Implementing and Interpreting Canonical Correspondence Analysis in SAS

Laxman Hegde, Frostburg State University, Frostburg, MD

ABSTRACT

The Canonical Correspondence Analysis (CCPA)[1] is a popular method among ecologists to study species environmental correlations using generalized singular value decomposition (GSVD) of a proper matrix. The CCPA is not so popular among researchers in other fields. Given two matrices **Y**(n by m) and **Z**(n by q), the CCPA involves computing GSVD of another matrix **A**(m by q) which can be thought of as a weighted averages for the columns of **Z** using the frequencies (the row totals) in **Y** matrix.

The SAS output for CCPA has several interesting parts (algebraically, numerically and graphically) that require interpretations to end users. In our presentation, we like to show how to perform CCPA in SAS/IML and interpret a few important results. The climax of this program is about constructing a Biplot of the **A** matrix.

Keywords: *Canonical Correspondence Analysis, Singular Value Decomposition, Biplot, Species Scores, Sample Scores, Species-Environmental Correlations*.

INTRODUCTION

The principal component analysis (PCA), the canonical correlation analysis(CCA), and the correspondence analysis (CA) are the well known multivariate data reduction techniques. The first two methods deal with an analysis of a correlation or variance-covariance matrix of a multivariate quantitative data set and the correspondence analysis deals with an analysis of data in a contingency table. It is well known that the general theory of all these methods is essentially based on the singular value decomposition (SVD) of a real rectangular data matrix that is being processed. For various applications of SVD in the correspondence analysis and other multivariate methods, see Greenacre (1984).

The canonical correspondence analysis (CCPA) can be thought of as a technique to analyze two sets of data, one in the form of a contingency table of species abundance at different sites and the other of external environmental variables (quantitative) on the same sites. One objective of CCPA is to develop relationships between these two different types of data sets.

LITERATURE

The paper by Ter Braak (1986) is considered a pioneering work of SVD applications in the field of species-environment relationship studies. Furthermore, Ter Braak (1988) has developed a very popular and frequently used Fortran program (CANOCO software) to perform canonical correspondence analysis. Following Ter Braak (1986,1988), Hegde and Naik (1999) developed a SAS Program to perform CCPA. Khattree and Naik (2000) have included a section on CCPA based on this work in their book on multivariate analysis. Furthermore, Hegde and Naik (2008) published a paper with detailed mathematical theory of CCPA and some new geometric interpretations.

WHY AGAIN IN NESUG?

Ever since the appearance of Hegde and Naik (1999) in the proceedings of SUGI 24, the authors have received more than 50 requests from ecologists all over the world for the reprints of this note and the SAS program. We have received more than 5 requests in the last 3 months. Hence, we have improved the SAS program with several comments for easy interpretations of SAS codes. More importantly, the modified SAS program is more flexible in constructing a Biplot.

We hope that NESUG conference proceedings will provide some additional publicity to the users of CCPA in different fields to our improved version of SAS program.

MAIN MATHEMATICAL THEORY OF CCPA

In a typical CCPA study in ecology, a researcher is encountered with two matrices **Y** and **Z** where **Y** is an n by m matrix containing m species abundance data and **Z** is an n by q matrix containing data on q environmental variables.

[1]Canonical Correspondence Analysis is popularly abbreviated as CCA in ecology community. However, CCA is popularly referred to as Canonical Correlation Analysis in statistics community. Hence we decided to call Canonical Correspondence Analysis as CCPA to avoid confusions.

Here n is the number sites where species abundance and environmental values are collected.

After standardizing the columns of **Z** the ecologists are generally interested in two vectors, an n by 1 vector **x** of site scores and an m by 1 vector **u** of species scores such that

$$\mathbf{u} \propto \mathbf{F}_1\mathbf{x}, \text{ and } \mathbf{x} \propto \mathbf{F}_2\mathbf{u}. \tag{1}$$

where $\mathbf{F}_1$ is an m by n matrix with each row representing a relative frequency distribution of counts of a species over different sites and $\mathbf{F}_2$ is an n by m matrix with each row representing a relative frequency distribution of counts of a species in a site over different species. Furthermore, in the equation (1), the vector **x** is restricted to satisfy the constraint $\mathbf{x} = \mathbf{Zb}$ where **b** is a q by 1 vector of regression coefficients. The equation (1) without the constraint $\mathbf{x} = \mathbf{Zb}$ is sometimes called reciprocal averaging in ecological research.

Solving (1) with constraint $\mathbf{x} = \mathbf{Zb}$ is equivalent to solving

$$\mathbf{u} \propto \mathbf{Ab}, \qquad \mathbf{b} \propto \mathbf{A}_2\mathbf{u}, \qquad \mathbf{b}'\Psi\mathbf{b} = 1, \tag{2}$$

where $\mathbf{A}$= $\mathbf{F}_1\mathbf{Z}$ and $\mathbf{A}_2 = \Psi^{-1}\mathbf{A}'w1$. Here Ψ is the correlation matrix of **Z** and $w1$ is an m by m diagonal matrix of column weights for **Y**. Solving the equation (2) involves performing the generalized singular value decomposition (GSVD) of **A** or the singular value decomposition (SVD) of the matrix $\mathbf{W} = w1^{1/2}\mathbf{A}\Psi^{\frac{1}{2}}$. See appendix for SVD and GSVD basics. For details of solving the equation (2), the readers may refer to Hegde and Naik (2008).

SOME RESEARCH OBJECTIVES OF CCPA

The the rows of the matrix **A** can be viewed as a cloud of m species points,each point is in R^q, a q-dimensional Euclidian space. Similarly the columns of the matrix **A** can be viewed as a cloud of q environ points,each point is n R^m, an m-dimensional Euclidian space. One objective of CCPA is to analyze variance-covariance structure in these clouds. This task is achieved through GSVD of **A** or SVD of **W**. A typical research interest lies in knowing the contributions of the first two eigenvalues of **W** to the sum of all the eigenvalues, the total variance in the row cloud or the column cloud. The total variance is defined as squared Frobenius norm of **W**.

Solving CCPA equation involves obtaining two matrices, a q by q matrix **B** of canonical coefficients and an m by q matrix **U** of species scores. Given these two matrices, a researcher is further interested in computing an n by q matrix **X**= **ZB** (linear combination of environmental values called LC scores) and a n by q matrix $\mathbf{X}^*$=$\mathbf{F}_2$**U** called sample scores. Then, a study of q by q correlation matrix Corr(**X**,$\mathbf{X}^*$) is an important goal in CCPA.

Finally, capturing the variance-covariance structure in the row cloud and in the column cloud of **A** into a two-dimensional plot (a Biplot) is a common objective in many CCPA studies.

SOME IMPORTANT STEPS OF CONDUCTING CCPA IN SAS/IML

Here we enumerate some important steps of carrying out CCPA in SAS/IML. For complete SAS program, the interested readers may contact the author. In addition to providing the SAS code, we will assist the users to interpret the results.[2]

1. Read data matrices **Y** and **Z** into SAS/IML.
2. Compute relative frequencies matrix **F**.

```
GT = Y[+,+]; /* Grand Total of all species */
F = Y/GT; /* Relative frequencies  */
```

3. Write a SAS module (Start .. Finish) to compute species weights and site weights. We wrote a simple module called Weights(mat) where mat is an input matrix. These weights are basically diagonal matrices containing relative frequencies for the species and the site totals observed in **Y**.

```
w1 = Weights(F); /* Species weights */
w2 = Weights(F`);  /*Site weights */
```

[2]Due to space restrictions, we are unable to display here the complete set of SAS codes.

4. Write a SAS module to standardize environmental variables in **Z** to have mean zero and standard deviation 1 using site weights w2. We wrote a simple module called Scale(mat,w) where the columns of mat are standardized with respects w.

```
Z = Scale(Z, w2); /* Standardize environmental variables */
```

5. Compute correlation matrix of **Z**.

```
Psi = Z`*w2*Z; /* Correlation of observed env variables */
w3 = inv(Psi);/* This matrix is called metric for environmental variables. */
```

6. Write a module to compute the square root of a matrix. We wrote a simple module SqrtMat(mat) using eigenvector-eigenvalue decomposition of a positive definite matrix. The square root of weight matrices are required in computing **W**, the weighted matrix **A**.

```
w1_hf = SqrtMat(w1); /* Square root of w1 */
w1_nhf = inv(w1_hf); /* Inverse of square root of w1 */
w3_hf = SqrtMat(w3);
w3_nhf = inv(w3_hf);
```

7. Create fundamental matrices of CCPA: **A** and **W**.

```
F1 = Inv(w1)*F`; /* Rel Freq dist of species over sites */
F2 = inv(w2)*F; /* Rel Freq dist of sites over species */
A = F1*Z; /* Weighted environmental (standardized) values */
W= w1_hf*A*w3_hf; /* Wighted A matrix to meet scaling conditions*/
```

8. Perform Singular Value Decomposition of **W**.

```
call svd(P,sv,Q,W);
```

9. Solutions to the equation (2)

```
umat = w1_nhf*P;
bmat = w3_nhf*Q;
```

10. Study singular values and eigenvalues

```
D=diag(sv);
Lambda=D*D;
```

11. Compute Species Scores and Regression Coefficients

```
power =1;
/* This number can change from -2 to 2.
Interpretation of some results in our analysis heavily depends on this number
The researchers may vary this number and get different sets of output.
Most popular choices are 0,0.5,1.
*/

DL= diag(sv ## power);  /* This scales umat, the left singular vector */
DR= diag(sv ## (1-power)); /* This scales bmat, the right singular vector */

U = umat*DL;  /* Species Scores. */
B = w3*bmat*inv(DR); /* Canonical Coefficients or Regression Coefficients */
```

12. Compute LC and Correlation between LC and **Z**.

```
X=Z*B; /* LC scores */
X = Scale(X,w2);
COEVO=Z`*w2*X;
```

13. Compute Correlation between **Z** and Sample Scores

```
X* =F2*U;
X* = Scale(X*,w2);/* Standardized Sample scores */
EOCORR=Z`*w2*X*;
```

14. Compute Species-Env Correlations

```
SECORR=X`*w2*X*;
SECORR=diag(SECORR);
```

15. Construct A Biplot

    Michael Friendly of York University has written a SAS macro (Biplot.sas) which can be downloaded easily from a google search. We have modified this macro suitably to meet the needs of our data set used in the example. The users of our program can also easily modify our macro to meet their respective needs.

AN EXAMPLE

1. Data Set

   Hunting spider data set from Van der Aart and Smeek-Enserink (1975) has been a very popular for illustrating CCPA computations. Ter Braak (1986, Table 3) adopted this data set after certain transformations for the illustration of CCPA.

   It may be of interest to some researchers to do some descriptive statistics/exploratory analysis using SAS procedures such as PROC UNIVARIATE, PROC MEANS, PROC CORR, and some other graphs using PROC SGPLOT.

   We will show only some important results in this example due to space restrictions.

	Site numbers																										
	15	19	20	16	17	18	2	8	21	5	6	14	4	7	13	3	1	9	12	25	11	10	28	23	22	27	24	26
Species																												
Arct lute										1	2	1	1	3	1	1												
Pard lugu	2	3	3	2	1	2	1	7	4	1		1	1	1	1	1				1				1				
Zora spin	1	1	1	2	1		3	1	1	4	5	5	5	4	4	1	2			2								
Pard nigr		1		1			3	1		9	5	3	5	9	7	4	3	1	1	2								
Pard pull							6	1	1	8	4	8	9	9	8	6	6	1	2		1							
Aulo albi							5	2		3	2	2	4	4	4	3	2			1	1							
Troc terr	5	4	4	5	4	5	8	5	4	9	7	9	9	9	9	8	7	1	3	4	2	1	1	1	1			1
Alop cune		1	1	1		1	1	3	1	4	2	1	2	2	6	4	3	1	3	1	1							
Pard mont							1	1	1	1	3	3	2	5	4	5	7	5	9	3	9	4	2	2	1	1	1	
Alop acce										1			1	1	1	3	5	1	4	3	3	1	3	4	2	5	3	1
Alop fabr															1	1				3	1	1	3	3	4	3	4	2
Arct peri																							1	2	1	2	2	4
Environmental variable																												
Water Content	9	7	8	8	9	8	8	6	7	8	9	8	6	8	9	6	5	5	5	3	4	4	0	0	1	0	2	0
Bare Sand	0	0	0	0	0	0	0	0	0	0	5	0	0	0	3	0	0	0	0	7	0	8	7	6	7	5	7	9
Cover Moss	1	3	1	1	1	0	2	2	1	0	5	4	5	1	1	5	7	9	8	2	9	7	8	9	9	8	9	4
Light Refl	1	0	0	0	2	2	3	1	0	5	1	2	6	5	7	8	8	7	8	5	8	8	8	9	8	8	9	9
Fallen Twigs	9	9	9	9	9	9	3	9	9	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cover Herbs	5	2	0	0	5	5	9	6	2	9	6	9	9	9	9	9	9	6	8	8	7	5	6	6	0	6	5	2

Hunting spider abundance data, with species (rows) and sites (columns)

2. Average Environmental Values

$$\mathbf{A} = \begin{pmatrix} 0.629769 & -0.110813 & -0.367615 & -0.223133 & -0.129843 & 0.457728 \\ 0.269121 & -0.416283 & -0.655741 & -1.10349 & 1.34925 & -0.954177 \\ 0.47921 & -0.153516 & -0.386301 & -0.494192 & 0.245792 & 0.0650301 \\ 0.435935 & -0.189168 & -0.379789 & -0.0744739 & -0.137867 & 0.407979 \\ 0.39163 & -0.362424 & -0.181087 & 0.00550386 & -0.260138 & 0.53032 \\ 0.351353 & -0.268658 & -0.290511 & -0.0702566 & -0.103988 & 0.514583 \\ 0.299329 & -0.220261 & -0.297999 & -0.372244 & 0.360045 & -0.142241 \\ 0.280868 & -0.26167 & -0.180473 & -0.040249 & 0.177195 & 0.0909022 \\ -0.350476 & 0.0712035 & 0.689378 & 0.517452 & -0.403042 & 0.0891841 \\ -1.1775 & 0.704692 & 0.982722 & 0.903484 & -0.608991 & -0.173264 \\ -1.77675 & 1.70812 & 1.0514 & 0.981226 & -0.596058 & -0.841195 \\ -2.27128 & 2.08766 & 1.06642 & 1.22656 & -0.630006 & -1.26214 \end{pmatrix}$$

- Each row corresponds to a species and each column corresponds to a standardized environmental variable.
- **Row Cloud**: 12 vectors, each with 6 coordinates and **Column Cloud**: 6 vectors,each with 12 coordinates.
- Each entry in this matrix is an average standardized environmental value for a given species.
- Note the signs. For example, in the first column, the last four values are negative indicating those species prefer dryness if we interpret the first column as water.
- CCPA in ecology analyzes this matrix given proper weights for species (rows) and proper weights for environmental variables (columns). Generally, species weights are frequencies or relative frequencies of their occurrences and env weights are in the inverse correlation matrix of **Z**.

3. Weighted Average Environmental Values

$$\mathbf{W} = \begin{pmatrix} 0.0905928 & 0.0221424 & -0.0188277 & -0.0339005 & -0.042086 & 0.043711 \\ -0.0790803 & -0.120468 & -0.122306 & -0.144538 & 0.262412 & -0.147858 \\ 0.0985968 & 0.00620758 & -0.048164 & -0.124552 & 0.00421431 & 0.02378 \\ 0.118231 & -0.00622075 & -0.0904924 & -0.0130824 & -0.0654825 & 0.0897482 \\ 0.11948 & -0.0775277 & -0.0222739 & -0.0163238 & -0.0955793 & 0.138474 \\ 0.0441676 & -0.0301789 & -0.0471713 & -0.0191927 & -0.00946398 & 0.122513 \\ 0.0698714 & -0.0782216 & -0.0843257 & -0.095609 & 0.109338 & -0.0502717 \\ 0.05573 & -0.0562631 & -0.0272998 & 0.0460913 & 0.0662339 & 0.0193313 \\ -0.0126899 & -0.0128693 & 0.224758 & 0.119222 & -0.072703 & 0.00209805 \\ -0.229019 & 0.10246 & 0.155209 & 0.145066 & -0.0648578 & -0.0300013 \\ -0.257685 & 0.27168 & 0.0779809 & 0.107134 & -0.0885281 & -0.151806 \\ -0.241603 & 0.212812 & 0.0113994 & 0.115348 & -0.0644476 & -0.166911 \end{pmatrix}$$

- This is a Fundamental Matrix in CCPA. This is weighted **A** matrix. Interpreting this matrix mat not be as simple as interpreting **A** matrix. The users who do not have matrix algebra background may simply view **W** matrix as weighted average environmental values matrix. Again, each row corresponds to a species.
- The CCPA analyzes species-environmental variations in this matrix.
- Also, because of how we construct **W**, the solutions to CCPA satisfy certain scaling conditions. Two CCPA programs may differ in how **W** is constructed.
- We have not shown the matrices $\mathbf{WW}'$ and $\mathbf{W}'\mathbf{W}$ in our program to save space. The total variation in **W** (Squared Frobenius norm) is the sum of all the diagonal elements of the matrix $\mathbf{W}'\mathbf{W}$ or $\mathbf{W}'\mathbf{W}$. Check that it is about 0.87. This sum also equals to the sum of all the eigenvalues in the SVD of **W**.
- The matrix $\mathbf{WW}'$may be called inter-species space (row cloud). A more serious researcher may be able to study inter-species distances and correlations using this matrix.
- Note that there are 12 dimensions in $\mathbf{WW}'$. One objective of CCPA is to find out what percent of the variations in the 12 dimensions is distributed in the first two dimensions, the first two eigenvectors of this matrix or the left singular vectors of **W** corresponding to the first two largest eigenvalues. The larger the percent, the better the quality of a Biplot.

- The matrix $\mathbf{W}'\mathbf{W}$ may be called inter-environ space (column cloud).
- A more serious researcher may be able to study inter-environ distances and correlations using this matrix. **A large diagonal value in this matrix also means a long arrow in a Biplot.**
- Note that there are 6 dimensions in $\mathbf{W}'\mathbf{W}$. One objective of CCPA is to find out what percent of the variations in the 6 dimensions is distributed in the first two dimensions, the first two eigenvectors of this matrix or the right singular vectors of **W** corresponding to the first two largest eigenvalues.

4. The SVD of **W**= **PDQ**$'$

$$\mathbf{P} = \begin{pmatrix} -0.091 & -0.14 \\ -0.161 & 0.759 \\ -0.174 & 0.016 \\ -0.166 & -0.204 \\ -0.197 & -0.338 \\ -0.134 & -0.137 \\ -0.209 & 0.275 \\ -0.101 & 0.019 \\ 0.181 & -0.319 \\ 0.434 & -0.09 \\ 0.575 & 0.1 \\ 0.496 & 0.185 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} 0.734 & 0. \\ 0. & 0.474 \end{pmatrix} \qquad \mathbf{Q} = \begin{pmatrix} -0.615 & -0.353 \\ 0.494 & -0.05 \\ 0.319 & -0.333 \\ 0.374 & -0.297 \\ -0.22 & 0.615 \\ -0.296 & -0.543 \end{pmatrix}$$

$$\Lambda = \mathbf{D}^2 = \begin{pmatrix} 0.539 & 0. \\ 0. & 0.224 \end{pmatrix}$$

- Here we have displayed information only for the first two axes from full SVD of **W** in which the matrix is **P** is 12 by 12 matrix, the matrix **D** is a 12 by 6 matrix, and the matrix **Q** is a 6 by 6 matrix. In applications,the researchers are interested to know what percent of the total variation in **W** is explained by the first two axes.
- In this example, of the total variation 0.87 in **W**, the first eigenvalues (the diagonal elements in Λ) add to 0.76 which is considered as very high.

5. GSVD of **A**= **MDN**$'$

This factorization of **A** is useful in generating a Biplot coordinates for the row and the column clouds. The first two columns of **M** provide a two dimensional orthonormal coordinate axes for the column cloud and the first two columns of **N** provide a two dimensional orthonormal coordinate axes for the row cloud. The matrix of singular values **D** is flexibly used to scale the axes and the respective Biplot coordinates as mentioned in the CCPA steps above.

$$\mathbf{M} = w1^{-1/2}\mathbf{P} = \begin{pmatrix} -0.684 & -1.048 \\ -0.662 & 3.123 \\ -0.626 & 0.057 \\ -0.528 & -0.649 \\ -0.558 & -0.955 \\ -0.552 & -0.565 \\ -0.44 & 0.579 \\ -0.381 & 0.071 \\ 0.513 & -0.901 \\ 1.582 & -0.328 \\ 2.665 & 0.466 \\ 3.384 & 1.262 \end{pmatrix}$$

$$\mathbf{N} = \Psi^{1/2}\mathbf{Q} = \begin{pmatrix} -0.937 & -0.079 \\ 0.748 & 0.057 \\ 0.673 & -0.317 \\ 0.633 & -0.57 \\ -0.399 & 0.79 \\ -0.335 & -0.775 \end{pmatrix}$$

$$\mathbf{A} = \mathbf{MDN}' \quad or \quad \mathbf{A}\Phi^{-1}\mathbf{N} = \mathbf{MD} \quad or \quad \mathbf{A}' = (\mathbf{N})(\mathbf{DM}')$$

In the above equation, let **x** be the first column of $\mathbf{A}'$. Note that **x** is the same the first row of **A**. Then check that $\mathbf{x} = \mathbf{Ny}$ where **y** is the first column of $\mathbf{DM}'$. We say **y** is the coordinate vector of **x** in the coordinate system provided by the columns of **N**.

Biplot coordinates for row cloud: 12 points, each with two coordinates. These points are represented by species labels in the Biplot.

$$\begin{pmatrix} -0.502 & -0.486 & -0.46 & -0.388 & -0.409 & -0.405 & -0.323 & -0.28 & 0.376 & 1.162 & 1.957 & 2.484 \\ -0.497 & 1.479 & 0.027 & -0.307 & -0.452 & -0.268 & 0.274 & 0.033 & -0.427 & -0.156 & 0.221 & 0.598 \end{pmatrix}$$

Biplot coordinates for column cloud: 6 points, each with two coordinates. These points are represented by arrows (labelled by environmental variables) in the Biplot.

$$\begin{pmatrix} -0.936753 & 0.748498 & 0.672746 & 0.6332 & -0.398647 & -0.334521 \\ -0.0790415 & 0.0572778 & -0.317131 & -0.569844 & 0.789517 & -0.775032 \end{pmatrix}$$

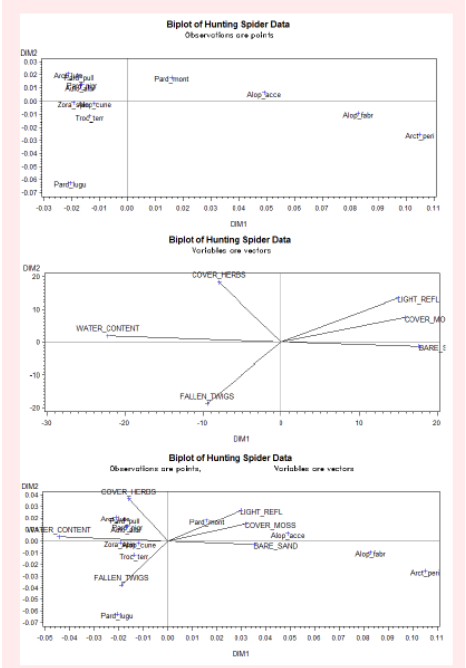
6. Biplot

In the figure below, we see three pictures. The first figure is a two-dimensional display for species space (row cloud), the second figure is a two-dimensional dispay for environ space (column cloud), and the third figure is a joint display of the first two figures.

Interpreting these pictures (specially the third figure) requires some understanding of innerproduct and projection concepts in a weighted Euclidian space. However, one must pay attention to some simple interpretations like

- the length of an arrow in a column cloud (magnitude and direction);
- the perpendicular projections of species points onto an environmental arrow;
- the angle between two arrows;
- the angle between an arrow and an axes;
- the position of species points, how far from the origin.

We believe that the following site is a good reference for understanding CCPA Biplots. http://www.umass.edu/landeco/teaching



7. Species Scores

$$\mathbf{U} = \begin{pmatrix} -0.502 & -0.497 & 0.446 & -0.264 & 0.017 & 0.095 \\ -0.486 & 1.479 & -0.243 & 0.062 & -0.089 & -0.03 \\ -0.46 & 0.027 & 0.277 & -0.238 & -0.135 & 0.05 \\ -0.388 & -0.307 & 0.307 & 0.08 & 0.082 & -0.022 \\ -0.409 & -0.452 & 0.059 & 0.078 & -0.054 & -0.097 \\ -0.405 & -0.268 & 0.118 & 0.261 & -0.142 & 0.1 \\ -0.323 & 0.274 & 0. & -0.059 & 0.04 & -0.012 \\ -0.28 & 0.033 & -0.171 & 0.146 & 0.242 & 0.105 \\ 0.376 & -0.427 & -0.458 & -0.163 & 0.034 & -0.011 \\ 1.162 & -0.156 & -0.287 & 0.17 & -0.146 & 0.018 \\ 1.957 & 0.221 & 0.402 & -0.108 & -0.032 & 0.062 \\ 2.484 & 0.598 & 0.583 & 0.105 & 0.214 & -0.136 \end{pmatrix}$$

8. Regression Coefficients or Canonical Coefficients

$$\mathbf{B} = \begin{pmatrix} -0.448 & -0.837 & 0.129 & -1.246 & 1.127 & 0.196 \\ 0.313 & -0.264 & 0.774 & -0.448 & 0.025 & 0.786 \\ 0.049 & -0.538 & -0.782 & -0.952 & -0.355 & 0.436 \\ 0.247 & -0.051 & -0.515 & 0.711 & 1.564 & 0.35 \\ -0.152 & 0.697 & -0.869 & 0.886 & 0.12 & 1.423 \\ -0.287 & -0.257 & -0.143 & 0.863 & -0.996 & 1.052 \end{pmatrix}$$

9. Four Important Correlation Structures in CCPA

   Generally it is a common practice to print the matrices **Z**, **X** (LC scores), and **X** (sample scores). However, these are large matrices in our example and hence we decided not display them to save space. More importantly, we have displayed the correlation structures between these matrices as shown below. For ecologists, a study of these correlations is an important objective.

   (a) Correlation between the columns of **Z**

$$\Psi = \begin{pmatrix} 1. & -0.56 & -0.688 & -0.573 & 0.401 & 0.266 \\ -0.56 & 1. & 0.286 & 0.27 & -0.088 & -0.312 \\ -0.688 & 0.286 & 1. & 0.558 & -0.396 & -0.099 \\ -0.573 & 0.27 & 0.558 & 1. & -0.772 & 0.348 \\ 0.401 & -0.088 & -0.396 & -0.772 & 1. & -0.592 \\ 0.266 & -0.312 & -0.099 & 0.348 & -0.592 & 1. \end{pmatrix}$$

   (b) Correlations between **Z** and LC scores

$$Corr(\mathbf{Z}, \mathbf{X}) = \begin{pmatrix} -0.937 & 0.748 & 0.673 & 0.633 & -0.399 & -0.335 \\ -0.079 & 0.057 & -0.317 & -0.57 & 0.79 & -0.775 \\ 0.142 & 0.46 & -0.579 & -0.195 & -0.094 & 0.062 \\ -0.162 & -0.176 & -0.262 & 0.389 & -0.256 & 0.486 \\ 0.242 & 0.017 & -0.2 & 0.289 & 0.092 & -0.194 \\ 0.106 & 0.44 & 0.053 & -0.039 & 0.368 & 0.096 \end{pmatrix}$$

   (c) Correlations Between **Z** and Sample Scores
   Correlation of an Environmental variable with an ordination axis

$$Corr(\mathbf{Z}, \mathbf{X}^*) = \begin{pmatrix} -0.898 & -0.075 & 0.094 & -0.115 & 0.151 & 0.042 \\ 0.717 & 0.054 & 0.306 & -0.125 & 0.01 & 0.176 \\ 0.645 & -0.301 & -0.385 & -0.186 & -0.125 & 0.021 \\ 0.607 & -0.54 & -0.129 & 0.277 & 0.18 & -0.016 \\ -0.382 & 0.748 & -0.062 & -0.182 & 0.057 & 0.147 \\ -0.321 & -0.735 & 0.041 & 0.346 & -0.121 & 0.038 \end{pmatrix}$$

(d) Correlations Between LC scores and Sample Scores
Species-Environment Correlations

$$Corr(\mathbf{X}, \mathbf{X}^*) = \begin{pmatrix} 0.958 & 0. & 0. & 0. & 0. & 0. \\ 0. & 0.948 & 0. & 0. & 0. & 0. \\ 0. & 0. & 0.665 & 0. & 0. & 0. \\ 0. & 0. & 0. & 0.711 & 0. & 0. \\ 0. & 0. & 0. & 0. & 0.624 & 0. \\ 0. & 0. & 0. & 0. & 0. & 0.399 \end{pmatrix}$$

CONCLUSION

In this paper we have attempted to introduce the basic ideas of conducting CCPA in SAS and interpreting some important CCPA results. We hope that SAS will implement in future one uniform procedure to handle all these data reduction techniques PCA, CCA, CA, and CCPA when a research involves two data matrices **Y** and **Z**.

REFERENCES

1. Van der Aart, P. J. M. and Smeek-Enseink, N. (1975). Correlations Between Distributions of Hunting Spiders and Environmental Characteristics in a Dune Area, *Netherlands Journal of Zoology,* **25**, 1-45.

2. Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis, *Academic Press*, London, England.

3. Ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis *Ecology*, **67**, 1167-1179.

4. Hegde, L. M. and Naik, D. N. (1999). Canonical correspondence analysis in SAS software. *Proceedings of the Twenty-Fourth Annual SAS Users Group International (SUGI) Conference,* paper **278**, 1607-1613.

5. Khattree, R. and Naik, D. N. (2000). *Multivariate Data Reduction and Discrimination with SAS Software*. SAS institute Inc., Cary, North Carolina, and J. Wiley and Sons, New York, USA.

6. Hegde, L. M. and Naik, D. N. (2008). Canonical Correspondence Analysis : Some New Interpretations And Computations Using SAS, *Journal of Statistics and Applications*,**Volume 3**,Issue 2,277-302

ACKNOWLEDGMENTS

The author would like to thank Dr. Dayanand Naik, Old Dominion University VA, for his role as a motivator in doing this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author:

Name:	Laxman Hegde
Enterprise:	Frostburg State University
Address:	101 Braddock St.
City, State, ZIP:	Frostburg, MD, 21532
Work Phone:	301-687-4777
E-mail:	lhegde2@frostburg.edu

APPENDIX

Singular Value Decomposition (SVD) and Generalized SVD (GSVD)

In this appendix we will provide the basics of SVD, GSVD, and some geometric interpretations. Suppose **E** is an m by q real matrix with rank r, $r \leq q$. Then **E** can always be represented (or factored) as

$$\mathbf{E} = \mathbf{P}_1 \mathbf{D}_1 \mathbf{Q}_1', \tag{3}$$

where $\mathbf{Q}_1$ and $\mathbf{P}_1$ are orthogonal matrices such that $\mathbf{P}_1'\mathbf{P}_1 = \mathbf{P}_1\mathbf{P}_1' = \mathbf{I}_m$, and $\mathbf{Q}_1'\mathbf{Q}_1 = \mathbf{Q}_1\mathbf{Q}_1' = \mathbf{I}_q$. The columns of $\mathbf{Q}_1$ (q by q) form a *full* set of eigenvectors of $\mathbf{E}'\mathbf{E}$ (called the right singular vectors of **E**) and the columns of $\mathbf{P}_1$ (m by m) form a *full* set of eigenvectors of $\mathbf{E}\mathbf{E}'$ (called the left singular vectors). Further, $\mathbf{D}_1$ is an m by q matrix such that the first r diagonal entries are positive and all other elements of the matrix are zero. In fact, the matrices $\mathbf{E}'\mathbf{E}$ and $\mathbf{E}\mathbf{E}'$ each share the same set of eigenvalues and the square root of the positive eigenvalues form the first r diagonal elements of $\mathbf{D}_1$. These elements of $\mathbf{D}_1$ are known as singular values of **E**.

It is a general practice to write equation (3) as

$$\mathbf{E} = \mathbf{P}\mathbf{D}\mathbf{Q}', \tag{4}$$

where **P** is the first r columns of $\mathbf{P}_1$, **Q** is the first r columns of $\mathbf{Q}_1$, and **D** is square and nonsingular matrix containing the first r positive diagonal elements of $\mathbf{D}_1$. Also it is conventional to order the elements of **D** from largest to smallest, and correspondingly reorder the columns of **P** and **Q**. Now the matrices **P** and **Q** are only column-wise orthogonal, meaning $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_r$. Sometimes the equation (4) is known as partial SVD. In fact, in most applications, it is sufficient to consider only the partial SVD.

However, in certain practical situations the interest may be in decomposing a matrix with weights attached to the rows or the columns or both. This leads to a generalized SVD (GSVD). Suppose **A** is an m by q matrix and interest is to factor **A** by attaching weights to the rows using an m by m positive definite matrix Ω and by scaling the columns using a positive definite matrix Φ^{-1}. To achieve this, we perform the SVD of **A** as a linear transformation from the weighted Euclidian space $(\Re^q, \Phi^{-1})$ to the weighted Euclidian space $(\Re^m, \Omega)$. The GSVD of **A** is computed as

$$\mathbf{A} = \mathbf{M}\mathbf{D}\mathbf{N}', \quad \text{where} \quad \mathbf{M} = \Omega^{-\frac{1}{2}}\mathbf{P}, \quad \mathbf{N} = \Phi^{\frac{1}{2}}\mathbf{Q},$$

and $\mathbf{M}'\Omega\mathbf{M} = \mathbf{N}'\Phi^{-1}\mathbf{N} = \mathbf{I}_r$. The matrices **P**, **D**, and **Q** are obtained by performing an SVD of the matrix $\Omega^{\frac{1}{2}}\mathbf{A}\Phi^{-\frac{1}{2}}$. Note that the matrices **M**, and **N** are column-wise orthonormal in their given metrics, Ω, and Φ^{-1}.