Principal Components Analysis

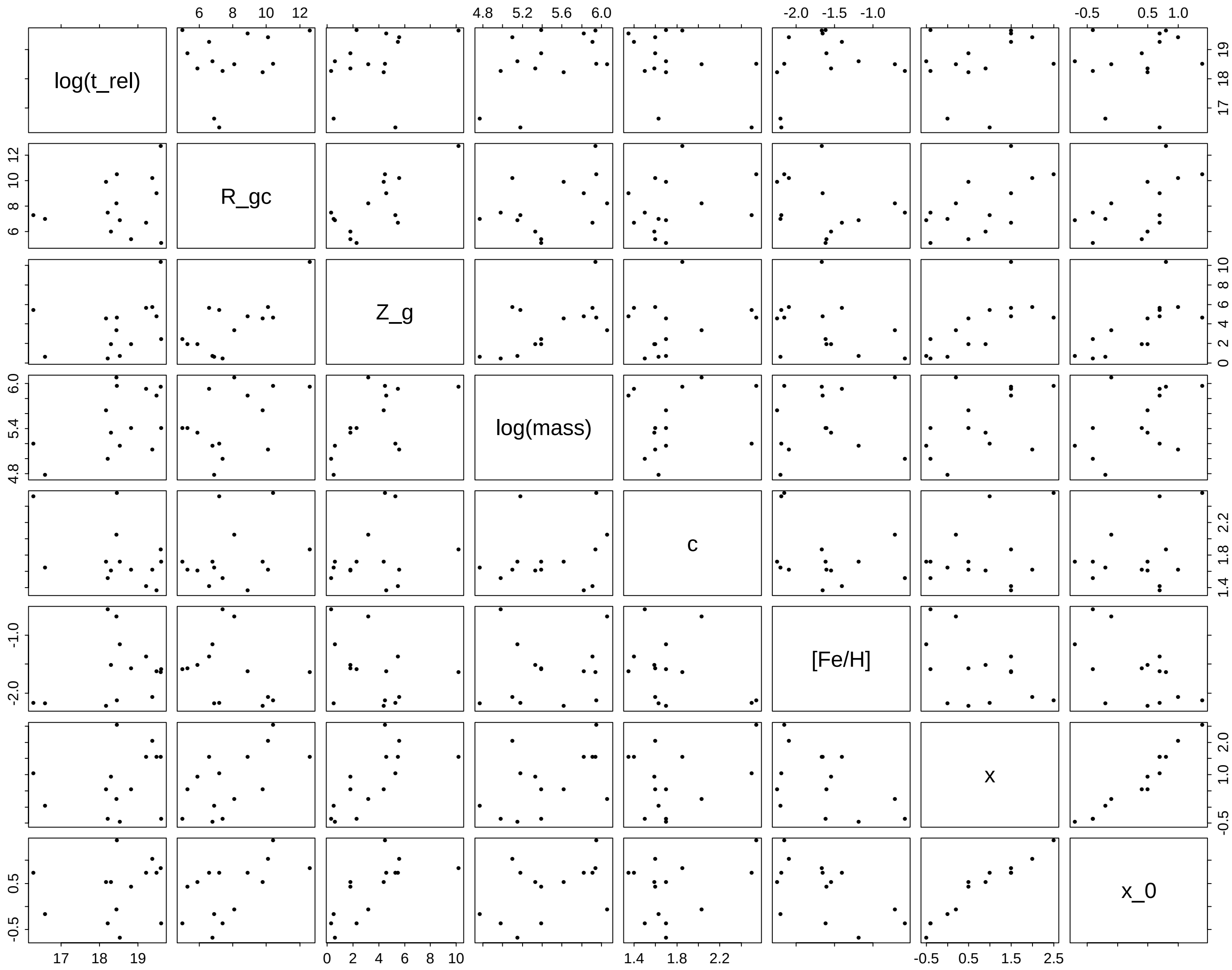
Topics:

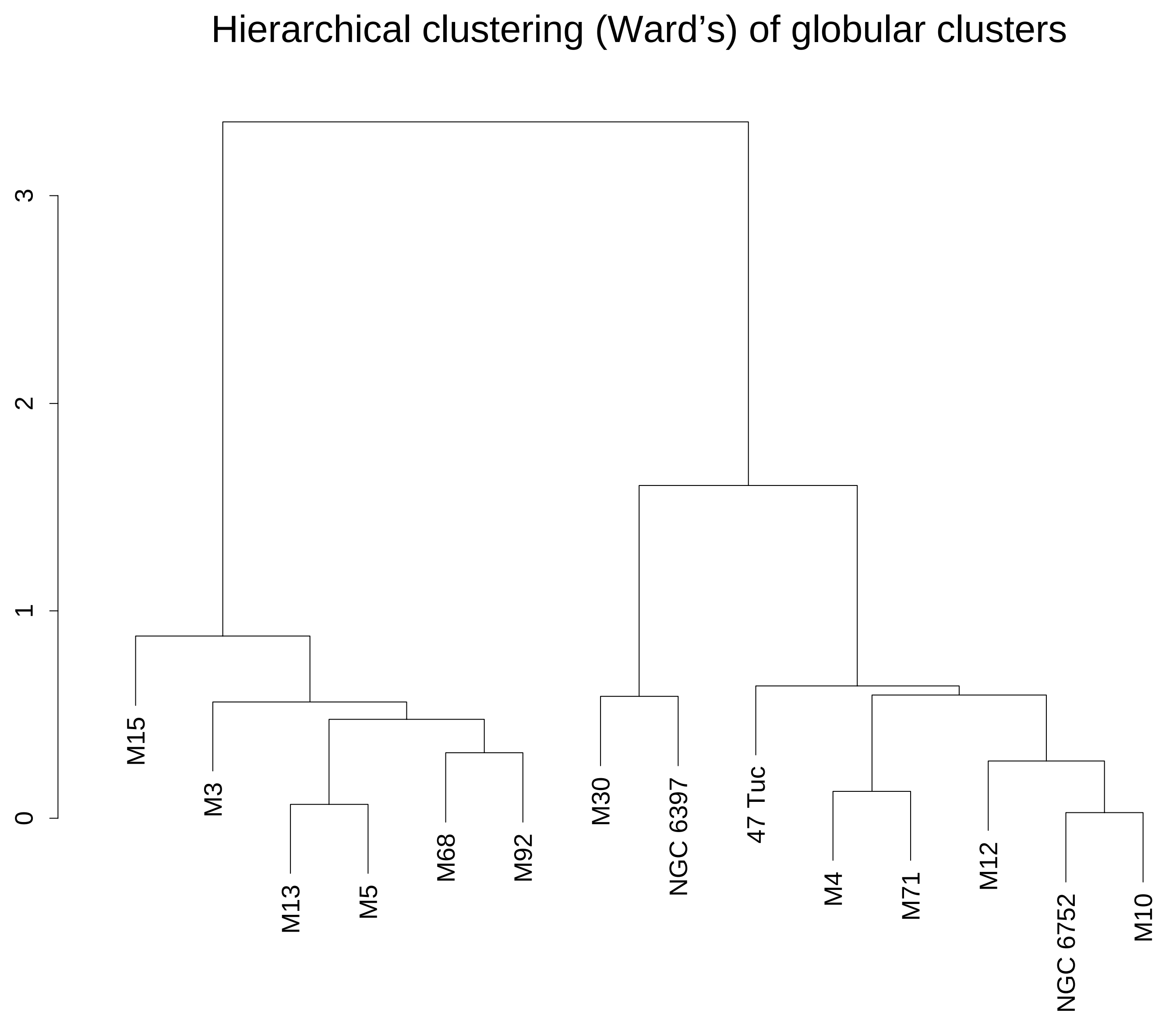
- Reference: F Murtagh and A Heck, Multivariate Data Analysis, Kluwer, 1987.
- Preliminary example: globular clusters.
- Data, space, metric, projection, eigenvalues and eigenvectors, dual spaces, linear combinations.
- Practical aspects – nonlinear terms, standardization, list of objectives, procedure followed.
- Image multiband compression, “eigen-faces”.
- Software: http://astro.u-strasbg.fr/∼fmurtagh/mda-sw

