Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A Short Demo Article: Regression Models for Count Data in **R**

Juan-Ramón González
ISGlobal

Dolors Pelegrí
ISGlobal

Isaac Subirana
CIBERESP

Abstract

.. to be completed ..

Keywords: JSS, style guide, comma-separated, not capitalized, R.

1. Introduction

In many biomedical studies, from genetics, clinical or to epidemiological ones, data are collected from recruited individuals, frequenlty having many variables of different types; they can be demographic, such as age, sex, race, etc., morphologic such as weight, height, waist, etc., lipidic such as cholesterol, HDL, LDL, tryglicerides, etc, or obtained from questionaries of nutrition, quality of live, physical activity, etc.

It may be interesting to integrate these sets of data in the analyses. There exists many statistical techniques to do so, from a standard Canonical Correlation Analyses (CCA) to relate two data sets, or Generalized Correlation Canonical Analyses (GCCA) which expands to more than two sets Tenenhaus and Tenenhaus (2011). In parallel, there are some tools to deal with sparse data, when many variables are involved and it is thought that most of them are not associated, see Tenenhaus, Tenenhaus, and Groenen (2017), or when variables are not normally distributed (Argelaguet, Arnol, and Bredikhin (2020)). This topics will not be covered either discussed in this paper.

The goal of CCA is find two canonical variables computed as a linear combination of variables of each data set with the highest correlation. When having more than two data sets, it can be computed several pairwise correlations between canonical variables. Therefore, there exists different criteria to obtain them, Tenenhaus *et al.* (2017). Here we focuse on the criteria that minimizes the distance between latent variables and canonical variables of each data set. This criteria has the advantage of providing common coordinates to represent the individuals as in a Principal Component Analyses (PCA). Also, weights for each variable of each data set is obtained as in ordinary CCA, to investigate which variables are most important in the analyses.

2. Methods

The method described in this paper was first introduced and formulated in Velden and Bijmolt (2006). It is an extension of canonical correlation analyses for two or more data sets (Generalized Canonical Correlation Analyses -GCCA-).

In this paper we want to focus on the strategy proposed in Velden and Bijmolt (2006) to deal with missing data in a whole row. They derive a closed form to compute the results inserting a dummy diagonal squared matrix indicating which individuals were missing for each data set. This non-iterative procedure speeds and simplify the calculus procedure.

Results are detailed in Velden and Bijmolt (2006). Shortly speaking, the method tries to find latent variables which minimizes the mean square error between them and a linear combination of each set of variables. This latent variables are orthogonal by construction and can be used to represent the individuals in a two dimensional space in order to distinguish underlying groups or outliers. It must be noted that the coordinates of the latent variables are obtained for all individuals, even if they have data in all sets (X_j) or not.

In order to introduce some elements that are discussed along this paper, the goal function of the method is

$$\min_{Y,B_j} \phi = \text{trace} \sum_{j=1}^{J} \left(Y - X_j B_j\right)^t K_j \left(Y - X_j B_j\right) \tag{1}$$

constrained to $Y^t K Y = \sqrt{J} I_L$, where I_L is the L by L identity matrix.

The elements in equation (1) are:

- X_j, $j = 1, \ldots, J$: n by p_j matrix representing the j-th set of observed variables. Note that all data sets, X_j, have the same number of rows, n, which represents the whole sample, i.e. individuals that are present in at least one data set. If a particular individual does not have data in X_j, his/her row is filled by an arbitrary value, for instance zero.

- K_j: n-square matrix with zeros outside the diagonal, and one in the i row and i column if the i-th individual is not missing for j-th data set, and zero otherwise.

- $K = \sum_{j=1}^{J} K_j$

- B_j: p_j by L matrix containing the coefficients of each variable for each data set.

- Y: n by L matrix with latent variables in columns.

The solution of equation (1) is obtained by computing the eigen values and eigen vectors of of the expression

$$K^{-\frac{1}{2}} \left(\sum_{j=1}^{J} K_j X_j \left(X_j^t K_j X_j \right)^{-1} X_j^t K_j \right) K^{-\frac{1}{2}} Y^* = Y^* \Lambda \tag{2}$$

where Λ is a diagonal matrix containing the eigen values λ_l, $l = 1, \ldots, L$, and Y^* is a n by L

orthonormal matrix with eigen vectors in columns. Finally, the latent variables matrix, Y, is obtained by

$$Y = \sqrt{J} K^{-\frac{1}{2}} Y^* \tag{3}$$

Note that when there are more columns than rows, $\left(X_j^t K_j X_j \right)$ in equation (2) becomes singular and the general inverse must be used such as Moore Ponrose-inverse.

