

Summer School in Statistics for  
Astronomers V

June 3, 2009

Multivariate Analysis

James L Rosenberger

Acknowledgements:

Donald Richards

Department of Statistics

Center for Astrostatistics

Penn State University

**Multivariate analysis:** The statistical analysis of data containing observations on two or more *variables* each measured on a set of *objects* or *cases*.

C. Wolf, K. Meisenheimer, M. Kleinheinrich, A. Borch, S. Dye, M. Gray, L. Wisotzki, E. F. Bell, H.-W. Rix, A. Cimatti, G. Hasinger, and G. Szokoly: "A catalogue of the Chandra Deep Field South with multi-colour classification and photometric redshifts from COMBO-17," *Astron. & Astrophys.*, 2004.

65 variables: Rmag, e.Rmag, ApDRmag, mu-max, Mcz, e.Mcz, MCzml, ..., IFD, e.IFD

63,501 objects: galaxies

<http://astrostatistics.psu.edu/datasets/COMBO17.dat>

Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
2	5	6	8	9	10	12	14
24.995	24.214	0.832	1.400	0.64	-17.67	-17.54	-17.76
25.013	25.303	0.927	0.864	0.41	-18.28	-17.86	-18.20
24.246	23.511	1.202	1.217	0.92	-19.75	-19.91	-20.41
25.203	24.948	0.912	0.776	0.39	-17.83	-17.39	-17.67
25.504	24.934	0.848	1.330	1.45	-17.69	-18.40	-19.37
23.740	24.609	0.882	0.877	0.52	-19.22	-18.11	-18.70
25.706	25.271	0.896	0.870	1.31	-17.09	-16.06	-16.23
25.139	25.376	0.930	0.877	1.84	-16.87	-16.49	-17.01
24.699	24.611	0.774	0.821	1.03	-17.67	-17.68	-17.87
24.849	24.264	0.062	0.055	0.55	-11.63	-11.15	-11.32
25.309	25.598	0.874	0.878	1.14	-17.61	-16.90	-17.58
24.091	24.064	0.173	0.193	1.12	-13.76	-13.99	-14.41
25.219	25.050	1.109	1.400	1.76	-18.57	-18.49	-18.76
26.269	25.039	0.143	0.130	1.52	-10.95	-10.30	-11.82
23.596	23.885	0.626	0.680	0.78	-17.75	-18.21	-19.11
23.204	23.517	1.185	1.217	1.79	-20.50	-20.14	-20.30
25.161	25.189	0.921	0.947	1.68	-17.87	-16.13	-16.30
22.884	23.227	0.832	0.837	0.20	-19.81	-19.42	-19.64
24.346	24.589	0.793	0.757	1.86	-18.12	-18.11	-18.58
25.453	24.878	0.952	0.964	0.72	-17.77	-17.81	-18.06
25.911	24.994	0.921	0.890	0.96	-17.34	-17.59	-18.11
26.004	24.915	0.986	0.966	0.95	-17.38	-16.98	-17.30
26.803	25.232	1.044	1.400	0.78	-16.67	-18.17	-19.17
25.204	25.314	0.929	0.882	0.64	-18.05	-18.68	-19.63
25.357	24.735	0.901	0.875	1.69	-17.64	-17.48	-17.67
24.117	24.028	0.484	0.511	0.84	-16.64	-16.60	-16.83
26.108	25.342	0.763	1.400	1.07	-16.27	-16.39	-15.54
24.909	25.120	0.711	1.152	0.42	-17.09	-17.21	-17.85
24.474	24.681	1.044	1.096	0.69	-18.95	-18.95	-19.22
23.100	24.234	0.826	1.391	0.53	-19.61	-19.85	-20.28
22.009	22.633	0.340	0.323	2.88	-17.49	-17.64	-18.17
.							
.							
.							

## **The goals of multivariate analysis:**

### Generalize univariate statistical methods

- Multivariate means, variances, and covariances

- Multivariate probability distributions

### Reduce the number of variables

- Structural simplification

- Linear functions of variables (principal components)

### Investigate the dependence between variables

- Canonical correlations

### Statistical inference

- Confidence regions

- Multivariate regression

- Hypothesis testing

### Classify or cluster “similar” objects

- Discriminant analysis

- Cluster analysis

### Prediction

## Organizing the data

$p$ : The number of variables

$n$ : The number of objects (cases) (the sample size)

$x_{ij}$ : the  $i^{\text{th}}$  observation on the  $j^{\text{th}}$  variable

Data array or data matrix

		Variables			
		1	2	...	$p$
Objects	1	$x_{11}$	$x_{12}$	...	$x_{1p}$
	2	$x_{21}$	$x_{22}$	...	$x_{2p}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	$n$	$x_{n1}$	$x_{n2}$	...	$x_{np}$

Data matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

We write  $\mathbf{X}$  as  $n$  row or as  $p$  column vectors

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p]$$

Matrix methods are essential to multivariate analysis

We will need only small amounts of matrix methods, e.g.,

$\mathbf{A}^T$ : The transpose of  $\mathbf{A}$

$|\mathbf{A}|$ : The determinant of  $\mathbf{A}$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

## Descriptive Statistics

The sample mean of the  $j^{\text{th}}$  variable:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

The sample mean vector:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

The sample variance of the  $j^{\text{th}}$  variable:

$$s_{jj} = \frac{1}{n-1} \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2$$

The sample covariance of variables  $i$  and  $j$ :

$$s_{ij} = s_{ji} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

**[Question:** Why do we divide by  $(n-1)$  rather than  $n$ ?]

The sample covariance matrix:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

The sample correlation coefficient of variables  $i$  and  $j$ :

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

Note that  $r_{ii} = 1$  and  $r_{ij} = r_{ji}$

The sample correlation matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

$\mathbf{S}$  and  $\mathbf{R}$  are *symmetric*

$\mathbf{S}$  and  $\mathbf{R}$  are *positive semidefinite*:  $\mathbf{v}^T \mathbf{S} \mathbf{v} \geq 0$   
for any vector  $\mathbf{v}$ .