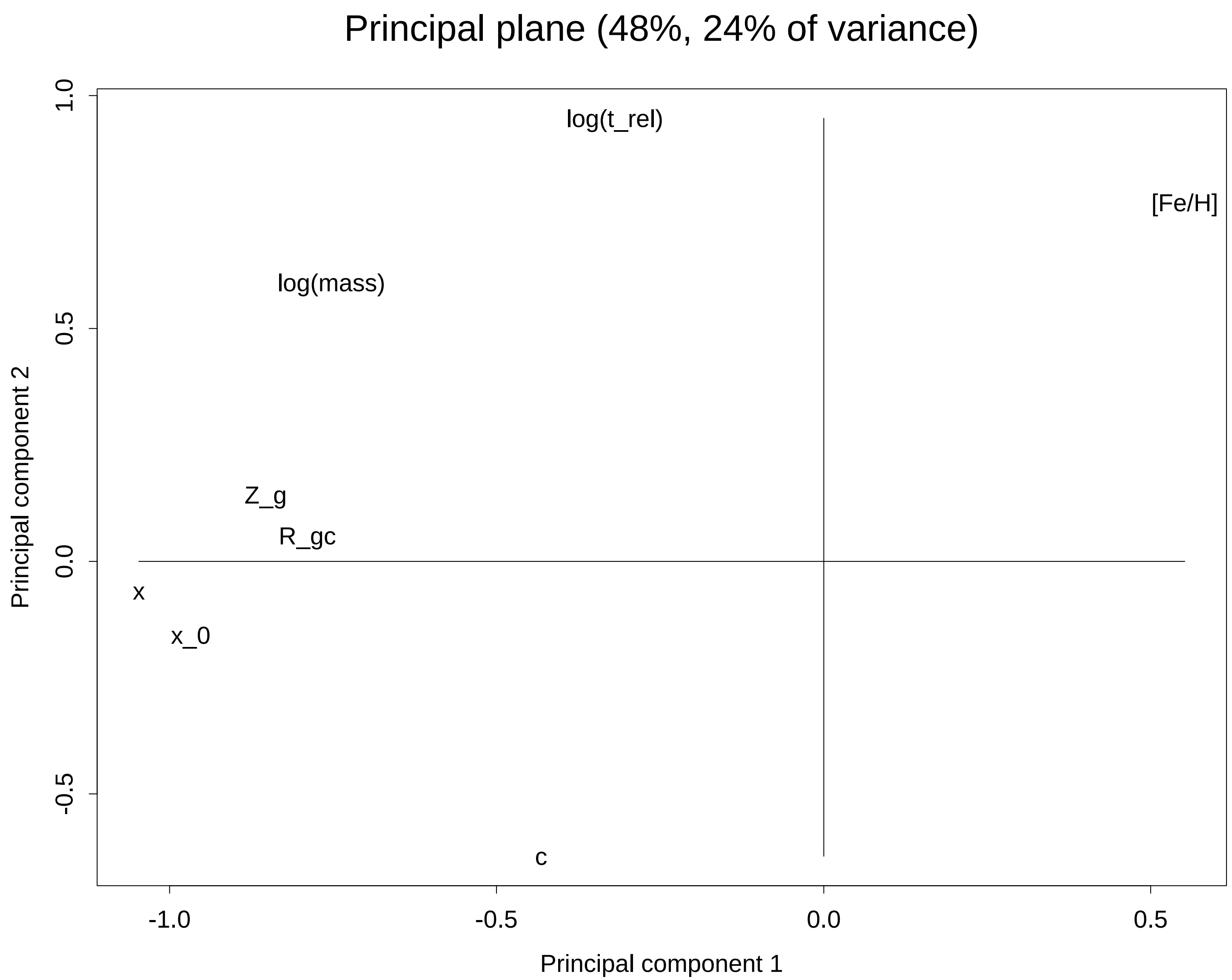
Example: analysis of globular clusters

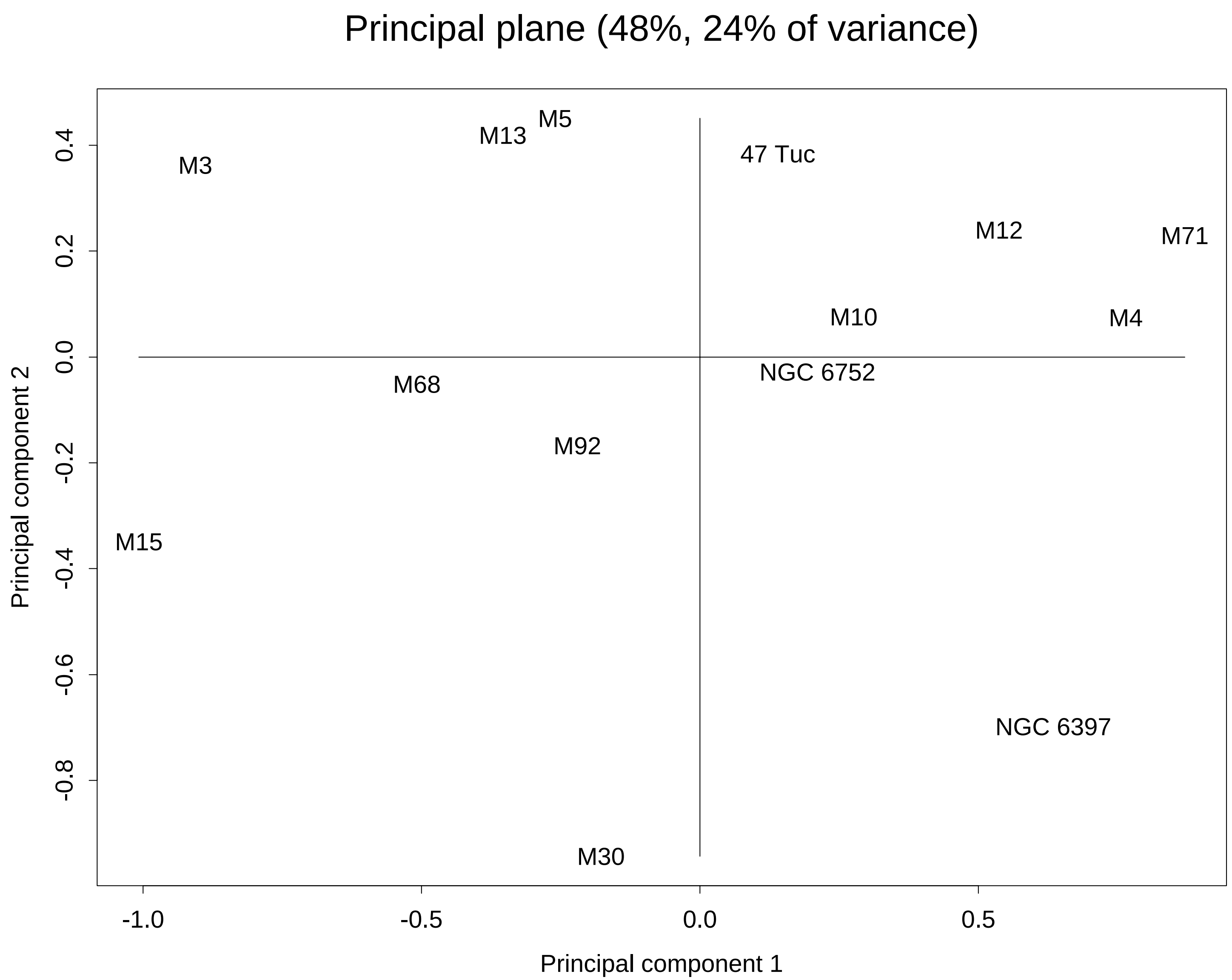
- M. Capaccioli, S. Ortolani and G. Piotto, "Empirical correlation between globular cluster parameters and mass function morphology", AA, 244, 298–302, 1991.
- 14 globular clusters, 8 measurement variables.
- Data collected in earlier CCD (digital detector) photometry studies.
- Pairwise plots of the variables.
- PCA of the variables.
- PCA of the objects (globular clusters).

Object	t_rlx years	Rgc Kpc	Zg Kpc	log(M/ M.)	c	[Fe/H]	x	x0
M15	1.03e+8	10.4	4.5	5.95	2.54	-2.15	2.5	1.4
M68	2.59e+8	10.1	5.6	5.1	1.6	-2.09	2.0	1.0
M13	2.91e+8	8.9	4.6	5.82	1.35	-1.65	1.5	0.7
M3	3.22e+8	12.6	10.2	5.94	1.85	-1.66	1.5	0.8
M5	2.21e+8	6.6	5.5	5.91	1.4	-1.4	1.5	0.7
M4	1.12e+8	6.8	0.6	5.15	1.7	-1.28	-0.5	-0.7
47 Tuc	1.02e+8	8.1	3.2	6.06	2.03	-0.71	0.2	-0.1
M30	1.18e+7	7.2	5.3	5.18	2.5	-2.19	1.0	0.7
NGC 6397	1.59e+7	6.9	0.5	4.77	1.63	-2.2	0.0	-0.2
M92	7.79e+7	9.8	4.4	5.62	1.7	-2.24	0.5	0.5
M12	3.26e+8	5.0	2.3	5.39	1.7	-1.61	-0.4	-0.4
NGC 6752	8.86e+7	5.9	1.8	5.33	1.59	-1.54	0.9	0.5
M10	1.50e+8	5.3	1.8	5.39	1.6	-1.6	0.5	0.4
M71	8.14e+7	7.4	0.3	4.98	1.5	-0.58	-0.4	-0.4


log(t_rel)
R_gc
Z_g
log(mass)
c
[Fe/H]
