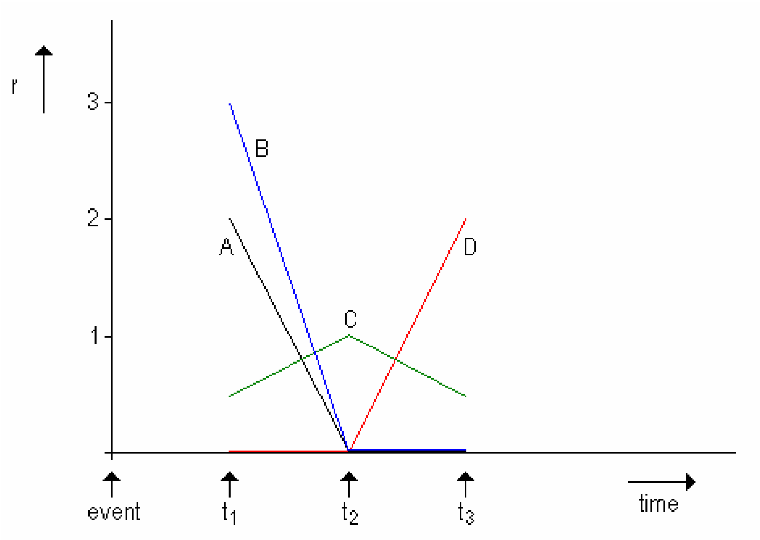
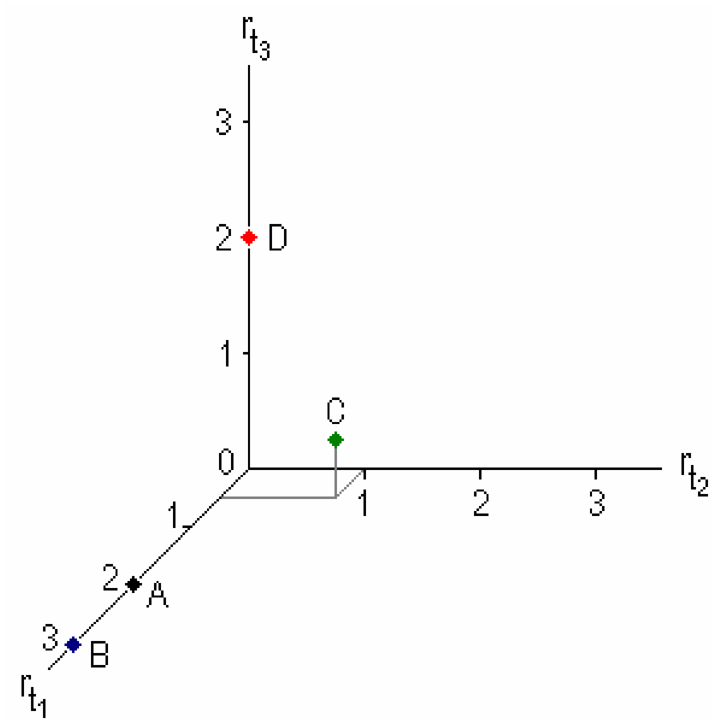
Principal Component Analysis

Some Mathematical Backgrounds

Arie van Erk,
BiGCaT
arie.vanerk@bigcat.unimaas.nl



We are working with this representation:



gene	t_1	t_2	t_3
A	2	0	0
B	3	0	0
C	0.5	1	0.5
D	0	0	2

Principal Component Analysis (PCA):

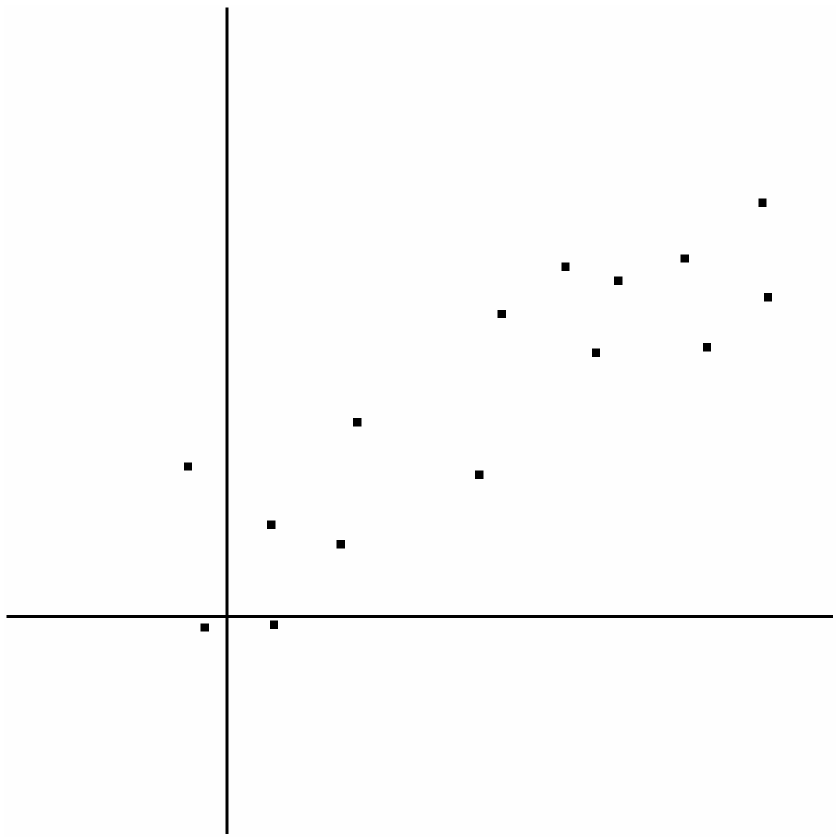
Tool for screening data:

- are there strange or unusual aspects?
- do the data have a normal distribution?
- are there outliers?

Transformation of the data:

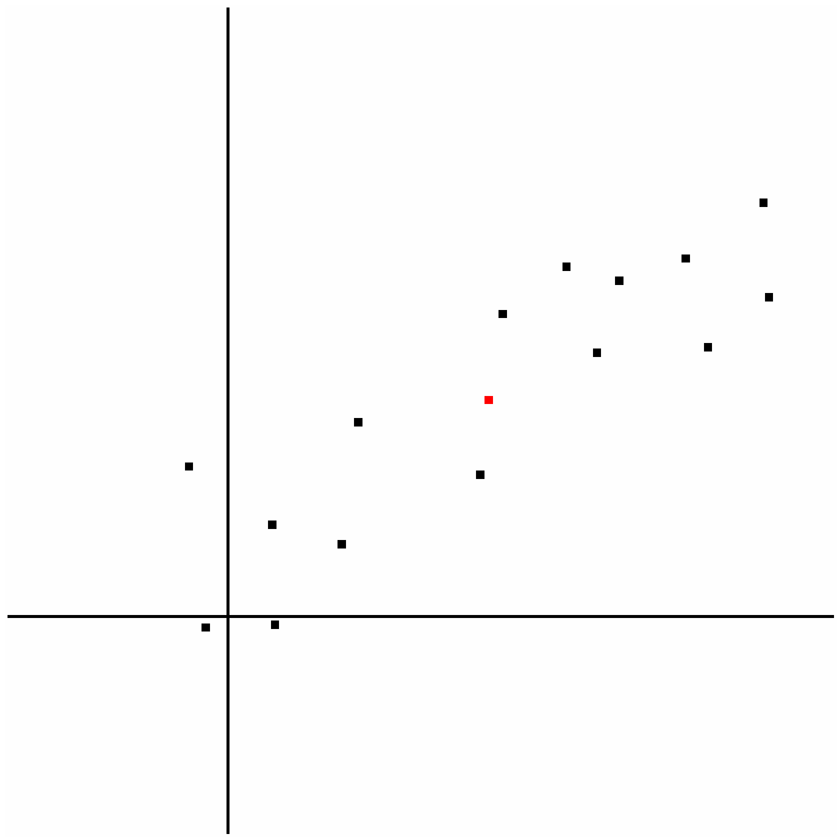
- into new set of variables (principal components)
- which are usually not interpretable
- first pc accounts for as much of the variability as possible
- and each succeeding pc as much of the remaining variability

Principal Component Analysis



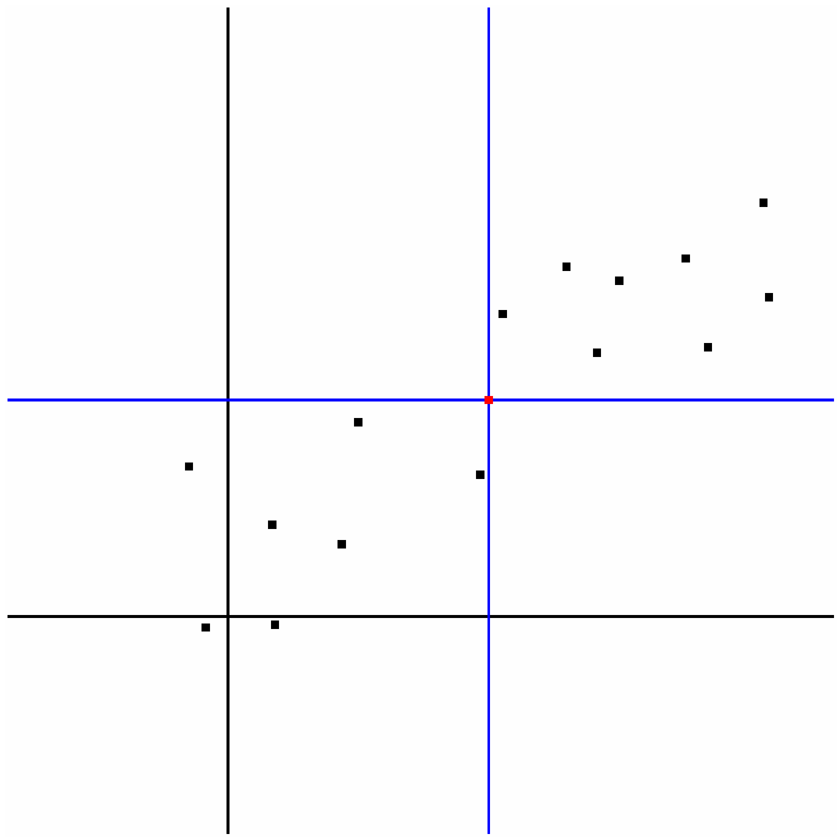
For a given data set …

Principal Component Analysis



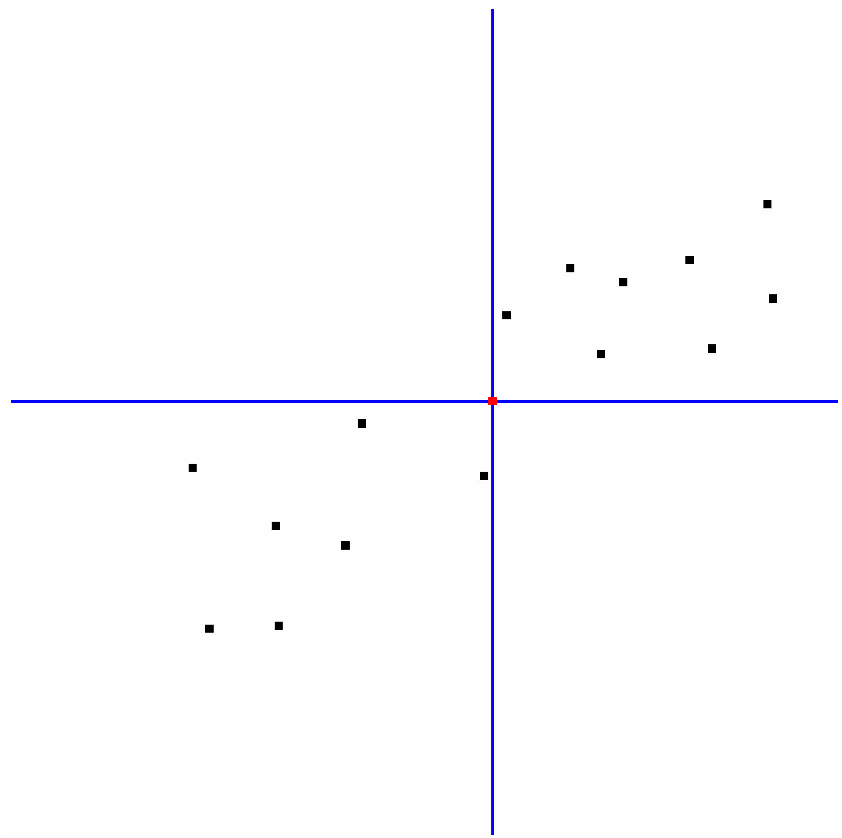
calculate the centroid (='mean in all directions') …

