Journal of Statistical Software

January 2008, Volume 23, Issue 12. *http://www.jstatsoft.org/*

CCA: An R Package to Extend Canonical Correlation Analysis

Ignacio González
Université de Toulouse

Sébastien Déjean
Université de Toulouse

Pascal G. P. Martin
Institut National de la Recherche Agronomique

Alain Baccini
Université de Toulouse

Abstract

Canonical correlations analysis (CCA) is an exploratory statistical method to highlight correlations between two data sets acquired on the same experimental units. The `cancor()` function in R (R Development Core Team 2007) performs the core of computations but further work was required to provide the user with additional tools to facilitate the interpretation of the results. We implemented an R package, **CCA**, freely available from the Comprehensive R Archive Network (CRAN, `http://CRAN.R-project.org/`), to develop numerical and graphical outputs and to enable the user to handle missing values. The **CCA** package also includes a regularized version of CCA to deal with data sets with more variables than units. Illustrations are given through the analysis of a data set coming from a nutrigenomic study in the mouse.

Keywords: canonical correlations, regularization, cross-validation.

1. Introduction

Canonical correlation analysis (CCA) is a multidimensional exploratory statistical method in the same vein as Principal Components Analysis (PCA): both methods lie on the same mathematical background (matrix algebra and eigen analysis) and results can be illustrated through similar graphical representations.

The main purpose of CCA is the exploration of sample correlations between two sets of quantitative variables observed on the same experimental units, whereas PCA deals with one data set in order to reduce dimensionality through linear combination of initial variables.

When performing CCA, mathematical arguments compel data to have more units than variables in each set. In practice, the number of units should be greater that the total amount of variables in both sets what is not always possible. In particular, in the context of high throughput biology, a set of variables may be one set of genes whose expression has been measured by means of microarray technology (see for example Muller and Nicolau 2004) on few experimental units. In this case, the number of genes can easily reach several hundreds or thousands whereas the number of units to be monitored cannot be so large. Note that similar situations can be found in other fields, for instance chemistry with recent developments in spectroscopy and chemometrics (Mullen and van Stokkum 2007). When other variables are acquired on these same units in order to highlight correlations, classical CCA cannot be performed. One solution consists in including a regularization step in the data processing (Bickel and Li 2006) to perform a regularized canonical correlation analysis (RCCA).

Another well known method can deal with the same kind of data: Partial Least Squares (PLS) regression (Mevik and Wehrens 2007). However, the object of PLS regression is to explain one or several response variables in one set, by way of variables in the other one. On the other hand, the object of CCA is to explore correlations between two sets of variables whose roles in the analysis are strictly symmetric. As a consequence, mathematical principles of both methods are fairly different.

Statistical softwares usually propose CCA computations (`cancor()` in R, `proc cancorr` in SAS, ...). We noticed that `cancor()` outputs are very limited and needed to be completed to be used in an efficient way. We developed an R package with three purposes: 1) to complete numerical as well as graphical outputs, 2) to enable the handling of missing values and 3) to implement RCCA.

Mathematical background of CCA and RCCA is presented in the second section: computations and graphical representations are described in both cases. In the third section, **CCA** is used on one data set available in the package.

2. Canonical correlation analysis

2.1. Notation

Let us consider two matrices X and Y of order $n \times p$ and $n \times q$ respectively. The columns of X and Y correspond to variables and the rows correspond to experimental units. The j^{th} column of the matrix X is denoted by X^j, likewise the k^{th} column of Y is denoted by Y^k. Without loss of generality it will be assumed that the columns of X and Y are standardized (mean 0 and variance 1). Furthermore, it is assumed that $p \leq q$ (in other words, the group which contains the least variables is denoted by X). We denote by S_{XX} and S_{YY} the sample covariance matrices for variable sets X and Y respectively, and by $S_{XY} = S_{YX}^\top$ the sample cross-covariance matrix between X and Y. The notation $A^\top$ means the transpose of a vector or a matrix A.

2.2. Principle

Classical CCA assumes first $p \leq n$ and $q \leq n$, then that matrices X and Y are of full column rank p and q respectively. In the following, the principle of CCA is presented as a problem

solved through an iterative algorithm.

