Program *Linear - Polynomial RDACCA* – User's notes

Vladimir Makarenkov and Pierre Legendre
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec H3C 3J7, Canada

May 2001

E-mail: **makarenv@makarenv.umontreal.ca** and **Pierre.Legendre@umontreal.ca**

What does the program Polynomial RDACCA do?

This program performs four forms of canonical analysis: linear or polynomial redundancy analysis (RDA) and linear or polynomial canonical correspondence analysis (CCA).

Classical linear redundancy analysis (Rao, 1964) and canonical correspondence analysis (ter Braak, 1986, 1987) are computed using multiple linear regression followed by direct eigenanalysis of the matrix of fitted values. The method of calculation is described in Chapter 11 of Legendre & Legendre (1998).

Polynomial RDA and CCA, which are generalizations of the linear forms, are implemented using a new approach proposed by Makarenkov and Legendre (1999, 2001). The polynomial methods are based on the use of multiple polynomial regression, during the first stage of RDA and CCA, instead of the multiple linear regression used in the linear forms. The explanatory variables are limited to their quadratic form in any term of the polynomial.

The program produces the output required to draw biplot diagrams for linear and polynomial RDA or CCA. In polynomial RDA or CCA, the explanatory variables can be represented in biplots in two different ways: (1) the individual terms of the polynomial equation can be represented as separate variables, or (2) one can choose to represent an explanatory variable using the multiple correlations (rescaled as required by the selected scaling method) of the canonical ordination axes against the linear and quadratic forms of the variable.

A permutation procedure allows one to test the significance of the two models (linear and polynomial) and of the difference between them.

Input files

The input data file is an ASCII text file which contains two data matrices. There is a matrix of response variables **Y**(nxp) and a matrix of explanatory variables **X**(nxm). The two matrices are written side by side in the input data file.

The data file is organized as follows:

• A first line contains three parameters: n, p, and m, separated by one or more spaces or by tabulators, where

n is the number of objects (sites), or rows of both the **Y** and **X** matrices.

p is the number of variables (columns) of the response matrix **Y**.

m is the number of variables (columns) of the explanatory matrix **X**.

• The two data matrices follow. They are written side by side on n successive rows. The first p columns of the data matrix correspond to matrix **Y** and the last m columns correspond to matrix **X**. See the example below.

The READDATA routine of the program reads the data in free field. This means that a row of data can take as many successive physical lines as needed. Values in the same line are separated by tabulators of by one or more spaces; the number of spaces does not matter.

Options of the program

The program can carry out one of the following 4 analyses:

• Classical linear redundancy analysis.

• Polynomial redundancy analysis.

• Classical linear canonical correspondence analysis.

• Polynomial canonical correspondence analysis.

Output file

The output is produced either on the monitor, or in a separate output file whose name is chosen by the user. The output consists either of the canonical analysis of **Y** by **X**, or the eigenanalysis of the residuals, or both, depending upon the user's request. See the example in Appendix.

Disclaimer

This program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers may use this program for scientific purposes, but the source code remains the property of Vladimir Makarenkov and Pierre Legendre. Publications should give proper credit to the method by referring to the published and accepted papers of Makarenkov and Legendre (1999, 2001). Users of the program should also refer to the present user's manual as follows:

Makarenkov, V. and Legendre, P. 2001. *Program Polynomial RdaCca – User's notes*. Département de sciences biologiques, Université de Montréal. 9 pages.

Technical notes

The program is distributed in a variety of formats from our base WWWeb site. WWWeb address: <http://www.fas.umontreal.ca/biol/legendre/>.

• FORTRAN source code for Windows 9x/ME/NT/2000 (file "RDA_CCA.f"), which can be compiled using a FORTRAN compiler (for example a free G77 Fortran compiler is available at the following URL: http://www.geocities.com/Athens/Olympus/5564/g77.htm). Users can modify the Parameter statement in the auxiliary file RDPARAM.f (provided) which fixes the size (*nmax*) of the largest data matrices **Y**(nxp) and **X**(nxm) that can be analyzed. Parameter *nmax* must be no less than the maximum of n, p and m+1. The default size (maximum size for the provided compiled versions) of the data matrix is 100 rows (objects), 100 columns (response variables) in matrix **Y**, and 99 columns (explanatory variables) in matrix **X**.

• A compiled version for 32-bit DOS (suitable for DOS sessions under Windows 9x/ME/NT/2000) (file " RDA_CCA.EXE"). To become effective, any modification to the file of parameters requires recompiling the source-code file "RDA_CCA.f".