2.1. Other common strategies to deal with missing rows/individuals

Other commonly used non-iterative approaches for dealing with missig rows are:

- **"IMPUTE"**: It consists of filling empty rows by the variable means from the rest of individuals for whom the data is available. This imputation is done variable by variable. This approach has the advantage of including all individuals in tha analyses regardless if they have information in all data sets or are missing in some of them. On contrary, it has the problem of not taking into account uncertainty since it assigns the same value to all missing individuals.
- **"COMPLETE"**: With this strategy, only individuals with available information in all data sets are included in the analyses. Unlike the "IMPUTE" strategy, it does not assign any value but sample size may be substancially reduced.

@@@ discarted strategies (multiple imputation, ¿¿single imputation taking into account other variables??)

3. Real data example

... to be completed ...

4. Simulation studies

4.1. Simulation methods

The proposed method (MGCCA) was validated and compared to the other common methods (IMPUTE and COMPLETE), in two simulation studies. The first one assesses how similar are the estimated scores of latent variables, Y, of each method compared to what would be obtained if all individuals were available ("full data"). Mean square distance, "MSD" is computed as follows as the mean of euclidan distances of each individual represented in the Y coordinates obtained with full data and for each method. In the second simulation study, data were simulated distinguishing two groups of individuals with different means, and each method is evaluated in terms of power to detect differences between the groups.

Simulated data were generated similarly to Velden and Bijmolt (2006). Detailed steps are listed below.

Simulation study I

- **Step 1:** generate a n by 2 matrix, Y, from a standardized normal distribution, which corresponds to the two latent variables.
- **Step 2:** generate two matrices, B_1 and B_2 with dimensions 2 by p_1 and 2 by p_2 of coefficients ranging from -1 to 1 under a uniform distribution.
- **Step 3:** Compute X_1 and X_2, of dimensions n by p_1 and n by p_2, respectively, post-multiplying Y by coefficient matrices B_1 and B_2.
- **Step 4:** Add noise to X_1 and X_2 by adding a gnerated normal value of zero mean and σ_2 standard deviation. At this point full data is obtained.
- **Step 5:** Randomly select a proportion of rows for X_1 and X_2 (not the same rows) to be declared as missing individuals.

Data were generated under the following scenarios:

- Fixing the number of individuals, n, to 500.
- Varying the number of variables to 50 and 100. In all scenarios, it has been considered the same number of variables for both data sets, X_1 and X_2, i.e. $p = q$.
- Varying the noise, σ, to 0.125 and 0.250.
- Varying the proportion of missing individuals, p, to 0.1, 0.2 and 0.3.

A total of 12 escenarios were simulated, and for each simulated scenario, 100 data sets were generated.

Generalized Canonical Correlation Analyses (GCCA) was performed for each generated data and two canonical latent variables were estimate ($\hat{Y}_1$, $\hat{Y}_2$), using full data ("FULL"). Then, the three methods (MGCCA, IMPUTE, COMPLETE) were applied to the data with missing rows. Finally, the Mean Square Distance (MSD) is computed as follows:

$$\text{MSD} = \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{y}_{\text{FULL}}[i1] - \hat{y}_{\text{METHOD}}[i1])^2 + (\hat{y}_{\text{FULL}}[i2] - \hat{y}_{\text{METHOD}}[i2])^2\right]$$

where $\hat{y}_{\text{METHOD}}[ij], \quad j = 1, 2$ are the latent variables coordinates obtained with each method (MGCCA, IMPUTE or COMPLETE) for the i individual, and $\hat{y}_{\text{FULL}}[ij], \quad j = 1, 2$ are the latent variables coordinates obtained with full data.

Note that when computing the MSD for COMPLETE method, n is the number of individuals with complete data, so rows of $\hat{y}_{\text{FULL}}$ and $\hat{y}_{\text{COMPLETE}}$ corresponds to these individuals.

Simulation study II

Another generation data process similar to the one described above was performed but now two groups are distinguished, and methods are assessed in terms of discriminate these groups.

All steps are the same except the step 1, where first $\frac{n}{2}$ rows (individuals) of Y matrix are generated under a normal distribution with mean equals to $\frac{\delta}{2}$ and the rest of the rows under a normal distribution with mean equals to $-\frac{\delta}{2}$.

Once the data were generated, a MANOVA anlyses was performed to test differences in means of estimated canonical variables scores among the two groups.