Equivalently,

$$s_{11} \geq 0, \begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix} \geq 0, \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} \geq 0,$$

etc.

If  $\mathbf{S}$  is singular so is  $\mathbf{R}$  and conversely.

If  $n \leq p$  then  $\mathbf{S}$  and  $\mathbf{R}$  will be *singular*:

$$|\mathbf{S}| = 0 \text{ and } |\mathbf{R}| = 0$$

Which practical astrophysicist would attempt a statistical analysis with 65 variables and a sample size smaller than 65?

$\mathbf{v}^T \mathbf{S} \mathbf{v} > 0$  is the variance of  $\mathbf{v}^T \mathbf{X}$

If  $n > p$  then, generally (*but not always*),  $\mathbf{S}$  and  $\mathbf{R}$  are strictly *positive definite*:

Then  $\text{Var}(\mathbf{v}^T \mathbf{X}) = \mathbf{v}^T \mathbf{S} \mathbf{v} > 0$  for any non-zero vector  $\mathbf{v}$

Equivalently,

$$s_{11} > 0, \begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix} > 0, \begin{vmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{vmatrix} > 0,$$

etc.

However, if  $n > p$  and  $|\mathbf{S}| = 0$  then for some  $\mathbf{v}$   $\text{Var}(\mathbf{v}^T \mathbf{X}) = 0$  implying  $\mathbf{v}^T \mathbf{X}$  is a constant and there is a linear relationship between the components of  $\mathbf{X}$

In this case, we can eliminate the dependent variables: **dimension reduction**

## The COMBO-17 data

Variables: Rmag,  $\mu_{\max}$ , Mcz, MCzml, chi2red, UjMAG, BjMAG, VjMAG

$p = 8$  and  $n = 3462$

The sample mean vector:

Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
23.939	24.182	0.729	0.770	1.167	-17.866	-17.749	-18.113

The sample covariance matrix:

	Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
Rmag	2.062	1.362	0.190	0.234	0.147	0.890	1.015	1.060
mumax	1.362	1.035	0.141	0.172	0.079	0.484	0.578	0.610
Mcz	0.190	0.141	0.102	0.105	-0.004	-0.438	-0.425	-0.428
MCzml	0.234	0.172	0.105	0.141	-0.009	-0.416	-0.414	-0.419
chi2red	0.147	0.079	-0.004	-0.009	0.466	0.201	0.204	0.221
UjMAG	0.890	0.484	-0.438	-0.416	0.201	3.863	3.890	3.946
BjMAG	1.015	0.578	-0.425	-0.414	0.204	3.890	4.500	4.219
VjMAG	1.060	0.610	-0.428	-0.419	0.221	3.946	4.219	4.375

## Advice given by some for Correlation Matrix:

- Use no more than two significant digits.
- Starting with the physically most important variable, reorder variables by descending correlations.
- Suppress diagonal entries to ease visual clutter.
- Suppress zeros before the decimal point.

### COMBO-17's correlation matrix

	Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
Rmag		.9	.4	.4	.2	.3	.3	.3
mumax	.9		.4	.5	.1	.2	.3	.3
Mcz	.4	.4		.9	-.0	-.7	-.6	-.6
MCzml	.4	.5	.9		-.0	-.6	-.5	-.5
chi2red	.2	.1	-.0	-.0		.2	.1	.2
UjMAG	.3	.2	-.7	-.6	.2		.9	1.0
BjMAG	.3	.3	-.6	-.5	.1	.9		1.0
VjMAG	.4	.3	-.6	-.5	.2	1.0	1.0	

**Reminder:** Correlations measure the strengths of linear relationships between variables *if* such relationships are valid

{UjMAG, BjMAG, VjMAG} are highly correlated; perhaps, two of them can be eliminated. Similar remarks apply to {Rmag, mumax} and {Mcz, Mczml}.

chi2red has small correlation with {mumax, Mcz, Mczml}; we would retain chi2red in the subsequent analysis

## Multivariate probability distributions

Find the *probability* that a galaxy chosen *at random* from the population of *all* COMBO-17 type galaxies satisfies

$$4 * Rmag + 3 * mumax + |Mcz-MCzml| - chi2red + (UjMAG + BjMAG)^2 + VjMAG^2 < 70?$$

$X_1$ : Rmag

$X_2$ : mumax

...