x
x_0


Hierarchical clustering (Ward's) of globular clusters
0
1
2
3
M15
M3
M13
M5
M68
M92
M30
NGC 6397
47 Tuc
M4
M71
M12
NGC 6752
M10


Principal plane (48%, 24% of variance)
log(t_rel)
[Fe/H]
log(mass)
Z_g
R_gc
x
x_0
c
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
Principal component 1
Principal component 2


Principal plane (48%, 24% of variance)
M5
M13
M3
47 Tuc
M12
M71
M10
M4
M68
NGC 6752
M92
M15
NGC 6397
M30
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Principal component 2
-1.0
-0.5
0.0
0.5
Principal component 1

Data

- Matrix X defines a set of n vectors in m-dimensional space: $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ for $1 \leq i \leq n$.
- We have: $x_i \in \mathbb{R}^m$
- Matrix X also defines a set of m column vectors in n-dimensional space: $x_j = \{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ for $1 \leq j \leq m$.
- We have: $x_j \in \mathbb{R}^n$
- By convention we usually take the space of row points, i.e. $\mathbb{R}^m$, as X; and the space of column points, i.e. $\mathbb{R}^n$, as the transpose of X, i.e. X' or X^t.
- The row points define a cloud of n points in $\mathbb{R}^m$.
- The column points define a cloud of m points in $\mathbb{R}^n$.

