Computational Statistics and Data Analysis 70 (2014) 88–100

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fast regularized canonical correlation analysis

Raul Cruz-Cano *, Mei-Ling Ting Lee

Department of Epidemiology and Biostatistics, University of Maryland, College Park, United States

ARTICLE INFO

Article history:
Received 4 July 2012
Received in revised form 20 September 2013
Accepted 23 September 2013
Available online 30 September 2013

Keywords:
Canonical correlation analysis
Regularization
Regularized canonical correlation
Shrinkage of covariance matrix
NCI-60 microRNA expression data

ABSTRACT

Canonical correlation analysis is a popular statistical method for the study of the correlations between two sets of variables. Finding the canonical correlations between these datasets requires the inversion of their corresponding sample correlation matrices. When the number of variables is large compared to the number of experimental units it is impossible to calculate the inverse of these matrices directly and therefore it is necessary to add a multiple of the identity matrix to them. This procedure is known as regularization. In this paper we present an alternative method to the existing regularization algorithm. The proposed method is based on the estimates of the correlation matrices which minimize the mean squared error risk function. The solution of this optimization problem can be found analytically and consists of a small set of computationally inexpensive equations. We also present material which shows that the proposed method is more stable and provides more accurate results than the standard regularized canonical correlation method. Finally, the application of our original method to NCI-60 microRNA cancer data proves that it can deliver useful insights in study cases which involve hundreds of variables.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Background

Canonical correlation analysis (CCA) is a popular exploratory statistical method which allows the analysis of the relationships that exist between two sets of variables. CCA has several advantages over other statistical techniques, e.g. its capability to limit the probability of committing Type I errors when more than two variables are being examined (Hair et al., 2009). Moreover, the complexity of many research studies involving biological data cannot be appropriately modeled if the variables are examined separately. On these occasions CCA reflects the reality of the research studies in a manner consistent with the reality of the problem. For these reasons over the years CCA has been applied to extremely diverse fields that range from ecological studies (Gittins, 1985) to human geography (Clark, 1975). In the last few years the capability of CCA has been challenged by new datasets that have appeared in fields such a Bioinformatics and Biostatistics. The problem is that finding the canonical correlations requires the inversion of sample correlation matrices. When the number of variables is large compared to the number of experimental units, as it is the case in many Bioinformatics problems, then it is impossible to calculate the inverse of these matrices directly. The most popular solution to this problem is a procedure called regularization. Regularization has several important disadvantages such as being computationally expensive and it can lead to different results depending on the initial values provided by the user. In this paper we present an alternative method to the existing regularization algorithm. The proposed method is based on the optimal estimates of the correlation matrices. The equations required to calculate these optimal estimates are computationally inexpensive and have analytical solutions. These characteristics allow the user to obtain consistent and accurate results in a fast manner.

* Corresponding author. Tel.: +1 301 405 0560.
E-mail address: raulcruz@umd.edu (R. Cruz-Cano).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.020

In the rest of Section 1, the theory behind the classical and regularized versions of CCA is presented. Section 2 is dedicated to describing in detail the original algorithm proposed in this paper (Section 2.1.), along with the significant contributions of our work, e.g. the calculation of the optimal shrinkage coefficients for the off-diagonal matrices (Section 2.2) and the formulas needed when the number of experimental units is small (Section 2.4). The advantages of the proposed algorithm over the regularized version of CCA are discussed in Section 2.3. Section 2 concludes with the presentation of an algorithm which, although not part of our original contributions, might be useful for researchers who desire to extend CCA to more than two sets of variables. In Section 3 we present several simulation studies which show that the proposed method is more stable and provide more accurate results than the regularized canonical correlation method. Finally, the application of our original method to NCI-60 microRNA cancer data proves that it can deliver useful insights in study cases which involve hundreds of variables.

