G E P BOX'S M-TEST FOR EQUALITY OF POPULATION COVARIANCE MATRICES

FINAL PROJECT : UIC STAT522 (SPRING 2023)

QAYUM KHAN

1. Executive Summary

When each of independent observations has several components, we are in the situation of multivariate statistics. When those responses are organized into finitely many known groups and we would like to test for similarity of the subpopulations (typically, of their intragroup mean vectors and intragroup covariance matrices), we would then like to apply multivariate analysis of variance (MANOVA). Typically the model is set up with a grand mean, then a fixed treatment-effect offset to each group's mean, possibly more offsets if studying multiway (multifactor) effects, and idiosyncratic errors. However, the mechanism of MANOVA has three assumptions (see textbook's two models [JW07, 6-38, 6-59]):

(1) the errors have intergroup independence (the subpopulations don't influence each other) as well as intragroup independence,
(2) per group the errors are identically multinormally distributed with population mean zero, and
(3) population covariance matrices between groups are identical.

The purpose of G E P Box's M-test is to check whether the latter condition (3) is satisfied, given that the former two conditions are satisfied. Thus, it is a critical tool for the correct setup of any MANOVA model.

Box's M-test works well when both the dimension of observations and the number of groups are each less than five, and the number of observations per group is more than twenty. Besides limited ranges, it is known to be sensitive to the normality assumption (2), whereas Wilks' Λ^* for testing equality of means is more robust to nonnormality.

In Sections 2–3, we consider the rigorous mathematics behind the statistical test, especially its weighting coefficient. In subsequent sections, we discuss the appendices in which R programs are run on various multivariate datasets. There is a prebuilt function `boxM()` in the CRAN library `heplots`. This library of Michael Friendly handles sum-of-squares-and-cross-products (SSP) matrices for linear hypotheses (H) and errors (E) using intervals/ellipses; graphics are drawn via `plot()`.

2. Historical Context and Scientific Statement

The main source of this exposition is Box's original paper [Box49]. We use the same notation, except we write g for the number of population groups rather than k, which we reserve for a summation index.

Consider g mutually independent groups (1) of random multivariate samples $x_{lj} \in \mathbb{R}^p$. Suppose that they are multinormally (2) distributed:

$$X_{l1}, \ldots, X_{ln_l} \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu_l, A_l) \quad \text{for all } 1 \leqslant l \leqslant g.$$

Here, the l-th group nonuniformly has n_l samples, $\mu_l \in \mathbb{R}^p$ is unknown population mean, and $A_l \in \mathbb{R}^{p\times p}$ is the unknown population covariance.

Write $\nu_l := n_l - 1$ for the l-th degree of freedom, $N := \nu_1 + \cdots + \nu_g$ for total number of degrees of freedom, and $\nu := N/g$ for average number of degrees of freedom. The (unbiased) **sample covariance matrix** is

$$S_l \;:=\; \nu_l^{-1} \sum_{j=1}^{n_l} (x_{lj} - \overline{x}_l)(x_{lj} - \overline{x}_l)' \;\in\; \mathbb{R}^{p\times p}$$

where $\overline{x}_l := n_l^{-1} \sum_{j=1}^{n_l} x_{lj} \in \mathbb{R}^p$ is the l-th sample mean. Abbreviate $n := N + g = \sum_{l=1}^{g} n_l$. The **pooled sample covariance matrix** is

$$S_0 \;:=\; N^{-1} \sum_{l=1}^{g} \nu_l S_l \;\in\; \mathbb{R}^{p\times p}.$$

Consider the null hypothesis that population covariances are equal

$$H_0 \;:\; A_1 = \ldots = A_g =: A.$$

The likelihood ratio (LR) statistic is introduced in [NP28, xix]. Our 3rd textbook [And03, 10.2:8] and Homework5#3 ($g = 2$) both calculate the maximum likelihood estimate (MLE) and LR in general and under H_0.

Result 2.1. *The likelihood ratio test (LRT) statistic of H_0 is*

$$\Lambda_0 \;=\; \prod_{l=1}^{g} \left(\frac{|\widehat{A}_l|}{|\widehat{A}_0|} \right)^{n_l/2} \;\in\; [0, 1]$$

where $\widehat{A}_l = \frac{\nu_l}{n_l} S_l$ is MLE of A_l and $\widehat{A}_0 = \frac{N}{n} S_0$ is MLE of A under H_0.

The ratio of likelihoods of null hypothesis over none is in [NP33, 12].

Theorem 2.2 (Neyman–Pearson 1933)**.** *For any significance level $0 < \alpha < 1$, most powerful (MP) is the hypothesis test that is defined by rejection of H_0 if LRT statistic $\Lambda_0 < \lambda_\alpha$, for appropriate constant λ_α.*

In the univariate case ($p = 1$), the above Λ_0 reduces to the so-called L_1-statistic of Neyman–Pearson (1931), which was proven an unbiased test by Brown [Bro39]. Here, a hypothesis test at level α with binary loss is **unbiased** if the **power function** $\beta(\theta) := \mathbb{P}[\text{reject } H_0|\theta]$ satisfies $\beta \leqslant \alpha$ under assumption H_0 is true and $\beta > \alpha$ assuming H_0 is false.

The constant λ_α can be determined asymptotically also, as follows. The following statement and a sketch is in [Wil38]; with appropriate hypotheses, a simple rigorous proof is provided by R M Dudley [Dud03].

Theorem 2.3 (Wilks 1938)**.** *Under suitable regularity hypotheses (such as for existence and uniqueness of an interior-point MLE), the statistic* $M_0 := -2\log\Lambda_0$ *is asymptotically distributed as* $\chi^2((p+1)(g-1)p/2)$.

A correction to the univariate case ($p = 1$) of Result 2.1 was offered in [Bar37, 15]. We state it for multivariate observations via determinants.

Definition 2.4 (Bartlett 1937)**.** Consider the non-MLE based statistic

$$\Lambda \;=\; \prod_{l=1}^{g} \left(\frac{|S_l|}{|S_0|}\right)^{\nu_l/2} \;\in\; [0,1].$$

Write $M := -2\log\Lambda$. The univariate M was used and studied by H O Hartley [Har40]. The unbiasedness of the multivariate Bartlett's Λ-test was eventually proven rigorously by M D Perlman [Per80, 2.1].

G E P Box is responsible for the development of multivariate M, so the test is named after him, and proposed a further correction [Box49].

Definition 2.5 (Box 1949, "Box's M-test")**.** Consider the constant

$$1-\rho \;:=\; \frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{l=1}^{g}\nu_l^{-1} - N^{-1}\right).$$

Reject the above H_0 if $\rho M > \chi^2_\alpha((p+1)(g-1)p/2)$, for a given level α.

Remark 2.6. In practice it seems to work well if $p, g < 5$ and all $n_l > 20$.

3. Partly Deriving the ρ-factor

The purpose of this section is to begin to reveal some of the mystery of the formula for ρ in Definition 2.5. The full derivation is extremely complicated because of asymptotic analysis of special functions, so we refer the reader to [Box49] for the details of the full tour-de-force story.

For motivation, consider the univariate case ($p = 1$). If $Z \sim \mathcal{N}(0,1)$ which has density $(2\pi)^{-1/2}\exp(-z^2/2)$ for all $z \in \mathbb{R}$, by change-of-variables its square Z^2 has density $\frac{(1/2)^{1/2}}{\Gamma(1/2)}x^{-1/2}e^{-x/2}$ for all $x > 0$, which is $\mathrm{Gamma}(1/2, 1/2) = \chi^2(1)$ with rate parameterization, hence

Z^2 has moment generating function $(1-2t)^{-1/2}$ for all $t < 1/2$. So if $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then their squares are also independent and identitically distributed, so the sum $Z_1^2 + \cdots + Z_d^2$ has moment generating function the product $[(1-2t)^{-1/2}]^d = (1-2t)^{-d/2}$, hence the sum has distribution Gamma$(d/2, 1/2) = \chi^2(d)$. Therefore, if $Y_1, \ldots, Y_{d+1} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then the scaled sample variance $s^2/\sigma^2 := \sum_{j=1}^{d+1} \left((Y_j - \overline{Y})/\sigma\right)^2$ has a $\chi^2(d)$ distribution, since we lose one degree of freedom due to $\overline{Y}$. The probability density function of $x > 0$ is $\frac{(1/2)^{d/2}}{\Gamma(d/2)} x^{d/2-1} e^{-x/2}$. Notice therein the occurence of the special function of Daniel Bernoulli (1729):

$$\Gamma(s) \ := \ \int_0^\infty t^{s-1} e^{-t}\, dt \quad \text{for all } s > 0,$$

where $\Gamma(n+1) = n!$. The special numbers B_n of Jacob Bernoulli (1713), where $B_0 = 1, B_1 = -1/2, B_2 = 1/6, B_3 = 0$, now occur because Stirling's approximation $n! \approx \sqrt{2\pi n}(n/e)^n$ has this generalization [Bar00].

Theorem 3.1 (Barnes 1900)**.** *For all complex $s \in \mathbb{C}$ and small $h \in \mathbb{C}$,*

$$\log \frac{\Gamma(s+h)}{\sqrt{2\pi}} = (s + h - \tfrac{1}{2}) \log s - s - \sum_{r=1}^{n} (-1)^r \frac{B_{r+1}(h)}{r(r+1)s^r} + \mathcal{O}(s^{-n-1})$$

where the Bernoulli function is $B_n(h) := (B+h)^n = \sum_{k=0}^n \binom{n}{k} B_{n-k} h^k$.

Now consider the multivariate case ($p \geqslant 1$). If $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} \mathcal{N}_p(0, A)$, then the sum $\sum_{j=1}^d Z_j Z_j'$ is a $p \times p$ random matrix that by definition has the Wishart distribution $\mathcal{W}_d(A)$ where $d \geqslant p$ is the degrees of freedom. For example, our setting has $\nu_l S_l \sim \mathcal{W}_{\nu_l}(A_l)$. Its density is in [Wis28].

Theorem 3.2 (Wishart 1928)**.** *The matrix-valued distribution $\mathcal{W}_d(A)$ has probability density function of positive definite $W \in \mathbb{R}^{p \times p}$ given by*

$$\frac{|W|^{(d-p-1)/2} \exp\left(-\operatorname{tr} A^{-1} W/2\right)}{2^{dp/2} |A|^{d/2} \cdot \pi^{p(p-1)/4} \prod_{k=0}^{p-1} \Gamma\left(\frac{d-k}{2}\right)}.$$

Box uses these two theorems as a starting point for intricate analysis of the characteristic exponent. We only jump in to show notation of his resulting algebra. He selects the multiplier ρ so that $\alpha_1 = 0$. Here, it lies in a family of constants with a particular definition [Box49, (43)]

$$\alpha_1 \ := \ -\frac{g}{3}(3D_1\beta + 2D_2)$$

where ρ occurs in $\beta := (1-\rho)\nu$, and $D_r := \delta_r \gamma_r$ with [Box49, (39–40)]

$$\begin{aligned}\delta_r &:= B_{r+1}\left(-\frac{B+p}{2}\right) - B_{r+1}\left(-\frac{B}{2}\right)\\ \gamma_r &:= g^{-1}\sum_{l=1}^{g}\left(\frac{\nu}{\nu_l}\right)^{r-1} - g^{-r}.\end{aligned}$$

Note $\alpha_1 = 0$ iff $1-\rho = -\frac{2D_2}{3D_1}/\nu = -\frac{2\delta_2}{3\delta_1}\frac{\gamma_2/\nu}{\gamma_1}$. Next, since $\nu = N/g$, note

$$\begin{aligned}\gamma_1 &= 1 - g^{-1}\\ \gamma_2 &= g^{-1}\sum_{l=1}^{g}\frac{\nu}{\nu_l} - g^{-2} = \nu g^{-1}\left(\sum_{l=1}^{g}\nu_l^{-1} - N^{-1}\right).\end{aligned}$$

Now, using the aforementioned special values of the Bernoulli numbers,

$$\begin{aligned}B_2(h) &= (B+h)^2 = B_2 + 2B_1 h + B_0 h^2 = \frac{1}{6} - h + h^2\\ B_3(h) &= (B+h)^3 = B_3 + 3B_2 h + 3B_1 h^2 + B_0 h^3 = \frac{1}{2}h - \frac{3}{2}h^2 + h^3.\end{aligned}$$

Then the formal variable "B" can be expanded to yield [Box49, (42)][1]

$$\begin{aligned}\delta_1 &= 4^{-1}\left[(B-p)^2 - B^2\right] = 4^{-1}\left[-(-p) + (-p)^2\right] = 4^{-1}p(p+1)\\ \delta_2 &= 8^{-1}\left[(B-p)^3 - B^3\right] = 8^{-1}\left[\frac{1}{2}(-p) - \frac{3}{2}(-p)^2 + (-p)^3\right]\\ &= -16^{-1}p(2p^2 + 3p - 1).\end{aligned}$$

Thus $-\frac{2\delta_2}{3\delta_1} = \frac{2p^2+3p-1}{6(p+1)}$ and $\frac{\gamma_2/\nu}{\gamma_1} = \frac{1}{g-1}\left(\sum_{l=1}^{g}\nu_l^{-1} - N^{-1}\right)$, as desired.

