

SAMSI Undergraduate Workshop
on Random Matrices
November 17-18, 2006

Random Matrices in
Multivariate Statistical Analysis

Donald Richards
Penn State University and SAMSI

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

Wolf, et al., “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

A portion of the COMBO-17 data

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 (the sample size)

x_{ij} : the i th observation on the j th variable

Data array

		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

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

We write X as n row or as p column vectors

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

Matrix methods are essential to multivariate analysis

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

A' : The transpose of A

$|A|$: The determinant of A

$$(AB)' = B'A'$$

Descriptive Statistics

The sample mean of the j 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 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 :

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

The sample covariance matrix:

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

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:

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

S and R are *symmetric*

S and R are *positive semidefinite*: $\mathbf{v}'S\mathbf{v} \geq 0$ 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 $n \leq p$ then S and R will be *singular* :

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

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

If $n > p$ then, most of the time (*but not always*), S and R are *positive definite*:

$\mathbf{v}'S\mathbf{v} > 0$ 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.

If $n > p$ and $|S| = 0$ then there is likely to be a linear relationship between your variables

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

The COMBO-17 data

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

$$p = 8 \text{ and } n = 3462$$

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

The sample correlation matrix:

	Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
Rmag	1.000	0.932	0.415	0.433	0.150	0.315	0.333	0.353
mumax	0.932	1.000	0.434	0.450	0.113	0.242	0.268	0.287
Mcz	0.415	0.434	1.000	0.871	-0.017	-0.698	-0.628	-0.642
MCzml	0.433	0.450	0.871	1.000	-0.035	-0.563	-0.519	-0.532
chi2red	0.150	0.113	-0.017	-0.035	1.000	0.150	0.141	0.155
UjMAG	0.315	0.242	-0.698	-0.563	0.150	1.000	0.933	0.960
BjMAG	0.333	0.268	-0.628	-0.519	0.141	0.933	1.000	0.951
VjMAG	0.353	0.287	-0.642	-0.532	0.155	0.960	0.951	1.000

Some books advise:

Use no more than two decimal places

Starting with the physically most important variable, reorder variables by descending correlations

Suppress diagonal entries to ease visual clutter

Suppress zeros appearing 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

$\{U_j\text{MAG}, B_j\text{MAG}, V_j\text{MAG}\}$ are highly correlated; perhaps, two of them can be eliminated. Similar remarks apply to $\{R_{\text{mag}}, \text{mumax}\}$ and $\{M_{\text{cz}}, M_{\text{czml}}\}$.

chi2red has small correlation with $\{\text{mumax}, M_{\text{cz}}, M_{\text{czml}}\}$; we would retain chi2red in the subsequent analysis

Homework: Compute the correlation matrix for all 65 variables in COMBO-17. Reorder the variables by (i) descending correlations and (ii) astrophysical importance. Use the correlation matrix to carry out a dimension reduction.

The *rank* of S :

The rank of S is the largest number of linearly independent rows or columns of S .

Here is an algorithm for calculating the rank:

Calculate $|S|$; if $|S| \neq 0$ then S is of rank p

If $|S| = 0$ then calculate the determinant of all $(p - 1) \times (p - 1)$ submatrices of S . If at least one of them is nonzero then S is of rank $p - 1$. If all of them are zero then go on to all $(p - 2) \times (p - 2)$ submatrices.

In general, the rank of S is the largest integer r such that at least one $r \times r$ subdeterminant of S is nonzero.

Alternatively, the rank of S is the number of nonzero eigenvalues of S .

The rank of any matrix is the largest number of *linearly independent* rows or columns

Recall (see p. 11) that

$$S = \frac{1}{n-1}(X - \mathbf{1}\bar{x}')'(X - \mathbf{1}\bar{x}')$$

Consequence: S has the same rank as $X - \mathbf{1}\bar{x}'$

Recall (see p. 12) that $R = D^{-1}SD^{-1}$ where

$$D = \text{diag}(s_{11}^{1/2}, s_{22}^{1/2}, \dots, s_{pp}^{1/2})$$

Consequence: R and S have the same rank

If $n > p$ then, usually, R will have rank p

If some variables are linearly related then the linear constraints usually will appear in the sample data, and then R will be singular

Conclusion: The rank of R is a good guide to the true dimensionality of the problem

Plotting multivariate data

Scatterplots between each pair of variables

$p = 65$: $\binom{p}{2} = 2080$ plots; is this practical?

Ink-blob plots for 3-D data: Represent each 3-D data vector (u, v, w) by a circular ink blob centered at (u, v) with area w

What if w can be negative?

Principal components: A way to reduce the dimension by using linear combinations of variables

Biplots: Plots of principal components

Lab assignment: Using astrophysical considerations *and* the correlation matrix, choose a subset of variables from COMBO-17. Use R to construct the corresponding biplot.