• FORTRAN source code for Macintosh (file "Polynomial RdaCca.f"), which can be compiled using a FORTRAN compiler. Users can modify the Parameter statement in the auxiliary file RDA_CCA-parameters.f (provided) which fixes the size (*nmax*) of the largest data matrices **Y**(nxp) and **X**(nxm) that can be analyzed. Parameter *nmax* must be no less than the maximum of n, p and m+1. The default size (maximum size for the provided compiled versions) of the data matrix is 100 rows (objects), 100 columns (response variables) in matrix **Y**, and 99 columns (explanatory variables) in matrix **X**.

• A compiled version of the program for MacOS, for PowerPC and 68xxx processors, (file "Polynomial RDACCA "). The program requires 5.6 Mb RAM for running with parameter *nmax* = 100; it only requires 0.5 Mb RAM, for example, if *nmax* = 20. To become effective, any modification to the file of parameters requires recompiling the source-code file "Polynomial RdaCca.f".

The present user's manual (in PDF), as well as sample data files, are also available.

References

Legendre, P. & L. Legendre. 1998. *Numerical ecology, 2nd English edition.* Elsevier Science BV, Amsterdam. <http://www.fas.umontreal.ca/biol/legendre/numecol.html>

Makarenkov, V. & P. Legendre. 1999. Une méthode d'analyse canonique non-linéaire et son application à des données biologiques. *Mathématiques, informatique et sciences humaines*, 37e année, n° 147: 135-147.

Makarenkov, V. & P. Legendre. 2001. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology* (accepted for publication).

Press, W. H., B. P. Flanery, S. A. Teukolsky & W. T. Vetterling. 1986. *Numerical recipes – The art of scientific computing.* Cambridge Univ. Press, Cambridge.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya A* 26: 329-358.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167-1179.

ter Braak, C. J. F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* 69: 69-77.

Appendix: Test run

The test data used in this test run are from Table 11.3 of Legendre and Legendre (1998).

Data file: `Coral_fish.txt`

```
10 6 3
  1   0   0   0   0   0   1   0   1
  0   0   0   0   0   0   2   0   1
  0   1   0   0   0   0   3   0   1
 11   4   0   0   8   1   4   0   0
 11   5  17   7   0   0   5   1   0
  9   6   0   0   6   2   6   0   0
  9   7  13  10   0   0   7   1   0
  7   8   0   0   4   3   8   0   0
  7   9  10  13   0   0   9   1   0
  5  10   0   0   2   4  10   0   0
```

Console output file for polynomial RDA

```
Français: tapez (1)
English:  type  (2)
2

Write results
  (1) To the screen
  (2) To an output file
2

Name of the output file?
Output file: Fish.out

Data file? Coral_fish.txt

Warning: The following rows of matrix Y only contain zeros:
2

Ordination method:
  (1) RDA (redundancy analysis), or PCA if m = 0
  (2) CCA (canonical correspondence analysis), or CA if m = 0
1

Compute the following:
  (1) Canonical analysis
  (2) Analysis of the residual matrix
  (3) Both
1

Scaling of rows and columns:
  (1) Distance biplot: distances among objects approximate Euclidean distances
  (2) Correlation biplot: angles reflect the correlations among variables
1
```

```
 Transformation of the binary variables:
  (1) Center the binary variables found in X
  (2) Do not centre the binary variables in X
1

Transform the Y variables before analysis ?
  (0)  No transformation
  (1)  Transformation Y = Log(Y + 1), natural log
  (2)  Transformation Y = exp(Y)-1
1

Multiple regression model:
  (1) Linear regression Y = XB + C
  (2) Polynomial regression Y = P(X)      (maximum degree for any term = 2)
2

Printing:
  (1) Print the polynomial coefficients
  (2) Do not print the polynomial coefficients
1

Tests of significance:
  (0) No permutation test
  (1) Permutation tests for the significance of the canonical analyses
1

How many permutations? E.g. 499, 999, ... (maximum 9999)
999

Test the significance of:
  (1) The linear model only?
  (2) The polynomial model only?
  (3) Both the linear and polynomial models,
      and also the difference between them?
3

Print the percentage of variance accounted for
by the chosen model after each iteration of the permutation loop?
  (1) Print the percentage of variance after each iteration
  (2) Do not print the percentage of variance after each iteration
2

Permutation tests started (999 permutations) ...

End of the program.
```

