

Principal Component Analysis (PCA) and other Multivariate Statistical Analysis Techniques in SAS

Dr. Haoyu Yu
Minnesota Supercomputer Institute
University of Minnesota

Dr. Michael Steinbach
Computer Science and Engineering
University of Minnesota

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



Outline

- introduction
 - multivariate data
 - overview of multivariate analysis techniques
- principal component analysis
 - basics and background
 - examples
 - regression
 - biplots

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



data characteristics ...

- characteristics of the data
 - will assume the data is a table (matrix) of real values
 - ex: revenue of each store from different products
 - ex: performance of athletes in a set of races
 - ex: height, weight, blood pressure, etc. for patients
 - data is a table of records (rows) which each have the same variables (columns)
 - no categorical data except class variable in some data
 - assume no missing data

ID	Income	Age	# years at present address	Credit Worthy
1	45k	35	5	Yes
2	32k	23	2	No
3	40k	27	1	No
4	50k	55	10	Yes
...

© 2010 Regents of the University of Minnesota. All rights reserved.

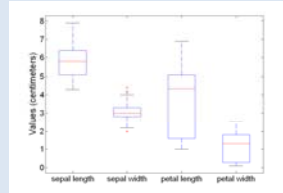
data characteristics (due to S. S. Stevens)

	Variable Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal variable values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
	Ordinal	Ordinal variable values also order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval variables, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

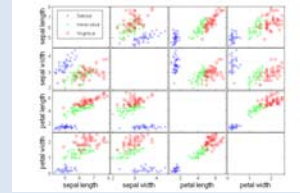
© 2010 Regents of the University of Minnesota. All rights reserved.

multivariate analysis techniques: visualization

- display data in a way that allows people to interpret the structure of the data
- Plots are from Fisher's Iris data set



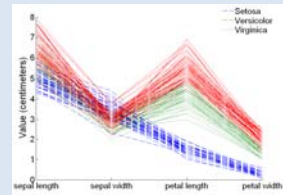
box plots



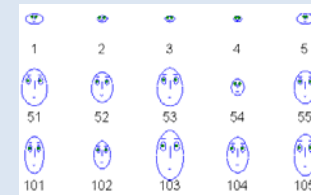
scatter plots



Virginia. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.



parallel plot



Chernoff faces

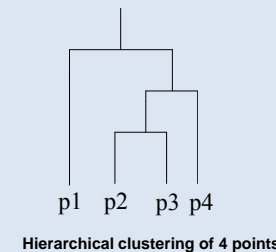
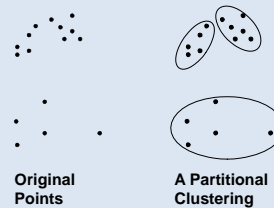
- Setosa
- Versicolour
- Virginia

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: clustering

- group data into meaningful groups
 - useful for understanding
 - useful for summarization
- important distinction between **hierarchical** and **partitional** clustering
 - partitional clustering is a division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
 - k-means is most common example
 - hierarchical clustering is a set of nested clusters organized as a hierarchical tree
 - common hierarchical techniques are single link, complete link, group average, Wards
 - many, many techniques



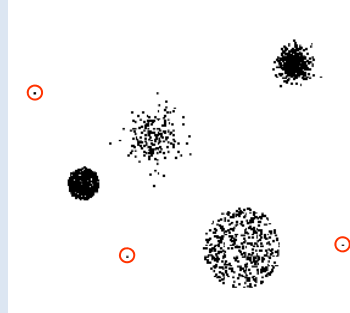
© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: anomaly detection

- **Anomaly:** object that is different from most other objects in the data set
- many statistical and non-statistical techniques have been developed
 - model based
 - distance based
 - density based
 - pattern based
 - graphical
- PCA and other dimensionality reduction approaches can make outliers more obvious
- Chandola, V., Banerjee, A., and Kumar, V. 2009. **Anomaly detection: A survey.** *ACM Comput. Surv.* 41, 3 (Jul. 2009), 1-58. DOI= <http://doi.acm.org/10.1145/1541880.1541882>



© 2010 Regents of the University of Minnesota. All rights reserved.

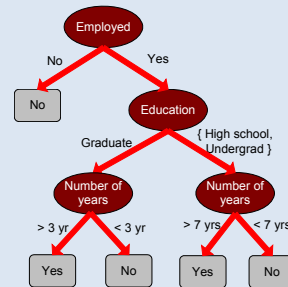
Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: classification

- given a set of training data with class labels for objects (observations), create a model to predict the class membership of a new object (observation)
- many approaches in machine learning and data mining
 - decision tree
 - rule-based methods
 - nearest-neighbor
 - neural networks
 - naïve Bayes and Bayesian networks
 - Support Vector Machines (SVM)
 - random forests
- in multivariate statistics discriminant analysis is traditionally used
- Linear Discriminant Analysis (LDA) finds a linear combination of features which separate two or more classes of objects

Decision tree for predicting credit worthiness



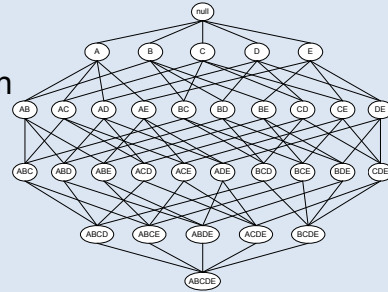
© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: association analysis

- find groups of variables that are strongly associated
- given two data sets, Canonical Correlation Analysis (CCA) is used for discovery and quantification of associations between two sets of variables
- in data mining, frequent pattern mining and association rule mining is used to find groups of related variables in binary transaction data



© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: dimensionality reduction

- purpose
 - capture key features of the data
 - avoid curse of dimensionality
 - reduce amount of time and memory required by data mining algorithms
 - allow data to be more easily visualized
 - may help to eliminate irrelevant features or reduce noise
- techniques
 - feature selection (manual or automatic)
 - Multidimensional Scaling (MDS)
 - **Principal Components Analysis (PCA)**
 - correspondence analysis is a similar technique for categorical data
 - Singular Value Decomposition (SVD)
 - factor analysis
 - reduces large number of variables to a smaller number of (hopefully) more interpretable and useful factors for modeling purposes
 - PCA can be regarded as one form of this
 - "Repairing Tom Swift's Electric Factor Analysis Machine", Preacher and MacCallum, 2003, Understanding Statistics, 2(1), 13–43
 - many other techniques, including supervised and non-linear approaches

© 2010 Regents of the University of Minnesota. All rights reserved.

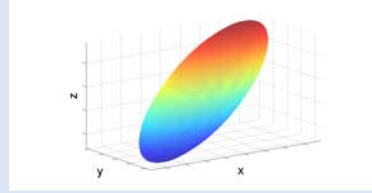
Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

multivariate analysis techniques: principal component analysis

- **basic concepts**

- temporarily, assume data is multivariate normal
- then it has an ellipsoidal distribution in multiple dimensions
- If the data has p dimensions, then PCA finds p new variables which are linear combinations of the old variables
- most simply, the new variables represent a removal of the mean and a rotation of the coordinate axes to match the ellipsoid axes
- mathematically, the linear combinations (components) are defined by the eigenvectors of the covariance matrix
- thus, the first new variable captures as much variability in the data as possible with a single linear projection
- additional variables (components) capture the maximum amount of variability subject to the constraint that they must be orthogonal (uncorrelated) to the previous variables
- dimensionality reduction is achieved by keeping only some of the components
- in practice the covariance matrix is estimated from the data and the coordinates (scores) of each data point is given in the new coordinate system



© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

principal component example



© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS for PCA: step by step

- a brief definition of PCA
 - assume we have p random variables, $\mathbf{x} = (x_1, \dots, x_p)'$, the variance-covariance matrix for \mathbf{x} is Σ
 - for the 1st principal component, we are looking for a vector, $\mathbf{a}_1 = (a_{11}, \dots, a_{1p})'$, such that
 - $\text{var}(\mathbf{a}_1' \mathbf{x} = a_{11}x_1 + \dots + a_{1p}x_p)$ is maximized among all linear combinations of \mathbf{x} and
 - $\mathbf{a}_1' \mathbf{a}_1 = 1$
 - in addition to the above conditions, any additional principal component, $\mathbf{a}_j' \mathbf{x}$, must also be uncorrelated with all previous $(j - 1)$ principal components

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS for PCA: step by step

- principal components are uncorrelated with each other
- the first principal component accounts for the maximum variation, the second principal component accounts for the highest variation among all the linear combinations uncorrelated with the 1st, etc.
- the first k principal components are the best linear predictors of the original variables among all possible sets of k variables
- the first k principal components provide the best possible fit to the data points as measured by the sum of squared perpendicular distances from each data point to the linear subspace spanned by these k principal components

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

Data set

- national track records for women from the *IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympics*
 - note: the dataset for women was selected for illustration purpose only
- seven race events are studied here:
 - 100, 200, 400, 800, 1500, 3000 meters, and Marathon
- first three events are recorded in seconds and the remaining four events are in minutes
- objective of the study:
 - overall performance in the track events

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



Data set

- 1st 35 records in the data set

	m100	m200	m400	m800	m1500	m3000	marathon	country
1	11.81	22.54	54.5	2.15	4.43	9.79	178.52	Argentina
2	11.2	22.35	51.08	1.98	4.13	9.08	152.37	Australia
3	11.43	23.09	59.62	1.99	4.22	9.34	159.37	Austria
4	11.41	23.04	52	2	4.14	8.08	157.05	Belgium
5	11.46	23.05	53.3	2.16	4.58	9.81	169.98	Bermuda
8	11.31	23.17	52.6	2.1	4.49	9.77	168.75	Brazil
7	12.14	24.47	55	2.18	4.45	9.51	191.02	Burma
6	11	22.25	50.06	2	4.08	8.81	149.45	Canada
9	12	24.52	54.9	2.05	4.23	9.37	171.38	Chile
10	11.95	24.41	54.97	2.08	4.33	9.31	168.48	China
11	11.6	24	53.26	2.11	4.35	9.46	165.42	Columbia
12	12.9	27.1	60.4	2.3	4.84	11.1	233.22	Cookis
13	11.96	24.6	55.25	2.21	4.05	10.43	171.8	Costa
14	11.09	21.97	47.99	1.89	4.14	8.92	158.05	Czech
15	11.42	23.52	53.6	2.03	4.18	8.71	151.75	Denmark
16	11.79	24.05	56.05	2.24	4.74	9.89	203.08	Dominica
17	11.13	22.39	50.14	2.03	4.1	8.52	154.23	Finland
18	11.15	22.59	51.73	2	4.14	8.98	155.27	France
19	10.81	21.71	48.16	1.93	3.96	8.75	157.68	GDR
20	11.01	22.39	49.75	1.95	4.03	8.59	148.53	FRG
21	11	22.13	50.46	1.98	4.03	8.62	149.72	GB&NI
22	11.79	24.08	54.93	2.07	4.35	9.87	182.2	Greece
23	11.84	24.54	56.09	2.28	4.86	10.54	215.08	Guatemala
24	11.45	23.06	51.5	2.01	4.14	8.98	156.37	Hungary
25	11.95	24.28	53.6	2.1	4.32	9.98	188.03	India
26	11.85	24.24	55.34	2.22	4.61	10.02	201.26	Indonesia
27	11.43	23.51	53.24	2.06	4.11	8.89	149.38	Ireland
28	11.45	23.57	54.9	2.1	4.25	9.37	160.49	Israel
29	11.29	23	52.01	1.96	3.98	8.63	151.62	Italy
30	11.73	24	53.73	2.09	4.35	9.2	150.5	Japan
31	11.73	23.88	52.7	2	4.15	9.2	181.05	Kenya
32	11.96	24.49	55.7	2.15	4.42	9.62	184.05	Korea
33	12.25	25.76	61.2	1.97	4.25	9.35	173.17	DR/Korea
34	12.03	24.96	56.1	2.07	4.38	9.64	174.68	Luxembourg
35	12.23	24.21	55.09	2.19	4.69	10.48	182.17	Malaysia

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



SAS code for PCA: step by step

```

options ls=64 nodate nonumber ;
data track ;
  infile 'womentrack.dat' firstobs = 1 ;
  input m100 m200 m400 m800 m1500 m3000 marathon country$ ;
title1 "1984 track data for women" ;
proc princomp data = track out = pctrack ;
  var m100 m200 m400 m800 m1500 m3000 marathon ;
  title2 "PCA on the 'time' variables" ;
run ;

```

- the procedure *princomp* analyzes the correlation matrix by default
- choose the option “cov” for using the covariance matrix
- if “cov” is used, make sure the variables are measured on comparable scales
- default PCs’ names (Prin1, ..., Prin7) can be changed by the “prefix” option with the proc statement, e.g. prefix=PC

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

Correlation Matrix							
	m100	m200	m400	m800	m1500	m3000	marathon
m100	1.0000	0.9528	0.8347	0.7277	0.7284	0.7417	0.6863
m200	0.9528	1.0000	0.8570	0.7241	0.6984	0.7099	0.6856
m400	0.8347	0.8570	1.0000	0.8984	0.7878	0.7776	0.7054
m800	0.7277	0.7241	0.8984	1.0000	0.9016	0.8636	0.7793
m1500	0.7284	0.6984	0.7878	0.9016	1.0000	0.9692	0.8779
m3000	0.7417	0.7099	0.7776	0.8636	0.9692	1.0000	0.8998
marathon	0.6863	0.6856	0.7054	0.7793	0.8779	0.8998	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.80568576	5.15204024	0.8294	0.8294
2	0.65364552	0.35376309	0.0934	0.9228
3	0.29988243	0.17440494	0.0428	0.9656
4	0.12547749	0.07166058	0.0179	0.9835
5	0.05381692	0.01476763	0.0077	0.9912
6	0.03904928	0.01660668	0.0056	0.9968
7	0.02244260		0.0032	1.0000

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
m100	0.368356	0.490060	0.286012	-.319386
m200	0.365364	0.536580	0.229819	0.083302
m400	0.381610	0.246538	-.515367	0.347377
m800	0.384559	-.155402	-.584526	0.042076
m1500	0.389104	-.360409	-.012912	-.429539
m3000	0.388866	-.347539	0.152728	-.363120
marathon	0.367004	-.369208	0.484370	0.672497

Eigenvectors			
	Prin5	Prin6	Prin7
m100	0.231169	0.619825	0.052177
m200	0.041455	-.710765	-.109225
m400	-.572178	0.190946	0.208497
m800	0.620324	-.019089	-.315210
m1500	0.030261	-.231248	0.692562
m3000	-.463355	0.009277	-.598359
marathon	0.130536	0.142281	0.069598

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- obtain principal components from the output:
 - let Σ be the variance-covariance or correlation matrix of p variables x_1, \dots, x_p
 - let $\lambda_1, \dots, \lambda_j, \dots, \lambda_p$ be the eigenvalues of Σ and $\mathbf{a}_1, \dots, \mathbf{a}_p$ be the corresponding eigenvectors
 - here we assume λ_1 is the largest eigenvalue, \dots , and λ_p is the smallest one
 - $\mathbf{a}_1' \mathbf{x} = a_{11} x_1 + \dots + a_{1p} x_p$

where $\mathbf{a}_1 = (a_{11}, \dots, a_{1p})'$, $\mathbf{a}_1' \mathbf{a}_1 = 1$, and $\text{var}(\mathbf{a}_1' \mathbf{x})$ is the maximum among all linear combination of \mathbf{x}

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- from the output:
 - the 1st principal component $\mathbf{a}'_1\mathbf{x}$ (i.e. PC₁) is
 $.3684x_1 + .3654x_2 + .3816x_3 + .3846x_4 + .3891x_5 + .3889x_6 + .367x_7$
 - the variance of $\mathbf{a}'_1\mathbf{x}$ is $\lambda_1 = 5.8057$
 - $\mathbf{a}'_1\mathbf{x}$ explains about $82.94\% = \frac{5.8057}{\text{total variance}} 100 = \frac{5.8070}{7} 100$ of the total variance
 - it measures a weighted average of all 7 races with about equal weight
 - the 2nd principal component (i.e. PC₂) is
 $.49x_1 + .5366x_2 + .2465x_3 - .1554x_4 - .3604x_5 - .3475x_6 + .3692x_7$
 - $\lambda_2 = 0.6536$ explains 9.34% of the total variance

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- observations from the result:
 - the 1st principal component (with variance λ_1) explains about 83% of the total variation in the data
 - the first three principal components all together actually explain about 97% of the total variation
 - all coefficient elements of the 1st eigenvector are positive, which can be considered as a measure of the average overall performance
 - the coefficients of the 2nd eigenvector measure the difference between the first three races (short distance) and the last four races (long distance)

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

```

options ls=64 nodate nonumber ;
title1 "1984 Women Track data" ;
proc princomp data=track out=pctrack ;
var m100 m200 m400 m800 m1500 m3000 marathon ;
title2 'PCA of the track data: Time' ;
run ;

proc sort data=pctrack ;
by prin1 ;
proc print ;
id country ;
var prin1 ;
title2 'Rankings by the First Principal Component: Time' ;
run ;

```

- all principal components are saved in the “out” file
- overall performance can be ordered for all countries based on the first principal component
- scores of principal components can be positive, negative, or zero because of the centering in the correlation or covariance matrix

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

1984 Women Track data
Rankings by the First Principal Component: Time

country	Prin1	country	Prin1	country	Prin1
GDR	-3.50602	Kenya	-0.43089	Malasiya	2.34054
USSR	-3.46469	Spain	-0.35565	Costa	2.61923
USA	-3.33581	Portugal	-0.22428	Guatemala	3.22730
Czech	-3.05380	Israel	-0.14297	Guinea	3.98086
FRG	-2.92578	Brazil	-0.11840	Mauritiu	4.23385
GB&NI	-2.78316	Mexico	-0.06348	Cookis	6.07728
Poland	-2.67210	Japan	-0.05923	WSamoa	8.33288
Canada	-2.60813	Columbia	0.14157		
Finland	-2.18184	Bermuda	0.38782		
Italy	-2.13954	DPRKorea	0.46230		
Australi	-2.09355	Argentin	0.52726		
Rumania	-2.02983	Chile	0.54783		
France	-1.89217	China	0.64127		
Sweden	-1.82775	Greece	0.81425		
Netherla	-1.79443	India	1.01454		
NZealand	-1.51126	Korea	1.23386		
Belgium	-1.50999	Luxembou	1.30174		
Norway	-1.48301	Turkey	1.60820		
Hungary	-1.47721	Philippi	1.64019		
Austria	-1.38044	Burma	1.68204		
Switzerl	-1.34665	Thailand	1.95318		
Ireland	-1.11735	Singapor	1.97013		
Denmark	-1.11638	Indonesi	2.11236		
Taipei	-0.50012	Dominica	2.29544		

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

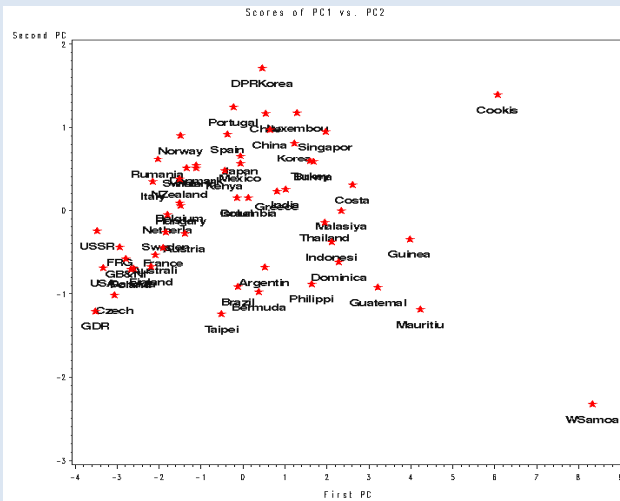
Rankings by the 2nd Principal Component (for time)

country	Prin1	Prin2
WSamoa	8.33288	-2.32698
Taipei	-0.50012	-1.23465
GDR	-3.50602	-1.20250
Mauritiu	4.23385	-1.18020
Czech	-3.05380	-1.01273
Bermuda	0.38782	-0.97641
Guatemal	3.22730	-0.91906
Brazil	-0.11840	-0.91152
Philippi	1.64019	-0.87604
Poland	-2.67210	-0.70232
Canada	-2.60813	-0.69529
USA	-3.33581	-0.68510
Argentin	0.52726	-0.67472
Finland	-2.18184	-0.67025
Dominica	2.29544	-0.61610
GB&NI	-2.78316	-0.57835
Australi	-2.09355	-0.53280
France	-1.89217	-0.44383
FRG	-2.92578	-0.43714
Indonesi	2.11236	-0.37827
Guinea	3.98086	-0.34024
Austria	-1.38044	-0.27500
Sweden	-1.82775	-0.25482
USSR	-3.46469	-0.24508

© 2010 Regents of the University of Minnesota. All rights reserved.

SAS code for PCA: step by step

- exploring clusters or other patterns



© 2010 Regents of the University of Minnesota. All rights reserved.

SAS code for PCA: step by step

- PCA starts with the correlation matrix of the data
 - based on a correlation matrix, PCA scores corresponding to i^{th} observations are defined as:

$$\mathbf{a}'_1 \mathbf{D}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \dots, \mathbf{a}'_j \mathbf{D}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \dots, \mathbf{a}'_p \mathbf{D}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}})$$

$i = 1, \dots, n$ where \mathbf{D} represents the diagonal elements of the covariance matrix; here $\mathbf{D}^{-1/2}$ is for standardization.
 \mathbf{x}_i represents the data vector and $\bar{\mathbf{x}}$ represents the p by 1 sample mean; $\mathbf{a}_j, j = 1, \dots, p$, are the p eigenvectors of the correlation matrix
 - this is equivalent to forcing all variables to have equal variance and the total variance to be equal to p

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- PCA starts with a covariance matrix of the data
 - PCA scores corresponding to the i^{th} observations in the covariance case are defined as:

$$\mathbf{a}'_1 (\mathbf{x}_i - \bar{\mathbf{x}}), \dots, \mathbf{a}'_j (\mathbf{x}_i - \bar{\mathbf{x}}), \dots, \mathbf{a}'_p (\mathbf{x}_i - \bar{\mathbf{x}}),$$

$i = 1, \dots, n$

where $\mathbf{a}'_1, \dots, \mathbf{a}'_j, \dots, \mathbf{a}'_p$ be the p transposed eigenvectors of the covariance matrix;
 \mathbf{x}_i represents the data vector and $\bar{\mathbf{x}}$ represents the p by 1 sample mean
 - in other words, scores are computed from the centered rather than standardized variables

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- correlation or covariance? which one should be used?
 - PCA results can be very different based on the two matrices
 - no obvious relationship between the eigenvalues and eigenvectors calculated from the two matrices
 - if the variables are on different scales and their variances are not on the same magnitude, using the correlation matrix would hide this fact
 - if the covariance matrix is analyzed, variables with large (small) variances to be more strongly associated with components with large (small) eigenvalues

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- re-examine the data “1984 national track records for women from 55 countries”
 - variables are the time they took to finish the races
 - short distance races (m100, m200, and m400) are recorded in seconds
 - long distance races (m800, m1500, m3000, and marathon) are recorded in minutes
 - the variances are very different (see the next a few slides)
 - the analysis that has done was using the sample correlation matrix (by default with the procedure)

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- covariance matrix on the “time” variables

Covariance Matrix, DF = 54							
	m100	m200	m400	m800	m1500	m3000	marathon
m100	0.204	0.479	1.011	2.137	6.570	16.589	566.662
m200	0.479	1.234	2.550	5.224	15.476	39.010	1390.718
m400	1.011	2.550	7.173	15.625	42.087	103.014	3449.548
m800	2.137	5.224	15.625	42.165	116.773	277.348	9238.944
m1500	6.570	15.476	42.087	116.773	397.824	956.094	31970.828
m3000	16.589	39.010	103.014	277.348	956.094	2446.306	81258.002
marathon	566.662	1390.718	3449.548	9238.944	31970.828	81258.002	3333445.934

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- simple statistics on the “time” variables

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
m100	55	11.61855	0.45221	11.60000	10.79000	12.90000
m200	55	23.64164	1.11106	23.54000	21.71000	27.10000
m400	55	53.40582	2.67834	53.30000	47.99000	60.40000
m800	55	124.58182	6.49345	123.00000	113.40000	139.80000
m1500	55	259.52727	19.94553	255.00000	232.20000	348.60000
m3000	55	566.85818	49.46015	560.40000	507.00000	782.40000
marathon	55	10395	1826	9879	8563	18360

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

```

data track ;
  infile 'womentrack.dat' firstobs=1 ;
  input m100 m200 m400 m800 m1500 m3000 marathon country$ ;
run ;

data speeds ; set track ;
  title1 "Speeds for women track races" ;
  /* convert the time unit from minutes to seconds */
  m800=m800*60 ;
  m1500=m1500*60 ;
  m3000=m3000*60 ;
  marathon=marathon*60 ;
  /* calculate the speed for these races */
  x1=100/m100 ;
  x2=200/m200 ;
  x3=400/m400 ;
  x4=800/m800 ;
  x5=1500/m1500 ;
  x6=3000/m3000 ;
  x7=42195/marathon ;
run ;

data speed ; set speeds ;
  keep country x1-x7 ;
run ;
proc export data = speed outfile = "wspeed.dat" dbms = dlm replace ;
run ;

```

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- covariance matrix on the “speed” variables

