Applied Multivariate Techniques

Daniele Zago

February 7, 2022

Contents

Lecture 3: Canonical correlation analysis

2022-01-27

Canonical correlation analysis (CCA) is a rather old technique which has seen a big resurgence of interest, especially in psychological and psychometric analysis. We consider the following problem: given n observation of two sets of variables,

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_{11} & \dots & y_{1q} \\ y_{21} & \dots & y_{2q} \\ \vdots & \dots & \vdots \\ y_{n1} & \dots & y_{nq} \end{pmatrix}$$

the goal is to find a linear combination $C_x = Xa$ and a linear combination $C_y = Yb$ such that

$$(a_1, b_1) = \underset{a,b}{\operatorname{argmax}} \operatorname{Corr}(Xa, Yb). \tag{1}$$

Notation The quantities C_x and C_y are called ***scores***.

Notation We define the following matrices:

$$\begin{aligned} \mathbb{V}[X] : \quad & S_{11\,p\times p} = \frac{1}{n} X^\top H^\top H X = \frac{1}{n} X^\top H X \\ \mathbb{V}[Y] : \quad & S_{22\,q\times q} = \frac{1}{n} Y^\top H Y \\ \operatorname{Cov}(X, Y) : \quad & S_{12\,p\times q} = \frac{1}{n} X^\top H Y \end{aligned}$$

The maximization problem in (1) thus becomes

$$(a_1, b_1) = \underset{a,b}{\operatorname{argmax}} \frac{a^\top S_{12} b}{\sqrt{a^\top S_{11} a \cdot b^\top S_{22} b}} = \frac{\operatorname{Cov}(C_x, C_y)}{\sqrt{\mathbb{V}[C_x] \cdot \mathbb{V}[C_y]}} \tag{2}$$

and if we define $C_X = HXa$, we have $S_{C_x C_x} = \frac{1}{n} a^\top X^\top H X a = a^\top S_{11} a$, and the same applies to $S_{C_y C_y} = b^\top S_{22} b$. Finally, $\operatorname{Cov}(C_x, C_y) = a^\top S_{12} b$, hence the final equality.

Since the solution is invariant under rescaling of vectors a and b, we can find an infinite number of solutions unless we impose some constraints on the maximization procedure. In this case, we impose the following constraints to Equation (2), which guarantee that the solution is unique:

$$a^\top S_{11} a = 1$$

$$b^\top S_{22} b = 1$$

After finding the first solution, we can proceed similarly to principal component analysis in order to find the second pair of canonical vectors, such that

$$(a_2, b_2) = \underset{\substack{a,b:\\ a^\top S_{11} a = 1\\ b^\top S_{22} b = 1\\ a_1^\top S_{11} a = 0\\ b_1^\top S_{22} b = 0}}{\operatorname{argmax}} \frac{a^\top S_{12} b}{\sqrt{a^\top S_{11} a \cdot b^\top S_{22} b}} = \frac{\operatorname{Cov}(C_x, C_y)}{\sqrt{\mathbb{V}[C_x] \cdot \mathbb{V}[C_y]}} \tag{3}$$

Theorem 1 (Canonical correlation analysis)
The k solutions to the canonical correlation problem can be found by defining the following matrix,

$$S_{11}^{-1/2} S_{12} S_{22}^{-1/2} \overset{SVD}{=} UDV^\top.$$

Then, the solution $A = (a_1 \ \cdots \ a_k)$ and $B = (b_1 \ \cdots \ b_k)$ is given by the first k eigenvectors of U and V, respectively.

Proof.
Let us start by considering $a^\top S_{12} b$ under the constraint that $a^\top S_{11} a = 1$ and $b^\top S_{22} b = 1$. Apply the following change of coordinates,

$$u_0 = S_{11}^{1/2} a \implies a = S_{11}^{-1/2} u_0$$

$$v_0 = S_{22}^{1/2} b, \implies b = S_{22}^{-1/2} v_0$$

then the problem (2) becomes

$$\underset{u_0, v_0}{\operatorname{argmax}}\, u_0^\top S_{11}^{-1/2} S_{12} S_{22}^{-1/2} v_0,$$

under the constraints $u_0^\top u_0 = 1$ and $v_0^\top v_0 = 1$. Hence, the solution is given by the first eigenvectors of the U and V matrices from the SVD of the matrix

$$S_{11}^{-1/2} S_{12} S_{22}^{-1/2} = UDV^\top.$$

Repeating the argument yields the following solutions to the canonical correlations problem.