$X_7$ : BjMAG

$X_8$ : VjMAG

We wish to make probability statements about random *vectors*

$p$ -dimensional random vector:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

where  $X_1, \dots, X_p$  are random variables

$\mathbf{X}$  is a *continuous random vector* if  $X_1, \dots, X_p$  all are continuous random variables

We shall concentrate on continuous random vectors

Each nice  $\mathbf{X}$  has a prob. density function  $f$

Three important properties of the p.d.f.:

1.  $f(\mathbf{x}) \geq 0$  for all  $\mathbf{x} \in \mathbb{R}^p$

2. The total area below the graph of  $f$  is 1:

$$\int_{\mathbb{R}^p} f(\mathbf{x}) d\mathbf{x} = 1$$

3. For all  $t_1, \dots, t_p$ ,

$$P(X_1 \leq t_1, \dots, X_p \leq t_p) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_p} f(\mathbf{x}) d\mathbf{x}$$

**Reminder:** “Expected value,” an average over the *entire* population

The *mean vector*:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

where

$$\mu_i = E(X_i) = \int_{\mathbb{R}^p} x_i f(\mathbf{x}) d\mathbf{x}$$

is the mean of the  $i$ th component of  $\mathbf{X}$

The *covariance* between  $X_i$  and  $X_j$ :

$$\begin{aligned} \sigma_{ij} &= E(X_i - \mu_i)(X_j - \mu_j) \\ &= E(X_i X_j) - \mu_i \mu_j \end{aligned}$$

The *variance* of each  $X_i$ :

$$\sigma_{ii} = E(X_i - \mu_i)^2 = E(X_i^2) - \mu_i^2$$

The *covariance matrix* of  $\mathbf{X}$ :

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

An easy result:

$$\Sigma = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T$$

Also,

$$\Sigma = E(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

To avoid pathological cases, we assume that  $\Sigma$  is nonsingular

## Theory vs. Practice

### Population vs. Random Sample

All galaxies of COMBO-17 type	A sample from the COMBO-17 data set
Random vector $\mathbf{X}$	Random sample $\mathbf{x}_1, \dots, \mathbf{x}_n$
Population Mean $\boldsymbol{\mu} = E(\mathbf{X})$	Sample mean $\bar{\mathbf{x}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$
Popn. cov. matrix $\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T$	Sample cov. matrix, $S = \frac{1}{n-1} \times \sum (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T$

Laws of Large Numbers: In a technical sense,  
 $\bar{\mathbf{x}} \rightarrow \boldsymbol{\mu}$  and  $S \rightarrow \boldsymbol{\Sigma}$  as  $n \rightarrow \infty$

## The Multivariate Normal Distribution

$\mathbf{X} = [X_1, \dots, X_p]^T$ : A random vector whose possible values range over all of  $\mathbb{R}^p$

$\mathbf{X}$  has a *multivariate normal distribution* if has a probability density function of the form

$$f(\mathbf{x}) = \text{const.} \times \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$

where

$$\text{const.} = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}}$$

Standard notation:  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Special case,  $p = 1$ : Let  $\boldsymbol{\Sigma} = \sigma^2$ ; then

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

Special case,  $\Sigma$  diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

$$|\Sigma| = \sigma_1^2 \sigma_2^2 \cdots \sigma_p^2$$

$$\Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_p^{-2} \end{bmatrix}$$

$$f(\mathbf{x}) = \prod_{j=1}^p \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x_j - \mu_j}{\sigma_j} \right)^2 \right]$$

Conclusion:  $X_1, \dots, X_p$  are mutually independent and normally distributed

Recall:  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  if its p.d.f. is of the form

$$f(\mathbf{x}) = \text{const.} \times \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

where

$$\text{const.} = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}}$$

Facts:

$$\boldsymbol{\mu} = E(\mathbf{X}),$$

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$$

$$\int_{\mathbb{R}^p} f(\mathbf{x}) d\mathbf{x} = 1$$

If  $A$  is a  $k \times p$  matrix then

$$A\mathbf{X} + \mathbf{b} \sim N_k(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$$

Proof: Use Fourier transforms

Special cases:

$\mathbf{b} = \mathbf{0}$  and  $A = \mathbf{v}^T$  where  $\mathbf{v} \neq \mathbf{0}$ :

$$\mathbf{v}^T \mathbf{X} \sim N(\mathbf{v}^T \boldsymbol{\mu}, \mathbf{v}^T \Sigma \mathbf{v})$$

Note:  $\mathbf{v}^T \Sigma \mathbf{v} > 0$  since  $\Sigma$  is positive definite

$\mathbf{v} = [1, 0, \dots, 0]^T$ :  $X_1 \sim N(\mu_1, \sigma_{11})$

Similar argument: Each  $X_i \sim N(\mu_i, \sigma_{ii})$

Decompose  $\mathbf{X}$  into two subsets,  $\mathbf{X} = \begin{bmatrix} \mathbf{X}_u \\ \mathbf{X}_l \end{bmatrix}$

Similarly, decompose

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_l \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{ul} \\ \boldsymbol{\Sigma}_{lu} & \boldsymbol{\Sigma}_{ll} \end{bmatrix}$$

Then

$$\boldsymbol{\mu}_u = E(\mathbf{X}_u), \quad \boldsymbol{\mu}_l = E(\mathbf{X}_l)$$

$$\boldsymbol{\Sigma}_{uu} = \text{Cov}(\mathbf{X}_u), \quad \boldsymbol{\Sigma}_{ll} = \text{Cov}(\mathbf{X}_l)$$

$$\boldsymbol{\Sigma}_{ul} = \text{Cov}(\mathbf{X}_u, \mathbf{X}_l)$$

The marginal distribution of  $\mathbf{X}_u$ :

$$\mathbf{X}_u \sim N_u(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu})$$