Metrics

- The notion of distance is crucial, since we want to investigate relationships between observations and/or variables.
- Recall: $x = \{3, 4, 1, 2\}, y = \{1, 3, 0, 1\}$, then: scalar product $\langle x, y \rangle = \langle y, x \rangle = x'y = xy' = 3 \times 1 + 4 \times 3 + 1 \times 0 + 2 \times 1.$
- Euclidean norm: $\|x\|^2 = 3 \times 3 + 4 \times 4 + 1 \times 1 + 2 \times 2.$
- Euclidean distance: $d(x, y) = \|x - y\|$. The squared Euclidean distance is: $3 - 1 + 4 - 3 + 1 - 0 + 2 - 1$
- Orthogonality: x is orthogonal to y if $\langle x, y \rangle = 0$.
- Distance is symmetric ($d(x, y) = d(y, x)$), positive ($d(x, y) \geq 0$), and definite ($d(x, y) = 0 \Longrightarrow x = y$).

Metrics (cont'd.)

- Any symmetric, positive, definite matrix M defines a generalized Euclidean space. Scalar product is $\langle x, y \rangle_M = x' M y$, norm is $\|x\|^2 = x' M x$, and Euclidean distance is $d(x, y) = \|x - y\|_M$.
- Classical case: $M = I_n$, the identity matrix.
- Normalization to unit variance: M is diagonal matrix with ith diagonal term $1/\sigma_i^2$.
- Mahalanobis distance: M is inverse variance-covariance matrix.
- Next topic: Scalar product defines orthogonal projection.

Metrics (cont'd.)

- Projected value, projection, coordinate: $x_1 = (x'Mu/u'Mu)u$. Here x_1 and u are both vectors.
- Norm of vector $x_1 = (x'Mu/u'Mu)\|u\| = (x'Mu)/\|u\|$.
- The quantity $(x'Mu)/(\|x\|\|u\|)$ can be interpreted as the cosine of the angle a between vectors x and u.

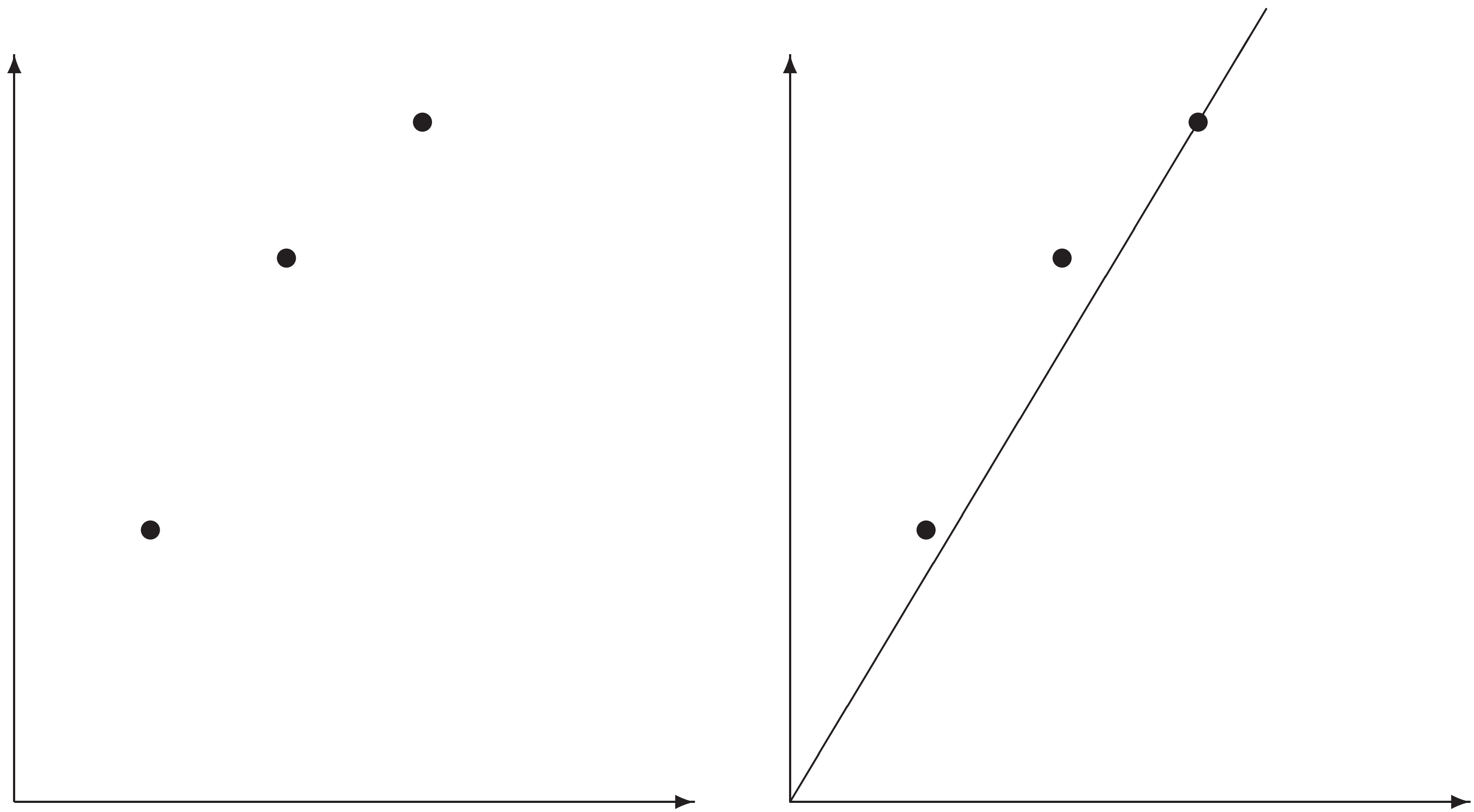
```
      +  x
     /|
    / |
   /  |
  /   |
 /a   |
+-----+-----  u
O     x1
```

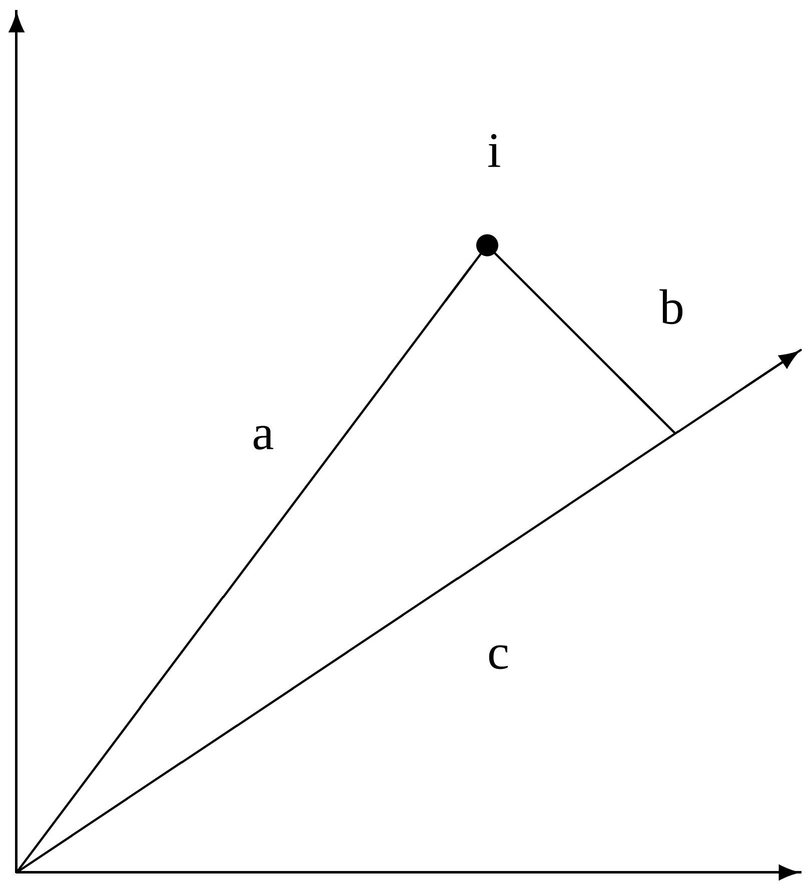
Least Squares Optimal Projection of Points

- Plot of 3 points in $\mathbb{R}^2$ (see following slides).
- PCA: determine best fitting axes.
- Examples follow.
- Note: optimization means either (i) closest axis to points, or (ii) maximum elongation of projections of points on the axis.
- This follows from Pythagoras's theorem: $x^2 + y^2 = z^2$. Call z the distance from the origin to a point. Let x be the distance of the projection of the point from the origin. Then y is the perpendicular distance from the axis to to the point.
- Minimizing y is the same as maximizing x (because z is fixed).

Examples of Optimal Projection

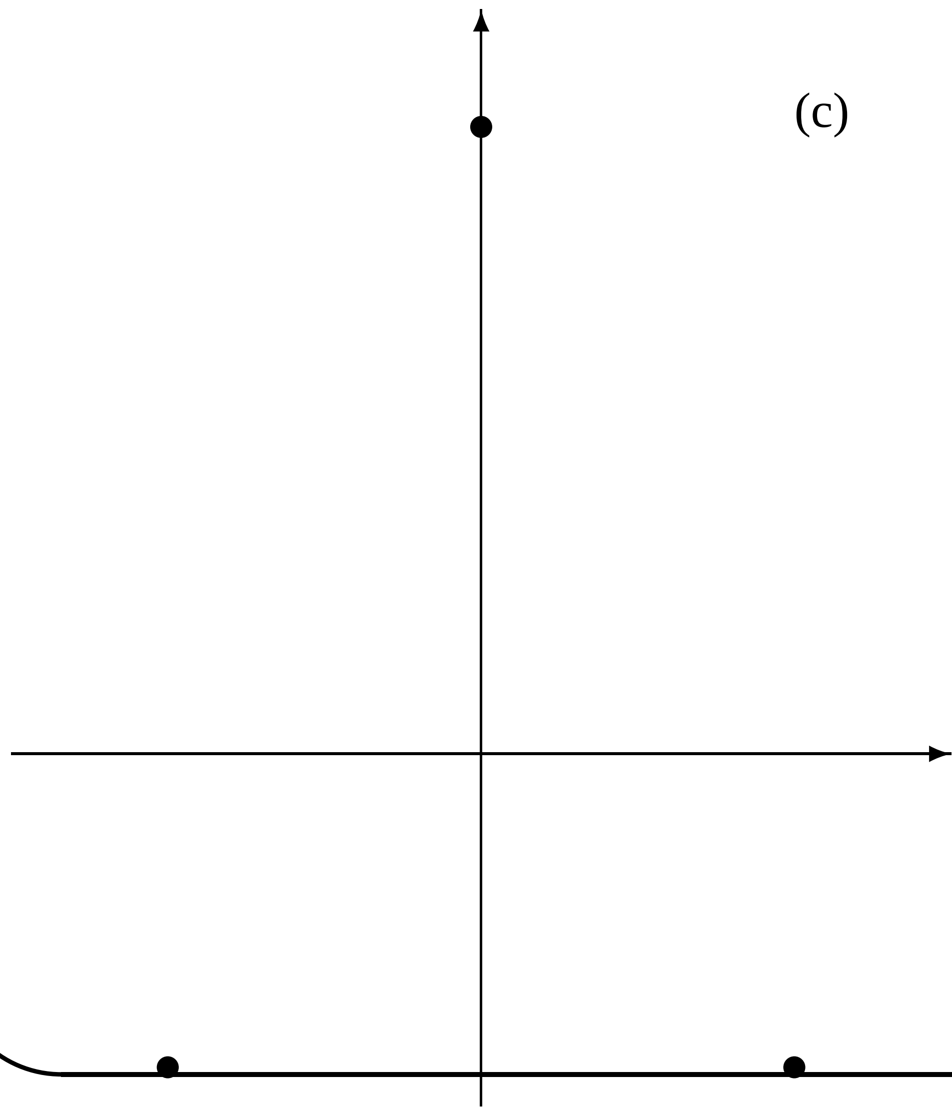
$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 5 \end{pmatrix}$$




i
b
a
c

(a)

(b)


(c)

Questions We Will Now Address

- How is the PCA of an $n \times m$ matrix related to the PCA of the transposed $m \times n$ matrix ?
- How may the new axes derived – the principal components – be said to be linear combinations of the original axes ?
- How may PCA be understood as a series expansion ?
- In what sense does PCA provide a lower-dimensional approximation to the original data ?

PCA Algorithm

- The projection of vector $\mathbf{x}$ onto axis $\mathbf{u}$ is $\mathbf{y} = \frac{\mathbf{x}' M \mathbf{u}}{\|\mathbf{u}\|_M} \ \mathbf{u}$
- I.e. the coordinate of the projection on the axis is $\mathbf{x}' M \mathbf{u} / \|\mathbf{u}\|_M$.
- This becomes $\mathbf{x}' M \mathbf{u}$ when the vector $\mathbf{u}$ is of unit length.
- The cosine of the angle between vectors $\mathbf{x}$ and $\mathbf{y}$ in the usual Euclidean space is $\mathbf{x}'\mathbf{y} / \|\mathbf{x}\| \|\mathbf{y}\|$.
- That is to say, we make use of the triangle whose vertices are the origin, the projection of $\mathbf{x}$ onto $\mathbf{y}$, and vector $\mathbf{x}$.
- The cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$ is then the coordinate of the projection of $\mathbf{x}$ onto $\mathbf{y}$, divided by the – hypotenuse – length of $\mathbf{x}$.
- The correlation coefficient between two vectors is then simply the cosine of the angle between them, when the vectors have first been centred (i.e. $\mathbf{x} - \mathbf{g}$ and $\mathbf{y} - \mathbf{g}$ are used, where $\mathbf{g}$ is the overall centre of gravity.

PCA Algorithm 2

- $X = \{x_{ij}\}$
- In $\mathbb{R}^m$, the space of objects, PCA searches for the best–fitting set of orthogonal axes to replace the initially–given set of m axes in this space.
- An analogous procedure is simultaneously carried out for the dual space, $\mathbb{R}^n$.
- First, the axis which best fits the objects/points in $\mathbb{R}^m$ is determined.
- If $\mathbf{u}$ is this vector, and is of unit length, then the product $X\mathbf{u}$ of $n \times m$ matrix by $m \times 1$ vector gives the projections of the n objects onto this axis.
- The sum of squared projections of points on the new axis, for all points, is $(X\mathbf{u})'(X\mathbf{u})$.
- Such a quadratic form would increase indefinitely if $\mathbf{u}$ were arbitrarily large, so $\mathbf{u}$ is taken to be of unit length, i.e. $\mathbf{u}'\mathbf{u} = 1$.
- We seek a maximum of the quadratic form $\mathbf{u}'S\mathbf{u}$ (where $S = X'X$) subject to

the constraint that $\mathbf{u}'\mathbf{u} = 1$.

- This is done by setting the derivative of the Lagrangian equal to zero.
- Differentiation of $\mathbf{u}'S\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1)$ where λ is a Lagrange multiplier gives $2S\mathbf{u} - 2\lambda\mathbf{u}$.
- The optimal value of $\mathbf{u}$ (let us call it $\mathbf{u}_1$) is the solution of $S\mathbf{u} = \lambda\mathbf{u}$.
- The solution of this equation is well–known: $\mathbf{u}$ is the eigenvector associated with the eigenvalue λ of matrix S.
- Therefore the eigenvector of $X'X$, $\mathbf{u}_1$, is the axis sought, and the corresponding largest eigenvalue, λ_1, is a figure of merit for the axis, – it indicates the amount of variance explained by the axis.
- The second axis is to be orthogonal to the first, i.e. $\mathbf{u}'\mathbf{u}_1 = 0$.
- The second axis satisfies the equation $\mathbf{u}'X'X\mathbf{u} - \lambda_2(\mathbf{u}'\mathbf{u} - 1) - \mu_2(\mathbf{u}'\mathbf{u}_1)$ where λ_2 and μ_2 are Lagrange multipliers.

- Differentiating gives $2S\mathbf{u} - 2\lambda_2\mathbf{u} - \mu_2\mathbf{u}_1$.
- This term is set equal to zero. Multiplying across by $\mathbf{u}_1'$ implies that μ_2 must equal 0.
- Therefore the optimal value of $\mathbf{u}$, $\mathbf{u}_2$, arises as another solution of $S\mathbf{u} = \lambda\mathbf{u}$.
- Thus λ_2 and $\mathbf{u}_2$ are the second largest eigenvalue and associated eigenvector of S.
- The eigenvectors of $S = X'X$, arranged in decreasing order of corresponding eigenvalues, give the line of best fit to the cloud of points, the plane of best fit, the three–dimensional hyperplane of best fit, and so on for higher–dimensional subspaces of best fit.
- $X'X$ is referred to as the *sums of squares and cross products* matrix.

Eigenvalues

- Eigenvalues are decreasing in value.
- $\lambda_i = \lambda_{i'}$? Then equally privileged directions of elongation have been found.
- $\lambda_i = 0$? Space is actually of dimensionality less than expected. Example: in 3D, points actually lie on a plane.
- Since PCA in $\mathbb{R}^n$ and in $\mathbb{R}^m$ lead respectively to the finding of n and of m eigenvalues, and since in addition it has been seen that these eigenvalues are identical, it follows that the number of *non-zero eigenvalues* obtained in either space is less than or equal to $\min(n, m)$.
- The eigenvectors associated with the p largest eigenvalues yield the best-fitting p-dimensional subspace of $\mathbb{R}^m$. A measure of the approximation is the percentage of variance explained by the subspace $\sum_{k \leq p} \lambda_k / \sum_{k=1}^{n} \lambda_k$ expressed as a percentage.

Dual Spaces

- In the dual space of attributes, $\mathbb{R}^n$, a PCA may equally well be carried out.
- For the line of best fit, $\mathbf{v}$, the following is maximized: $(X'\mathbf{v})'(X'\mathbf{v})$ subject to $\mathbf{v}'\mathbf{v} = \mathbf{1}$.
- In $\mathbb{R}^m$ we arrived at $X'X\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$.
- In $\mathbb{R}^n$, we have $XX'\mathbf{v}_1 = \mu_1 \mathbf{v}_1$.
- Premultiplying the first of these relationships by X yields $(XX')(X\mathbf{u}_1) = \lambda_1(X\mathbf{u}_1)$.
- Hence $\lambda_1 = \mu_1$ because we have now arrived at two eigenvalue equations which are identical in form.
- Relationship between the eigenvectors in the two spaces: these must be of unit length.

- Find: $\mathbf{v}_1 = \frac{1}{\sqrt{\lambda_1}} X \mathbf{u}_1$.
- $\lambda > 0$ since if $\lambda = 0$ eigenvectors are not defined.
- For λ_k: $\mathbf{v}_k = \frac{1}{\sqrt{\lambda_k}} X \mathbf{u}_k$
- And: $\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} X' \mathbf{v}_k$
- Taking $X \mathbf{u}_k = \sqrt{\lambda_k}\, \mathbf{v}_k$, postmultiplying by $\mathbf{u}'_k$, and summing gives:
  $X \sum_{k=1}^{n} \mathbf{u}_k \mathbf{u}'_k = \sum_{k=1}^{n} \sqrt{\lambda_k}\, \mathbf{v}_k \mathbf{u}'_k$.
- LHS gives the identity matrix (due to orthogonality of eigenvectors). Hence:
- $X = \sum_{k=1}^{n} \sqrt{\lambda_k}\, \mathbf{v}_k \mathbf{u}'_k$
- This is termed: Karhunen-Loève expansion or transform.
- We can approximate the data, X, by choosing some eigenvalues/vectors only.

Linear Combinations

- The variance of the projections on a given axis in $\mathbb{R}^m$ is given by $(X\mathbf{u})'(X\mathbf{u})$, which by the eigenvector equation, is seen to equal λ.
- In some software packages, the eigenvectors are rescaled so that $\sqrt{\lambda}\,\mathbf{u}$ and $\sqrt{\lambda}\,\mathbf{v}$ are used instead of $\mathbf{u}$ and $\mathbf{v}$. In this case, the *factor* $\sqrt{\lambda}\,\mathbf{u}$ gives the new, rescaled projections of the points in the space $\mathbb{R}^n$ (i.e. $\sqrt{\lambda}\,\mathbf{u} = X'\mathbf{v}$).
- The coordinates of the new axes can be written in terms of the old coordinate system. Since $\mathbf{u} = \frac{1}{\sqrt{\lambda}} X'\mathbf{v}$ each coordinate of the new vector $\mathbf{u}$ is defined as a linear combination of the initially–given vectors:
$u_j = \sum_{i=1}^n \frac{1}{\sqrt{\lambda}} v_i x_{ij} = \sum_{i=1}^n c_i x_{ij}$ (where $i \leq j \leq m$ and x_{ij} is the $(i,j)^{th}$ element of matrix X).
- Thus the j^{th} coordinate of the new vector is a *synthetic* value formed from the j^{th} coordinates of the given vectors (i.e. x_{ij} for all $1 \leq i \leq n$).

Finding Linear Combinations in Practice

- Say $\lambda_k = 0$.
- Then $X\mathbf{u} = \lambda\mathbf{u} = \mathbf{0}$
- Hence: $\sum_j u_j \mathbf{x}_j = 0$
- This allows redundancy in the form of linear combinations to be found.
- PCA is a linear transformation analysis method.
- But let's say we have three variables, y_1, y_2, and y_3.
- We would also input the variables y_1^2, y_2^2, y_3^2, $y_1 y_2$, $y_1 y_3$, and $y_2 y_3$.
- If the linear combination $y_1 = c_1 y_2^2 + c_2 y_1 y_2$ exists, then we would find it using PCA.
- Similarly we could feed in the logarithms or other functions of variables.

Finding Linear Combinations: Example

Thirty objects were used, and 5 variables defined as followsq.

$$y_{1j} = -1.4, -1.3, \ldots, 1.5$$

$$y_{2j} = 2.0 - y_{1j}^2$$

$$y_{3j} = y_{1j}^2$$

$$y_{4j} = y_{2j}^2$$

$$y_{5j} = y_{1j} y_{2j}$$

```
COVARIANCE MATRIX FOLLOWS.
  22.4750
  -2.2475   13.6498
   2.2475  -13.6498   13.6498
  -2.9262   28.0250  -28.0250   62.2917
  14.5189    0.5619   -0.5619    0.7316   17.3709
```

Finding Linear Combinations: Example

```
EIGENVALUES FOLLOW.
Eigenvalues        As Percentages    Cumul. Percentages
-----------        --------------    ------------------
  88.3852             68.2842              68.2842
  34.5579             26.6985              94.9828
   5.2437              4.0512              99.0339
   1.2505              0.9661             100.0000
   0.0000              0.0000             100.0000
```

The fifth eigenvalue is zero.

Finding Linear Combinations: Example

EIGENVECTORS FOLLOW.

VBLE.	EV-1	EV-2	EV-3	EV-4	EV-5
1	-0.0630	0.7617	0.6242	-0.1620	0.0000
2	0.3857	0.0067	-0.1198	-0.5803	0.7071
3	-0.3857	-0.0067	0.1198	0.5803	0.7071
4	0.8357	0.0499	0.1593	0.5232	0.0000
5	0.0018	0.6460	-0.7458	0.1627	0.0000

Since we know that the eigenvectors are centred, we have the equation:

$0.7071\mathbf{y}_2 + 0.7071\mathbf{y}_3 = 0.0$

Normalization or Standardization

- Let r_{ij} be the original measurements.
- Then define: $x_{ij} = \frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}}$
- $\overline{r}_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}$
- $s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)^2$
- Then te matrix to be diagonalized, $X'X$, is of $(j,k)^{th}$ term:
  $\rho_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)(r_{ik} - \overline{r}_k)/s_j s_k$
- This is the correlation coefficient between variables j and k.
- Have distance
  $d^2(j,k) = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 = \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - 2 \sum_{i=1}^{n} x_{ij} x_{ik}$
- First two terms both yield 1. Hence:
- $d^2(j,k) = 2(1 - \rho_{jk})$

- Thus the distance between variables is directly proportional to the correlation between them.
- For row points (objects, observations):
  $d^2(i,h) = \sum_j (x_{ij} - x_{hj})^2 = \sum_j (\frac{r_{ij} - r_{hj}}{\sqrt{n} s_j})^2 = (\mathbf{r}_i - \mathbf{r}_h)' M (\mathbf{r}_i - \mathbf{r}_h)$
- $\mathbf{r}_i$ and $\mathbf{r}_h$ are column vectors (of dimensions $m \times 1$) and M is the $m \times m$ diagonal matrix of j^{th} element $1/ns_j^2$.
- Therefore d is a Euclidean distance associated with matrix M.
- Note that the row points are now centred but the column points are not: therefore the latter may well appear in one quadrant on output listings.

Implications of Standardization

- Analysis of the matrix of $(j,k)^{th}$ term ρ_{jk} as defined above is PCA on a *correlation* matrix.
- The row vectors are centred and reduced.
- Centring alone used, and not the rescaling of the variance: matrix of $(j,k)^{th}$ term $c_{jk} = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)(r_{ik} - \overline{r}_k)$
- In this case we have PCA of the *variance-covariance* matrix.
- If we use no normalization, we have PCA of the *sums of squares and cross-products* matrix. That was what we used to begin with.
- Usually it is best to carry out analysis on correlations.

Iterative Solution of Eigenvalue Equations

- Solve: $A\mathbf{u} = \lambda\mathbf{u}$
- Choose some trial vector, $\mathbf{t}_0$: e.g. $(1, 1, \ldots, 1)$.
- Then define $\mathbf{t}_1, \mathbf{t}_2, \ldots$:
- $$\begin{array}{ll} A\mathbf{t}_0 = \mathbf{x}_0 & \mathbf{t}_1 = \mathbf{x}_0/\sqrt{\mathbf{x}_0'\mathbf{x}_0} \\ A\mathbf{t}_1 = \mathbf{x}_1 & \mathbf{t}_2 = \mathbf{x}_1/\sqrt{\mathbf{x}_1'\mathbf{x}_1} \\ A\mathbf{t}_2 = \mathbf{x}_2 & \mathbf{t}_3 = \ldots \end{array}$$
- Halt when there is convergence.
- $|\mathbf{t}_n - \mathbf{t}_{n+1}| \leq \epsilon$
- At convergence, $\mathbf{t}_n = \mathbf{t}_{n+1}$
- Hence: $A\mathbf{t}_n = \mathbf{x}_n$
- $\mathbf{t}_{n+1} = \mathbf{x}_n/\sqrt{\mathbf{x}_n'\mathbf{x}_n}$.

- Substituting for $\mathbf{x}_n$ in the first of these two equations gives:
- $A\mathbf{t}_n = \sqrt{\mathbf{x}_n'\mathbf{x}_n}\ \mathbf{t}_{n+1}$.
- Hence $\mathbf{t}_n = \mathbf{t}_{n+1}$, $\mathbf{t}_n$ is the eigenvector, and the associated eigenvalue is $\sqrt{\mathbf{x}_n'\mathbf{x}_n}$.
- The second eigenvector and associated eigenvalue may be found by carrying out a similar iterative algorithm on a matrix where the effects of $\mathbf{u}_1$ and λ_1 have been *partialled out*:
- $A_{(2)} = A - \lambda_1 \mathbf{u}_1 \mathbf{u}_1'$.
- Let us prove that $A_{(2)}$ removes the effects due to the first eigenvector and eigenvalue.
- We have $A\mathbf{u} = \lambda\mathbf{u}$.
- Therefore $A\mathbf{u}\mathbf{u}' = \lambda\mathbf{u}\mathbf{u}'$;
- Or equivalently, $A\mathbf{u}_k\mathbf{u}_k' = \lambda_k\mathbf{u}_k\mathbf{u}_k'$ for each eigenvalue.
- Summing over k gives: $A\sum_k \mathbf{u}_k\mathbf{u}_k' = \sum_k \lambda_k\mathbf{u}_k\mathbf{u}_k'$.

- The summed term on the left hand side equals the identity matrix.
- Therefore $A = \lambda_1 \mathbf{u}_1 \mathbf{u}_1' + \lambda_2 \mathbf{u}_2 \mathbf{u}_2' + \ldots$
- From this *spectral decomposition* of matrix A, we may successively remove the effects of the eigenvectors and eigenvalues as they are obtained.
- See Press et al., Numerical Recipes, Cambridge Univ. Press, for other (better!) algorithms.

Objectives of PCA

- dimensionality reduction;
- the determining of linear combinations of variables;
- feature selection: the choosing of the most useful variables;
- visualization of multidimensional data;
- identification of underlying variables;
- identification of groups of objects or of outliers.

Indicative Procedure Followed

- Ignore principal components if the new axes retained explain $> 75\%$ of the variance.
- Look at projections of rows, or columns, in planes (1,2), (1,3), (2,3), etc.
- Projections of correlated variables are close (if we have carried out a PCA on correlations).
- PCA is sometimes motivated by the search for latent variables: i.e. characterization of principal components.
- Highest or lowest projection values may help with this.
- Clusters and outliers can be found using planar projections.

PCA with Multiband Data

- Consider a set of image bands (from a multiband or multispectral or hyspectral) data set, or frames (from video). Say we have p images, each of dimensions $n \times m$.
- We define the "eigen-images" as follows.
- Each pixel can be considered as associated with a vector of dimension p. We can take this as defining a matrix for analysis of number of rows $= n.m$, and number of columns = p.
- Carry out a PCA. The row projections define a matrix with $n.m$ rows and $p' < p$ columns. If we keep just the first eigenvector, then we have a matrix of dimensions $n.m \times 1$.
- Say $n = 512, m = 512, p = 6$. The eigenvalue/vector finding is carried out on a $p \times p$ correlation matrix. Eigenvector/value finding has computational cost $O(p^3)$.

- For just one principal component, $p' = 1$, convert the matrix of dimensions $n.m \times 1$ back to an image of dimensions $n \times m$ pixels.
- Applications: finding typical or "eigen" face in face recognition; or finding typical or "eigen" galaxy in galaxy morphology.
- What are the conditions for such a procedure to work well?

Some References

- A. Bijaoui, “Application astronomique de la compression de l’information”, *Astronomy and Astrophysics*, **30**, 199–202, 1974.
- P. Brosche, “The manifold of galaxies: Galaxies with known dynamical properties”, *Astronomy and Astrophysics*, **23**, 259–268, 1973.
- P. Brosche and F.T. Lentes, “The manifold of globular clusters”, *Astronomy and Astrophysics*, **139**, 474–476, 1984.
- V. Bujarrabal, J. Guibert and C. Balkowski, “Multidimensional statistical analysis of normal galaxies”, *Astronomy and Astrophysics*, **104**, 1–9, 1981.
- R. Buser, “A systematic investigation of multicolor photometric systems. I. The UBV, RGU and *uvby* systems.”, *Astronomy and Astrophysics*, **62**, 411–424, 1978.
- C.A. Christian and K.A. Janes, “Multivariate analysis of spectrophotometry”. *Publications of the Astronomical Society of the Pacific*, **89**, 415–423, 1977.

- C.A. Christian, "Identification of field stars contaminating the colour–magnitude diagram of the open cluster Be 21", *The Astrophysical Journal Supplement Series*, **49**, 555–592, 1982.
- T.J. Deeming, "Stellar spectral classification. I. Application of component analysis", *Monthly Notices of the Royal Astronomical Society*, **127**, 493–516, 1964. G. Efstathiou and S.M. Fall, "Multivariate analysis of elliptical galaxies", *Monthly Notices of the Royal Astronomical Society*, **206**, 453–464, 1984.
- M. Fofi, C. Maceroni, M. Maravalle and P. Paolicchi, "Statistics of binary stars. I. Multivariate analysis of spectroscopic binaries", *Astronomy and Astrophysics*, **124**, 313–321, 1983.
- M. Fracassini, L.E. Pasinetti, E. Antonello and G. Raffaelli, "Multivariate analysis of some ultrashort period Cepheids (USPC)", *Astronomy and Astrophysics*, **99**, 397–399, 1981.
- P. Galeotti, "A statistical analysis of metallicity in spiral galaxies", *Astrophysics and Space Science*, **75**, 511–519, 1981.

- S. Okamura, K. Kodaira and M. Watanabe, "Digital surface photometry of galaxies toward a quantitative classification. III. A mean concentration index as a parameter representing the luminosity distribution", *The Astrophysical Journal*, **280**, 7–14, 1984.