Principal Component Analysis



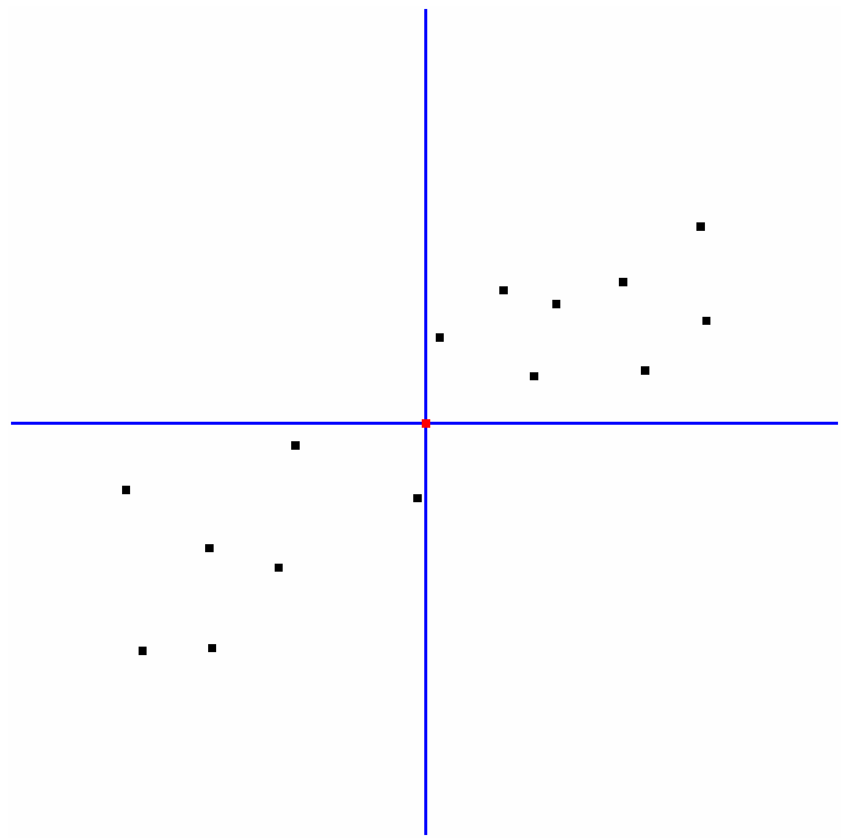
shift the grid to the centroid …

Principal Component Analysis

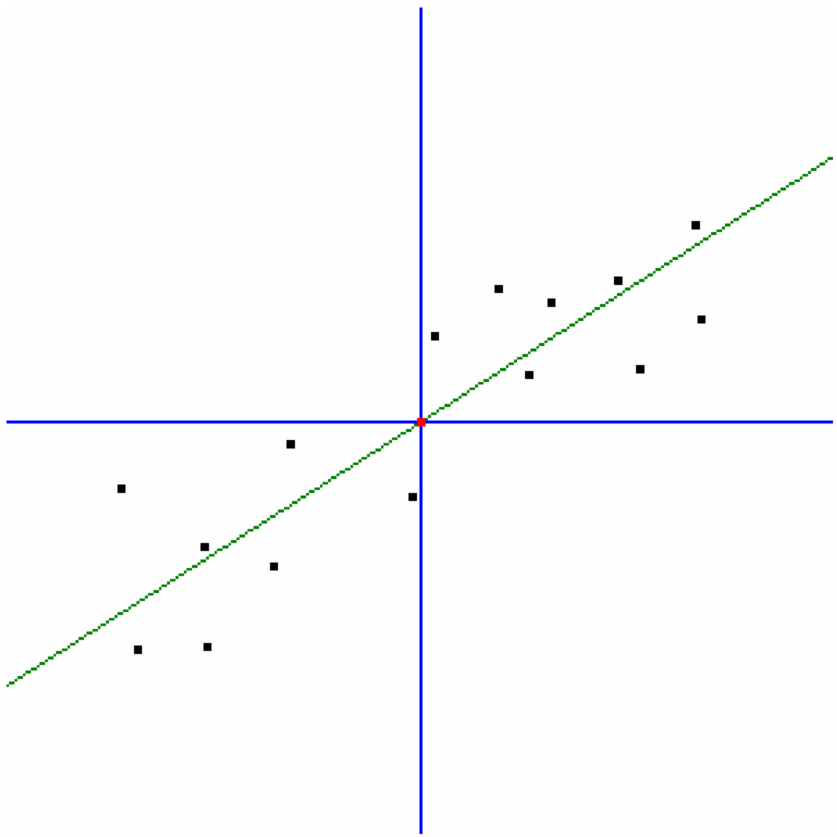


take this as our new coordinate system …

Principal Component Analysis

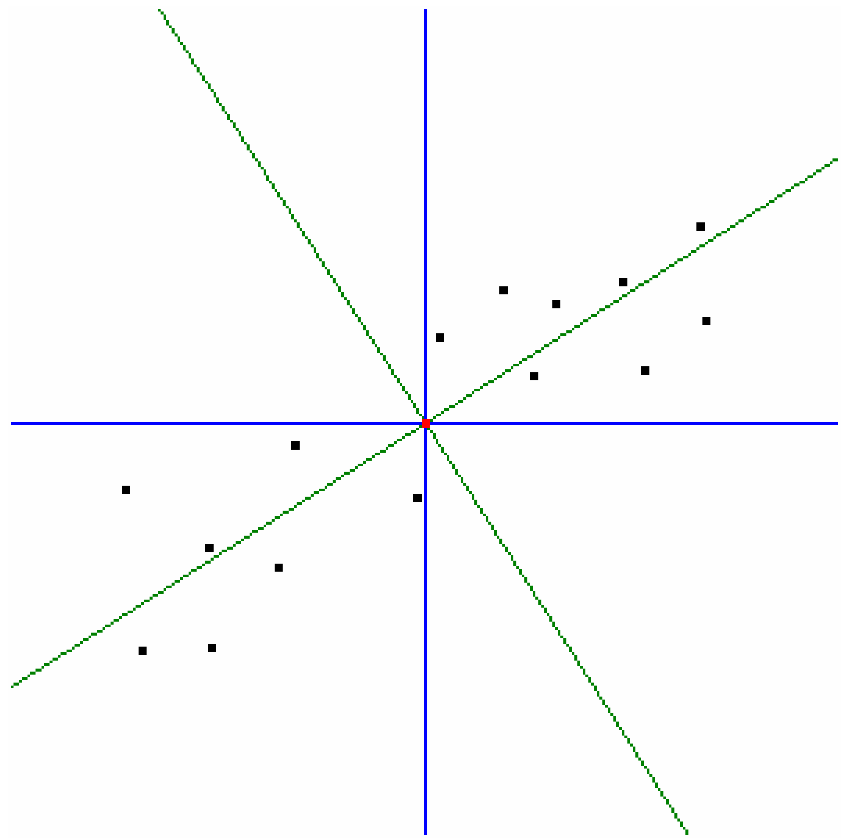


Principal Component Analysis



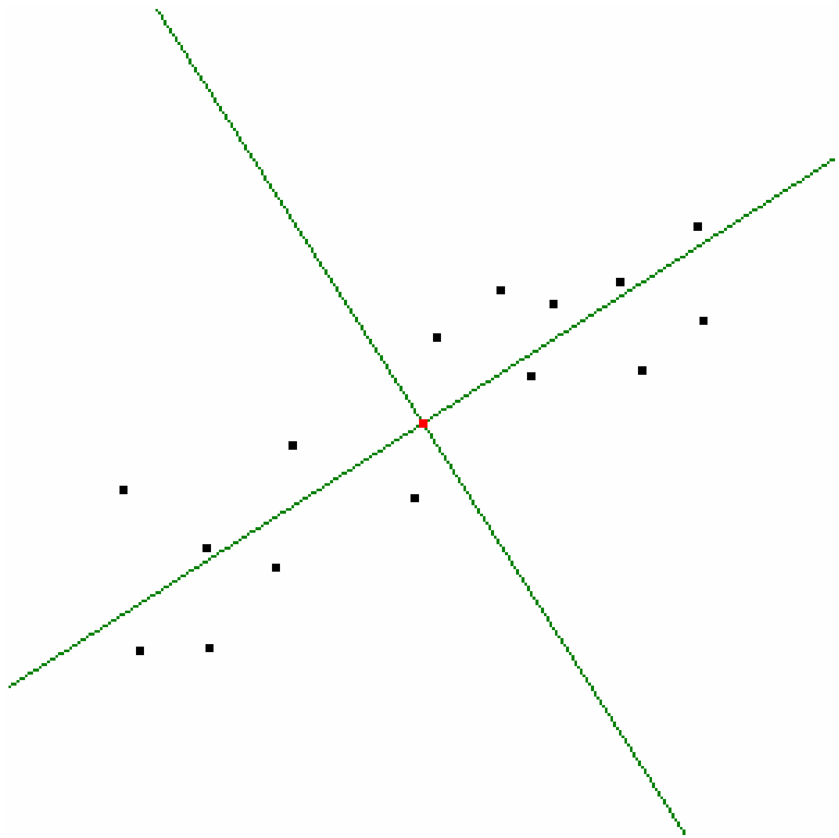
calculate the direction in which the variance is maximal …

Principal Component Analysis



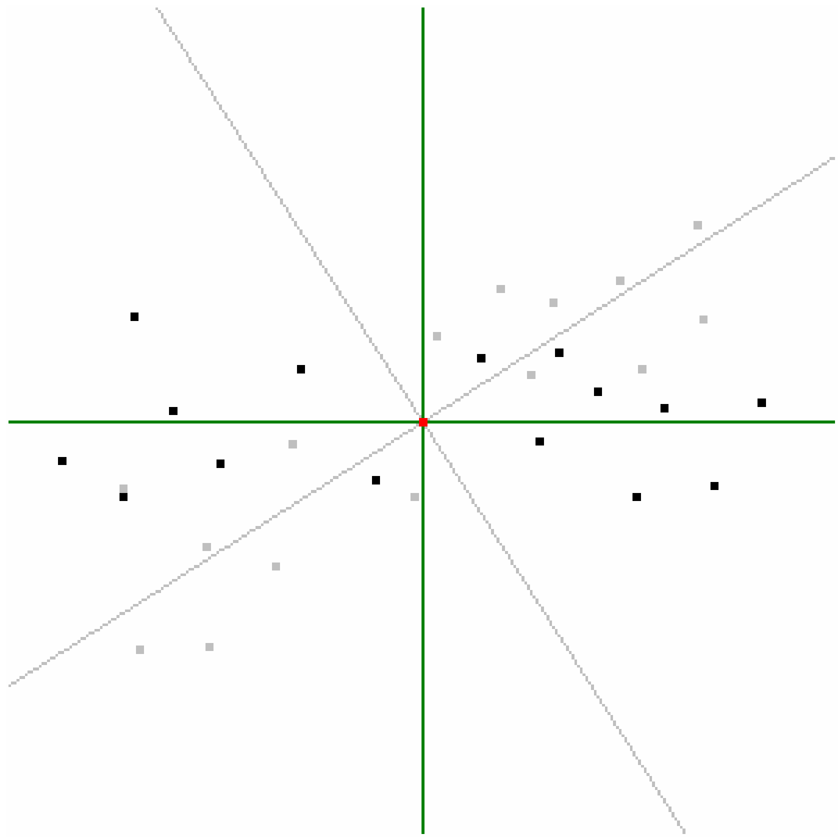
and repeat this for each next perpendicular axis …

Principal Component Analysis



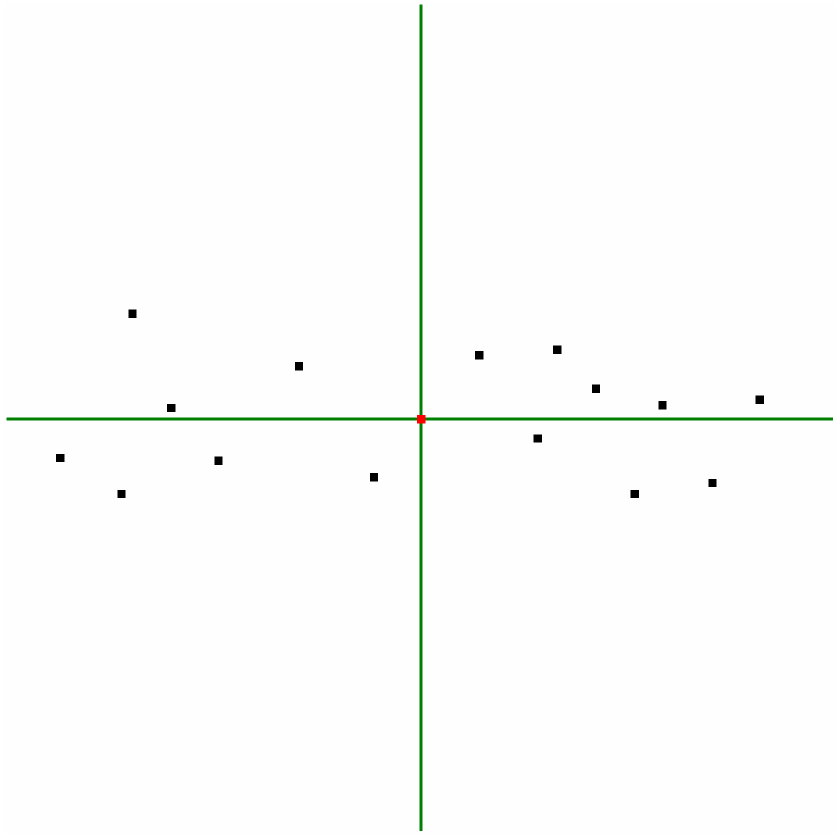
leaving us with a rotated grid …

Principal Component Analysis



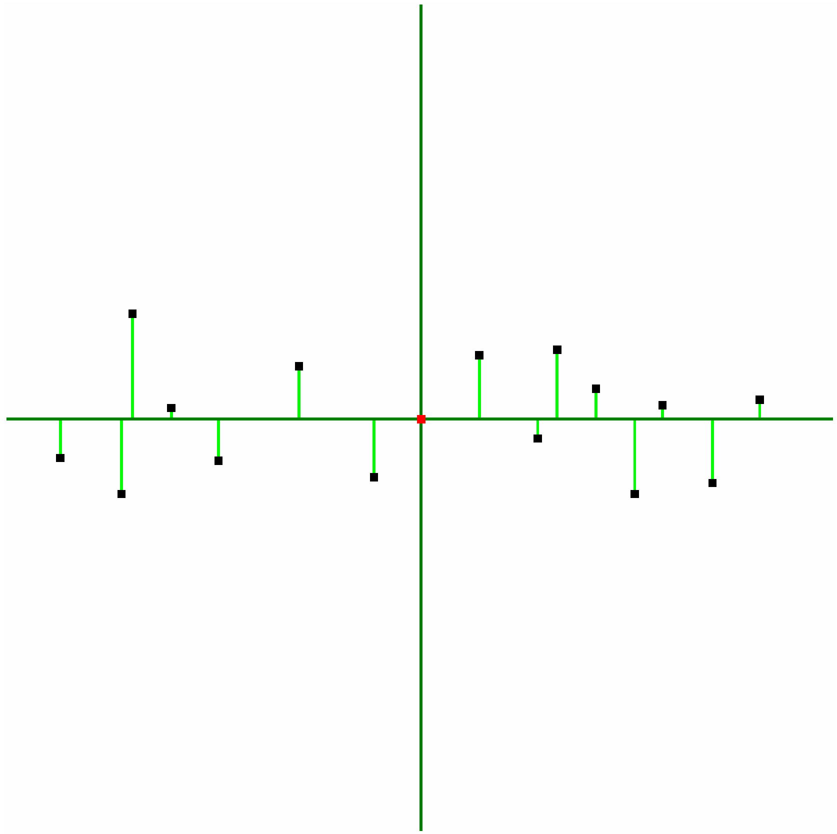
which we can rotate to a 'normal' position …

Principal Component Analysis



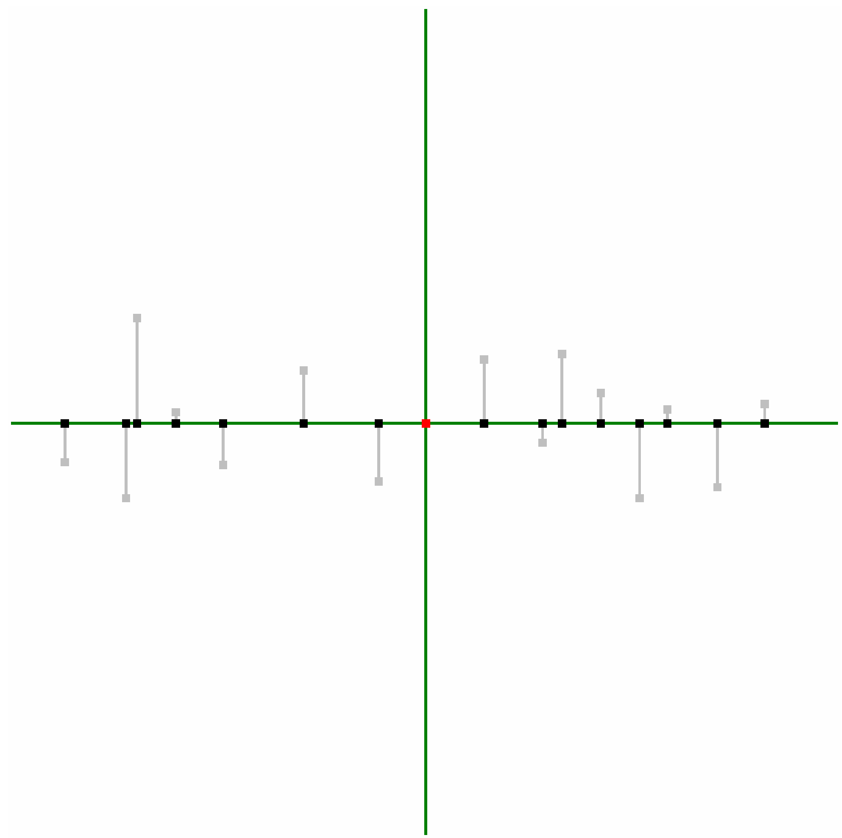
showing us maximal variance.