The first stage of CCA consists in finding two vectors $a^1 = (a_1^1, \ldots, a_p^1)^\top$ and $b^1 = (b_1^1, \ldots, b_q^1)^\top$ that maximize the correlation between the linear combinations

$$U^1 = Xa^1 = a_1^1 X^1 + a_2^1 X^2 + \cdots + a_p^1 X^p$$

and

$$V^1 = Yb^1 = b_1^1 Y^1 + b_2^1 Y^2 + \cdots + b_q^1 Y^q \,,$$

assuming that vectors a^1 and b^1 are normalized so that

$$\mathrm{var}(U^1) = \mathrm{var}(V^1) = 1 \,.$$

In other words, the problem consists in solving

$$\rho_1 = \mathrm{cor}(U^1, V^1) = \max_{a,b} \mathrm{cor}(Xa, Yb) \,,$$

subject to the constraint

$$\mathrm{var}(Xa) = \mathrm{var}(Yb) = 1 \,.$$

The resulting variables U^1 and V^1 are called the first canonical variates and ρ_1 is referred as the first canonical correlation.

Higher order canonical variates and canonical correlations can be found as a stepwise problem. For $s = 1, \ldots, p$, we can successively find positive correlations $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_p$ with corresponding vectors $(a^1, b^1), \ldots, (a^p, b^p)$, by maximizing

$$\rho_s = \mathrm{cor}(U^s, V^s) = \max_{a^s, b^s} \mathrm{cor}(Xa^s, Yb^s) \quad \text{subject to} \quad \mathrm{var}(Xa^s) = \mathrm{var}(Yb^s) = 1 \,,$$

under the additional restriction

$$\mathrm{cor}(U^s, U^t) = \mathrm{cor}(V^s, V^t) = 0 \quad \text{for} \quad 1 \leq t < s \leq p \,.$$

2.3. Mathematical aspects

From a geometrical point of view, let us define

$$P_X = X(X^\top X)^{-1} X^\top = \frac{1}{n} X S_{XX}^{-1} X^\top \quad \text{and} \quad P_Y = Y(Y^\top Y)^{-1} Y^\top = \frac{1}{n} Y S_{YY}^{-1} Y^\top$$

the orthogonal projectors onto the linear spans of the columns of X and Y respectively.

It is well known (Mardia *et al.* 1979) that:

Proposition 2.1.

- *canonical correlations* ρ_s *are the positive square roots of the eigenvalues* λ_s *of* $P_X P_Y$ *(which are the same as* $P_Y P_X$*):* $\rho_s = \sqrt{\lambda_s}$*;*
- *vectors* $U^1, \ldots, U^p$ *are the standardized eigenvectors corresponding to the decreasing eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_p$ *of* $P_X P_Y$*;*

- *vectors* $V^1, \ldots, V^p$ *are the standardized eigenvectors corresponding to the same decreasing eigenvalues of* $P_Y P_X$.

2.4. Regularized CCA

CCA cannot be performed when the number of experimental units is less than the greatest amount of variables in both data set ($n \leq \max(p, q)$). Actually, when the number of variables increases, greatest canonical correlations are nearly 1 because of recovering of canonical subspaces that do not provide any meaningful information. Therefore, a standard condition usually advocated for CCA (Eaton and Perlman 1973) is $n \geq p + q + 1$.

Furthermore, when variables $X^1, \ldots, X^p$ and/or $Y^1, \ldots, Y^q$ are highly correlated, *i.e.* nearly collinear, matrices S_{XX} and/or S_{YY} respectively, tend to be ill-conditioned and their inverses unreliable.

One way to deal with this problem consists in including a regularization step in the calculations. Such a regularization in the context of CCA was first proposed by Vinod (1976), then developed by Leurgans *et al.* (1993). A recent reference deal with the application of the regularization to linear discriminant analysis (Guo *et al.* 2007). A survey about regularization in statistics is proposed in Bickel and Li (2006).

In this framework, S_{XX} and S_{YY} are replaced respectively by $\Sigma_{XX}(\lambda_1)$ and $\Sigma_{YY}(\lambda_2)$ defined by

$$\Sigma_{XX}(\lambda_1) = S_{XX} + \lambda_1 I_p \quad \text{and} \quad \Sigma_{YY}(\lambda_2) = S_{YY} + \lambda_2 I_q.$$

This method rises a new problem: how to set "good" values for the regularization parameters? This problem is addressed in the next section through a rather standard cross-validation procedure.

2.5. Cross-validation for tuning regularization parameters

Let us denote $\lambda = (\lambda_1, \lambda_2)$. For a given value of λ, denote by ρ_λ^{-i} the first canonical correlation of CCA computed from the units with rows X_i and Y_i removed. Let $a_\lambda^{(-i)}$ and $b_\lambda^{(-i)}$ be the corresponding vectors defining the first canonical variates. We do this for $i = 1, \ldots, n$ and obtain n pairs of vectors $\left(a_\lambda^{(-1)}, b_\lambda^{(-1)}\right), \ldots, \left(a_\lambda^{(-n)}, b_\lambda^{(-n)}\right)$.