□

Remark Note that if $k = \operatorname{rank}\left(S_{11}^{-1/2} S_{12} S^{-1/2}\right)$, then we have that in most cases

$$k \approx \min\left\{\operatorname{rank} X, \operatorname{rank} Y\right\},$$

hence we can find at most k canonical vectors

$$U = (a_1, a_2, \ldots, a_k), \quad V = (b_1, b_2, \ldots, b_k).$$

As always, this solution is unique up to a change in sign of the eigenvectors.

Partial least squares CCA has connection to the Partial Least Squares (PLS) estimator, which

Consider the SVD applied to the residualized matrices,

$$HX = U_X D_X V_X^\top$$

$$S_{11} = V_X D_X^2 V_X^\top$$

$$HY = U_Y D_Y V_Y^\top$$

$$S_{22} = V_Y D_Y V_Y^\top$$

$$S_{12} = V_X D_X U_X^\top U_Y D_Y V_Y^\top$$

then, if we write the matrix solution in terms of the above SVD, we have

$$\begin{aligned} S_{11}^{-1/2} S_{12} S_{22}^{-1/2} &= V_X D_X^{-1} \cancel{V_X^\top V_X} D_X U_X^\top U_Y D_Y \cancel{V_Y^\top V_Y} D_Y^{-1} V_Y^\top \\ &= V_X U_X^\top U_Y V_Y^\top, \end{aligned}$$

and we have that $U_Y V_Y^\top$ is the SVD of the normalized data, i.e. all variances are equal. Hence, we conclude that this solution is invariant under any linear transformation of the data (unlike the PLS).

Lecture 4: Closed-testing framework

2022-02-02

In this lecture we will consider the problem of performing multiple tests while controlling the overall Type I error at the specified α level. We will do so by casting the usual multiple comparison adjustments into the closed-testing framework (Goeman and Solari, 2011). This framework offers a unified view of multiple testing and is the *de-facto* standard for hypothesis testing.

4.1 Multiple testing

Consider two groups y_1 and y_2, which we assume are drawn from two densities,

$$y_1 \sim P_1, \quad y_2 \sim P_2.$$

Our goal is to compare the two groups and see if the samples come from the same distribution. Consider for example when we assume a parametric form for P_i, for instance $P_1 = \mathcal{N}(\mu_i, \sigma^2)$, then the hypothesis would become

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

With the usual t-test, we consider the test statistic

$$t_{\text{obs}} = \frac{\bar{y}_1 - \bar{y}_2}{\widehat{\sigma}_{\bar{y}_1 - \bar{y}_2}} \sim t_{n-2},$$

and we define the ***p-value*** as the probability under the null hypothesis of observing a result as extreme as the observed statistic,

$$p = \mathbb{P}(|T| \geq t_{\text{obs}} | H_0), \quad T \sim t_{n-2}.$$

The ***statistical test*** is an object which yields a binary outcome, either 1 for a rejection and 0 for a non-rejection, depending on the limit L that we choose,

$$\varphi = \begin{cases} 1 & \text{if } p \leq L \\ 0 & \text{if } p \geq L \end{cases} \tag{4}$$

We do have different types of errors, for instance

$$\begin{aligned} \textsc{Type-I error} \quad & \mathbb{P}(\varphi = 1 | H_0) = \mathbb{P}(p \leq L | H_0) \leq \alpha. \\ \textsc{Power} \quad & \mathbb{P}(\varphi = 1 | H_1) \geq \alpha \\ \textsc{Type-II error} \quad & 1 - \textsc{power} = \beta \end{aligned}$$

if $(1-\beta) \geq \alpha$, the test is called ***unbiased***, whereas if $1 - \beta \to 1$, the test is ***consistent***.

We have that the p-value of a continuous statistic t is uniformly distributed in $[0, 1]$ under the null hypothesis (Murdoch et al., 2008), i.e.

$$P | H_0 \sim U(0, 1),$$

whereas if the test is consistent, then under H_1 the p-value is more skewed towards 0.

