, J.R. Stat. Soc. B 43, No. 2, 231–239.

[G] Tiao and Box (1981). Modelling multiple time series with applications. Journal of the American Statistical Association, 76. 802 – 816.

[H] Liu, Aue, Paul (2015). On the Marchenko-Pastur Law for Linear Time Series. Annals of Statistics Vol. 43, No. 2, 675–712.

[I] Liu, Aue, Paul (2017). Spectral analysis of sample autocovariance matrices of a class of linear time series in moderately high dimensions. Bernouilli 23(4A), 2181–2209.

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

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