The leave-one-out cross validation score for $\lambda = (\lambda_1, \lambda_2)$ is then defined by Leurgans *et al.* (1993):

$$CV(\lambda_1, \lambda_2) = \text{cor}\left(\left\{X_i a_\lambda^{(-i)}\right\}_{i=1}^n, \left\{Y_i b_\lambda^{(-i)}\right\}_{i=1}^n\right).$$

Then we choose the value of λ_1 and λ_2 that maximizes this correlation:

$$\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2) = \arg\max_{\lambda_1, \lambda_2} CV(\lambda_1, \lambda_2).$$

Note that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are chosen with respect to the first canonical variates and are then fixed for higher order canonical variates.

There are two tuning parameters in the regularized CCA, so the cross-validation is performed on a two-dimensional surface. Directly searching for a maximum on the two-dimensional parameter surface ensure to obtain the optimal value for λ, but may require intensive computing. An alternative consists in building a relatively small grid of "reasonable" values for λ_1 and λ_2, to evaluate the cross validation score at each point of the grid, and then to choose the value of $\lambda = (\lambda_1, \lambda_2)$ that gives the largest CV-score (Friedman 1989; Guo *et al.* 2007).

2.6. Graphical representations

As in PCA, two kinds of graphical representations can be displayed to visualize and interpret the results of CCA: scatter plots for the initial variables X^j and Y^k and scatter plots for the experimental units. If d $(1 \leq d \leq p)$ is the selected dimension for results of CCA, graphical representations can be drawn for every pair (s, t) of axes such that $1 \leq s < t \leq d$. For a given pair (s, t), variables plots and units plots can be considered with respect either to U^s and U^t or to V^s and V^t. If the canonical correlations are close to one, then the graphical representations on the axes defined by (U^s, U^t) and (V^s, V^t) are similar.

Choosing the dimension

Like in PCA, it is advocated to choose a small value for dimension d $(1 \leq d \leq p)$. In practice, this value is very often 2, 3 or 4. Note that small canonical correlations are not relevant: they do not express linear relationships between columns of X and Y and can be neglected.

For great values of p, we suggest an empirical approach for choosing the dimension based on the joint examination of two graphical representations: the scree graph of canonical correlations and the scatter plots of variables. The scree graph is the plot of canonical correlations versus the dimension; a clear gap between two successive values suggest to select for d the rank of the greatest one. On the other hand, we consider the scatter plot of variables according to axes (U^s, U^{s+1}) for the first values of s and we neglect axes such that almost all points representing either X-variables or Y-variables are within the circle of radius 0.5 (that is correlations between variables X^j or Y^k and canonical variates U^s or V^s are less than 0.5).

Representations of the variables

The variables plot is of interest because it allows to discern the structure of correlation between the two sets of variables X and Y. Coordinates of variables X^j and Y^k on the axis defined by U^s are Pearson correlations between these variables and U^s. As variables X^j and Y^k are assumed to be of unit variance, their projections on the plane defined by the axes (U^s, U^t) are inside a circle of radius 1 centered at the origin, called the correlation circle. On this graphic two circumferences are plotted corresponding to the radius 0.5 and 1 to reveal the most salient patterns in the ring defined between these two circumferences. Variables with a strong relation are projected in the same direction from the origin. The greater the distance from the origin the stronger the relation. The principle and interpretation are similar as the "correlation loadings" plot provided by the **pls** package in the context of PLS regression (Mevik and Wehrens 2007).

Representations of the units

The representation of the units can be useful to clarify the interpretation of the correlation