1.2. Classical canonical correlation analysis (CCA)

Canonical correlation analysis is a multivariate data analysis tool that studies the extent and nature of the correlation between two sets of variables, $X = X_1, X_2, \ldots, X_p$ and $Y = Y_1, Y_2, \ldots, Y_q$ (Hotelling, 1936). As in Clark (1975), without loss of generality it is assumed that $p \geq q$ and that each and every variable in X and Y has been standardized to have mean zero and variance equal to one. The main purpose of the CCA is the exploration of sample correlations between two sets of variables observed on the same experimental units by analyzing the coefficients $A_1 = (a_1, a_2, \ldots, a_p)^T$ and $B_1 = (b_1, b_2, \ldots, b_q)^T$ which maximize the correlation between linear combinations of the variables X and Y while being subject to having the variances $V[XA_1] = V[YB_1] = 1$. These linear combinations $U_1 = XA_1$ and $V_1 = YB_1$ are called the first canonical variates. Their correlation ρ_1 is the first canonical correlation and the coefficients A_1 and B_1 are the first canonical weights or first canonical coefficients.

The process can be duplicated looking now for the weights A_2 and B_2 which maximize the correlation ρ_2 subject to $V[XA_2] = V[YB_2] = 1$ and to the additional constraint that the new canonical variates $U_2 = XA_2$ and $V_2 = YB_2$ are uncorrelated with the first pair of canonical variates. Actually, this process can be repeated q times.

For the sample cross-correlation matrix

$$\mathbf{S} = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix} \tag{1}$$

the canonical correlations can be calculated by finding the descending-ordered square roots of the eigenvalues of the matrix:

$$M = S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}. \tag{2}$$

The canonical weights for Y are the corresponding q eigenvectors, i.e. the elements of eigenvector i are the canonical weights B_i. The vector of canonical weights for X corresponding to the ith canonical correlation is:

$$A_i = (S_{XX}^{-1} S_{XY} B_i)/\varepsilon_i, \quad 1 \leq i \leq q \tag{3}$$

where ε_i is the ith eigenvalue of the matrix in Eq. (2) and B_i is the corresponding eigenvector. In this manuscript the significance of the canonical correlations is calculated using Wilk's Lambda. The R package *CCP* provides a set of functions which, given a set of canonical correlations, provide the appropriate p-values (Menzel, 2011).

The correlations between the canonical variates and the original variables are called canonical factor loadings and can also be used to analyze the intra and inter-set relationships. They offer some advantages over the canonical weights such as: a smaller standard error and greater stability in replicate samples (Meredith, 1964). It can be proved that the formula for the ith interset canonical factor loadings of X is $S_{YX}A_i$ while for Y is $S_{XY}B_i$. The calculation of the intraset canonical factor loadings is not part of this research. The set of all the values mentioned so far in this sub-section (canonical correlations, p-values, canonical weights and canonical factor loadings) is known as the canonical structure.

The variables in X and Y can be drawn in a two-dimensional plane in which the axes are a couple of selected canonical factor loadings. In this space, variables with a strong positive relation are projected in the same direction from the origin, while if it is a negative correlation they are placed on the opposite side of the origin. The variables with the strongest relationships will lie further away from the origin; hence this measure (distance from the origin) is commonly applied to determine which variables are the most relevant. In this research only the variables which lie outside a .5 radius circle are labeled.

For example, assume that variables $X = \{X_1, X_2, X_3\}$ and $Y = \{Y_1, Y_2, Y_3\}$ are measured for 40 individuals and the results of applying CCA are drawn in Fig. 1. The X variables are represented by gray triangles while the Y variables are seen as black circles. In this case Y_1 and Y_3 have a very positive relationship among themselves and opposite to X_1 and Y_2. X_2 is also negatively correlated with X_1 and even more so with Y_2 which lies further away of the inner circle. X_1 and Y_2 are positively associated with each other, although not as significantly as Y_1 and Y_3. The same can be said about X_2 and X_3.

Another common analysis tool is the canonical variates units plot. In this plot the experimental units are drawn in a plane in which the canonical variates U_i and V_i are the axes. This type of graph allows seeing the strength of the canonical correlation i and whether groups of units are clustered due to a common characteristic. In Fig. 1(b) the clustered position of the experimental units 1 to 10 indicate that they have common characteristics compared to the cluster composed of units 11 to 20.

Download English Version:

https://daneshyari.com/en/article/415470

Download Persian Version:

https://daneshyari.com/article/415470

Daneshyari.com