The conditional distribution of  $\mathbf{X}_u | \mathbf{X}_l$ :

$$\mathbf{X}_u | \mathbf{X}_l \sim N_u(\dots, \dots)$$

If  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then  $\mathbf{v}^T \mathbf{X}$  has a 1-D normal distribution for every vector  $\mathbf{v} \in \mathbb{R}^p$

Conversely, if  $\mathbf{v}^T \mathbf{X}$  has a 1-D normal distribution for every  $\mathbf{v}$  then  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Proof: Fourier transforms again

(The assumption that an  $\mathbf{X}$  is normally distributed is very strong)

Let us use this result to construct an exploratory test of whether some COMBO-17 variables have a multivariate normal distribution

Choose several COMBO-17 variables, e.g.,

Rmag, mumax, Mcz, MCzml, chi2red, UjMAG,  
BjMAG, VjMAG

Use  $R$  to generate a “random” vector  $\mathbf{v} = [v_1, v_2, \dots, v_8]^T$

For each galaxy, calculate

$$v_1 * R_{\text{mag}} + v_2 * m_{\text{umax}} + \dots + v_8 * V_{\text{jMAG}}$$

This produces 3,462 such numbers ( $\mathbf{v}$ -scores)

Construct a Q-Q plot of all these  $\mathbf{v}$ -scores against the standard normal distribution

Study the plot to see if normality seems plausible

Repeat the exercise with a new random  $\mathbf{v}$

Repeat the exercise  $10^3$  times

Note: We need only those vectors for which  $v_1^2 + \dots + v_8^2 = 1$  (why?)

## Mardia's test for multivariate normality

If the data contain a substantial number of outliers then it goes against the hypothesis of multivariate normality

If one COMBO-17 variable is not normally distributed then the full set of variables does not have a multivariate normal distribution

In that case, we can try to transform the original variables to produce new variables which are normally distributed

Example: Box-Cox transformations, log transformations (a special case of Box-Cox)

For data sets arising from a multivariate normal distribution, we can perform accurate inference for the mean vector and covariance matrix

Variables (random vector):  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are unknown

Data (measurements):  $\mathbf{x}_1, \dots, \mathbf{x}_n$

Problem: Estimate  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$

$\bar{\mathbf{x}}$  is an unbiased and consistent estimator of  $\boldsymbol{\mu}$

$\bar{\mathbf{x}}$  is the MLE of  $\boldsymbol{\mu}$

The MLE of  $\boldsymbol{\Sigma}$  is  $\frac{n-1}{n}S$ ; this is not unbiased

The sample covariance matrix,  $S$ , is an unbiased estimator of  $\boldsymbol{\Sigma}$

Since  $S$  is close to being the MLE of  $\boldsymbol{\Sigma}$ , we estimate  $\boldsymbol{\Sigma}$  using  $S$

A confidence region for  $\mu$

Naive method: Using only the data on the  $i$ th variable, construct a confidence interval for each  $\mu_i$

Use the collection of confidence intervals as a confidence region for  $\mu$

Good news: This can be done using elementary statistical methods

Bad news: A collection of 95% confidence intervals, one for each  $\mu_i$ , does not result in a 95% confidence region for  $\mu$

Starting with individual intervals with lower confidence levels, we can achieve an overall 95% confidence level for the combined region

Bonferroni inequalities: Some difficult math formulas are needed to accomplish that goal

Worse news: The resulting confidence region for  $\mu$  is a rectangle

This is not consonant with a density function of the form

$$f(\mathbf{x}) = \text{const.} \times \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$

The contours of the graph of  $f(\mathbf{x})$  are ellipsoids, so we should derive an ellipsoidal confidence region for  $\mu$

Fact: Every positive definite symmetric matrix has a unique positive definite symmetric square root

$\Sigma^{-1/2}$ : The p.d. square-root of  $\Sigma^{-1}$

Recall (see p. 31): If  $A$  is a  $p \times p$  nonsingular matrix and  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  then

$$A\mathbf{X} + \mathbf{b} \sim N_p(A\boldsymbol{\mu} + \mathbf{b}, A\boldsymbol{\Sigma}A^T)$$

Set  $A = \Sigma^{-1/2}$ ,  $\mathbf{b} = -\Sigma^{-1/2}\boldsymbol{\mu}$

Then  $A\boldsymbol{\mu} + \mathbf{b} = \mathbf{0}$ ,  $A\boldsymbol{\Sigma}A^T = \Sigma^{-1/2}\boldsymbol{\Sigma}\Sigma^{-1/2} = I_p$

$$\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, I_p)$$

$I_p = \text{diag}(1, 1, \dots, 1)$ , a diagonal matrix

# Methods of Multivariate Analysis

Reduce the number of variables

- Structural simplification

- Linear functions of variables (Principal Components)

Investigate the dependence between variables

- Canonical correlations

Statistical inference

- Estimation

- Confidence regions

- Hypothesis testing

Classify or cluster “similar” objects

- Discriminant analysis

- Cluster analysis

Predict

- Multiple Regression

- Multivariate regression

## Principal Components Analysis (PCA)

COMBO-17:  $p = 65$  (wow!)

Can we reduce the dimension of the problem?