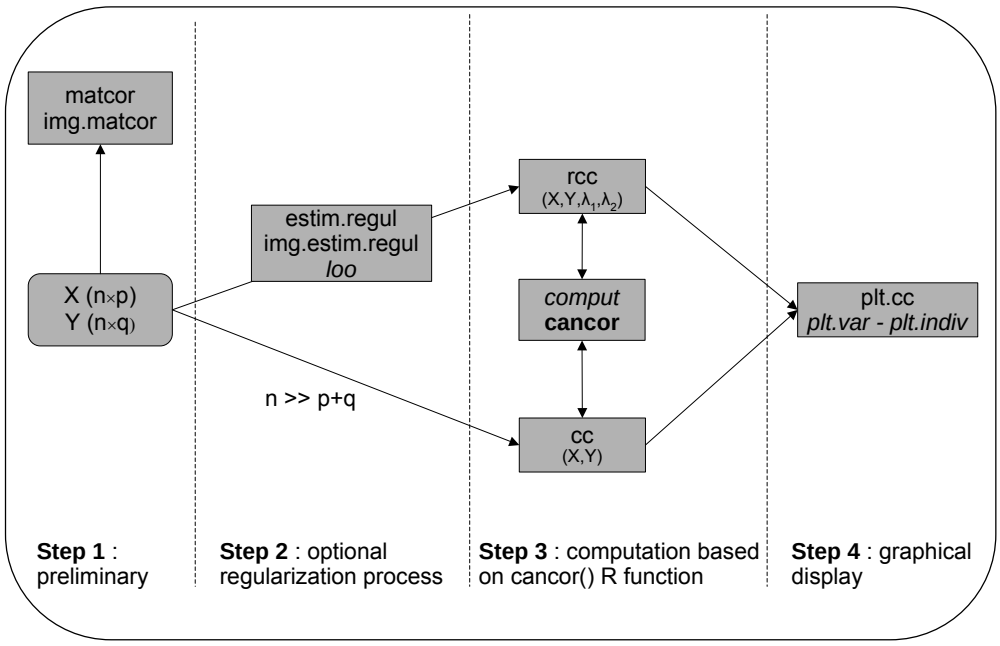

Figure 1: Schematic view of the canonical correlation analysis process using **CCA**. Functions in italic font are internal functions the user does not have to call. In bold face, the `cancor()` R function around which the package is built.

between variables. This representation of units is possible by using the axes defined by (U^s, U^t): the coordinate of the i^{th} unit on the axis U^s is U_i^s (the i^{th} coordinate of the s^{th} canonical variate).

The relationships between the two plots (variables and units) drawn on the matching axes can reveal associations between variables and units.

3. Using CCA

CCA is freely available from the Comprehensive R Archive Network (CRAN, `http://CRAN.R-project.org/`). Once loaded into R, **CCA** provides the user with functions to perform CCA and one data set for illustration purpose.

3.1. Implementation issues

Figure 1 provides a schematic view of the canonical correlation analysis process, from the data to graphical displays, using the **CCA** package.

Each step is illustrated in the following sections using the `nutrimouse` data set.

3.2. `nutrimouse` data set

The `nutrimouse` data set comes from a nutrition study in mouse (Martin *et al.* 2007). Forty mice were studied; two sets of variables were acquired:

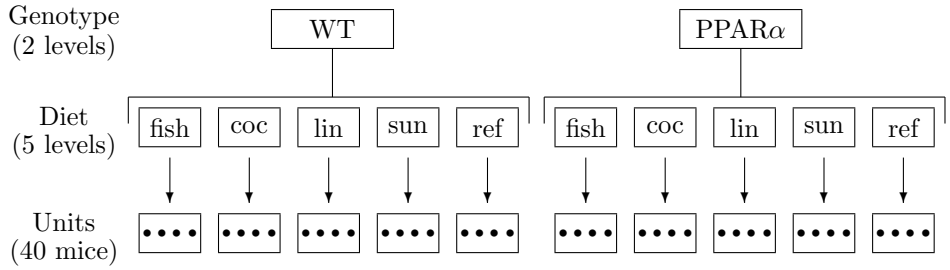

Figure 2: Experimental design of the nutrition study.

- expressions of 120 genes measured in liver cells, selected (among about 30000) as potentially relevant in the context of the nutrition study. These expressions were acquired thanks to microarray technology (Muller and Nicolau 2004);
- concentrations of 21 hepatic fatty acids (FA) measured by gas chromatography.

Biological units (mice) are cross-classified according to two factors (Figure 2):

- Genotype: study were done on wild-type (WT) and PPARα deficient (PPARα) mice.
- Diet: Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (REF), hydrogenated coconut oil for a saturated FA diet (COC), sunflower oil for an ω6 FA-rich diet (SUN), linseed oil for an ω3 FA-rich diet (LIN) and corn/colza/enriched fish oils for the FISH diet (42.5/42.5/15).

This data set was used to present a survey of statistical methods applied in the context of microarray data analysis (Baccini *et al.* 2005). Exploratory methods (PCA, multidimensional scaling, hierarchical clustering), modeling (ANOVA, mixed models, multiple testing), learning methods (random forests, Breiman 2001) were used to highlight a small sets of "interesting" genes on which CCA were performed to explore correlations with the fatty acid data set.

The data set can be loaded into the R workspace by `data("nutrimouse")`. The description of this data set is also available by calling `help(nutrimouse)`.

To avoid problems in the computations, the user should convert the data into the `matrix` format before performing CCA.

```
R> X <- as.matrix(nutrimouse$gene)
R> Y <- as.matrix(nutrimouse$lipid)
```

3.3. Preliminary

Canonical correlations analysis aims at highlighting correlations between two data sets. A preliminary step may consist in visualizing the correlation matrices. The package **CCA** proposes two ways to obtain this kind of representation: either the whole matrix concatenating X and Y or by expanding the 3 correlation matrices as it is done in Figure 3 with the following code.