In these simulation study, number of generated individuals and variables were fixed to $n = 500$, and $p = q = 50$, respectively. Noise standard deviation was fixed to $\sigma = 0.2$. While the varying parameters where the difference in means of groups $\delta = \{0, 0.25, 0.5\}$ and proportion of missing individuals $p = \{0.1, 0.2, 0.3\}$

Finally, 500 data sets were generated and p-values were computed for each of them. Power was computed as the proportion of times the p-value was lower than significance level that was set to 5%.

4.2. Simulation results

Simulation study I

From simulation study I, the method with best performance was IMPUTE method (see Figure 1) under all scenarios, since it provides the lowest MSD and therefore the estimated latent variables coordinates was more similar to the ones that would be obtained if all data was available (no missing individuals). While the worst method by far was the COMPLETE one, which analyse only complete cases, i.e. indivivuals with information in all data sets. Results are similar regardless the number of variables (rows) and noise variable (columns). And the larger of missing individuals (x-axis), the larger MSD, specially for COMPLETE method. On the other side, IMPUTE and MGCCA method is more robust when proportion of missing individuals increases.

Additionally, consistency of results between generated data (replicates) has been described using boxplots within each scenario (Figure 2). It can be seen that MGCCA method is very consistent, i.e. simulated results are very similar between replicates, compared to other two methods. Therefore, while, on average the IMPUTE method provides better results in terms of MSD, for some data results can be much worse than the ones obtained with MGCCA.

Simulation study II

From simulation study II results, it can be seen that MGCCA is the method that provides better power overcoming the other two methods in all scenarios (see Figure 3), specially when proportions of missing individuals is low (0.1) or moderate (0.2). The COMPLETE method is the least powerfull in all scenarios. When proportion of missings is high (0.3) the three methods perform similar in terms of power, but COMPLETE that performs much worse than the other two when difference between groups is high. Finally, when data is generated under no difference between groups ("Difference=0" on x-axis), all three methods estimate the same power to significance level, showing that there is no excess of false posive rate.

5. Summary and discussion

... to be completed ...

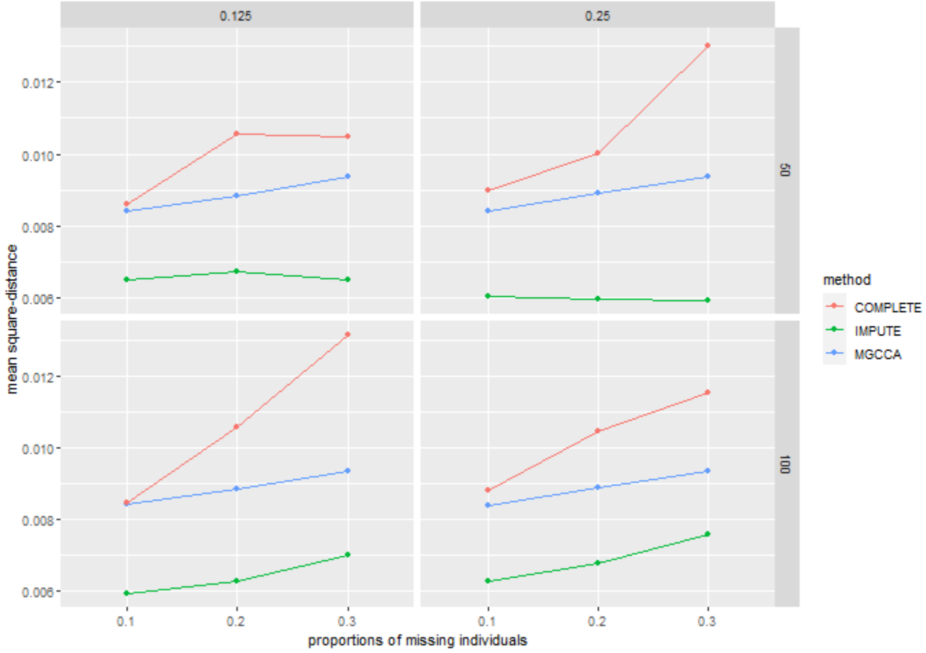

Figure 1: Average of Mean Square Differencs of all 100 replicates for each scenario and method by proportion of missing individuals, stratified by number of variables, $p = q$ (rows) and noise standard deviation σ (columns).