Output file

```
Canonical redundancy analysis (RDA)
and canonical correspondence analysis (CCA)
using the linear and polynomial models,
with permutation tests.

Vladimir Makarenkov and Pierre Legendre
Département de sciences biologiques
Université de Montréal.
© V. Makarenkov & P. Legendre, 2000
-------------------------------------------

 Input file:                Coral_fish.txt
    10 objects
     6 response variables (matrix Y)
     3 explanatory variables (matrix X)

 Maximum number of canonical eigenvalues =  6
 Maximum number of non-canonical eigenvalues =  6

 Total inertia =        55.09067   Total variance =        6.12119

              Polynomial coefficients

 All coefficients printed are computed between the
 centred matrix X (using weights in CCA) and the
 centred (RDA) or transformed to Q-bar (CCA) matrix Y.

 Numbers in parentheses indicate the columns (variables) of X that are
 combined during an iteration to form a new column (variable) of X.
 The new column replaces the old ones in further iterations.

 In each row of six coefficients,
 - the first one is the coefficient of the first combined variable;
 - the second is the coefficient of the second combined variable;
 - the third is coefficient of the product of the two variables;
 - the fourth is the coefficient of the square of the first variable;
 - the fifth is the coefficient of the square of the second variable;
 - the sixth is the intercept.

 Approximation of column number  1 of matrix Y

  (  3  )     +     (  1  )
   -2.568068   -0.427657   -0.054921   -0.264293   -1.207865    0.000000

  (  2  )     +     (  3  ,  1  )
    0.117643    1.022913   -0.025810   -0.090304    0.000000    0.000000


 Approximation of column number  2 of matrix Y

  (  3  )     +     (  1  )
   -1.084654    0.354568    0.050548    0.169600    1.032773    0.000000

  (  2  )     +     (  3  ,  1  )
    0.009516    1.006417   -0.005197   -0.018665    0.000000    0.000000
```

```
Approximation of column number  3 of matrix Y

 (  2  )     +     (  1  )
   2.642441    0.093470   -0.022607   -0.112982    0.092725    0.000000

 (  2  ,  1  )     +     (  3  )
   1.004983   -0.431931   -0.300257    0.000000    0.000000    0.000000


 Approximation of column number  4 of matrix Y

 (  2  )     +     (  1  )
   2.372131   -0.111012    0.025876    0.129038   -0.098329    0.000000

 (  2  ,  1  )     +     (  3  )
   0.992231   -0.151198   -0.348278    0.000000    0.000000    0.000000


 Approximation of column number  5 of matrix Y

 (  2  )     +     (  1  )
  -1.712796   -0.526995    0.039550    0.183688    0.222577    0.000000

 (  2  ,  1  )     +     (  3  )
   0.995599   -2.601751    0.397099    0.000000    0.000000    0.000000


 Approximation of column number  6 of matrix Y

 (  2  )     +     (  1  )
  -1.196873    0.353151   -0.029651   -0.141670   -0.057875    0.000000

 (  2  ,  1  )     +     (  3  )
   1.000022   -0.979689   -0.000907    0.000000    0.000000    0.000000


 Mean coefficient of multiple determination R^2  =    0.99522

 (RDA) Percentage of the total variance of Y accounted for =  99.62535

 Canonical analysis:
 Total inertia =        54.88427   Total variance =         6.09825

 *** Canonical redundancy analysis ***

 Canonical eigenvalues
   3.65278   2.15430   0.22166   0.05236   0.01713   0.00002

 % of total variance of PCA
  59.67444  35.19423   3.62124   0.85533   0.27987   0.00025

 Cumulative % of total variance of PCA
  59.67444  94.86866  98.48990  99.34523  99.62510  99.62535

 Cumulative % of canonical variance
  59.89885  95.22543  98.86028  99.71882  99.99975 100.00000

 Sum of all canonical eigenvalues        6.09825
```

Scaling = 1: distance biplot

Normalized eigenvectors (matrix U): Species scores

0.28695	-0.54256	-0.44755	0.39017	0.35092	0.38419
0.24599	-0.49924	0.61773	-0.29602	-0.27418	0.38192
0.66340	0.09482	-0.14344	0.29598	-0.60524	-0.27644
0.59597	0.08485	0.07720	-0.43808	0.58045	-0.32065
-0.20438	-0.54909	-0.43642	-0.47529	-0.24511	-0.42460
-0.14177	-0.37243	0.44845	0.50469	0.19585	-0.58908