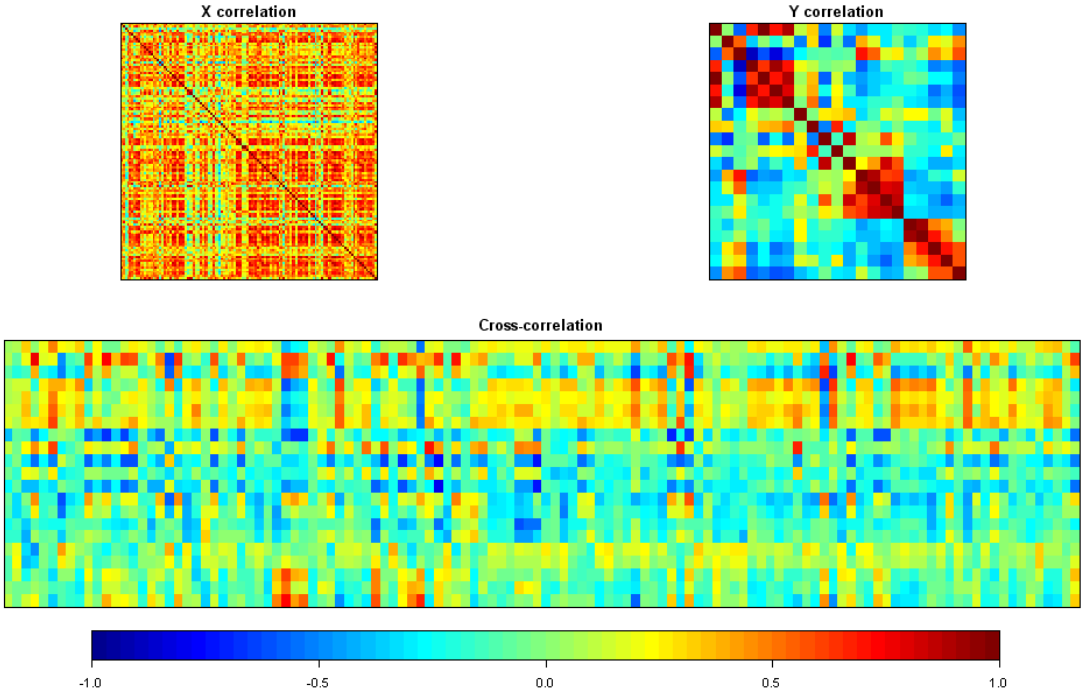

Figure 3: Correlation matrices for: X variables (upper-left), Y variables (upper-right), cross-correlation X × Y (bottom). Increasing values are translated into colors from blue (negative correlation) to red (positive correlation).

```
R> correl <- matcor(X, Y)
R> img.matcor(correl, type = 2)
```

Figure 3 highlights some significant correlations not only within each set of variables (squared matrices 120×120 and 21×21 at the top) but also between both sets (the rectangular matrix 21×120 at the bottom) *i.e.* between gene expression and fatty acids concentration.

The work must be stopped here if images obtained are uniformly in light green color corresponding to nearly null correlation.

3.4. Performing CCA

Classical CCA

When dealing with data sets in which the number of experimental units is greater than the number of variables, the classical CCA can be performed with the `cc()` function. The following command lines illustrate its use with a restricted number of genes from the X data set.

```
R> Xr <- as.matrix(nutrimouse$gene[, sample(1:120, size = 10)])
R> res.cc <- cc(Xr, Y)
```

```
R> barplot(res.cc$cor, xlab = "Dimension",
+ ylab = "Canonical correlations", names.arg = 1:10, ylim = c(0,1))
R> plt.cc(res.cc)
```

The explanation of the above command lines are:

1. first, we randomly choose 10 genes among the 120 and the data are stored as a matrix in the object `Xr`;
2. canonical correlations analysis is performed with `Xr` (40×10) and `Y` (40×21);
3. the barplot of canonical correlations is displayed;
4. variables and units are plotted on the first two canonical variates (default values for `d1` and `d2` arguments of `plt.cc`).

Similar plots will be presented and discussed in the following section when dealing with the complete case.

Regularized CCA

When dealing with the whole data set ($X_{(40\times120)}$ and $Y_{(40\times21)}$), classical CCA cannot be performed and the regularization step must be included in the data processing.

Choice of regularization parameters As mentioned in section 2.5, we opted to build a grid around reasonable values for λ on which we detect the cell where the CV-score reached its maximum.