Covariance Matrix, DF = 54							
	x1	x2	x3	x4	x5	x6	x7
x1	0.1095994742	0.1237832898	0.1038905488	0.0795460124	0.0991434510	0.1031927674	0.1348335245
x2	0.1237832898	0.1533183476	0.1264920245	0.0939865908	0.1136836024	0.1174087274	0.1582593575
x3	0.1038905488	0.1264920245	0.1408367011	0.1111527230	0.1217076143	0.1222224807	0.1518052167
x4	0.0795460124	0.0939865908	0.1111527230	0.1085419120	0.1220487003	0.1199176569	0.1468255042
x5	0.0991434510	0.1136836024	0.1217076143	0.1220487003	0.1624693091	0.1617652421	0.1963723872
x6	0.1031927674	0.1174087274	0.1222224807	0.1199176569	0.1617652421	0.1734350606	0.2097175807
x7	0.1348335245	0.1582593575	0.1518052167	0.1468255042	0.1963723872	0.2097175807	0.3215983865

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- simple statistics on the “speed” variables

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
x1	55	8.61956	0.33106	8.62069	7.75194	9.26784
x2	55	8.47768	0.39156	8.49618	7.38007	9.21234
x3	55	7.50826	0.37528	7.50469	6.62252	8.33507
x4	55	6.43832	0.32946	6.50407	5.72246	7.05467
x5	55	5.80989	0.40307	5.88235	4.30293	6.45995
x6	55	5.32765	0.41646	5.35332	3.83436	5.91716
x7	55	4.15434	0.56710	4.27118	2.29820	4.92748

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 **UNIVERSITY OF MINNESOTA**
Driven to Discover™

SAS code for PCA: step by step

```
proc princomp data = speeds cov out = pcspeeds ;
  var x1-x7 ;
  title2 "PCA on speed variables for the track data" ;
run ;

proc sort data = pcspeeds ;
  by prin1 ;
run ;

proc print ;
  id country ;
  var prin1 ;
  title1 "Rankings by the 1st Principal Component on the speed variables" ;
run ;

data labels ; set pcspeeds ;
  retain xsys '2' ysys '2' ;
  length text $12 function $8 ;
  text=country ;
  style='swissb' ;
  color='black' ;
  y=prin2 ;
  x=prin1 ;
  size=1.2 ;
  position='8' ;
  function = 'LABEL' ;
run ;

options hsize=8in vsize=8in ;
title1 h=1.2 "first two principal components of speed variables" ;
title2 h=1.2 "scatter plot of the principal components' scores" ;
proc gplot data=pcspeeds ;
  plot prin2*prin1/annotate=labels ;
  label prin1='First PC'
        prin2='Second PC' ;
  symbol font=marker value=V color=gold ;
run ;
```

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 **UNIVERSITY OF MINNESOTA**
Driven to Discover™

SAS code for PCA: step by step

PCA on speed variables for the track data

The PRINCOMP Procedure

Total Variance 1.1697991911

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	0.97910541	0.88053887	0.8370	0.8370
2	0.09856654	0.04551016	0.0843	0.9212
3	0.05305638	0.02981429	0.0454	0.9666
4	0.02324210	0.01603734	0.0199	0.9865
5	0.00720475	0.00198981	0.0062	0.9926
6	0.00521494	0.00180586	0.0045	0.9971
7	0.00340907		0.0029	1.0000

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

PCA on speed variables for the track data

The PRINCOMP Procedure

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
x1	0.290843	0.426967	-.250363	0.329132
x2	0.341929	0.558190	-.320174	0.132018
x3	0.338593	0.381781	0.320836	-.537041
x4	0.305423	0.007925	0.475251	-.309251
x5	0.385868	-.197142	0.372461	0.362247
x6	0.399608	-.253973	0.214598	0.474764
x7	0.531023	-.506889	-.566770	-.365473

Eigenvectors

	Prin5	Prin6	Prin7
x1	0.155822	0.727883	-.089472
x2	0.099659	-.628763	0.215159
x3	-.476833	0.053809	-.343489
x4	0.504602	0.137220	0.558303
x5	0.364538	-.225856	-.598658
x6	-.590741	0.023273	0.393500
x7	0.043953	0.039572	-.052738

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research



UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- PCA from the speed variables on the track data
 - now the 1st principal component is

$$\mathbf{a}_1' \mathbf{x} = a_{11} x_1 + \dots + a_{1p} x_p$$

$$= .29x_1 + .34x_2 + .34x_3 + .31x_4 + .39x_5 + .4x_6 + .53x_7$$
 - its variance is $\lambda_1 = 0.9791$ and explains about 83.7% variation, which is:

$$0.9791/(\text{total variance}) = 0.9791/1.1697991 = 0.8369813$$
 - it is still a weighted average of the speeds except
 - Marathon contributes slightly more to the weighted average
 - 100 meter race contributes relatively less

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- PCA from the speed variables on the track data
 - now the 2nd principal component is

$$\mathbf{a}_2' \mathbf{x} = a_{21} x_1 + \dots + a_{2p} x_p$$

$$= .43x_1 + .56x_2 + .38x_3 + .008x_4 - .2x_5 - .25x_6 - .51x_7$$
 - its variance is $\lambda_2 = 0.0986$ and explains about 8.4% variation, which is:

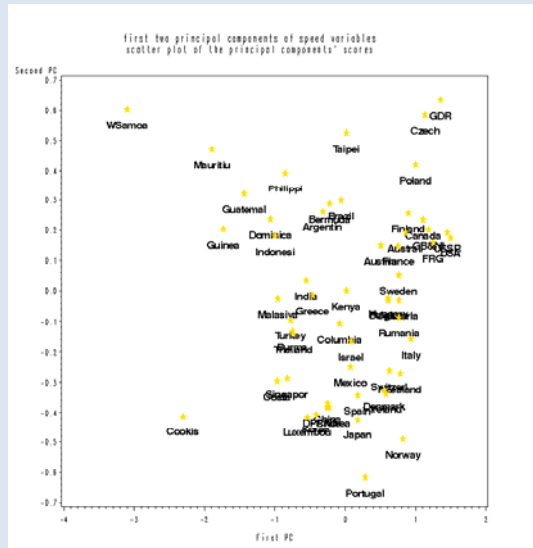
$$0.0986/(\text{total variance}) = 0.0986/1.1697991 = 0.08428797$$
 - 2nd principal component measures a difference between short distance and long distance speeds
 - 800 meter speed contributes very little to the 2nd PC
 - the first two PCs together explain more than 92% of the total variation

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step



© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

Rankings by the 1st Principal Component on the speed variables

country	Prin1	country	Prin1	country	Prin1
WSamoa	-3.09597	Brazil	-0.05736	Canada	1.11069
Cookis	-2.30587	Kenya	0.01762	Czech	1.14187
Mauritiu	-1.88921	Taipei	0.01992	GB&NI	1.18965
Guinea	-1.72851	Mexico	0.07458	FRG	1.24939
Guatemala	-1.43046	Israel	0.08567	GDR	1.35933
Dominica	-1.05942	Japan	0.17758	USSR	1.45505
Indonesi	-0.99198	Spain	0.18103	USA	1.49962
Costa	-0.96112	Portugal	0.28669		
Malasiya	-0.95346	Austria	0.50540		
Philippi	-0.84551	Denmark	0.55867		
Singapor	-0.82385	Ireland	0.58710		
Turkey	-0.77728	Belgium	0.61260		
Burma	-0.74651	Hungary	0.61361		
Thailand	-0.73757	Switzerl	0.63133		
India	-0.55369	France	0.75901		
Luxembou	-0.53930	Sweden	0.76182		
Greece	-0.45989	Netherla	0.76689		
Korea	-0.41383	Rumania	0.76788		
Argentina	-0.31551	NZealand	0.77987		
DPRKorea	-0.26088	Norway	0.82482		
China	-0.24364	Australi	0.86067		
Chile	-0.23314	Finland	0.89759		
Bermuda	-0.22015	Italy	0.93873		
Columbia	-0.07435	Poland	1.00439		

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code for PCA: step by step

- correlations between variables and principal components:
 - recall the variance-covariance matrix Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ and the corresponding eigenvectors are $\mathbf{a}_1, \dots, \mathbf{a}_p$
 - principal components are $\mathbf{a}'_1 \mathbf{x}, \mathbf{a}'_2 \mathbf{x}, \dots, \mathbf{a}'_p \mathbf{x}$
 - variance of the principal components are $\text{var}(\mathbf{a}'_i \mathbf{x}) = \lambda_1, \dots, \text{var}(\mathbf{a}'_p \mathbf{x}) = \lambda_p$
 - covariance between the i^{th} variable x_i and the j^{th} principal component is $\lambda_j a_{ji}$
 - correlation coefficient $\text{corr}(x_i, y_j) = a_{ji} \sqrt{\lambda_j / \text{var}(x_i)}$

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS PCA and linear regression

- PCA and regression:
 - consider a linear regression problem:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$
 - y is the dependent variable and x_1, \dots, x_p are the p independent variables
 - the problem is to predict y by estimating the β s
 - assume the random error ε from observations has mean zero and constant variance σ^2 and uncorrelated
 - the estimated function $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$ is the estimated value for the response y for a given set of predictor values for x_1, \dots, x_p

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code and linear regression

- with n observations, the n equations can be written in a matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
 - the vector \mathbf{y} contains the observed values for the response variable y and $\boldsymbol{\beta}$ is the vector of parameters
 - the least-squares estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ is obtained by solving the normal equation: $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$
 - if $\mathbf{X}'\mathbf{X}$ is of full rank, the unique solution to $\boldsymbol{\beta}$ from the normal equation is: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
 - otherwise, a generalized inverse or some kind of pseudoinverse may be used. However there is no unique solution to the normal equation

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code and linear regression

- what about if some of the eigenvalues of $\mathbf{X}'\mathbf{X}$ are not exactly zero, but close to?

in other words, if $\mathbf{X}'\mathbf{X}$ is ill-conditioned, this creates a problem of multicollinearity in regression analysis

 - the estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are unstable and inflate the estimates $\hat{\boldsymbol{\beta}}$
 - this inflates predictions of the response variable \mathbf{y} for given sets of predictor \mathbf{X} values
- to handle multicollinearity, some of the techniques are
 - ridge regression
 - principal component regression

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code and principal component regression

- ideas behind the approach:
 - look for estimators with a smaller mean square error than the usual estimator by allowing a small bias in the estimation
 - ignore the last few principal components in the case of the principal component regression
- procedure of principal component regression
 - compute the principal components for x_1, \dots, x_p
 - ignore the last a few principal components that explain negligible percentage of variability
 - perform linear regression with the remaining PCs

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

SAS code and linear regression

- the linear regression model on the PCs is

$$y = \xi_0 + \xi_1 PC_{1i} + \dots + \xi_{ri} PC_{ri} + \varepsilon_i, i = 1, \dots, n$$
 where $PC_{1i}, \dots, PC_{ri}, i = 1, \dots, n$ are the first r principal component scores
- principal components are not correlated so for the regression using these PCs, there is no multicollinearity
- all original predictor variables are still in this model even though only the first r principal components are used
- some selection techniques may be used to decide which among the first r PCs should be included in the model
 - selection: rsquare, cp, or step-type procedures if many PCs are considered initially

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

Biplot

- biplot for multivariate data matrix
 - represent all observations and variables in one plot
 - $X_{n \times p}$ is a standardized (either with zero column mean or column mean zero and unit σ) data matrix with n rows of observations on p variables
 - consider SVD on X: $X = USV'$
where S is the r by r diagonal matrix of all positive singular values: $\lambda_1 > \lambda_2 > \dots > \lambda_r > 0$
 - $X = US^\alpha S^{1-\alpha} V' = (US^\alpha)(VS^{(1-\alpha)})' = GH'$
 - some implementations of biplot use $\alpha = 1, 0.5, 0$
 - $X \approx \lambda_1 u_1 v_1' + \lambda_2 u_2 v_2'$

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™

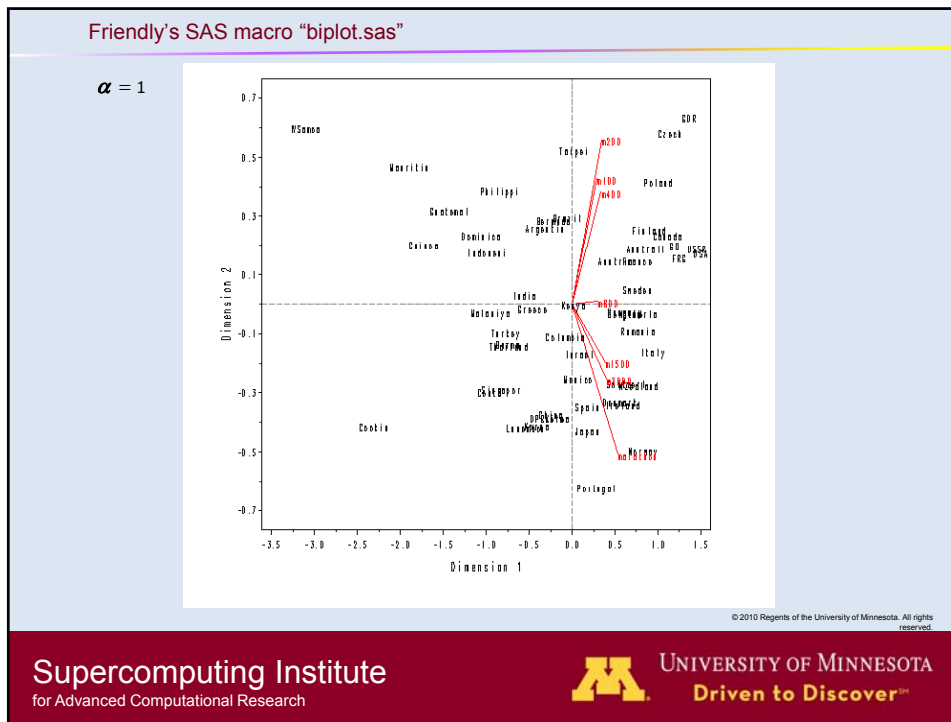
Biplot for PCA

- biplot for principal component analysis
 - using the first two PCs: $X_{n \times p} \approx G_{n \times 2} H'_{2 \times p}$
 - $G_{n \times 2}$ is the n by 2 matrix of scores on the first 2 principal components
 - $H'_{2 \times p}$ is the 2 by p matrix and its two rows are the first 2 eigenvectors
 - under the representation with $\alpha = 1$
 - locations in the biplot of the n points are the same as the score plot of the first 2 principal components
 - the Euclidean distance between the i^{th} and j^{th} rows of G approximates the Euclidean distance between the i^{th} and the j^{th} rows in the data matrix X

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

 UNIVERSITY OF MINNESOTA
Driven to Discover™



Summary

- PCA can be used to summarize data and detect linear relationships
- PCA can also be used to reduce the number of variables in multivariate analysis like in a regression or in a clustering analysis
- Plots from PCA can be very useful in exploring data and assist further analysis on the data
- Very easy to perform a PCA with SAS
- PCA is just one of number of multivariate analysis techniques

The University of Minnesota is an equal opportunity educator and employer. This PowerPoint is available in alternative formats upon request. Direct requests to Minnesota Supercomputing Institute, 599 Walker Library, 117 Pleasant St. SE, Minneapolis, Minnesota, 55455, 612-624-0528.

© 2010 Regents of the University of Minnesota. All rights reserved.

Supercomputing Institute
for Advanced Computational Research

UNIVERSITY OF MINNESOTA
Driven to Discover™