Principal Component Analysis



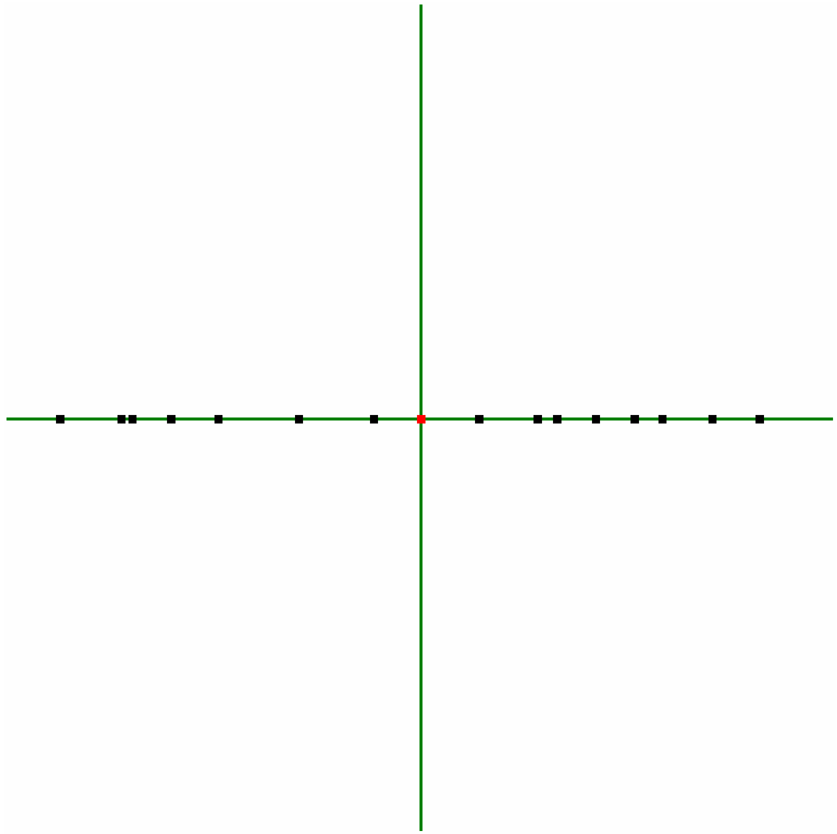
We can also use this to reduce the complexity of the data set …

Principal Component Analysis



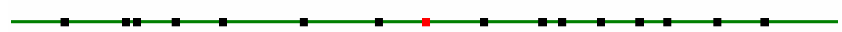
by eliminating a number of axis by projection of the points.

Principal Component Analysis



in this example moving from two …

Principal Component Analysis

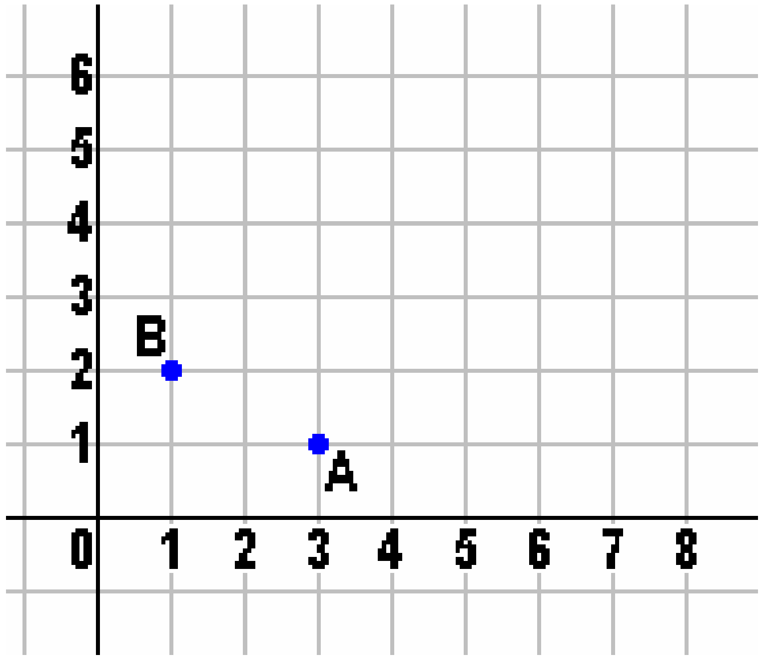


to one dimensional data points.

Principal Component Analysis

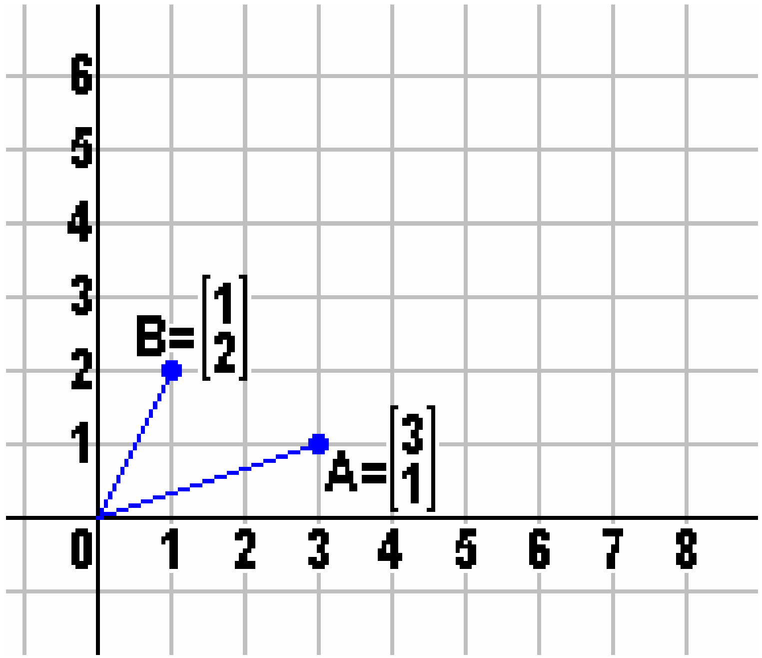
Now, what has happened?

Principal Component Analysis

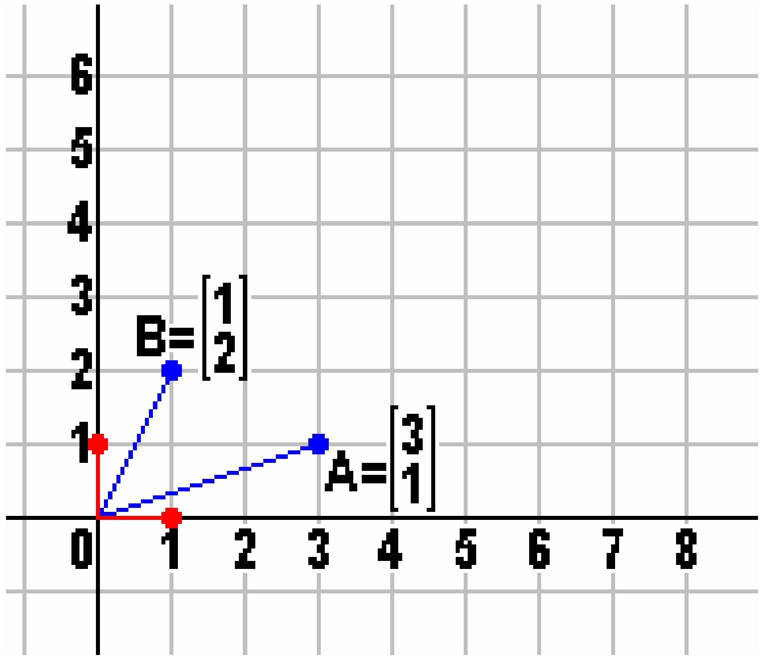


Remember the notations from linear algebra…

Principal Component Analysis

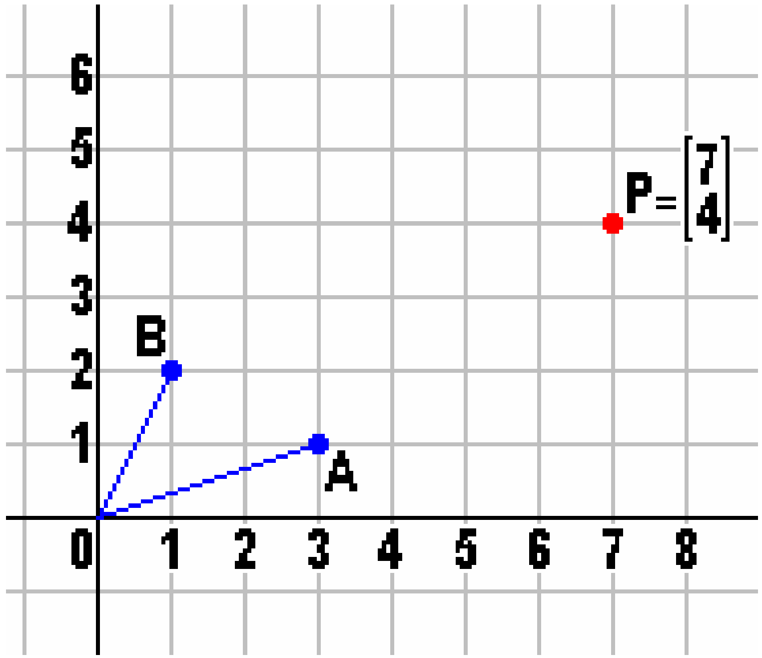


Principal Component Analysis

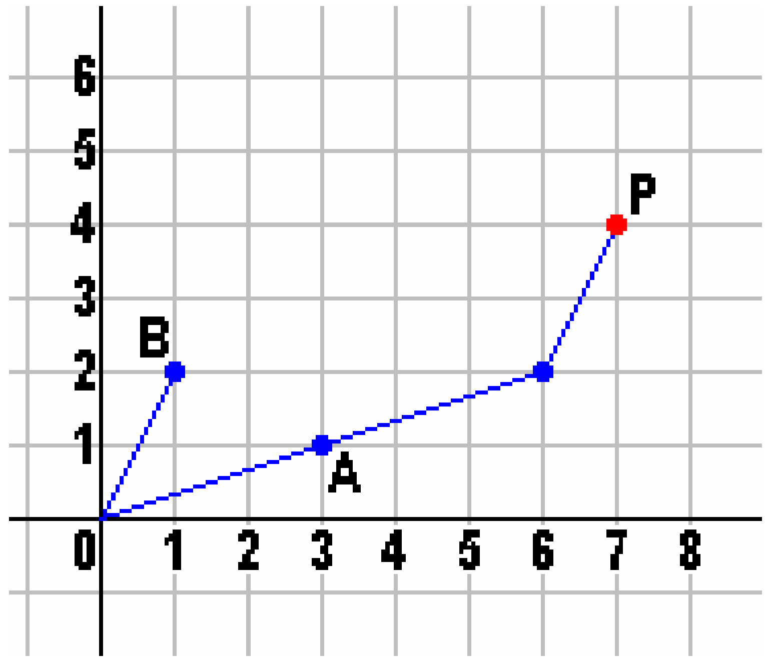


$$A = 3 \bullet \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 1 \bullet \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Principal Component Analysis

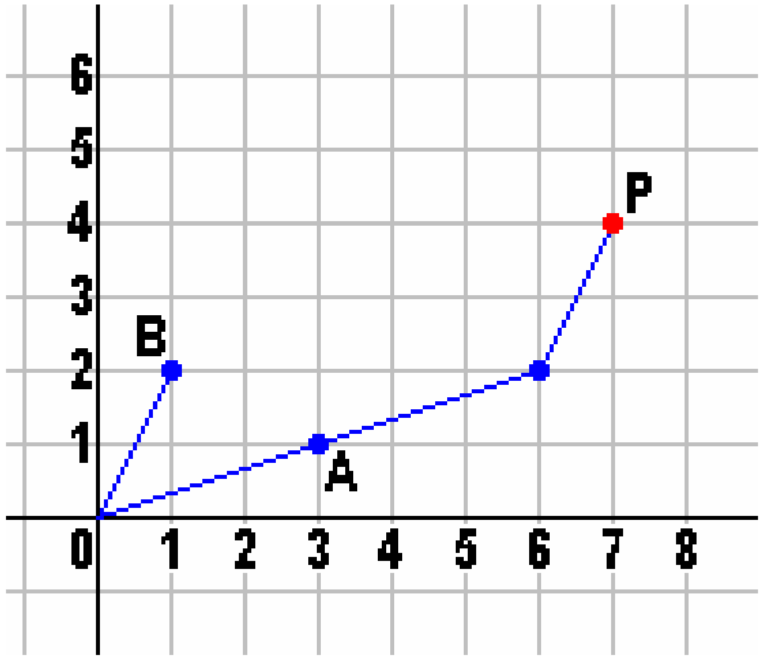


Principal Component Analysis



$$\mathbf{p} = 2 \bullet \mathbf{a} + 1 \bullet \mathbf{b}$$

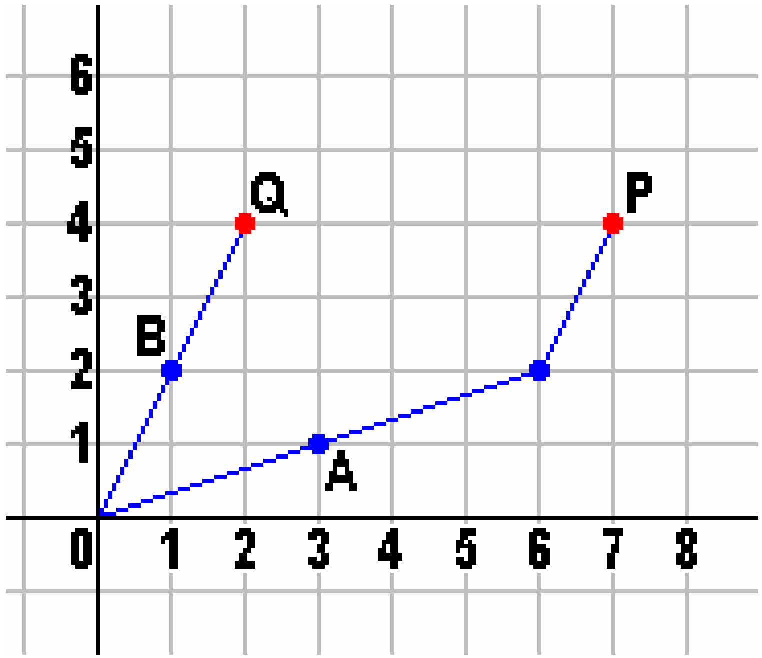
Principal Component Analysis



$$\mathbf{p} = 2 \bullet \mathbf{a} + 1 \bullet \mathbf{b}$$

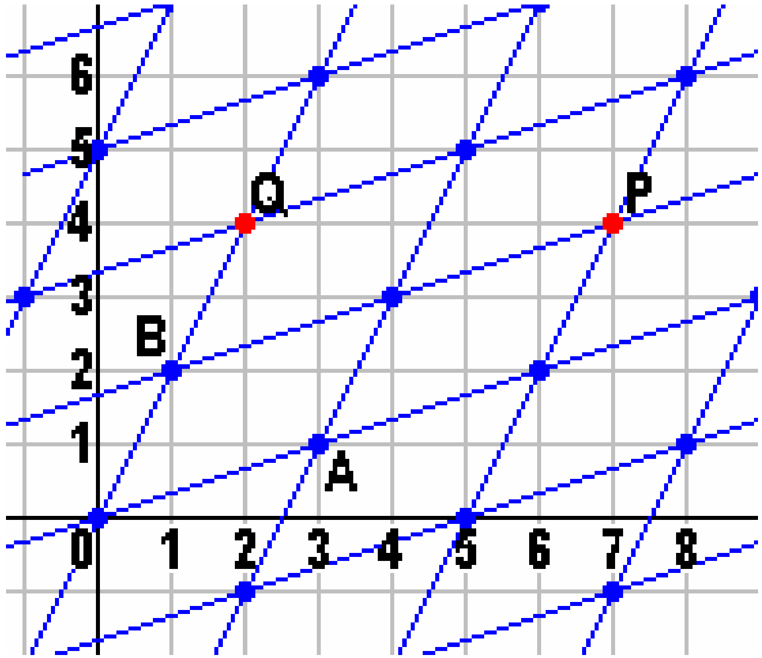
$$\mathbf{p} = 2 \bullet \begin{bmatrix} 3 \\ 1 \end{bmatrix} + 1 \bullet \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \bullet 3 + 1 \bullet 1 \\ 2 \bullet 1 + 1 \bullet 2 \end{bmatrix} = \begin{bmatrix} 7 \\ 4 \end{bmatrix}$$

Principal Component Analysis



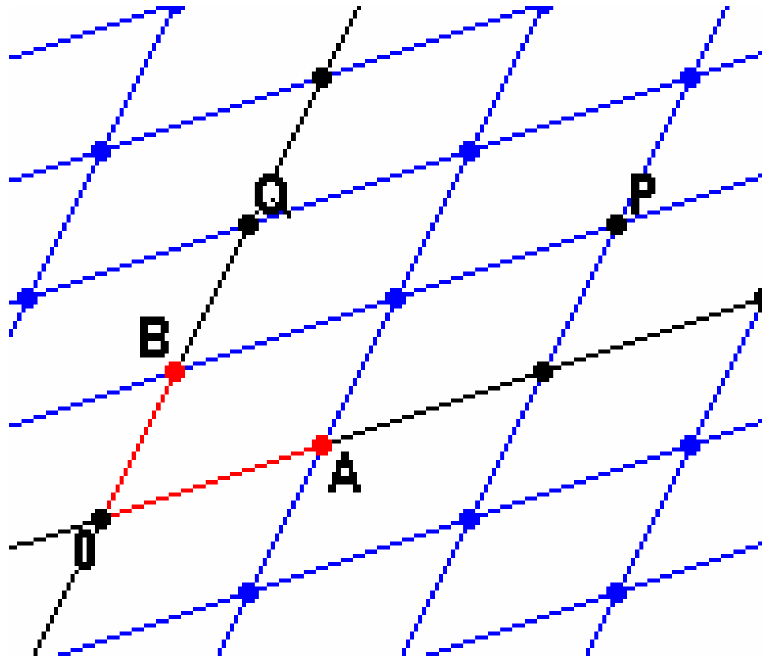
$$\mathbf{q} = 0 \bullet \mathbf{a} + 2 \bullet \mathbf{b}$$

Principal Component Analysis



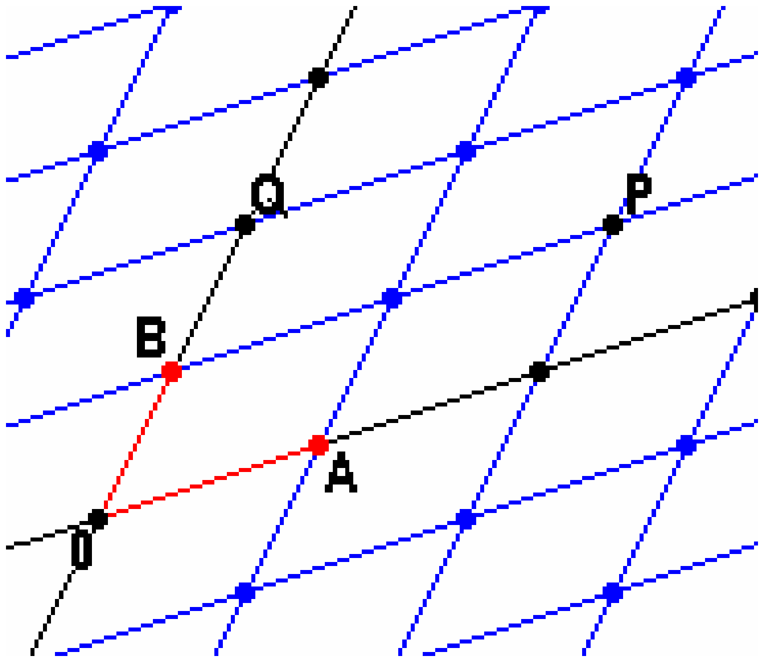
So, we can express each point in terms of A and B …