The leave-one-out cross-validation process is implemented in the function `estim.regul()`. The default grid on which the CV-criterion is calculated is regular with 5 equally-spaced discretization points of the interval [0.001, 1] in each dimension.

The experience can guide the user to refine the discretization-grid but one can also use the `estim.regul()` function in a recursive way. First, use the default grid and locate an area where the optimal value for λ_1 and λ_2 could be reached and then, determine a new grid around these first optimal values.

Figure 4 was obtained by calculating the CV-criterion on the 51 × 51 grid by using the `estim.regul()` function:

```
R> res.regul <- estim.regul(X, Y, plt = TRUE,
+ grid1 = seq(0.0001, 0.2, l=51),
+ grid2 = seq(0, 0.2, l=51))
```

It enables to evaluate the optimal values for λ_1 and λ_2 at respectively 0.008096 and 0.064. These values are returned by the `estim.regul()` function with the value of the CV-criterion.

```
  lambda1 =  0.008096
  lambda2 =  0.064
 CV-score =  0.8852923
```

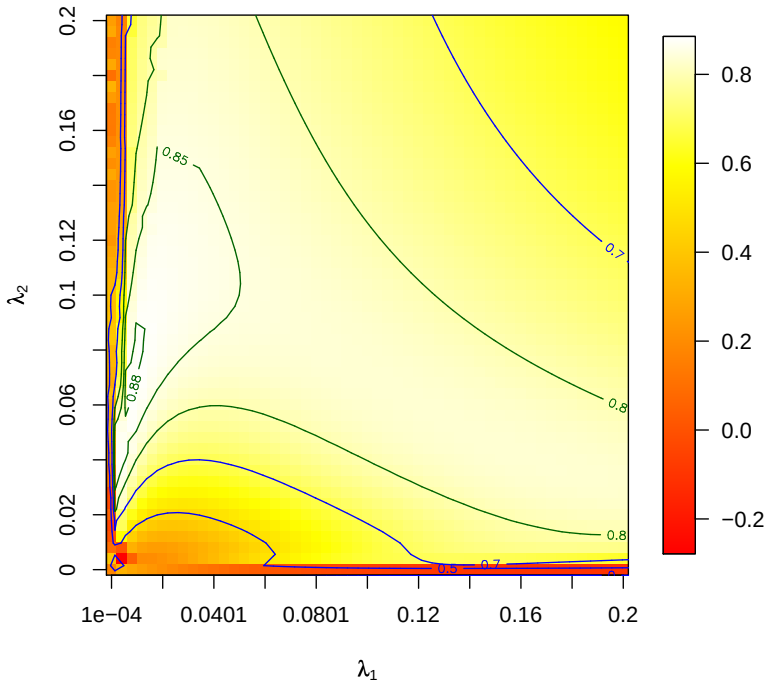

Figure 4: Image representing the CV-score for λ_1 and λ_2 on a 51×51 grid defined by equally-spaced discretization points on the region: $0.0001 \leq \lambda_1 \leq 0.2$ and $0 \leq \lambda_2 \leq 0.2$. Two kinds of contour plots are also displayed for values equal to {0-0.5-0.7} (in blue) and to $\{0.8\text{-}0.85\text{-}0.88^{(*)}\}$ (in green). (*) maximal value reached on the grid.

The computation is not very demanding. It lasted less than one hour and half on a "current use" computer for the 51×51 grid.

Figure 4 were completed by adding contour plots on the image with:

```
R> contour(res.regul$grid1, res.regul$grid2, res.regul$mat, add = TRUE,
+ levels = c(0,0.5,0.7), col = "blue")

R> contour(res.regul$grid1, res.regul$grid2, res.regul$mat, add = TRUE,
+ levels = c(0.8,0.85,0.88), col = "darkgreen")
```

Regularized CCA computations Once regularization parameters are fixed, the user calls the `rcc()` function with values for the parameters `lambda1` and `lambda2` instead of `cc()`.

```
R> res.rcc <- rcc(X, Y, 0.008096, 0.064)
```

The scree graph of canonical correlations can be plotted as a barplot.

```
R> barplot(res.rcc$cor, xlab = "Dimension",
+ ylab = "Canonical correlations", names.arg = 1:21, ylim = c(0,1))
```

