Building Bridges: Implementing Redundancy Analysis by means of a Neural Network

Anonymous Author(s)
Affiliation
Address
email

Abstract

When analyzing data, there are situations where it is convenient to divide the analyzed variables into two subsets: a $\boldsymbol{X}$ block, of range p, that plays an independent or predictor role and a $\boldsymbol{Y}$ block, of range q, whose role is dependent, each of them containing the measurements of n individuals. However, knowing *a priori* which one is independent and which one is dependent is not trivial.

The Redundancy Analysis was developed by van den Wollenberg (1977) as a particular case of CCA (Hotelling, 1936). Given two standarized data sets $\boldsymbol{X}$ (orthogonalizable) and $\boldsymbol{Y}$, it consists in determining a set of orthonormal redundant components $\boldsymbol{w}_i$ ($i = 1, ..., min(p,q)$) such that the squared correlation of each of the redundant variables $\boldsymbol{X}\boldsymbol{w}_i$ with all the $\boldsymbol{Y}$ variates (Eq. 1) is maximized subject to unitary variance of the redundant variable. Thus, the Redundancy Index (RI) of $\boldsymbol{X}$ over $\boldsymbol{Y}$, proposed by Stewart and Love (1968), is maximized (Eq. 2). Such index is an asymmetrical measurement for the amount of explanation one set has over the other. By comparing $\boldsymbol{R_x}$ and $\boldsymbol{R_y}$, it can be determined which set of variables is explanatory and which one is the response.

$$cor^2(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{w}_i) = n^{-2}\boldsymbol{w}_i^t\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{X}\boldsymbol{w}_i \tag{1}$$

$$\begin{aligned} \max \quad \boldsymbol{R_y} \quad &= \max_{w_i} \sum_{i=1}^{min(p,q)} n^{-2}\boldsymbol{w}_i^t\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{X}\boldsymbol{w}_i \\ \text{s.t.} \quad & n^{-1}\boldsymbol{w}_i^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{w}_i = 1 \end{aligned} \tag{2}$$

It is important to mention that RDA is mainly applied in the vast field of Numerical Ecology (Legendre and Legendre, 1998), where it is studied how particular environmental/abiotic variables influence over some species or biotic variables. Also, such analysis has great versatility as Israëls (1992) presents how to perform it for different types of variables: quantitative, binary and qualitative.

On the other hand, it has been shown that artificial intelligence techniques are strongly related to statistical models; they might be equivalent in some cases. In fact, Cheng and Titterington (1994), state that many ideas and familiar actions for statisticians can be expressed in the notation of Neural Networks. In this order, Sarle (1994) establishes that RDA is equivalent to a linear NN with a fully-connected hidden layer. However, the author does not include in his article the demonstration of his statement. Thus, it is of interest to establish a base that shows the reason for such parallelism and generate a powerful alternative capable of performing this technique of multivariate data analysis. In the current era of information we are living in, the tool presented in this work provides a substitute for the machine learning community by enabling to perform an asymmetric rank reduction over two data sets, without the need to study the complex mathematical concepts that underlie standard RDA. [1]

[1] This article states the case in which RDA seeks to maximize $\boldsymbol{R_y}$ (i.e. when the variables x's explain the y's). Otherwise, an analogous procedure to the one described here must be carried out. Moreover, and without loss of generality, $min(p,q)=q$ is taken regarding the number of redundant components.

Muller (1981) presents a method to perform RDA by means of a model similar to multivariate multiple regression. In addition, Kroonenberg and van der Kloot (1984) suggest that Muller's proposal is based on finding the estimator $\widehat{\boldsymbol{Y}}$ by means of the least squares. Then, they demonstrate that this procedure is equivalent to maximize the RI. Nonetheless, this equivalence depends fundamentally on both the orthonormality condition of the $\boldsymbol{W}$ matrix formed by the q redundant components and on the estimate of $\boldsymbol{B}$. Thus, in order to show the equivalence and find the matrix of interest using a NN, we propose to reform Sarle's model (1994) following the steps of Kroonenberg and van der Kloot (1984) with respect to the estimator $\widehat{\boldsymbol{B}}$ (Eq. 3).