$\mathbf{X}$ : A  $p$ -dimensional random vector

Covariance matrix:  $\Sigma$

Solve for  $\lambda$ :  $|\Sigma - \lambda I| = 0$

Solutions:  $\lambda_1, \dots, \lambda_p$ , the *eigenvalues* of  $\Sigma$

Assume, for simplicity, that  $\lambda_1 > \dots > \lambda_p$

Solve for  $\mathbf{v}$ :  $\Sigma \mathbf{v} = \lambda_j \mathbf{v}$ ,  $j = 1, \dots, p$

Solution:  $\mathbf{v}_1, \dots, \mathbf{v}_p$ , the *eigenvectors* of  $\Sigma$

Scale each eigenvector to make its length 1

$\mathbf{v}_1, \dots, \mathbf{v}_p$  are orthogonal

The first PC: The linear combination  $\mathbf{v}^T \mathbf{X}$  such that

(i)  $\text{Var}(\mathbf{v}^T \mathbf{X})$  is maximal, and

(ii)  $\mathbf{v}^T \mathbf{v} = 1$

Maximize  $\text{Var}(\mathbf{v}^T \mathbf{X}) = \mathbf{v}^T \Sigma \mathbf{v}$  subject to  $\mathbf{v}^T \mathbf{v} = 1$

Lagrange multipliers

Solution:  $\mathbf{v} = \mathbf{v}_1$ , the first eigenvector of  $\Sigma$

$\mathbf{v}_1^T \mathbf{X}$  is the first principal component

The second PC: The linear combination  $\mathbf{v}^T \mathbf{X}$  such that

(i)  $\text{Var}(\mathbf{v}^T \mathbf{X})$  is maximal,

(ii)  $\mathbf{v}^T \mathbf{v} = 1$ , and

(iii)  $\mathbf{v}^T \mathbf{X}$  has zero correlation with the first PC

Maximize  $\text{Var}(\mathbf{v}^T \mathbf{X}) = \mathbf{v}^T \Sigma \mathbf{v}$  with  $\mathbf{v}^T \mathbf{v} = 1$  and  
 $\text{Cov}(\mathbf{v}^T \mathbf{X}, \mathbf{v}^T \mathbf{X}) \equiv \mathbf{v}^T \Sigma \mathbf{v} = 0$

Lagrange multipliers

Solution:  $\mathbf{v} = \mathbf{v}_2$ , the second eigenvector of  $\Sigma$

The  $k$ th PC: The linear combination  $\mathbf{v}^T \mathbf{X}$  such that

- (i)  $\text{Var}(\mathbf{v}^T \mathbf{X})$  is maximal,
- (ii)  $\mathbf{v}^T \mathbf{v} = 1$ , and
- (iii)  $\mathbf{v}^T \mathbf{X}$  has zero correlation with all prior PCs

Solution:  $\mathbf{v} = \mathbf{v}_k$ , the  $k$ th eigenvector of  $\Sigma$

The PCs are random variables

Simple matrix algebra:  $\text{Var}(\mathbf{v}_k^T \mathbf{X}) = \lambda_k$

$p$ -dimensional data:  $\mathbf{x}_1, \dots, \mathbf{x}_n$

$S$ : the sample covariance matrix

$\tilde{\lambda}_1 > \dots > \tilde{\lambda}_p$ : The eigenvalues of  $S$

Remarkable result:

$$\tilde{\lambda}_1 + \dots + \tilde{\lambda}_p = s_{11} + \dots + s_{pp}$$

$\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_p$ : The corresponding eigenvectors

$\tilde{\mathbf{v}}_1^T \mathbf{X}, \dots, \tilde{\mathbf{v}}_p^T \mathbf{X}$ : The sample PCs

$\tilde{\lambda}_1, \dots, \tilde{\lambda}_p$ : The estimated variances of the PCs

Basic idea: Use the sample PCs instead of  $\mathbf{X}$  to analyze the data

Example: (Johnson and Wichern)

$$S = \begin{bmatrix} 4.31 & 1.68 & 1.80 & 2.16 & -.25 \\ 1.68 & 1.77 & .59 & .18 & .17 \\ 1.80 & .59 & .80 & 1.07 & -.16 \\ 2.16 & .18 & 1.07 & 1.97 & -.36 \\ -.25 & .17 & -.16 & -.36 & .50 \end{bmatrix}$$

The sample principal components:

$$Y_1 = .8X_1 + .3X_2 + .3X_3 + .4X_4 - .1X_5$$
$$Y_2 = -.1X_1 - .8X_2 + .1X_3 + .6X_4 - .3X_5$$

etc.

$$\tilde{\lambda}_1 = 6.9, \tilde{\lambda}_2 = 1.8, \dots; \tilde{\lambda}_1 + \dots + \tilde{\lambda}_5 = 8.4$$

$X_1$ : Rmag  
 $X_2$ : mumax  
etc.

The PCs usually have no physical meaning, but they can provide insight into the data analysis

$\tilde{\lambda}_1 + \dots + \tilde{\lambda}_p$ : A measure of total variability of the data

$\frac{\tilde{\lambda}_k}{\tilde{\lambda}_1 + \dots + \tilde{\lambda}_p}$ : The proportion of total variability of the data “explained” by the  $k$ th PC

How many PC's should we calculate?

