Canonical Correlation Analysis (I)

Carsi González, Ana

ana.carsi_gonzalez@uni-heidelberg.de

July 2024

I. Introduction

Dimensionality reduction remains a cornerstone in the field of data analysis and machine learning. For example, in the case of a multivariate regression model, estimating a large number of regression coefficients becomes impractical, especially when the number of observations n is smaller than the number of parameters. One example of this is Gudmundsson's [1] analysis of economic variables, which includes 42 regression coefficients with only 36 data points available.

To address the problem of high-dimensional parameter estimation, reduced-rank models are often employed, one of which is canonical correlation analysis (CCA). This approach reduces dimensionality and improves interpretability of the data by focusing on the linear combinations that exhibit the highest correlation between two classes[2].

Moreoever, CCA can be utilized at different stages in a machine learning pipeline, such as feature-extraction, noise reduction and post-processing. This can provide insights into which features are most predictive and how different sets of features relate to each other. This report serves as a brief summary of the mathematical framework involved in the CCA setup, from its classical definition to the adaptations for numerical solvability.

A. Classical CCA Formulation

Canonical correlation analysis was introduced by Hotelling [3] as a method of summarizing relationships between two sets of variables. The objective is to find the linear combination of one set of variables which is most correlated with any linear combination of a second set of variables.

In this report we will deal with a $n \times (p_1 + p_2)$ data block-matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ which we assume for simplicity is centered. We name the covariance matrices of the first and second set $\mathbf{C}_{11} = \mathbf{X}_1^T \mathbf{X}_1$, $\mathbf{C}_{22} = \mathbf{X}_2^T \mathbf{X}_2$, which are PSD since:

$$u^T \mathbf{C}_{11} u = u^T (\mathbf{X}_1^T \mathbf{X}_1) u = (\mathbf{X}_1 u)^T (\mathbf{X}_1 u) \geq 0$$

$$v^T \mathbf{C}_{22} v = v^T (\mathbf{X}_2^T \mathbf{X}_2) v = (\mathbf{X}_2 v)^T (\mathbf{X}_2 v) \geq 0$$

and the classical assumption that $\max\{p_1, p_2\} < n$. Plus, the cross-covariance matrix $\mathbf{C}_{12}$ will then fulfill: $\operatorname{rank}(\mathbf{C}_{12}) \leq \min\{p_1, p_2\}$.

Let now $y_1 = \mathbf{X}_1 u$ and $y_2 = \mathbf{X}_2 v$ be two linear combinations. The Pearson correlation coefficient between them is given by

$$\rho = \frac{u^T \mathbf{C}_{12} v}{\sqrt{u^T \mathbf{C}_{11} u} \sqrt{v^T \mathbf{C}_{22} v}}. \tag{1}$$

which can be interpreted as the *canonical correlation coefficient* [4]. Maximizing this coefficient is the objective of CCA.

Thus, CCA can also be formulated making use of the Rayleigh Quotient [5], where the goal is to maximize the ratio of the covariance between the linear combinations u and v to the product of their standard deviations:

$$\max_{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}} \frac{u^T \mathbf{C}_{12} v}{\sqrt{u^T \mathbf{C}_{11} u} \sqrt{v^T \mathbf{C}_{22} v}} \tag{2}$$

Analogically, this can be rewritten as a constrained optimization problem:

$$\begin{cases} \max_{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}} u^T \mathbf{C}_{12} v \\ \\ \text{subject to } u^T \mathbf{C}_{11} u = 1 \text{ and} \\ v^T \mathbf{C}_{22} v = 1 \end{cases} \tag{3}$$

The constraints ensure that the linear combinations u and v are normalized, which is analogous to finding the eigenvectors associated with the largest eigenvalues in the traditional Rayleigh quotient problem.

A.1 Eigenproblem

The Rayleigh quotient for a symmetric matrix $\mathbf{C}$ is given by:

$$R(x) = \frac{x^T \mathbf{C} x}{x^T x}$$

The maximum value of this quotient corresponds to the largest eigenvalue of the matrix $\mathbf{C}$. Therefore, CCA can be formulated by means of an eigenvalue problem, where the eigenvalues obtained from it correspond to the squares of the canonical correlations. Indeed,

- The largest eigenvalue corresponds to the strongest canonical correlation.
- Subsequent eigenvalues provide orthogonal linear relationships between the two set of variables.