4.2 Multivariate framework

Consider now a setting in which we perform a statistical test on a multiple variable, i.e.

$$y_1 \sim P_1, \quad y_2 \sim P_2, \quad P_i \in \mathbb{R}^n,$$

then the null hypothesis becomes

$$\begin{cases} H_1 : \mu_{11} = \mu_{21} \\ H_2 : \mu_{12} = \mu_{22} \\ \dots \\ H_n : \mu_{1n} = \mu_{2n} \end{cases} \implies H_0 : \bigcap_{i=1}^{n} H_i$$

We can solve the problem using Hotelling's T, i.e.

$$T^2 = (\bar{y}_1 - \bar{y}_2)^\top \Sigma^{-1} (\bar{y}_1 - \bar{y}_2),$$

which has a χ^2 distribution if Σ does not have to be estimated. Whenever Σ has to be estimated by a $\widehat{\Sigma}$, the T^2 statistic has a Hotelling's T distribution. If $p < L$ we conclude that there is a difference between the distributions, but we do not know *where* this difference lies.

The concept is that there is a true set $\tau \subseteq \{1, 2, \dots, n\}$ that collect the true variables which differ between he populations. Hence, the true null hypothesis is

$$H_0 : \bigcap_{i \in \tau} H_i.$$

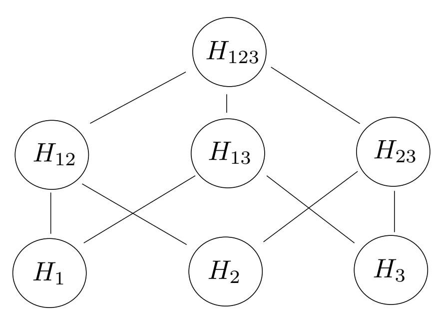

Figure 1: Graph of the hierarchical relationship between the null hypotheses.

We want a testing procedure such that all cases depicted in Figure 1 are considered and the rejection happens at the α level. This is an extension of the Type-I error given by the ***family-wise error rate***, which can be loosely defined as

$$\text{FWER} = \{\text{at least 1 error among all hypotheses}\}$$

We can apply a Hotelling's T test for any of the above situations, however we do not know which of the $i = 1, \dots, 2^3$ null hypotheses is actually true.

A good solution to the above problem is provided by the ***closed testing*** procedure, which has been proven to be the only admissible procedure (Goeman and Solari, 2011), i.e. if there is another procedure which controls the FWER then it must be a closed testing procedure.

Closed-testing procedure Consider p_{123} to be the p-value which tests H_{123}, p_{12} the p-value which tests H_{12}, and so on. Suppose that we want to test individual hypotheses H_1 and H_2. We reject H_1 if we reject all hypotheses H_{ij}, H_{ijk} which contain the subscript 1, and the same applies for H_2. Then,

$$H_1 \text{ rejected } \iff p_1, p_{12}, p_{13}, p_{123} \leq \alpha$$

$$H_2 \text{ rejected } \iff p_2, p_{12}, p_{23}, p_{123} \leq \alpha$$

In general, the adjusted test using the above procedure for a general subset of null hypotheses $S \subseteq \{1, 2, \ldots, n\}$, denoted by $\tilde{\varphi}_S$, is

$$\tilde{\varphi} = \min_{\mathcal{S} \supseteq S} \varphi_{\mathcal{S}},$$

You can check using the definition (4) of statistical test that this indeed is the correct definition of the closed testing procedure. Hence if $\tilde{\varphi}_S = 1 \implies$ we reject H_1. This closed-testing procedure has been first described by Marcus et al. (1976) and the proof of the fact that the FWER is controlled by α is very simple.

Proof.
Consider $H_0 : \bigcap_{i \in \tau} H_i$ and the following sets,

$$A = \{\text{at least 1 false rejection}\}$$

$$B = \{\varphi_\tau = 1\}$$

and observe that $A \cap B = A$ by construction of the closed-testing procedure. We know that

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) \leq \mathbb{P}(B) \leq \alpha,$$

since B is a proper test. Hence, the probability of making *any* false rejection is bounded by α.

□

4.3 Bonferroni correction

The most frequent approach to multiple testing is the Bonferroni procedure, which can be shown to be a special case of the closed-testing procedure. For $i \in \{1, \ldots, m\}$, the statistical test for the i-th hypothesis is

$$\tilde{\varphi}_i = \mathbb{1}_{p_i \leq \frac{\alpha}{m}} = \mathbb{1}_{m \cdot p_i \leq \alpha},$$

hence we usually talk about ***adjusted p-values*** instead of adjusted limit.

Proof.
Assume that the set of true null hypotheses is τ, then the FWER for the Bonferroni procedure is

$$\mathbb{P}\Big(\bigcup_{i \in \tau} p_i \leq \frac{\alpha}{m} | H_0\Big) \leq \sum_{i \in \tau} \mathbb{P}\Big(p_i \leq \frac{\alpha}{m} | H_0\Big) = |\tau| \cdot \frac{\alpha}{m} \leq m \cdot \frac{\alpha}{m} = \alpha.$$

□

Remark This is a very powerful result which does not assume any type of dependence between the p-values. However, when the dependence is very high we have an extremely conservative test which tends to be too strict.

4.4 Bonferroni-Holm

The Bonferroni-Holm procedure uses ordered p-values, and starts computing

$$\begin{aligned} p_{(1)} \cdot m \leq \alpha &\implies \text{reject } H_1\text{, otherwise stop} \\ p_{(2)} \cdot (m-1) \leq \alpha &\implies \text{reject } H_2\text{, otherwise stop} \\ &\vdots \\ p_{(m)} \cdot 1 \leq \alpha &\implies \text{reject } H_m\text{, otherwise stop} \end{aligned}$$

We will now see whether Bonferroni and Bonferroni-Holm procedures can be seen as special cases of the closed-testing procedure. Suppose that we want to test the global null hypothesis H_{123}, then using Bonferroni we would test

$$\begin{aligned} \text{Reject } H_{123} &\iff \min p_i \cdot 3 = p_{(1)} \cdot 3 \leq \alpha \\ \text{Reject } H_{12} &\iff \min\{p_1, p_2\} \cdot 2 = p_{(1)} \cdot 2 \leq \alpha \end{aligned}$$

hence, if we reject for H_{123} we automatically reject all the connected null hypotheses. Consider now rejecting H_2, by the closed testing procedure we now only have to check for H_{23} if $p_2 \cdot 2 \leq \alpha$, and we get a rejected H_2 for free. Finally, we only need to check for H_3, which can be done by only checking if $p_3 \leq \alpha$.

Hence, by applying the closed-testing procedure using the minimum function we are employing the Bonferroni-Holm procedure.

In conclusion, the closed-testing procedure only needs the definition of

1. A hierarchical multiple testing setting.
2. Any kind of statistical testing procedure to put on each node (likelihood ratio, permutations, bootstrap, ...).

Issues Given m tests, we have a total graph consisting of $2^m - 1$ nodes, hence we need to find shortcuts in order to compute the overall procedure. In the Bonferroni case, we only need to sort the p-values and we have a complexity of $\mathcal{O}(m)$.

Multiple testing procedures often tried to maximize the power in univariate leaf tests $H_1, H_2, \ldots, H_m$. However, it is often the case that we can reject H_{12} under the closed testing procedure but neither H_1 nor H_2 can be rejected. As a consequence, we get some information in which combinations yield the difference between distributions.

Therefore, we can define a ***upper bound*** for the number of null hypotheses

$$\overline{m}_0(S = H_{123}) = \max_k \{|k| : \tilde{\varphi}_k = 0\}.$$

As a consequence, the ***lower bound*** on the number of alternative hypotheses

$$\underline{\mathrm{m}}_1(S) = \min_k \{|k| : \tilde{\varphi}_k = 1\} = |S| - \overline{m}_0.$$

For instance, rejecting H_{123}, H_{12} and H_{13} means that among H_1, H_2, H_3 we're not able to judge whether we have H_1, H_2 or H_3 alternative hypotheses, but we are able to tell that two of them are alternative.

Conclusion

With the closed-testing procedure, we are calculating confidence intervals in the number of null hypotheses.

References

Goeman, J. J. and Solari, A. (2011). «Multiple Testing for Exploratory Research». In: *Statistical Science* 26.4.

Marcus, R. et al. (1976). «On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance». In: *Biometrika* 63.3, 655–660.

Murdoch, D. J. et al. (2008). «P-Values Are Random Variables». In: *The American Statistician* 62.3, 242–245.