Stop when

$$\frac{\tilde{\lambda}_1 + \dots + \tilde{\lambda}_k}{\tilde{\lambda}_1 + \dots + \tilde{\lambda}_p} \geq 0.9$$

*Scree plot*: Plot the points  $(1, \tilde{\lambda}_1), \dots, (p, \tilde{\lambda}_p)$  and connect them by a straight line. Stop when the graph has flattened.

*Other rule*: Kaiser's rule; rules based on tests of hypotheses, ...

Some feel that PC's should be calculated from correlation matrices, not covariance matrices

Argument for correlation matrices: If the original data are rescaled then the PCs and the  $\tilde{\lambda}_k$  all change

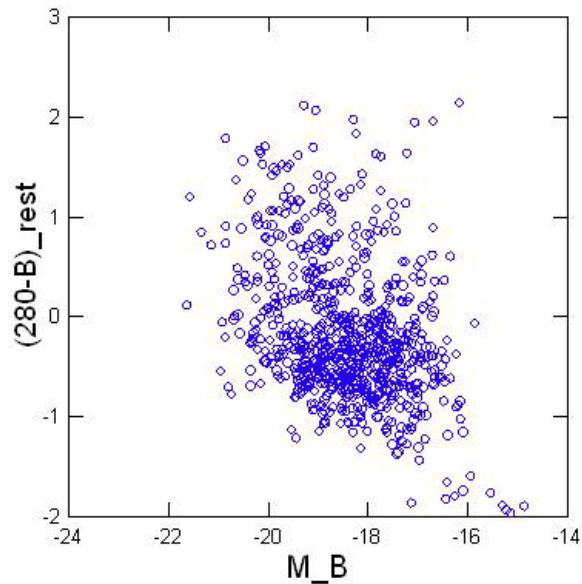
Argument against: If some components have significantly smaller means and variances than others then correlation-based PCs will give all components similarly-sized weights

## COMBO-17 data:

Two classes of galaxies, redder and bluer, but with overlapping distributions

Dataset: galaxy brightnesses in 17 bands—detailed view of "red" and "blue" for each galaxy

The following figure of  $M_B$  (BjMag) vs  $(280-B)$  (S280MAG-BjMag) for restricted range 0.7-0.9 of  $z$  (McZ) shows two cluster ("blue" below and "red" above), similar to the one in the website (also Wolf et al., 2004)



## **An exercise:**

We investigate the relationship of these colors to the brightness variables by multivariate analysis.

From combo17 dataset collected the even-numbered columns (30, 32, ..., 54).

Normalized each to (say) the value in column 40 (W640FE) for each galaxy. These are called “colors” .

Removed variable W640FE from the dataset

We added to this dataset Bjmag ( $M_B$ ). Also kept Mcz.

Modified “W” variables have been renamed with an “R” in the beginning.

Table 1 of Wolf et al. (2005, <http://arxiv.org/pdf/astro-ph/0506150v2>)

mean locations in multidimensional parameter space for "dust-free old" (= "red") and "blue cloud" (= "blue") galaxies

red galaxies has a mean value of  $(U - V) = 1.372$

blue galaxies has a mean  $(U - V) = 0.670$ —which are widely separated values

redshift  $z$  is a scientifically (very!) interesting variable denoting age of galaxy

We classify as "red" if  $(U - V) > 0.355$  and as "blue" if  $(U - V) \leq 0.355$ —color variable

This is the dataset. Data for the first few galaxies with the first few "R" readings:

RW420FE	RW462FE	RW485FD	RW518FE	RW571FS	RW604FE	BJMAG	MCZ	U-V	COLOR
-0.018	-0.006	0.000	-0.001	-0.004	-0.002	-17.540	0.832	0.090	1
-0.003	0.002	-0.000	-0.002	0.007	0.006	17.860	0.927	-0.080	1
-0.010	-0.003	0.002	-0.007	-0.000	0.000	-19.910	1.202	0.660	2
0.006	0.010	-0.005	-0.004	-0.005	0.003	-17.390	0.912	-0.160	1
0.002	0.005	0.002	0.010	0.004	0.007	-18.400	0.848	1.680	2
0.004	0.004	0.005	0.002	0.005	0.005	-18.110	0.882	-0.520	1
-0.004	-0.009	-0.008	-0.011	-0.008	-0.011	-16.060	0.896	-0.860	1
-0.002	-0.005	-0.006	-0.000	-0.004	0.002	-16.490	0.930	0.140	1
0.018	0.017	0.008	0.020	0.011	0.015	-17.680	0.774	0.200	1
0.006	0.007	0.001	-0.004	-0.004	-0.000	-11.150	0.062	-0.310	1
-0.009	-0.007	-0.010	-0.009	-0.009	-0.008	-16.990	0.874	-0.030	1
-0.032	-0.021	-0.018	-0.024	-0.019	-0.020	-13.990	0.173	0.650	2
-0.015	-0.009	-0.013	-0.006	-0.013	-0.014	-18.490	1.109	0.190	1
0.002	-0.002	0.002	0.002	0.012	0.002	-10.300	0.143	0.870	2
-0.028	-0.023	-0.020	-0.020	-0.025	-0.017	-18.210	0.626	1.360	2
0.011	0.015	-0.002	-0.003	0.002	0.009	-20.140	1.185	-0.200	1
0.010	0.007	0.012	0.010	0.010	0.015	-16.130	0.921	-1.570	1
0.001	0.004	0.004	0.001	0.002	0.003	-19.420	0.832	-0.170	1
0.005	0.013	-0.002	0.008	0.007	0.007	-18.110	0.793	0.460	2
-0.007	-0.002	-0.009	-0.002	0.000	-0.008	-17.810	0.952	0.290	1
-0.004	-0.004	-0.007	-0.009	-0.007	-0.002	-17.590	0.921	0.770	2
-0.007	-0.008	-0.014	-0.004	-0.003	-0.002	-16.980	0.986	-0.080	1
0.008	-0.004	0.003	-0.001	-0.001	0.007	-18.170	1.044	2.500	2
-0.000	0.002	0.004	0.000	0.004	0.001	-18.680	0.929	1.580	2
0.002	0.003	0.008	-0.003	0.001	0.001	-17.480	0.901	0.030	1
0.020	0.013	0.009	0.009	0.018	0.026	-16.600	0.484	0.190	1
0.016	0.008	0.019	0.019	0.014	0.010	-16.390	0.763	-0.730	1
0.001	0.001	0.006	0.004	0.003	0.002	-17.210	0.711	0.760	2
0.003	-0.001	-0.008	0.004	0.002	-0.001	-18.950	1.044	0.270	1
0.007	0.007	0.006	0.008	0.007	0.011	-19.850	0.826	0.670	2
-0.030	-0.013	-0.017	-0.001	0.021	0.025	-17.640	0.340	0.680	2
-0.058	-0.031	-0.037	-0.026	-0.015	-0.012	-17.600	0.365	0.390	2
0.004	0.006	0.008	0.013	0.018	0.021	-20.040	0.898	0.080	1
-0.005	-0.004	-0.006	-0.006	0.001	0.005	-19.540	0.878	0.290	1
-0.009	0.003	-0.009	-0.006	0.001	-0.007	-12.970	0.082	0.510	2