Reframing the maximization objetive of CCA as an eigenvalue problem allows us to use power iteration or the QR algorithm, simplifying the complexity of the problem. Plus, the eigenvalue problem ensures global optimality for symmetric matrices, whereas maximizing the correlations in CCA directly often involves non-convex optimization and a more complicated procedure.

To solve the eigenvalue problem, we can rewrite (3) in terms of the Lagrangian:

$$\begin{aligned} \mathcal{L}(u, v, \lambda, \mu) = \quad & u^T \mathbf{C}_{12} v - \frac{\lambda}{2}(u^T \mathbf{C}_{11} u - 1) \\ & - \frac{\mu}{2}(v^T \mathbf{C}_{22} v - 1) \end{aligned} \tag{4}$$

Taking derivatives, we get

$$\begin{aligned} \frac{\partial L}{\partial u} &= \mathbf{C}_{12} v - \lambda \mathbf{C}_{11} u = 0 \\ \frac{\partial L}{\partial v} &= u^T \mathbf{C}_{12} - \mu \mathbf{C}_{22} v = 0 \end{aligned} \tag{5}$$

which yields

$$\begin{cases} \mathbf{C}_{12} v = & \lambda \mathbf{C}_{11} u \\ \mathbf{C}_{12}^T u = & \mu \mathbf{C}_{22} v \end{cases} \tag{6}$$

Since the canonical correlation coefficient ρ is a single value and correlations are symmetric, we can assume $\lambda = \mu$ and we get

$$\begin{cases} \mathbf{C}_{12}^T \mathbf{C}_{11}^{-1} \mathbf{C}_{12} v = & \lambda^2 \mathbf{C}_{22} v \\ \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T u = & \lambda^2 \mathbf{C}_{11} u \end{cases} \tag{7}$$

In these equations, u and v are the eigenvectors corresponding to the generalized eigenvalues λ^2. The matrices U and V are then the matrices containing these eigenvectors as columns:

$$\begin{aligned} \mathbf{C}_{12}^T \mathbf{C}_{11}^{-1} \mathbf{C}_{12} V = \mathbf{C}_{22} V \Lambda^2 \\ \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T U = \mathbf{C}_{11} U \Lambda^2 \end{aligned} \tag{8}$$

where Λ is the diagonal matrix of eigenvalues from the eigenvalue decomposition. We can refer to the pairs of matrices $(\mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T, \mathbf{C}_{11})$ and $(\mathbf{C}_{12}^T \mathbf{C}_{11}^{-1} \mathbf{C}_{12}, \mathbf{C}_{22})$ as two matrix pencils. $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ are both symmetric positive definite by definition of the rank of $\mathbf{C}_{12}$. $\mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T$ and $\mathbf{C}_{12}^T \mathbf{C}_{11}^{-1} \mathbf{C}_{12}$ are symmetric positive definite.

A.2 Generalized Eigenvalue Problem Solution

Given the setup:

$$\begin{aligned} \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T u = \lambda^2 \mathbf{C}_{11} u \\ \mathbf{C}_{12}^T \mathbf{C}_{11}^{-1} \mathbf{C}_{12} v = \lambda^2 \mathbf{C}_{22} v \end{aligned} \tag{9}$$

one could compute the Cholesky decomposition [4]:

$$\mathbf{C}_{11} = \mathbf{L}_{11} \mathbf{L}_{11}^T \text{ , } \mathbf{C}_{22} = \mathbf{L}_{22} \mathbf{L}_{22}^T.$$

Where $\mathbf{L}_{11}$ and $\mathbf{L}_{22}$ are lower triangular matrices. This transformation helps to convert the generalized eigenvalue problem into a standard eigenvalue problem. Then the first equation can be transformed:

$$\mathbf{C}_{11}^{-1/2} \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{12}^T u = \lambda^2 \mathbf{C}_{11}^{-1/2} \mathbf{C}_{11} u$$

since $\mathbf{C}_{11}^{-1} = \mathbf{L}_{11}^{-T} \mathbf{L}_{11}^{-1}$ ($\mathbf{L}_{11}^{-T}$ and $\mathbf{L}_{11}^{-T}$ exist thanks to positive definiteness of $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$)

and because $\mathbf{C}_{11}^{-1/2}\mathbf{C}_{11}\mathbf{C}_{11}^{-1/2} = \mathbb{I}_{p_1}$ we can write

$$\mathbf{C}_{11}^{-1/2}\mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{12}^{T}\mathbf{C}_{11}^{-1/2} z = \lambda^2 z.$$

