Inference and covariance matrix estimation in large dimension

Aurélien RIBES, Serge PLANTON

CNRM / GAME, Météo France / CNRS

Let ψ_i, for $i = 1, \ldots, n$, be n random variables in $\mathbb{R}^p$, independent and normally distributed, with mean 0 and covariance C. The $p \times p$ matrix C is assumed to be unknown. We will consider the following problem: given a $\psi_{n+1} \in \mathbb{R}^p$, a vector $g \in \mathbb{R}^p$, and assuming that ψ_{n+1} follows a $N(\mu g, C)$ distribution, where μ is an unknown real coefficient, we search for an efficient test of

$$H_0 : \text{“}\mu = 0\text{”} \quad \text{vs} \quad H_1 : \text{“}\mu > 0\text{”}.$$

A solution of this problem could be given when the matrix C is known. Let us consider the family $(T_f)_{f \in \mathbb{R}^p}$, whose element T_f is defined by the rejection region

$$W_f = \left\{ \psi_{n+1}, d_f = \langle \psi_{n+1}, f \rangle \geq \sqrt{f^T C f}\, \Phi^{-1}(1-\alpha) \right\},$$

where Φ denotes the cumulative distribution function of the standard normal distribution. In the case where C is known, the likelihood ratio test (LRT) can be easily computed, and shown to be $T_{C^{-1}g}$. This test can also be shown to be the most powerful among the family $(T_f)_{f \in \mathbb{R}^p}$.

Our strategy to investigate the case C unknown is to approximate the test $T_{C^{-1}g}$, using the so-called learning sample $(\psi_i)_{i=1,\ldots,n}$. Two difficulties are involved: the optimal direction $f_o = C^{-1}g$ has to be approximated, and the null distribution of the resulting test variable need to be evaluated, in order the level to be nominal. Consequently, the study is divided into two steps.

As a first step, the estimation of the direction f_o requires an estimate of the covariance matrix C. The empirical estimate $\widehat{C}$ of C can be easily computed from the $(\psi_i)_{i=1,\ldots,n}$. However, in a large dimension context, the quality of this estimator is not ensured. In particular, in the case where $C = I$, and n and p goes to the infinity together, the Marčenko-Pastur distribution (Marčenko and Pastur, 1967) gives an illustration of the deformation of the spectrum due to the empirical estimation. An equivalent phenomenon occurs for other values of C, that can deteriorate the estimation $\widehat{C}$ of C. When a quantity involving C^{-1} is needed, like in our case, this deterioration may be particularly sensitive, due to the lack of control on the small eigenvalues.

This problem can be solved using a well-conditioned covariance estimator introduced by Ledoit and Wolf (2004). This estimator $\widehat{C}_I$ can be written as a combination

$$\widehat{C}_I = \gamma \widehat{C} + \rho I,$$

where I is the identity matrix, and where γ and ρ are real coefficients. Ledoit and Wolf (2004) proposed, in particular, some values for the coefficients γ and ρ, depending on the $(\psi_i)_{i=1,\dots,n}$, and that are asymptotically optimal. Optimality is meant with respect to a quadratic loss function, asymptotically as n and p goes to the infinity together.

For the problem studied here, we shown via Monte-Carlo simulations that a test based on the variable $d_{\widehat{C}_I^{-1}g} = \langle \psi_{n+1}, \widehat{C}_I^{-1} g \rangle$ is relatively efficient in the sense that it is more powerful than the rather naive test T_g.

As a second step, a suitable evaluation of the H_0-distribution is needed for providing correct p-values, and for ensuring the test to have a nominal level. Once the estimation method of the direction f_o is chosen, several resampling methods could be proposed for this task. The main difficulty is that the same sample is used both for estimating the direction f_o and for estimating the variance in that direction, that can bias the results.

We will describe a bootstrap procedure that allows to take into account the dependencies between those two estimates. Then we will compare this procedure to other resampling methods (particularly validation and cross-validation), through numerical simulations, in order to show its efficiency. We finally show that the resulting test is still more powerful than the naive test T_g.

This statistical model is useful for climate study, and particularly climate change detection. For example, it can be used considering that the $(\psi_i)_{i=1,\dots,n}$ are n years of observed temperatures at p different places, and g is a spatial guess-pattern of investigated climate change, provided by physically-based climate models. We will give an illustration of the testing procedure with such a climate change detection study, based on a real temperature dataset covering the Mediterranean area.

Finally, a very similar problem, also useful for climate study, regards the regression in large dimension. Consider now that $\psi_{n+1} \in \mathbb{R}^p$ and K real vectors $g_k \in \mathbb{R}^p$, for $k = 1, \dots, K$, are given. Assuming that ψ_{n+1} follows a $N(\sum_{k=1}^K \mu_k g_k, C)$, where $\mu = (\mu_1, \dots, \mu_K)^T$ is an unknown real vector, we search for an efficient estimation of μ. This problem can be identified as a multivariate linear regression problem, and the best linear unbiased estimator of μ is known to involve the quantity C^{-1}. Like previously, we will see that the use of the well-conditioned estimator $\widehat{C}_I$ could be relatively efficient when C is unknown and in a large dimension context.

References

Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2):365–411

Marčenko V, Pastur L (1967) Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR - Sbornik 1(4):457–483