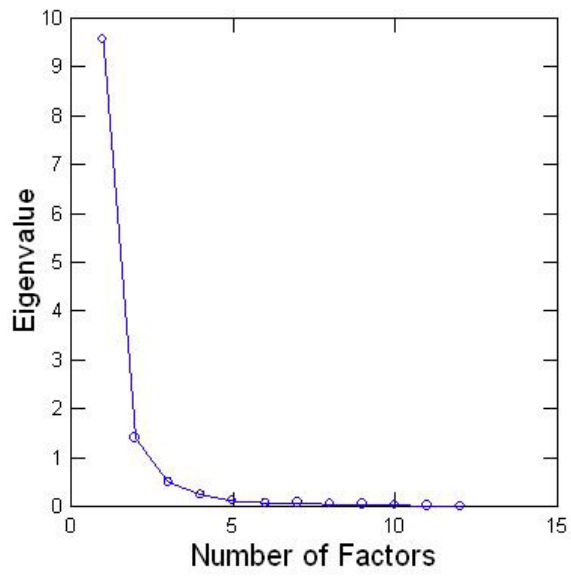
## PCA of Combo17 data:

PCA of the 12 color variables RW420FE RW462FE  
.... RW856FD RW914FD

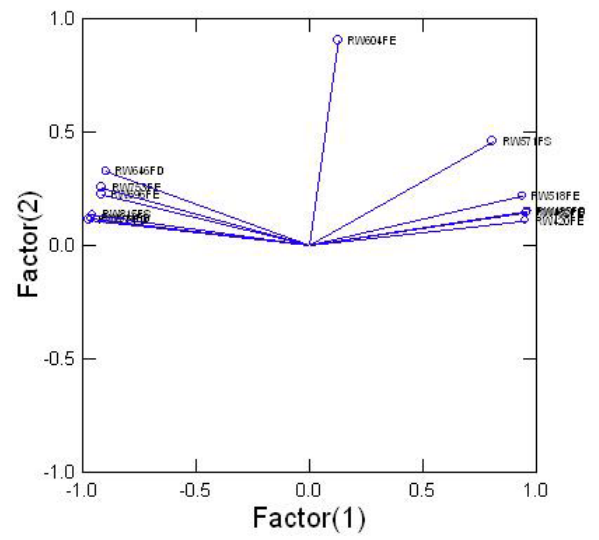
The scree plot suggests that two components are adequate.

Variable	PC1 weight	PC2 weight
RW420FE	0.954	0.107
RW462FE	0.957	0.144
RW485FD	0.960	0.149
RW518FE	0.938	0.218
RW571FS	0.810	0.456
RW604FE	0.128	0.902
RW646FD	-0.897	0.326
RW696FE	-0.914	0.223
RW753FE	-0.913	0.252
RW815FS	-0.953	0.134
RW856FD	-0.970	0.110
RW914FD	-0.961	0.117
Variance explained	9.547	1.386
% Variance explained	79.555	11.553

Scree Plot



Factor Loadings Plot



Two components explain most of the variation (about 91%)

### **Interpretation:**

#### **Principal Component 1:**

Weights are nearly the same in magnitude (except for RW604FE—insignificant)

RW4... and RW5... vs RW6... RW7.. RW8..  
RW9..