Applying the same to the second equation, we have two standard eigenvalue problems:

$$\mathbf{M}_1 z = \lambda^2 z \tag{10}$$

$$\mathbf{M}_2 w = \lambda^2 w \tag{11}$$

where $\mathbf{M}_{11} = \mathbf{C}_{11}^{-1/2}\mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{12}^{T}\mathbf{C}_{11}^{-1/2}$ and $\mathbf{M}_{22} = \mathbf{C}_{22}^{-1/2}\mathbf{C}_{12}^{T}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22}^{-1/2}$.

Numerically, this could be solved with a command as

```
    # Cholesky decomposition
    L_1 = np.linalg.cholesky(C_1)
    # Transform matrices
    M1 = np.linalg.inv(L_1).T @
        C_12 @ np.linalg.inv(C_2) @
         C_12.T @ np.linalg.inv(L_1
        )
    # Solve the generalized
        eigenvalue problem
    eigvals1, eigvecs1 = eigh(M1)
    # Get the canonical variates
    u = np.linalg.inv(L_x) @
        eigvecs1
```

using the python function *eigh* from the package *scipy.linalg* and *numpy*.

B. Simultaneous Multiple Projections

In cases where we want to derive multiple pairs of Canonical Variates (CVs), we can extend the CCA formulation to find p pairs of u and v. This is referred to as a **fixed rank p-problem**, since it is assumed that both the U and U matrices will have p columns. Instead of maximizing individual ratios, we can reformulate the problem into a matrix form that captures the contributions of all p pairs simultaneously [5]. Indeed, the trace of a product of matrices can be interpreted as a form of generalized sum of eigenvalues:

$$\text{tr}(U^T\mathbf{C}_{12}V) = \sum_{j=1}^{p} R(u_j, v_j)$$

where R stands for the Rayleigh quotient subject to the constraints $u_j^T\mathbf{C}_{11}u_j = \mathbb{I}_p$, $v_j^T\mathbf{C}_{22}v_j = \mathbb{I}_p$. Thus, the maximization objective of the overall variance explained by the projections becomes:

$$\max_{U\in\mathbb{R}^{p_1\times p}, V\in\mathbb{R}^{p_2\times p}} \text{tr}(U^T\mathbf{C}_{12}V) \tag{12}$$

It is important to note that (13) and (2) are not equivalent, since the total variation of the canonical correlations is maximized in (13), rather than their individual successive ones as in (2). Every solution of (2) is solution to (13), but the opposite is not true.

C. Alternative CCA Definitions

Applying the Cholesky decomposition, we could introduce the unknowns $\tilde{u} = \mathbf{L}_{11}u$ and $\tilde{v} = \mathbf{L}_{22}v$ and the CCA problem becomes

$$\max_{\substack{\tilde{u}^T\tilde{u}=1\\ \tilde{v}^T\tilde{v}=1}} \tilde{u}^T(\mathbf{L}_{11}^{-T}\mathbf{C}_{12}\mathbf{L}_{22}^{-1})\tilde{v} \tag{13}$$

which allows us to consider the problem as optimization on a product of two unit spheres. However, the objective function remains non-symmetric. After the solution is found, one should not forget to multiply $\tilde{u}$ and $\tilde{v}$ by $\mathbf{L}_{11}^{-1}$ and $\mathbf{L}_{22}^{-1}$ respectively.

C.1 Singular Scatter $\mathbf{C}_{11}$ and/or $\mathbf{C}_{22}$

When the number of variables p_1 and p_2 is far bigger than the number of observations n, $\mathbf{C}_{11}$ or $\mathbf{C}_{22}$ may not be PD. Furthermore, $\mathbf{C}_{11}$ or $\mathbf{C}_{22}$ will be singular in case of collinearity between variables. A way through tackle this problem is through *regularization*. Here, a term is added to stabilize the solution, leading to the following GEVD:

$$\mathbf{C}_{12}(\mathbf{C}_{22} + \tau_2\mathbb{I}_{p_2})^{-1}\mathbf{C}_{21}u = \lambda^2(\mathbf{C}_{11} + \tau_{11}\mathbb{I}_{p_1})u$$

where τ_1 and τ_2 are regularization parameters.

D. Sparse CCA

The CVs from CCA can be difficult to interpret due to the involvement of multiple variables. To address that, methods to produce sparse solutions have been developed, which not only increase the interpretability but also make the solutions more robust against overfitting [4]. However, sparse CVs do not retain the optimal property of the standard CVs, i.e. sparse CVs tend to be correlated among the groups of variables. Indeed, although sparse CCA methods still impose the constraints $A_1^T \mathbf{C}_{11} A_1 = \mathbb{I}_{r_1}$ and $A_2^T \mathbf{C}_{22} A_2 = \mathbb{I}_{r_2}$, they fail to satisfy $u^T C_{12} v = 0_{r_1 \times r_2}$ since these CVs are no longer result of the eigenvalue problem. Here, r_1 and r_2 denote the number of canonical variates retained in each set after imposing sparsity constraints and A_1 and A_2 are $p_1 \times r_1$ and $p_2 \times r_2$ matrices representing the coefficients of the first and second set of variables.

D.1 Sparse CCA using GEVD

GEVD involves solving a generalized eigenvalue problem of the form:

$$Ua = \mu Va$$

where U and V are symmetric matrices and V is PD. For CCA, the formulation has matrices:

$$U = \begin{pmatrix} 0 & \mathbf{X}_1^T \mathbf{X}_2 \\ \mathbf{X}_2^T \mathbf{X}_1 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} \mathbf{X}_1^T \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2^T \mathbf{X}_2 \end{pmatrix}$$

A sparse GEVD problem can be formulated as:

$$\begin{cases} \max\left(a^T U a - \lambda P(a)\right) \\ \text{subject to } a^T V a \leq 1 \end{cases} \tag{14}$$

whew $P(a)$ is a penalty term that induces sparsity in a. A commonly used penalty term is

$$P(a) = \sum_{i=1}^{p} \log(\epsilon + |a_i|)$$

where ϵ is a small non-negative number to avoid issues with zero elements.

D.2 CCA as a LS problem

Exercise. Show that the CCA problem is equivalent to minimizing the distance between the canonical variates y_1 and y_2:

$$\min_{\substack{a_1^T C_{11} a_1 = 1 \\ a_2^T C_{22} a_2 = 1}} ||\mathbf{X}_1 a_1 - \mathbf{X}_2 a_2||_2$$

i.e. CCA becomes a LS fitting problem.
Solution We have:

$$\begin{aligned} ||\mathbf{X}_1 a_1 - \mathbf{X}_2 a_2||^2 = \\ (\mathbf{X}_1 a_1 - \mathbf{X}_2 a_2)^T (\mathbf{X}_1 a_1 - \mathbf{X}_2 a_2) = \\ = a_1^T \mathbf{C}_{11} a_1 - 2 a_1^T \mathbf{C}_{12} a_2 + a_2^T \mathbf{C}_{22} a_2 \end{aligned} \tag{15}$$

with the constraints $a_1^T \mathbf{C}_{11} a_1 = 1$ and $a_2^T \mathbf{C}_2 a_2 = 1$. The term $a_1^T \mathbf{C}_{12} a_2$ is the cross-correlation term that the CCA objective maximizes, which is equivalent to minimizing $-2 a_1^T \mathbf{C}_{12} a_2$. Therefore, minimizing the distance under the constraints is equivalent to solving the CCA problem.

When $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ are PD the problem is a constrained optimization problem on the product of unit spheres defined by them.

II. Conclusion

This report aimed to summarize the fundamental concepts of linear algebra related to canonical correlation analysis. Further discussion on numerical applications will be done in the following chapter.

References

[1] G. Gudmundsson. Multivariate analysis of economic variables. *Applied Statistics*, 26:48–59, 1977.

[2] Kun Chen Gregory C. Reinsel, Raja P. Velu. *Multivariate Reduced-Rank Regression*. Springer, 2022.

[3] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[4] N. Trendafilov and M. Gallo. *Multivariate Data Analysis on Matrix Manifolds*. Springer, 2021.

[5] Florian Yger et al. Adaptive canonical correlation analysis based on matrix manifolds. *Proceedings of the 29th International Conference on Machine Learning*, 2012.