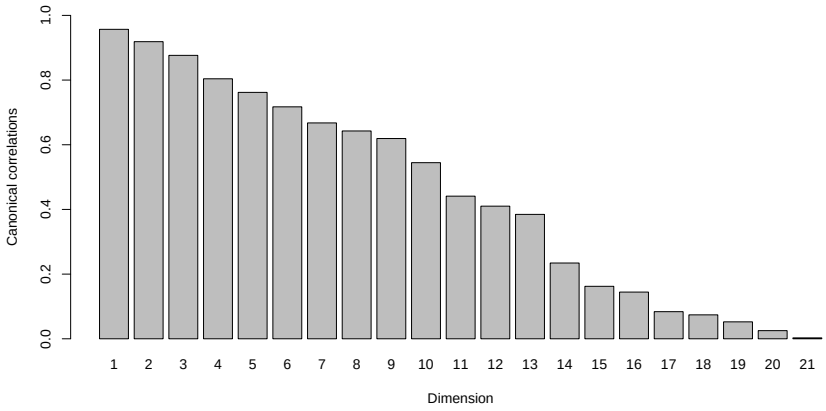

Figure 5: Barplot of canonical correlations.

Figure 5 can lead to several arguable choices for d: 3, 6 and 9 reveal larger gaps between two successive canonical correlations.

3.5. Graphical displays

In Figure 6, we focus on the first two dimensions for illustration purpose. Plots for larger dimensions could be displayed and interpreted in a more detailed study.

```
R> plt.cc(res.rcc, var.label = TRUE,
+ ind.names = paste(nutrimouse$genotype, nutrimouse$diet, sep = "-"))
```

Arguments `var.label` and `ind.names` of the `plt.cc()` function are given to make easier the interpretation of variables and units representation.

Some biological arguments are given on the basis of Figure 6 to focus on the relevancy of graphical outputs provided by RCCA.

First, RCCA provides an integrated graphical representation of two data sets (gene expression and fatty acid composition) which interpretation is fully in accordance with all the conclusions drawn in Martin *et al.* (2007) by way of standard statistical methods applied separately on each data set. The following two examples illustrate well this concordance:

1. the stronger effect of the genotype versus the dietary effects (see the clear separation of the genotypes along the first canonical variate) which is mostly due to the accumulation of linoleic acid (C18.2n.6) and CAR transcript in PPARα mice and the lower expression of fatty acid metabolism genes in these animals (genes with high positive coordinates on the first dimension) was previously underlined and discussed;

2. specific effects of the diets on FA composition and gene expression were reported and are also well illustrated by RCCA through the second canonical variate. In particular, the altered response of PPARα deficient mice to diet-induced changes in gene expression is well illustrated by the less clear separations of the units corresponding to different diets for PPARα mice (see the units with negative coordinates on the first dimension, right panel of Figure 6).

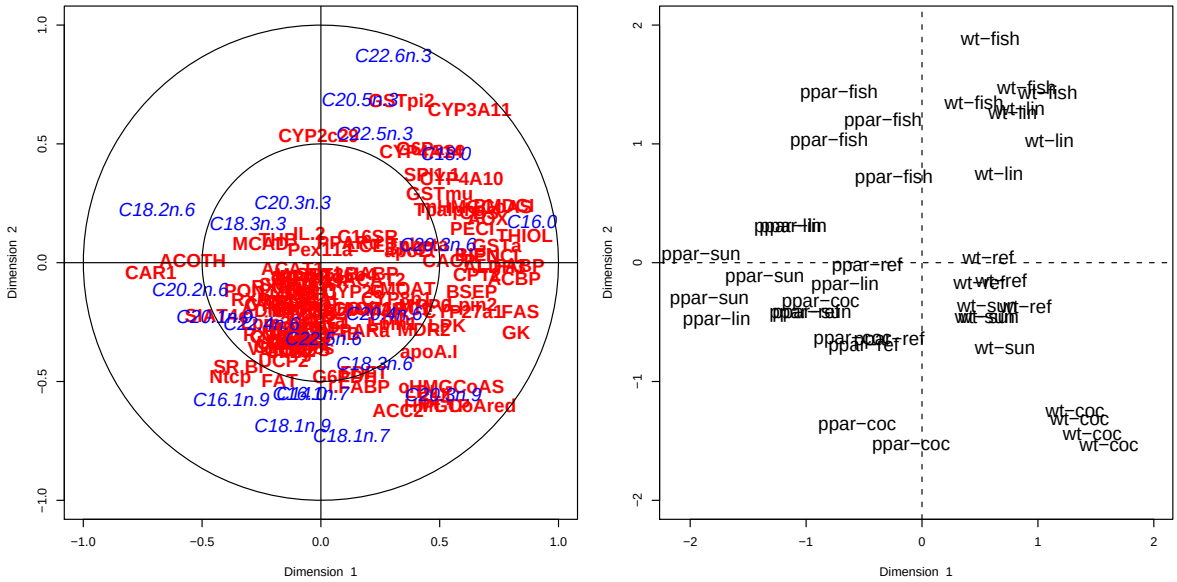

Figure 6: Variables (on the left) and units (on the right) representations on the plane defined by the first two canonical variates.