#### **Principal Component 2:**

RW604E the main component

Rest are nearly equal and small

Two components complement each other

Plot of PC scores of galaxies can be used for classification

Will see this in the Cluster Analysis chapter

## Classification Methods:

Two distinct types of classification problems—unsupervised and supervised

Unsupervised classification: Cluster Analysis:  
to find groups in the data objects  
objects within a group are similar

Example: what kinds of celestial objects are there—stars, planets, moons, asteroids, galaxies, comets, etc.

Multivariate (qualitative and quantitative) data on objects used

Characterize each type of object by these variables

Example: C. Wolf, M. E. Gray, and K. Meisenheimer (2008): Red-sequence galaxies with young stars and dust: The cluster Abell 901/902 seen with COMBO-17. *Astronomy & Astrophysics* classify galaxies into three classes with properties in the following table by cluster analysis

Mean properties of the three galaxy SED class samples.

Property	Dust-free old	Dusty red-seq	Blue cloud
$N_{\text{galaxy}}$	294	168	333
$N_{\text{fieldcontamination}}$	6	7	49
$N_{\text{spectra}}$	144	69	36
$z_{\text{spec}}$	0.1646	0.1646	0.1658
$\sigma_{cz}/(1+z)/(\text{km/s})$	939	1181	926
$z_{\text{spec},N}$	0.1625	0.1615	N/A
$z_{\text{spec},S}$	0.1679	0.1686	N/A
$\sigma_{cz,N}/(1+z)/(\text{km/s})$	589	597	N/A
$\sigma_{cz,S}/(1+z)/(\text{km/s})$	522	546	N/A
$\log(\Sigma_{10}(\text{Mpc}/h)^2)$	2.188	1.991	1.999
$EW_e(OII)/\overset{\circ}{A}$	N/A	$4.2 \pm 0.4$	$17.5 \pm 1.5$
$EW_a(H\delta)/\overset{\circ}{A}$	$2.3 \pm 0.5$	$2.6 \pm 0.5$	$4.5 \pm 1.0$
age/Gyr	6.2	3.5	1.2
$E_{B-V}$	0.044	0.212	0.193
$(U - V)_{\text{rest}}$	1.372	1.293	0.670
$M_{V,\text{rest}}$	-19.31	-19.18	-18.47
B - R	1.918	1.847	1.303
V - I	1.701	1.780	1.290
R - I	0.870	0.920	0.680
U - 420	0.033	-0.079	-0.377
420 - 464	0.537	0.602	0.560
464 - 518	0.954	0.827	0.490
604 - 646	0.356	0.339	0.238
753 - 815	0.261	0.274	0.224

## **Supervised Learning or Discriminant Analysis**

Know that there are these three types of galaxies

Have **Training Samples** where an expert (supervisor) classifies units in the sample

Multivariate observations on the sample units available

A new object is seen on which multivariate observations made

Problem: Classify it in one or other of the groups

In discriminant Analysis we develop a formula for such classification

Formula arrived at by performing discriminant analysis of training data

Some assumptions are often made

Multivariate normality in each group with a common covariance matrix

Find a classification rule that minimizes misclassification

This leads to **Linear Discriminant Function**, a linear combination of observed variables

## Discriminant Analysis Example

Use “R” data to develop a formula for classification into color 1 or 2

The linear discriminant function is

$$\begin{aligned} &0.345 + RW420FE*14.277 - RW462FE*0.844 \\ &- RW485FD*36.890 + RW518FE*6.541 \\ &+ RW571FS*2.249 + RW604FE*25.670 \\ &+ RW646FD*18.331 + RW696FE*15.123 - RW753FE*29.072 \\ &- RW815FS*16.970 - RW856FD*16.467 + RW914FD*2.024 \end{aligned}$$

If this value is  $> 0$  we classify a galaxy as 1 (red); else 2 (blue)

Using the formula on the training sample, we get an idea of the performance of the classification rule as follows:

Actual	Classified		
Group	Group		
	1	2	%correct
-----+-----			
1	2,111	45	98
2	1,020	286	22
-----			
Total	3,131	331	69

This is not a very good classification rule—the chosen variables do not provide adequate separation between blue and red

## Multiple Regression

If a supervisor had used the value of  $U - V$  to classify the galaxies into red and blue, and if values of  $U - V$  are indeed available, then why not use them rather than the red-blue classification?

$U - V$  data rather than color data in training sample

Leads to Multiple Regression Analysis

Develop a formula for prediction of  $U - V$  in a new galaxy from "R" values.

Results of such a multiple (linear) regression analysis:

Multiple correlation: a measure of how good the regression is: 0.344

Not very good—much as in Discriminant Analysis

Table below shows which "R" variables are useful for prediction of  $U - V$ : those with small  $p$ -values.

Regression Coefficients and their significance

Effect	Coefficient	Standard Error	t	p-value
CONSTANT	0.175	0.012	15.010	0.000
RW420FE	-1.624	0.839	-1.936	0.053
RW462FE	0.895	1.371	0.653	0.514
RW485FD	5.072	1.664	3.049	0.002
RW518FE	-1.921	1.199	-1.602	0.109
RW571FS	-1.126	1.178	-0.956	0.339
RW604FE	-4.636	1.456	-3.184	0.001
RW646FD	-2.345	1.340	-1.750	0.080
RW696FE	-2.729	0.917	-2.977	0.003
RW753FE	3.943	1.020	3.866	0.000
RW815FS	3.394	0.902	3.761	0.000
RW856FD	3.059	0.961	3.182	0.001
RW914FD	0.036	0.740	0.049	0.961