Sums of squares

1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Lengths

1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Site scores (matrix F)

-1.37802	1.67256	-0.36895	0.24332	0.22672	0.08990
-1.57692	2.04863	-0.05873	-0.02713	-0.01652	-0.17640
-1.40642	1.70259	0.36945	-0.23232	-0.20657	0.08833
-1.01532	-1.56769	-0.82472	-0.22851	0.01140	0.05169
2.73363	0.25640	-0.31807	0.35653	-0.17812	-0.00320
-0.99099	-1.64976	-0.24376	-0.07517	-0.00382	-0.02199
2.77514	0.21490	0.00187	-0.01366	0.01598	0.00399
-0.96522	-1.57654	0.28720	0.06849	-0.01222	-0.03835
2.74974	0.22216	0.29279	-0.34381	0.16243	-0.00718
-0.92564	-1.32325	0.86292	0.25225	0.00072	0.01321

Variances

3.65558	2.16337	0.22552	0.05396	0.01700	0.00576

Site scores (matrix Z)/polynomial combinations of explanatory variables

-1.44063	1.79497	-0.38966	0.23271	0.21812	-0.00013
-1.45286	1.80608	-0.01784	-0.00601	0.00062	0.00200
-1.46908	1.82505	0.35074	-0.24421	-0.21634	-0.00169
-1.02585	-1.62319	-0.77288	-0.13782	-0.02660	-0.00641
2.74642	0.26122	-0.30479	0.36750	-0.16931	-0.00067
-0.98983	-1.60964	-0.28807	-0.12387	-0.01403	0.00490
2.75307	0.20826	-0.03254	-0.04444	-0.00374	0.00317
-0.95584	-1.52368	0.26425	0.01364	0.00639	0.00562
2.75846	0.22624	0.30671	-0.33220	0.17022	-0.00252
-0.92388	-1.36529	0.88409	0.27471	0.03467	-0.00427

A biplot can be drawn using either matrices U and F, or matrices U and Z

Variances

3.65278	2.15430	0.22166	0.05236	0.01713	0.00002

Either of the two following sets can be used to represent the explanatory variables X in the biplot

FIRST WAY OF REPRESENTING EXPLANATORY VARIABLES IN BIPLOT:
Biplot scores of explanatory variables; multiple linear correlations between the explanatory variables X and the site scores Z

X1	0.41312	-0.43293	0.14654	-0.02619	0.03454	0.00045
X2	0.76776	0.06468	-0.00285	-0.00085	-0.00026	0.00000

```
X3   -0.40560   0.50447  -0.00528  -0.00163   0.00022   0.00002

 SECOND WAY OF REPRESENTING EXPLANATORY VARIABLES IN BIPLOT:
 Biplot scores of explanatory variables X,their squares X^2 and
 products Xi*Xj (i,j=1,...,m); linear correlations between the explanatory
 variables X, X^2, and Xi*Xj and the site scores Z


X1        0.32853  -0.39027   0.11526  -0.00614   0.00648   0.00001
X2        0.76776   0.06468  -0.00285  -0.00085  -0.00026   0.00000
X3       -0.40560   0.50447  -0.00528  -0.00163   0.00022   0.00002
X1*X1    -0.25047   0.18739   0.09049   0.02546   0.03392  -0.00045
X2*X2     0.76776   0.06468  -0.00285  -0.00085  -0.00026   0.00000
X3*X3    -0.40560   0.50447  -0.00528  -0.00163   0.00022   0.00002
X1*X2     0.23112   0.35538  -0.04625  -0.05166   0.02179  -0.00016
X1*X3     0.33573  -0.43227  -0.05437  -0.01139  -0.02241  -0.00010
X2*X3    -0.33877  -0.53240   0.00760   0.00232   0.00004  -0.00001

 Biplot scores of centroids of binary explanatory variables

X2        2.75265   0.23191  -0.01021  -0.00305  -0.00094  -0.00001
X3       -1.45419   1.80870  -0.01892  -0.00584   0.00080   0.00006

 Percentage of variance associated with matrix Y (unpermuted)

 Percentage of variance (linear regression) =          98.24874
 Percentage of variance (polynomial regression) =      99.62535
 Difference of variance accounted for (Pol-Lin) =       1.37661


 Permutation tests stareted (999 permutations) ...


     P(Lin)   0.00100
     P(Pol)   0.00100
 P(Pol-Lin)   0.03100


 Time spent:     59.76 seconds
```