4. Conclusion

In this article we have proposed an efficient way to perform canonical correlation analysis in R. The functions provided in the **CCA** package outperform the `cancor()` R function according to several points:

- the ability to handle missing values;
- the processing of data sets with more variables than units through the regularized extension of the CCA;
- integrated solutions for graphical outputs.

These solutions were assessed on a real data set coming from a recent nutrition study.

We maintain a web page about canonical correlation analysis. It will always contain the latest release of the CCA package: `http://www.lsp.ups-tlse.fr/CCA/`. Future versions of **CCA** will aim at providing the user with additional tools. In particular, the cross-validation procedure can be computationally improved on larger data sets by performing a K-fold version instead of the leave-one-out one already implemented. Furthermore, other multidimensional methods such as linear discriminant analysis will be included as specific cases of CCA.

Acknowledgments

The authors are grateful to Dr. Thierry Pineau for the availability of the data and for interesting discussions about biological interpretation of the results. This work was partially supported by grants from ACI IMPBio and ANR PNRA.

References

Baccini A, Besse P, Déjean S, Martin PG, Robert-Granié C, SanCristobal M (2005). "Stratégies pour l'analyse statistique de données transcriptomiques." *Journal de la Société Française de Statistique*, **146**(1-2), 5–44.

Bickel PJ, Li B (2006). "Regularization in Statistics." *Sociedad de Estadistica e Investigacion Operativa, Test*, **15**(2), 271–344.

Breiman L (2001). "Random forests." *Machine Learning*, **45**(1), 5–32.

Eaton ML, Perlman MD (1973). "The Non-Singularity of Generalized Sample Covariance Matrices." *The Annals of Statictics*, **1**(4), 710–717.

Friedman JH (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, **84**(405), 165–175.

Guo Y, Hastie T, Tibshirani R (2007). "Regularized Linear Discriminant Analysis and its Application in Microarrays." *Biostatistics*, **8**(1), 86–100.

Leurgans SE, Moyeed RA, Silverman BW (1993). "Canonical Correlation Analysis when the Data are Curves." *Journal of the Royal Statistical Society B*, **55**(3), 725–740.

Mardia KV, Kent JT, Bibby JM (1979). *Multivariate Analysis*. Academic Press.

Martin PG, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi JM, SanCristobal M, Legrand P, Besse P, Pineau T (2007). "Novel Aspects of PPARα-mediated Regulation of Lipid and Xenobiotic Metabolism Revealed through a Nutrigenomic Study." *Hepatology*, **54**(2), 767–777.

Mevik BH, Wehrens R (2007). "The **pls** Package: Principal Component and Partial Least Squares Regression in R." *Journal of Statistical Software*, **18**(2). ISSN 1548-7660. URL http://www.jstatsoft.org/v18/i02/.

Mullen KM, van Stokkum IHM (2007). "An Introduction to the 'Special Volume Spectroscopy and Chemometrics in R'." *Journal of Statistical Software*, **18**(1). ISSN 1548-7660. URL http://www.jstatsoft.org/v18/i01/.

Muller UR, Nicolau DV (2004). *Microarray Technology and its Applications*. Springer-Verlag.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Vinod HD (1976). "Canonical Ridge and Econometrics of Joint Production." *Journal of Econometrics*, **4**(2), 147–166.

Affiliation:

Ignacio González
Institut de Mathématiques
Université de Toulouse et CNRS (UMR 5219)
Université Paul Sabatier
F-31062 Toulouse Cedex 9, France
and
Departamento de Matemática
Universidad de Carabobo, Venezuela
E-mail: ignacio.gonzalez@math.univ-toulouse.fr

Sébastien Déjean
Institut de Mathématiques
Université de Toulouse et CNRS (UMR 5219)
Université Paul Sabatier
F-31062 Toulouse Cedex 9, France
E-mail: sebastien.dejean@math.ups-tlse.fr
URL: http://www.math.univ-toulouse.fr/~sdejean/

Pascal G. P. Martin
Laboratoire de Pharmacologie et Toxicologie
Institut National de la Recherche Agronomique (UR 66)
180 Chemin de Tournefeuille, BP 3
F-31931 Toulouse, Cedex 9, France
E-mail: pascal.martin@toulouse.inra.fr

Alain Baccini
Institut de Mathématiques
Université de Toulouse et CNRS (UMR 5219)
Université Paul Sabatier
F-31062 Toulouse Cedex 9, France
E-mail: alain.baccini@math.univ-toulouse.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 23, Issue 12 *Submitted:* 2007-03-20
January 2008 *Accepted:* 2007-12-04