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{W}^t\boldsymbol{X}^t\boldsymbol{Y}) \tag{3}$$

We formulate a cost function that not only takes into account the least squares as proposed by Kroonenberg and van der Kloot (1984), but also the orthonormality of $\boldsymbol{W}$. The cost function for each of the n individuals is shown in Eq. 4.

$$E_l = ||\boldsymbol{y}_l - \widehat{\boldsymbol{y}_l}|| + ||\boldsymbol{W}_l{}^t\boldsymbol{W}_l - \boldsymbol{I}||_F^2 \tag{4}$$

For the proposed architecture to model RDA, unlike the NNs with their usual supervised training, the data set cannot be divided into training and testing sets because the model seeks to obtain as much information as possible from the sample studied, as done by RDA. Therefore, overestimation is necessary and the entire sample represents the training set for the network.

However, when carrying out the development of the proposed NN, we observed that, although the RI was maximized for each execution, the vectors that make up the $\boldsymbol{W}$ matrix were not unique, clearing the notion presented by Kroonenberg and van der Kloot (1984) that maximizing the RI by extracting all of the redundant components at once is equivalent to performing RDA. Although this analysis seeks to find the vectors that maximize the RI, such optimization is the consequence of adding the individual maximizations of the correlations between each redundant variable $\boldsymbol{X}\boldsymbol{w}_i$ with the set of variables $\boldsymbol{Y}$. This restriction is not intrinsic to the cost function established since it maximizes the sum, but not each of the addends.

This is how we propose a method for the progressive search of the vectors $\boldsymbol{w}_i$, which make up the weight matrix $\boldsymbol{W}$. The method consists of finding the first vector that minimizes the cost function, thus, maximizing the correlation inherent to it. The following vectors also intent to minimize the cost function, but with the particularity that they must be orthogonal to all those already found, while maintaining their unit norm. The cost function for the i-th redundant component on the l-th individual is given by Eq. 5.

$$E_{il} = ||\boldsymbol{y}_l - \widehat{\boldsymbol{y}}_{il}|| + \xi_N|\boldsymbol{w}_{il}{}^t\boldsymbol{w}_{il} - 1| + \xi_O \sum_{1\leq j<i} \boldsymbol{w}_{il}{}^t\boldsymbol{w}_{jl} \tag{5}$$

In this cost function, ξ_N and ξ_O are coefficients that penalize respectively the non-normality and the non-orthogonality, since it is necessary to give greater importance to these restrictions, than to maximazig the correlation, in order to obtain the expected results.

With this modification to the cost function, we have the supervised learning for the i-th vector on the l-th individual given by Eq. 6 and Eq. 7. [2]

$$\boldsymbol{w}_{il+1} = \boldsymbol{w}_{il} + \Delta\boldsymbol{w}_{il} \tag{6}$$

$$\Delta\boldsymbol{w}_{il} = -\eta\frac{\partial E_{il}}{\partial \boldsymbol{w}_i} \tag{7}$$

The architecture suggested by Sarle (1994) is equivalent to the one that emerges from the results of Kroonenberg and van der Kloot (1984), maximizing the RI but not the partial correlations. With this research, we determine that in order to model RDA through a NN, modifications of great importance to the original proposal must be made.

[2]We emphazise that we implement the learning process updating the weight matrix after each of the individuals in oder to give more emphasis to the orthonormality of $\boldsymbol{W}$.

References

Cheng, B. & Titterington, D. (1994) Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, **9**(1):2–54.

D'Ambra, L. & Lauro, N. (1984) L'Analyse non symetrique des Correspondences. *Elsevier*, 433–446.

Hotelling, H. (1936) Relations Between Two Sets of Variates. *Biometrika*, **28**(3/4):321–377.

Israëls, A.Z. (1968) Redundancy Analysis for Various Types of Variables. *Statistica Applicata*, **4**(4):531–542.

Izenman, A. (1975) Reduced-Rank Regression for the Multivariate Linear Model. *Journal of Multivariate Analysis*, (5):248–264.

Kroonenberg, P. & van der Kloot, W. (1984) External three-mode Principal Component Analysis and Redundancy. Leiden.

Legendre, P. & Legendre, L. (1998) Numerical Ecology. *Elsevier*, 579–595.

Muller, K. (1981) Relationships between Redundancy Analysis, Canonical Correlation, and Multivariate Regression. *Phychometrika*, **46**:139–142.

Rao, C. (1964) The use and interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics*, **26**(4):329–358.

Sarle, W. (1994) Neural Networks and Statistical Models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. New York.

Stewart, S. & Love, W. (1968) A general canonical correlation index. *Psychological Bulletin*, **70**:160–163.

van den Wollenberg, A. (1977) Redundancy analysis, an alternative for canonical correlation analysis. *Psychometrika*, **42**(2):207–219.