Andrews plots: Transform each data vector x into a sinusoidal function

Example: The components of the observation vectors

Rmag	mumax	Mcz	MCzml	chi2red	UjMAG	BjMAG	VjMAG
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

etc.

are assigned coefficients

$2^{-1/2} \sin(t) \cos(t) \sin(2t) \cos(2t) \sin(3t) \cos(3t) \sin(4t)$

Construct the function

$$24.995/\sqrt{2} + 24.214 \sin t + 0.832 \cos t + 1.400 \sin 2t \\ + 0.64 \cos 2t - 17.67 \sin 3t - 17.54 \cos 3t - 17.76 \sin 4t$$

Repeat this procedure for each vector

Plot all n functions over the range $[-\pi, \pi]$

Observations which are “close” in p -dimensional space will have similar wave patterns

The Euclidean distance between two observations is directly related to the squared area (L^2 -distance) between their sinusoidal curves

These plots are useful for locating outliers and clusters

Drawback: If n is very large then there are too many sinusoidal curves, so it becomes difficult to discern features of the data set

Here are some more examples

Chernoff faces: How to display 13-D data on a 2-D surface

Discretize each variable into, say, 10 possible values

Represent each variable by a facial characteristic

- 1 eyebrow slant
- 2 eye size
- 3 nose length
- 4 head eccentricity
- 5 eye size
- 6 eye spacing
- 7 eye eccentricity
- 8 pupil size
- 9 eyebrow slant
- 10 nose size
- 11 mouth shape
- 12 mouth size
- 13 mouth opening

Chernoff faces

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:

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

Also,

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

To avoid pathological cases, we assume that Σ 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 $\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$	Sample cov. matrix, $S = \frac{1}{n-1} \times \sum (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})'$

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

$\mathbf{X} = [X_1, \dots, X_p]'$: 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

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

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, Σ 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

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

Proof: Use Fourier transforms

Special cases:

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

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

Note: $\mathbf{v}'\Sigma\mathbf{v} > 0$ since Σ is positive definite

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

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 $\mathbf{v}'\mathbf{X}$ has a 1-D normal distribution for every vector $\mathbf{v} \in \mathbb{R}^p$

Conversely, if $\mathbf{v}'\mathbf{X}$ 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 $\mathbf{v} = [v_1, v_2, \dots, v_8]'$

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 μ

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

Use the collection of confidence intervals as a confidence region for μ

Good news: This can be done using elementary statistical methods

Bad news: A collection of 95% confidence intervals, one for each μ_i , does not result in a 95% confidence region for μ

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 μ is a rectangle

This is not consonant with a density function of the form

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

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

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

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

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

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

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

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

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

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

Recall (see p. 30): Since the covariance matrix is diagonal then the normal random vector is a set of independent normal variables

Conclude: The components of $\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ are independent $N(0, 1)$, random variables

Elementary results (see Bevington):

If $W \sim N(0, 1)$ then $W^2 \sim \chi_1^2$, and

A sum of independent χ^2 random variables is again a χ^2 random variable, with the degrees of freedom added together

Add the squared components of $\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$

$$\left(\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\right)' \left(\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\right) \sim \chi_p^2$$

or

$$(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2$$

A similar argument applies to a random sample

If $\mathbf{X}_1, \dots, \mathbf{X}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
then $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$

Conclude: $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi_p^2$

$$P\left(n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \chi_{p,0.05}^2\right) = 0.95$$

Assuming that $\boldsymbol{\Sigma}$ is known (which is unlikely),
a 95% confidence region for $\boldsymbol{\mu}$ is the region

$$\left\{ \boldsymbol{\nu} : n(\bar{\mathbf{X}} - \boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\nu}) \leq \chi_{p,0.05}^2 \right\}$$

This is an elliptical region centered at $\bar{\mathbf{X}}$

This region generally has smaller volume than
the rectangular componentwise region

What do we do if Σ is unknown?

Motivation: In 1-D, we construct confidence intervals for μ when σ^2 is unknown using

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Note that

$$t^2 = \frac{(\bar{X} - \mu)^2}{s^2/n} = n(\bar{X} - \mu)(s^2)^{-1}(\bar{X} - \mu)$$

In p dimensions, we use Hotelling's T^2 -statistic:

$$T^2 = \frac{n-p}{(n-1)p} \times n(\bar{\mathbf{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$$

has an F -distribution with $(p, n-p)$ degrees of freedom

T^2 measures the “statistical distance” between $\boldsymbol{\mu}$ and $\bar{\mathbf{X}}$

The resulting elliptical confidence region:

$$\{\boldsymbol{\nu} : n(\bar{\mathbf{X}} - \boldsymbol{\nu})'S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\nu}) \leq \text{cutoff}\}$$

“cutoff” is found in the tables of the F -distn.

Test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

The likelihood ratio test statistic is equivalent to Hotelling's T^2 -statistic

So many topics ... so little time

Inference for Σ when the data are drawn from $N_p(\boldsymbol{\mu}, \Sigma)$

We require the distribution of S , the sample covariance matrix

In 1-D, $(n - 1)s^2/\sigma^2 \sim \chi_{n-1}^2$

In p -D, $(n - 1)S$ has a “Wishart” distribution

Discriminant analysis

Multivariate regression

Multivariate analysis of variance

Factor analysis

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: Σ

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

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

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 Σ

Scale each eigenvector to make its length 1

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

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

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

(ii) $\mathbf{v}'\mathbf{v} = 1$

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

Lagrange multipliers

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

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

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

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

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

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

Maximize $\text{Var}(\mathbf{v}'\mathbf{X}) = \mathbf{v}'\Sigma\mathbf{v}$ with $\mathbf{v}'\mathbf{v} = 1$ and $\text{Cov}(\mathbf{v}'\mathbf{X}, \mathbf{v}'_1\mathbf{X}) \equiv \mathbf{v}'\Sigma\mathbf{v}_1 = 0$

Lagrange multipliers

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

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

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

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

The PCs are random variables

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

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

$\tilde{\mathbf{v}}_1\mathbf{X}, \dots, \tilde{\mathbf{v}}_p\mathbf{X}$: 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.

$$\tilde{\lambda}_1 = 6.9, \tilde{\lambda}_2 = 0.8, \dots; \tilde{\lambda}_1 + \dots + \tilde{\lambda}_5 = 8.4$$

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

Carry out a PCA of the COMBO-17 data