4. Example : weight gain in rats

In Figure 4, three treatment groups of rats' weight gains are measured in Weeks 1–4 (i.e. once differenced timeseries) [Box50, Table D].

5. Example : various from CRAN library candisc

Michael Friendly's "HE plot MANOVA examples" : `https://cran.r-project.org/web/packages/heplots/vignettes/HE_manova.html`

[1] Our last equality of δ_2 actually yields "+1" instead of "−1", so $2p^2 + 3p + 1 = (2p+1)(p+1)$. Is this a typo carried forward in the literature and computer code?

INITIAL WEIGHT AND WEEKLY GAINS IN WEIGHT FOR 27 RATS

Group 1. Control						Group 2. Thyroxin						Group 3. Thiouracil					
Rat	y_0	y_1	y_2	y_3	y_4	Rat	y_0	y_1	y_2	y_3	y_4	Rat	y_0	y_1	y_2	y_3	y_4
1	57	29	28	25	33	11	59	26	36	35	35	18	61	25	23	11	9
2	60	33	30	23	31	12	54	17	19	20	28	19	59	21	21	10	11
3	52	25	34	33	41	13	56	19	33	43	38	20	53	26	21	6	27
4	49	18	33	29	35	14	59	26	31	32	29	21	59	29	12	11	11
5	56	25	23	17	30	15	57	15	25	23	24	22	51	24	26	22	17
6	46	24	32	29	22	16	52	21	24	19	24	23	51	24	17	8	19
7	51	20	23	16	31	17	52	18	35	33	33	24	56	22	17	8	5
8	63	28	21	18	24							25	58	11	24	21	24
9	49	18	23	22	28							26	46	15	17	12	17
10	57	25	28	29	30							27	53	19	17	15	18

y_0 represents initial weight of rat
y_1 gain in 1st week
y_2 gain in 2nd week
y_3 gain in 3rd week
y_4 gain in 4th week

Figure 1. weight gains of rats per treatment group

References

[And03] Theodore W Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[Bar00] Ernest W Barnes. The theory of the Gamma function. *Messenger Math.*, 29:64–128, 1900. URL: http://resolver.sub.uni-goettingen.de/purl?PPN599484047_0029.

[Bar37] Maurice S Bartlett. Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, 160:268–282, 1937. doi:10.1098/rspa.1937.0109.

[Box49] George E P Box. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36:317–346, 1949. doi:10.1093/biomet/36.3-4.317.

[Box50] George E P Box. Problems in the analysis of growth and wear curves. *Biometrics*, 6(4):362–389, 1950. doi:10.2307/3001781.

[Bro39] George W Brown. On the power of the L_1 test for equality of several variances. *Ann. Math. Statist.*, 10(2):119–128, 1939. doi:10.1214/aoms/1177732210.

[Dud03] Richard M Dudley. Wilks' theorem; a likelihood ratio test for nested composite hypotheses, 2003. OpenCourseWare lecture note from Mathematical Statistics (18.466). URL: https://math.mit.edu/~rmd/650/wilks.pdf.

[Har40] Herman O Hartley. Testing the homogeneity of a set of variances. *Biometrika*, 31:249–255, 1940. doi:10.1093/biomet/31.3-4.249.

[JW07] Richard A Johnson and Dean W Wichern. *Applied multivariate statistical analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, sixth edition, 2007.

[NP28] Jerzy Neyman and Egon S Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A(1–2):175—240, 1928. doi:10.2307/2331945.

[NP33] Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A*, 231:289–337, 1933. doi:10.1098/rsta.1933.0009.

[Per80] Michael D Perlman. Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *Ann. Statist.*, 8(2):247–263, 1980. doi:10.1214/aos/1176344951.

[Wil38] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 1938. doi:10.1214/aoms/1177732360.

[Wis28] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1–2):32–52, 1928. doi:10.1093/biomet/20A.1-2.32.

Email address: qayum@uchicago.edu