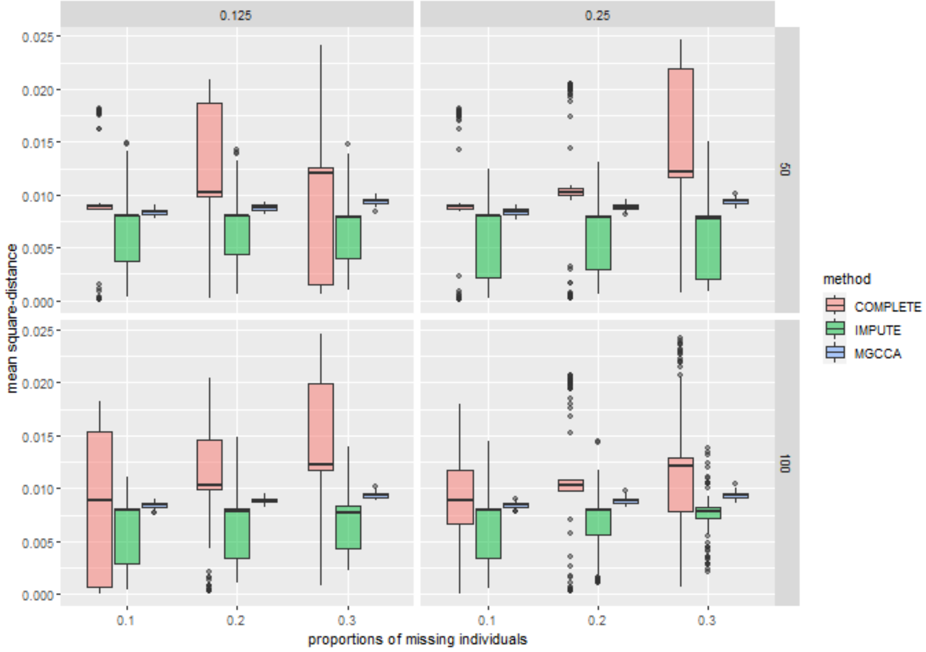

Figure 2: Boxplot of Mean Square Differencs within 100 replicates for each scenario and method by proportion of missing individuals, stratified by number of variables, $p = q$ (rows) and noise standard deviation σ (columns).

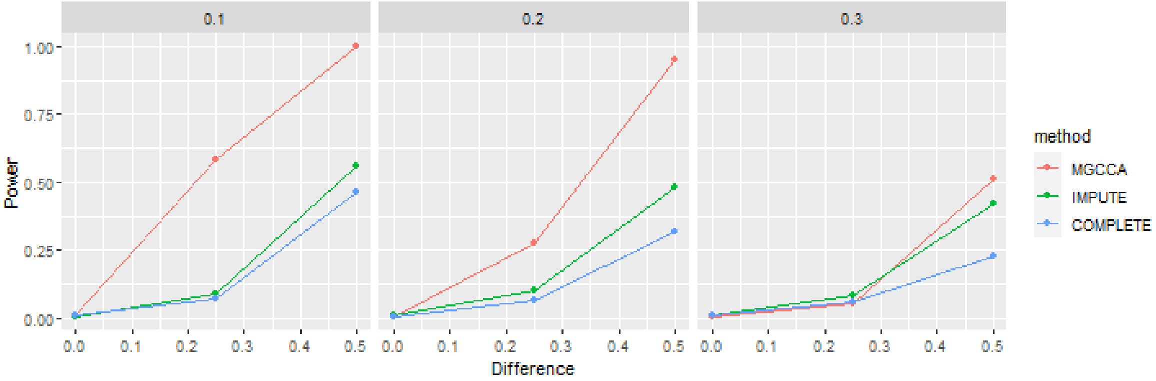

Figure 3: Power for each method depending on the difference between groups, and stratified by proportions of missing individuals.

Computational details

... specify R version and used packages ...

... consumed time for simulation. ? analyses of real data ?

Acknowledgments

References

Argelaguet R, Arnol D, Bredikhin D (2020). "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data." *Genome Biol.*, **21**(1). doi:10.1186/s13059-020-02015-1.

Tenenhaus A, Tenenhaus M (2011). "Regularized Generalized Canonical Correlation Analysis." *Psychometrika*, **76**(2). doi:10.1007/s11336-011-9206-8.

Tenenhaus M, Tenenhaus A, Groenen P (2017). "Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods." *Psychometrika*, **in press**. doi:10.1007/s11336-017-9573-x.

Velden M, Bijmolt T (2006). "Generalized canonical correlation analysis of matrices with missing rows: a simulation study." *Psychometrika*, **71**(2), 323–331. doi:10.1007/s11336-004-1168-9.

A. More technical details

... not sure if necessary to include an appendix...

Affiliation:

Juan-Ramón González
ISGlobal
and
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: `Achim.Zeileis@R-project.org`
URL: `https://eeecon.uibk.ac.at/~zeileis/`

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd