

Using Linear Multivariate Methods in SAS to Compare Treatment Responses



Paul Watson

Alberta Research Council

Paul.watson@arc.ab.ca

780-632-8218

Background

What are MV Methods?

- Numerical strategies used to represent multi-dimensional data structures in a lower (2) -dimensional space, while maximally preserving trended variation present in the data.

Background

My Research

- **Agriculture & Agronomy**
- **Determine if treatments differ from each other (e.g. weed communities)**
- **Treatments based on agricultural production practices**
 - **Tillage (\pm)**
 - **Crop Sequence (Rotation)**
 - **Fertility**

Background

Three-step approach to analysis

– Data exploration

- Relative importance of factors (PCA)

– Hypothesis testing

- CDA on statistically sig. axes from step 1
 - Broken stick, bar chart of eigenvalues
- If difference are found:

– Association of weed species to treatments

- CDA on weed species

Background

Why?

- **PCA (covariance) maps data into new coordinate frame but leaves distances between objects unchanged**
- **CDA on significant PCA axes removes noise from non-significant axes**
- **CDA on weed species tells us about associations**

Background

Assumptions

- “Linear” data
 - Not too many zeros (<60-70%)
 - Can delete rare variables (weed species)
- COV - Data are on same scale
 - not like ph (0-14) versus soil characteristics like sand, silt, clay (0-100)

Typical data

Plot	tmt	CropCode	rot	rep	year	Wdsp1	Wdsp2	Wdsp3	Wdsp4
501	14	P	cfwwp	1	2000	0	1.2	5.6	132.8
502	6	Cs	spring	1	2000	1.6	4	3.2	60
503	1	Ws	spring	1	2000	0	0.8	34	184
504	3	Ws	cfwp	1	2000	0	0	59.2	138
505	12	P	cfwp	1	2000	0	0	60	34.8
506	15	Cf	cfpww	1	2000	0	0	12.8	0
507	7	Cs	cwwp	1	2000	0	1.6	17.6	2.8
508	2	Ww	cwwp	1	2000	0	0	0	0
509	4	Ww	cfwwp	1	2000	0	0	0	0
510	9	Cf	cfwwp	1	2000	0	0	2	0.4
511	11	P	spring	1	2000	0	0.8	16.8	0.4
512	13	P	cwwp	1	2000	0	0.4	12.8	0
513	10	Cf	cfwp	1	2000	0	0	3.2	0
514	8	Ww	cfpww	1	2000	0	0	0	0
515	5	P	cfpww	1	2000	0	0	28.8	1.2
601	7	Cs	cwwp	2	2000	0	0	29.2	489.2
602	5	P	cfpww	2	2000	6.4	0.4	21.6	225.6
603	10	Cf	cfwp	2	2000	30.4	0	7.6	4.4
604	11	P	spring	2	2000	0	0.4	25.6	26.4

Click to edit

Click to edit

SAS Code - PCA

PROC PRINCOMP

- COVARIANCE

- N=5 (*reduces data output*)

- data=F1

- outstat=F2

- out=PS001sco (*arbitrary name: data goes to CDA*)

- (keep= plot *tmt rot Tillage* NumWintr rep *prin1 prin2 prin3 prin4 prin5*);

PCA Output

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	<u>Proportion</u>	<u>Cumulative</u>
1	5.76193434	2.13598527	0.3540	0.3540
2	3.62594907	1.39191676	0.2228	0.5768
3	2.23403231	1.00549257	0.1373	0.7141
4	1.22853974	0.33771578	0.0755	0.7895
5	0.89082396		0.0547	0.8443

Broken-stick method, Legendre and Legendre, 1998

Click to edit

Click to edit

PCA Output – Variable Scores

Eigenvectors

	Prin1	Prin2
AMAXX	0.067731	0.132030
ARTBI	0.003942	0.005470
AVEFA	0.574916	0.032084
CAPBP	-.143142	0.701099
CHEAL	0.001833	0.004644
CIRAR	0.200314	0.244019
ECHCG	0.444304	0.110055
EPHSP	0.014071	0.005756
EPIGL	0.009379	-.000484
HORJU	0.023560	0.027551
LAMAM	0.082835	0.042781
MELNO	0.059427	0.129860
MEUXX	-.004518	0.009644
MUSSP	-.388888	-.188752
POLCO	0.171175	0.109453
POLXX	-.002160	0.055322
SETVI	0.111616	-.311645
SINAR	0.005157	0.008857

Click to edit

Click to edit

PCA Output – Object Scores

Obs	Plot	tmt	rep	Prin1	Prin2
1	501	14	1	1.59879	-1.47839
2	502	6	1	0.79116	-2.93344
3	503	1	1	3.85631	-1.73028
4	504	3	1	4.35466	-1.13610
5	505	12	1	1.90190	-0.00513
6	506	15	1	-1.74230	0.89775
7	507	7	1	-0.73150	0.18407
8	508	2	1	-3.64023	-0.52440
9	509	4	1	-3.62591	-0.67862
10	510	9	1	-2.57072	0.71997
11	511	11	1	0.74099	3.62033
12	512	13	1	0.25227	3.62474
13	513	10	1	-2.11807	0.62225
14	514	8	1	-3.95664	-0.80578

Plot (object) and variable scores need not be on the same scale and can be rescaled to make them fit well on the same biplot

Click to edit

Click to edit

PCA Output – Biplot

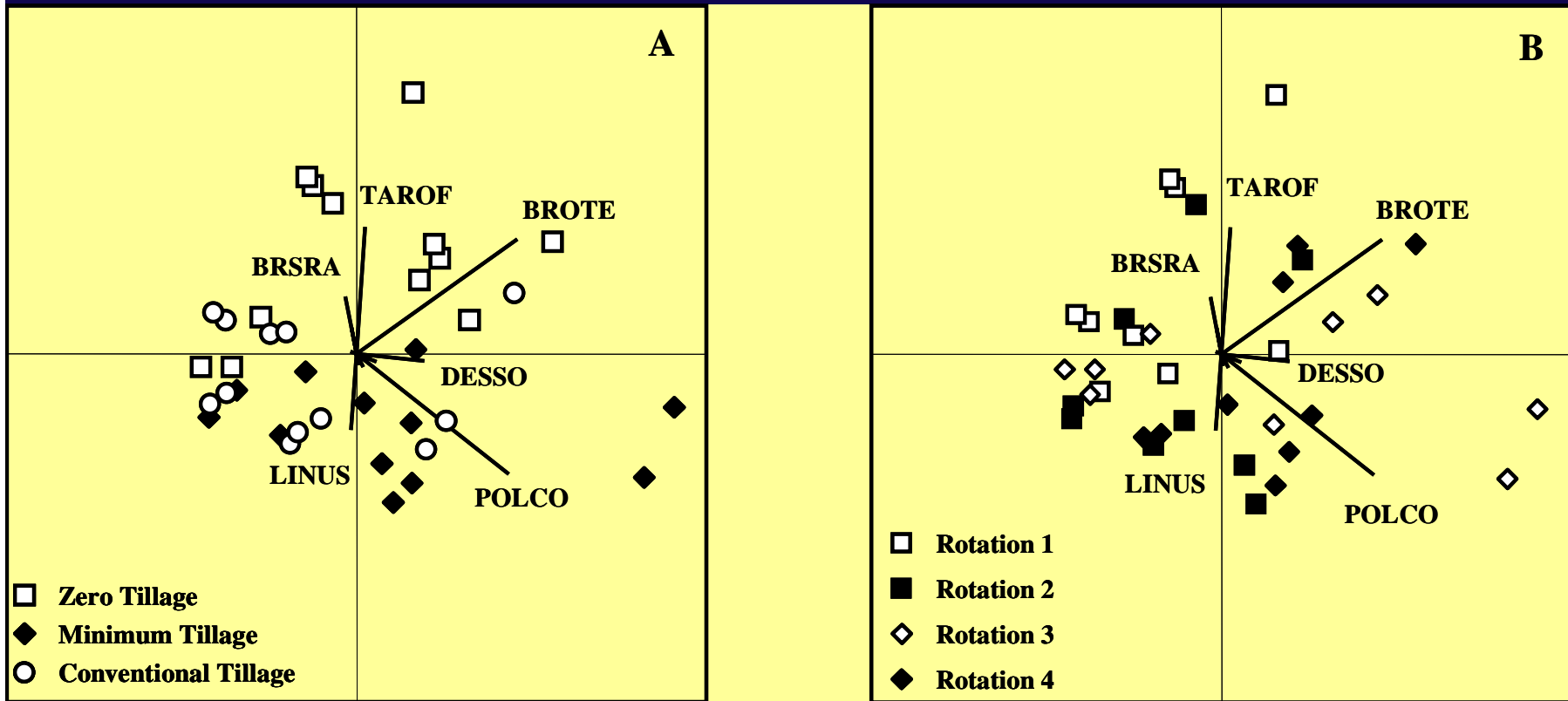


Figure 1. PCA using all weed species as variables (Step 1) biplot with plots labelled by: (A) tillage system, and (B) rotation. Species vectors accounting for < 25% of the total vector length on the first 2 axes were not labelled.

SAS Code – Send to CDA

```
DATA F12 (KEEP=plot tmt rot CropCode NumWintr  
prin1 prin2);  
SET PS001sco;
```

```
DATA F13 (KEEP=plot tmt rot CropCode NumWintr  
prin1 prin2 prin3);  
SET PS001sco;
```

Do this for as many iterations as required

N=1, 2, 3, 4, 5, ...

SAS Code – CDA on PCA scores

```
PROC CANDISC uni distance data=F12 out=PS001a  
(Keep=plot tmt rot CropCode NumWintr Can1);  
CLASS tmt;  
VAR Prin1;
```

```
PROC CANDISC uni distance data=F13 out=PS001b  
(Keep=plot tmt rot CropCode NumWintr Can1 Can2);  
CLASS tmt;  
VAR Prin1 Prin2;
```

Do this for as many iterations as required

N=1, 2, 3, 4, 5, ...

SAS Code – Output

Table 6. Prob > Mahalanobis Distance for Squared Distance to tmt based on 2 significant PCA axes.

From tmt	1	2	3	4	5	6	7	8
1	1.0000	0.0736	<.0001	0.0008	0.0096	0.6553	0.7366	0.2919
2	0.0736	1.0000	<.0001	<.0001	<.0001	0.3680	0.3015	0.7457
3	<.0001	<.0001	1.0000	0.7800	0.2984	<.0001	<.0001	<.0001
4	0.0008	<.0001	0.7800	1.0000	0.6655	<.0001	<.0001	<.0001
5	0.0096	<.0001	0.2984	0.6655	1.0000	0.0010	0.0014	0.0002
6	0.6553	0.3680	<.0001	<.0001	0.0010	1.0000	0.9902	0.7991
7	0.7366	0.3015	<.0001	<.0001	0.0014	0.9902	1.0000	0.7210
8	0.2919	0.7457	<.0001	<.0001	0.0002	0.7991	0.7210	1.0000
9	0.2595	0.7803	<.0001	<.0001	0.0001	0.7542	0.6733	0.9954
10	0.2693	0.6434	<.0001	<.0001	0.0003	0.7134	0.6421	0.9095
11	0.1523	0.9112	<.0001	<.0001	<.0001	0.5708	0.4919	0.8939
12	0.2005	0.8526	<.0001	<.0001	<.0001	0.6639	0.5835	0.9428
13	0.1805	0.8857	<.0001	<.0001	<.0001	0.6297	0.5485	0.9354
14	0.0744	0.8658	<.0001	<.0001	<.0001	0.3527	0.2939	0.6526
15	0.2092	0.8566	<.0001	<.0001	<.0001	0.6821	0.5986	0.9779

Click to edit

Click to edit

SAS Code – CDA on original data

```
PROC CANDISC uni distance data=F1 out=PS001tmt  
(Keep=plot tmt rot CropCode NumWintr Can1 Can2);  
CLASS tmt;
```

Usually 2 CDA axes are enough, and it is as much as can easily be represented anyway

SAS Code – CDA Output

Total-Sample Standardized Canonical Coefficients (Variable scores)

Variable	Can1	Can2
AGRRE	0.820496712	0.242018986
AMAXX	0.499441241	0.188406138
AVEFA	0.452147467	0.691248447
CAPBP	2.872977548	0.299048866

Click to edit



Click to edit

SAS Code – CDA Output

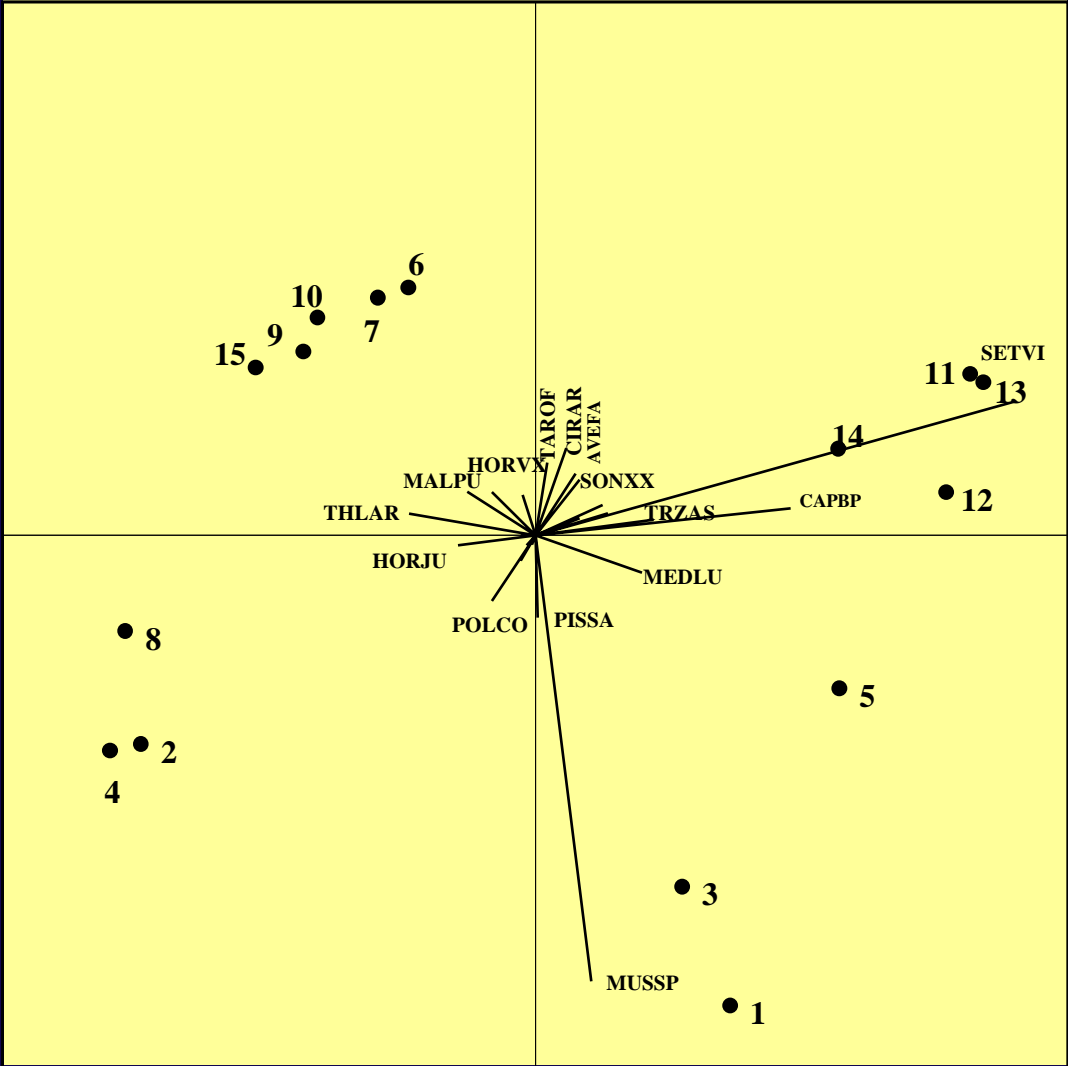
Class Means on Canonical Variables

tmt	Can1	Can2	Can3	Can4	Can5
1	4.39341486	-10.59281765	4.53584174	-1.83192671	1.59151203
2	-8.88550050	-4.70978307	-3.75007038	0.28776479	0.09869085
3	3.31475184	-7.92261215	3.38036017	-0.25619362	-2.23239153
4	-9.57991518	-4.85004439	-4.70345043	-0.36608239	-0.29470353
5	6.84999740	-3.45569833	3.36680042	3.15347595	0.32120895
6	-2.85848962	5.57057730	2.63619698	-3.64250766	-2.17730810
7	-3.54561971	5.34506628	3.61346968	-2.48358092	0.03864653
8	-9.23967689	-2.16231519	-4.13675960	-0.92820097	0.53998895
9	-5.22518388	4.13216755	1.88584989	0.20091347	1.35201404
10	-4.90515711	4.89790113	3.41249322	-0.13392401	1.46285002
11	9.80354939	3.62786554	-2.47693116	-0.57586259	0.24523121
12	9.26241694	0.96116256	-4.24474848	-0.16034495	-1.08204779
13	10.09414940	3.44293463	-3.38629825	-0.80177011	0.91435685
14	6.82116487	1.94021123	-2.05110028	1.65070217	0.22030064
15	-6.29990181	3.77538459	1.91834649	5.88753754	-0.99834912

Click to edit

Click to edit

SAS Code – CDA Biplot



Click to edit



Click to edit

Using Linear Multivariate Methods in SAS to Compare Treatment Responses



Paul Watson

Alberta Research Council

Paul.watson@arc.ab.ca

780-632-8218