Principal Component Analysis



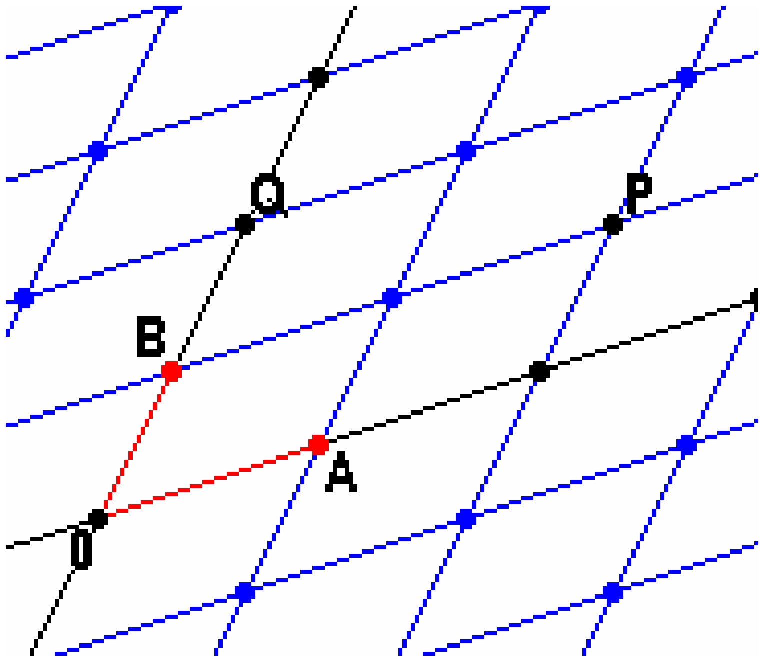
which form another base for our grid.

Principal Component Analysis



$$\mathbf{v}' = \begin{bmatrix} u \\ w \end{bmatrix} \rightarrow \mathbf{v} = u \bullet \begin{bmatrix} 3 \\ 1 \end{bmatrix} + w \bullet \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \bullet \begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} 3 \bullet u + 1 \bullet w \\ 1 \bullet u + 2 \bullet w \end{bmatrix}$$

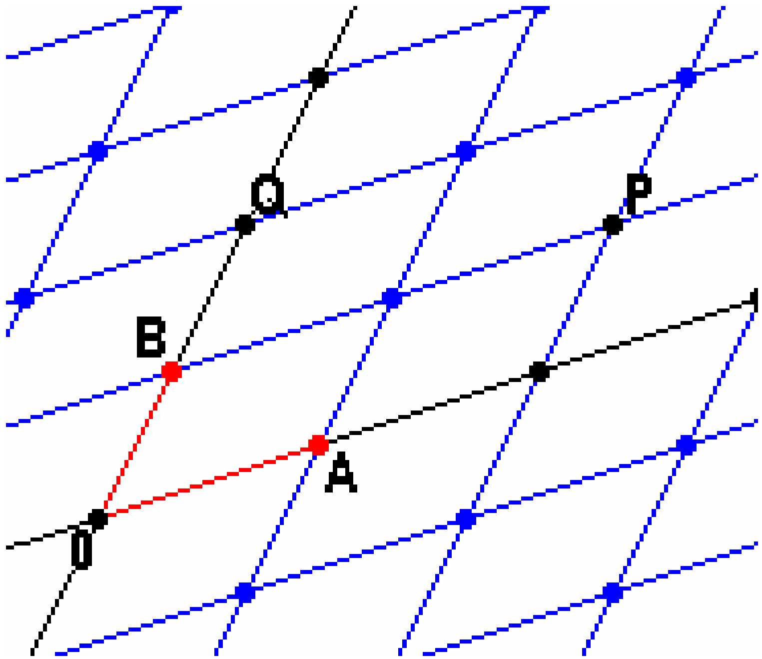
Principal Component Analysis



$$\mathbf{v}' = \begin{bmatrix} u \\ w \end{bmatrix} \rightarrow \mathbf{v} = u \bullet \begin{bmatrix} 3 \\ 1 \end{bmatrix} + w \bullet \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \bullet \begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} 3 \bullet u + 1 \bullet w \\ 1 \bullet u + 2 \bullet w \end{bmatrix}$$

$$\mathbf{v} = \mathbf{A} \bullet \mathbf{v}'$$

Principal Component Analysis



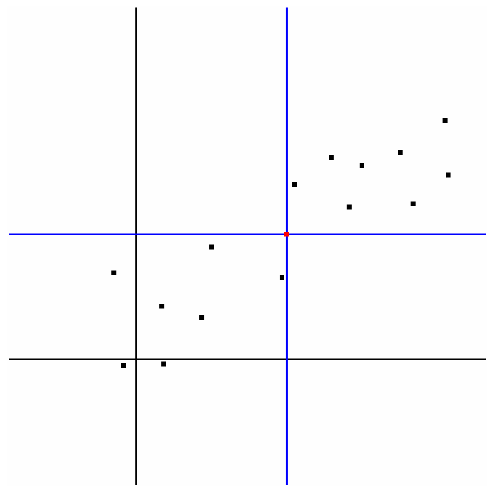
$$\mathbf{v}' = \begin{bmatrix} u \\ w \end{bmatrix} \rightarrow \mathbf{v} = u \bullet \begin{bmatrix} 3 \\ 1 \end{bmatrix} + w \bullet \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \bullet \begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} 3 \bullet u + 1 \bullet w \\ 1 \bullet u + 2 \bullet w \end{bmatrix}$$

$$\mathbf{v} = \mathbf{A} \bullet \mathbf{v}' \qquad \mathbf{v}' = \mathbf{A}^{-1} \bullet \mathbf{v}$$

And now for PCA:

Principal Component Analysis

PCA step 1: moving the grid



Principal Component Analysis

	x = time1	y = time2
gene A	4	2
gene B	0	1
gene C	8	7
gene D	2	2
gene E	6	3

Principal Component Analysis

	x	y
gene A	4	2
gene B	0	1
gene C	8	7
gene D	2	2
gene E	6	3
	μ_1=4	μ_2= 3

Principal Component Analysis

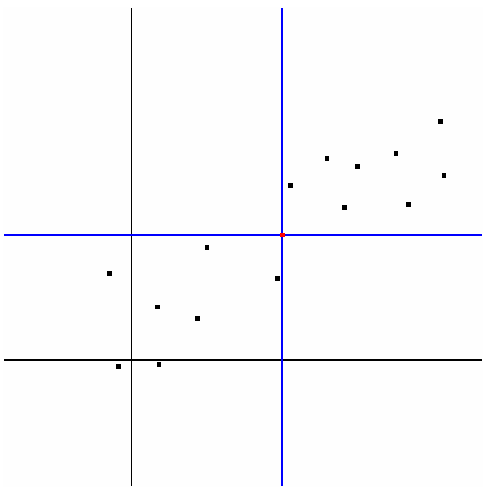
	$x-\mu_1$	$y-\mu_2$
gene A	4-4	2-3
gene B	0-4	1-3
gene C	8-4	7-3
gene D	2-4	2-3
gene E	6-4	3-3
	$\mu_1=4$	$\mu_2= 3$

Principal Component Analysis

	$x-\mu_1$	$y-\mu_2$
gene A	0	-1
gene B	-4	-2
gene C	4	4
gene D	-2	-1
gene E	2	0
	μ_1=4	μ_2= 3

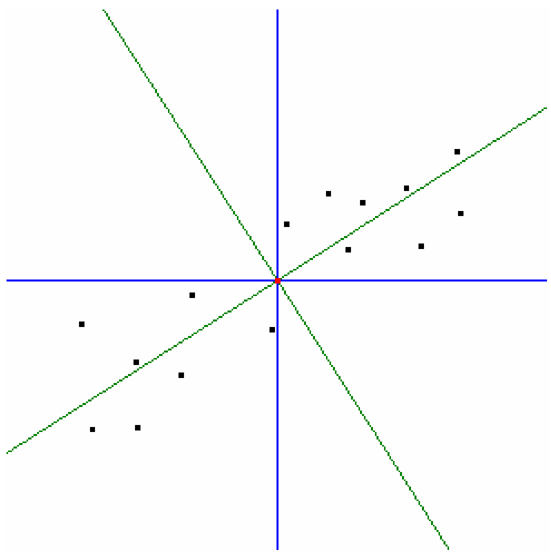
Principal Component Analysis

	$x'=x-\mu_1$	$y'=y-\mu_2$
gene A	0	-1
gene B	-4	-2
gene C	4	4
gene D	-2	-1
gene E	2	0



Principal Component Analysis

PCA step 2: rotating the grid, based on variance



Principal Component Analysis

$$\sigma_p^2 = \frac{\sum_{i=1}^{n}(p_i - \mu)^2}{n} = \frac{\sum p_i^{\,2}}{n} - \left(\frac{\sum p_i}{n}\right)^2 \qquad \text{variance}$$

$$\sigma^2_p = E(p^2) - [E(p)]^2$$

$p = x + y$:

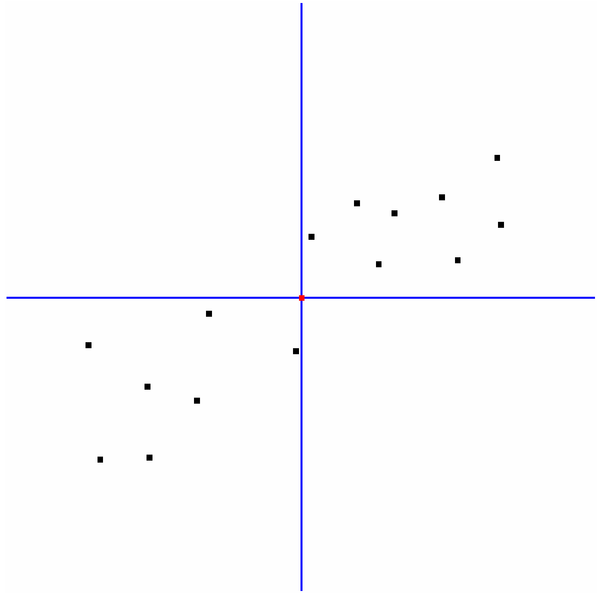
$$\begin{aligned}
\sigma^2_p &= E((x+y)^2) - [E(x+y)]^2 \\
&= E(x^2)+2E(xy)+E(y^2) - [E(x)+E(y)]^2 \\
&= E(x^2)+2E(xy)+E(y^2) - [E(x)]^2-2E(x)E(y)-[E(y)]^2 \\
&= E(x^2)-[E(x)]^2 + E(y^2)-[E(y)]^2 + 2E(xy)-2E(x)E(y) \\
&= \sigma^2_x + \sigma^2_y + 2\sigma^2_{x,y}
\end{aligned}$$

$$\sigma^2_{x,y} = E(xy) - E(x)E(y) \qquad \text{covariance}$$

Principal Component Analysis

$\sigma^2_{x,y} = E(xy) - E(x)E(y)$ covariance

	x_1 (=x')	x_2 (=y')
gene A	0	-1
gene B	-4	-2
gene C	4	4
gene D	-2	-1
gene E	2	0



$\sigma^2_{x1,x2} = E(x_1x_2) = (0+8+16+2+0)/5 = 5.2$

$\sigma^2_{x2,x1} = \sigma^2_{x1,x2} = 5.2$

$\sigma^2_{x1,x1} = E(x_1x_1) = (0+16+16+4+4)/5 = 8$

$\sigma^2_{x2,x2} = E(x_2x_2) = (1+4+16+1+0)/5 = 4.4$

Principal Component Analysis

$\sigma^2_{1,2} = \sigma^2_{x1,x2} = 5.2$
$\sigma^2_{2,1} = \sigma^2_{1,2} = 5.2$
$\sigma^2_{1,1} = 8$
$\sigma^2_{2,2} = 4.4$

Covariance Matrix:

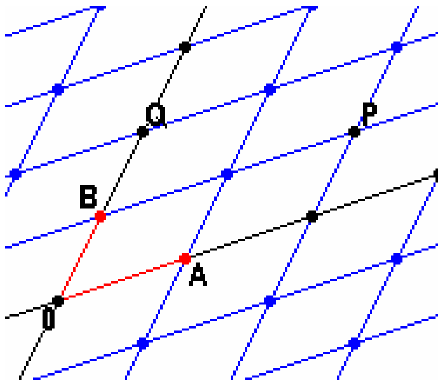
$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} & \dots & \sigma^2_{1,m} \\ \sigma^2_{2,1} & \sigma^2_{2,2} & \dots & \sigma^2_{2,m} \\ \dots & \dots & \dots & \dots \\ \sigma^2_{m,1} & \sigma^2_{m,2} & \dots & \sigma^2_{m,m} \end{bmatrix}$$

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix}$$

Principal Component Analysis

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix}$$

$$\sigma^2_p = \sigma^2_x + \sigma^2_y + 2\sigma^2_{x,y} \longrightarrow \sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$



$$\mathbf{p} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \bullet \mathbf{p'} = \mathbf{A} \bullet \mathbf{p'}$$

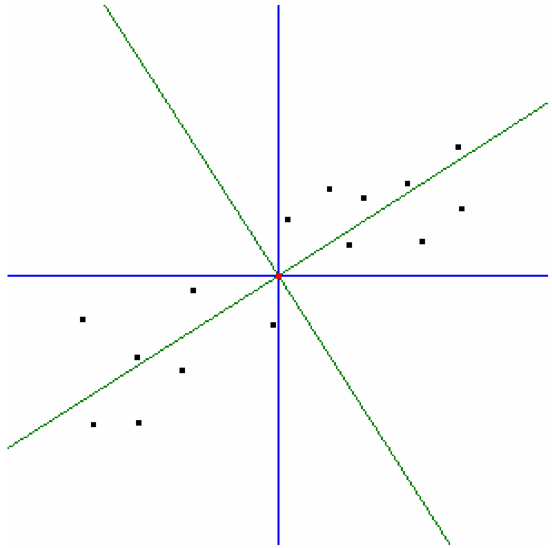
Principal Component Analysis

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix}$$

$$\sigma^2{}_p = \sigma^2{}_x + \sigma^2{}_y + 2\sigma^2{}_{x,y} \xrightarrow[\mathbf{p} = X \bullet \mathbf{p'}]{} \sigma^2{}_{p'} = \sigma^2{}_{x'} + \sigma^2{}_{y'}$$

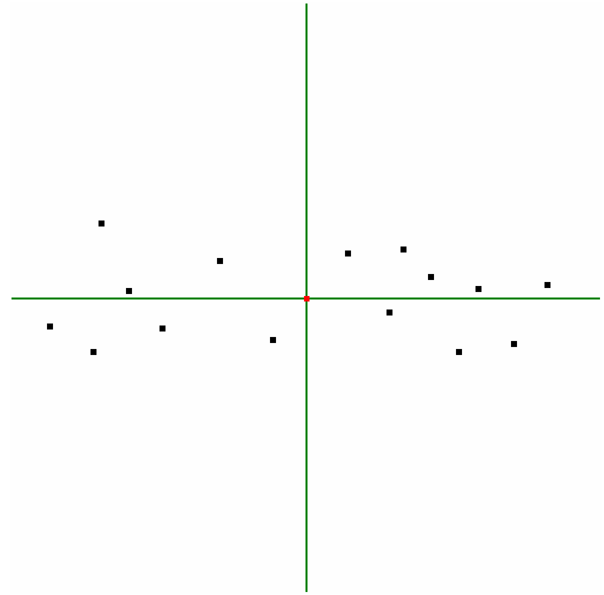
$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} \rightarrow C' = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix}$$



Principal Component Analysis

$$\sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$



$$C' = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix}$$

$$X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix} = \begin{bmatrix} 1' & 0' \\ 0' & 1' \end{bmatrix}$$

for each **v'** on x'-axis: $\mathbf{v'} = \begin{bmatrix} v' \\ 0 \end{bmatrix}$

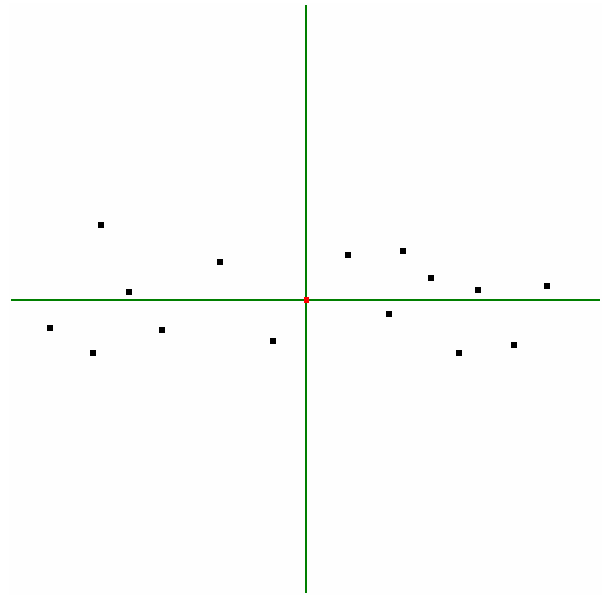
$$\mathrm{cov}(\mathbf{v'}) = C' \bullet \mathbf{v'} = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix} \bullet \begin{bmatrix} v' \\ 0 \end{bmatrix} = \begin{bmatrix} v' \bullet e'_1 \\ 0 \end{bmatrix} = e'_1 \bullet \begin{bmatrix} v' \\ 0 \end{bmatrix} = e'_1 \bullet \mathbf{v'}$$

$$C' \bullet \mathbf{v'} = \lambda_1 \bullet \mathbf{v'}$$

Principal Component Analysis

v_1' on x-axis: $C' \bullet \mathbf{v_1'} = \lambda_1 \bullet \mathbf{v_1'}$

v_2' on y-axis: $C' \bullet \mathbf{v_2'} = \lambda_2 \bullet \mathbf{v_2'}$



λ_i = eigenvalue of C'
v_i = eigenvector corresponding to λ_i

The value of λ_i corresponds to the variance on the x_i-axis

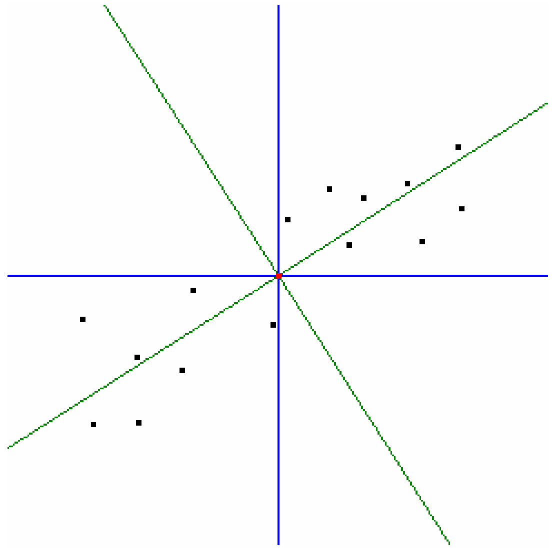
Principal Component Analysis

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix}$$

$$\sigma^2_{p} = \sigma^2_{x} + \sigma^2_{y} + 2\sigma^2_{x,y} \xrightarrow[\mathbf{p} = X \bullet \mathbf{p'}]{} \sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$

$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} \rightarrow C' = \begin{bmatrix} e'_1 & 0 \\ 0 & e'_2 \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix}$$



Principal Component Analysis

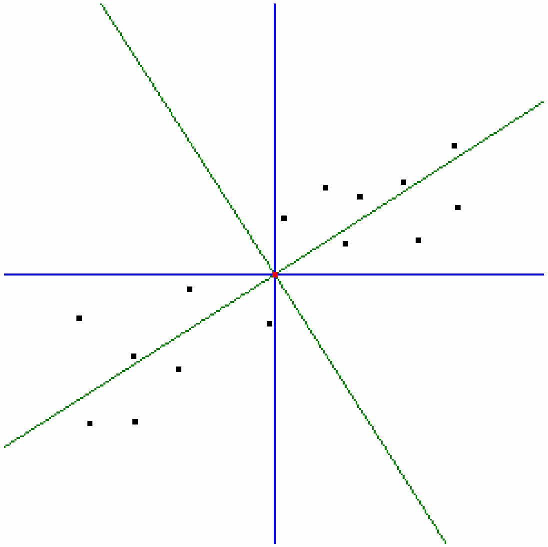
$$C = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix} \qquad X = \begin{bmatrix} ev_{1,1} & ev_{1,2} \\ ev_{2,1} & ev_{2,2} \end{bmatrix}$$

We have to solve: $C \bullet \mathbf{x} = \lambda \bullet \mathbf{x}$ for all λ and $\mathbf{x}$ (with $|\mathbf{x}_i| \equiv 1$)

$$\begin{bmatrix} 8 & 5.2 \\ 5.2 & 4.4 \end{bmatrix} \bullet \mathbf{x} = \lambda \bullet \mathbf{x} = \lambda \bullet \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \bullet \mathbf{x} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \bullet \mathbf{x}$$

$$\Leftrightarrow \begin{bmatrix} 8 - \lambda & 5.2 \\ 5.2 & 4.4 - \lambda \end{bmatrix} \bullet \mathbf{x} = 0 \Leftrightarrow \begin{cases} 8\mathbf{x}_1 - \lambda \mathbf{x}_1 + 5.2\mathbf{x}_2 = 0 \\ 5.2\mathbf{x}_1 + 4.4\mathbf{x}_2 - \lambda \mathbf{x}_2 = 0 \\ \mathbf{x}_1^2 + \mathbf{x}_2^2 = 1 \end{cases}$$

Principal Component Analysis



$$\lambda_1 \approx 73.59 \qquad \mathbf{x}_1 \approx \begin{bmatrix} 0.8428 \\ 0.5383 \end{bmatrix}$$

$$\lambda_2 \approx 4.33 \qquad \mathbf{x}_2 \approx \begin{bmatrix} -0.5383 \\ 0.8428 \end{bmatrix}$$

$$\sigma^2_{p'} = \sigma^2_{x'} + \sigma^2_{y'}$$

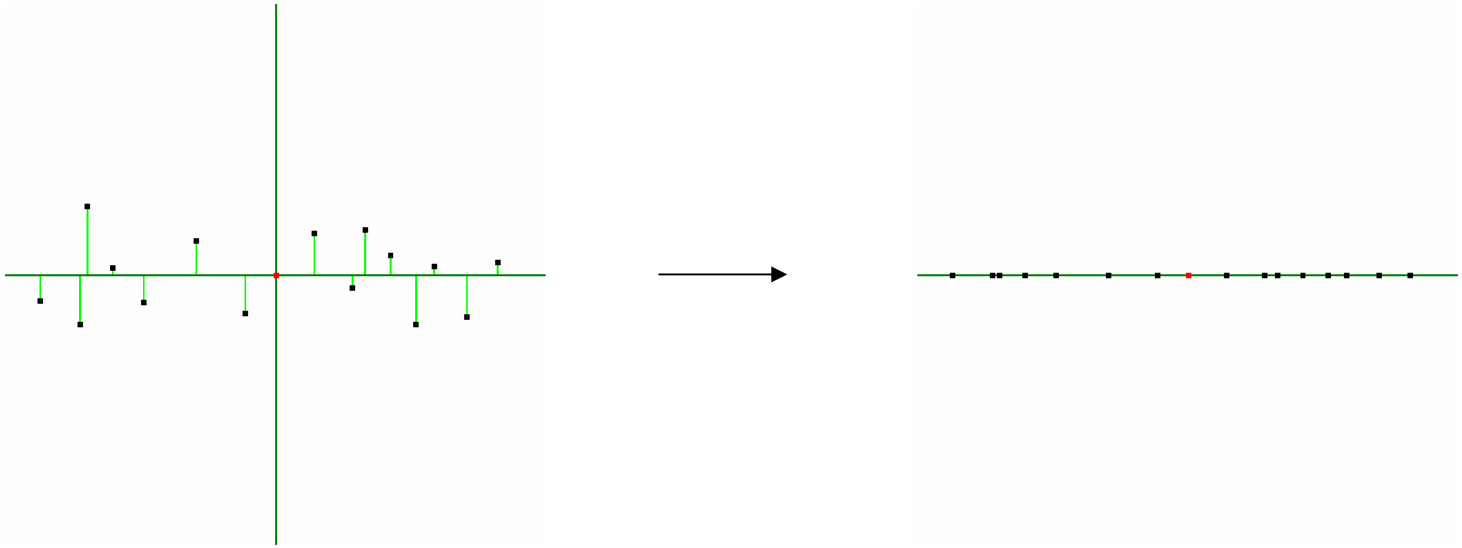
$\lambda_1 \approx 73.59 \cong 94.44\%$ of the total variance

$\lambda_2 \approx 4.33 \cong 5.56\%$ of the total variance

Principal Component Analysis

PCA step 3: reducing complexity

Principal Component Analysis



Reducing complexity = removing dimensions:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \rightarrow \mathbf{v'} = \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \cong \begin{bmatrix} v_1 \end{bmatrix}$$

Principal Component Analysis

PCA: an example:

8 arrays,
1120 genes

covariation:

0.082	0.085	0.067	0.085	0.065	0.049	0.053	0.035
0.085	0.137	0.106	0.132	0.089	0.070	0.066	0.056
0.067	0.106	0.132	0.166	0.104	0.085	0.074	0.058
0.085	0.132	0.166	0.261	0.161	0.133	0.111	0.087
0.065	0.089	0.104	0.161	0.138	0.098	0.098	0.072
0.049	0.070	0.085	0.133	0.098	0.094	0.076	0.056
0.053	0.066	0.074	0.111	0.098	0.076	0.091	0.060
0.035	0.056	0.058	0.087	0.072	0.056	0.060	0.061

Eigenvalue:	%-variance:	som%-variance:
0.7788:	78.2862%	78.2862%
0.0830:	8.3442%	86.6304%
0.0588:	5.9087%	92.5391%
0.0238:	2.3914%	94.9306%
0.0153:	1.5408%	96.4713%
0.0149:	1.4954%	97.9667%
0.0119:	1.1925%	99.1592%
0.0084:	0.8408%	100.0000%

mean:

0.041	0.010	-0.035	-0.006	0.031	-0.025	0.030	-0.028

Eigenvalues:

0.7788	0.0830	0.0588	0.0238	0.0153	0.0149	0.0119	0.0084

Eigenvectors (matrix X):

0.2372	-0.5089	0.3112	-0.5889	-0.0258	-0.0140	0.3536	-0.3395
0.3437	-0.6736	0.0588	0.3954	0.1540	-0.2464	-0.2782	0.3265
0.3780	-0.1457	-0.3817	0.1185	-0.2134	0.7361	-0.1585	-0.2521
0.5482	0.2153	-0.5569	-0.1363	-0.0477	-0.3681	0.3816	0.2025
0.3882	0.2877	0.2753	-0.1770	-0.4124	-0.2872	-0.6085	-0.1911
0.3121	0.2596	0.0977	-0.1033	0.8603	0.1105	-0.1866	-0.1668
0.2905	0.2158	0.5081	-0.0343	-0.1130	0.3924	0.2214	0.6277
0.2230	0.1517	0.3168	0.6489	-0.0699	-0.1143	0.4155	-0.4640

Jacobi max error: 0.00000016

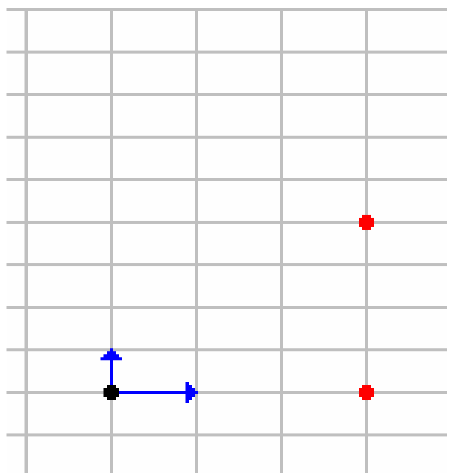
Jacobi iterations: 4

PCA and clustering

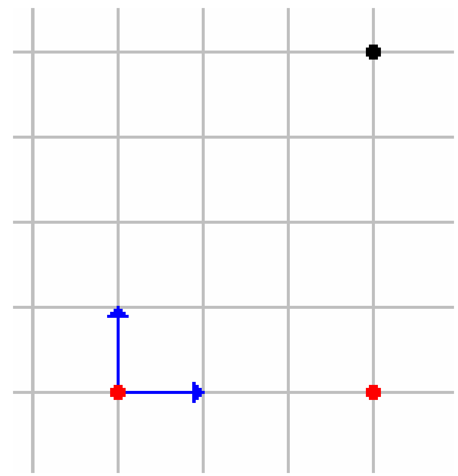
PCA and clustering:
- we are able to reduce the dimensionality of the data set, and
- identify variables with strong relationships with a component
- PCA helps to evaluate the quality of a given clustering
- all distances between data points remain the same

Demand: variables have to be "on an equal footing":
should be measured in the same (or comparable) units

Transformation of axis with different factors
can change a clustering:

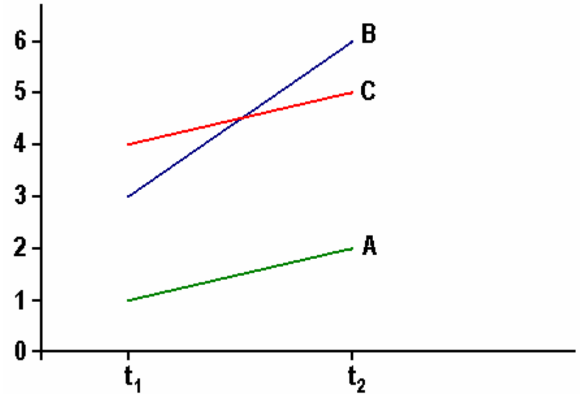


vs

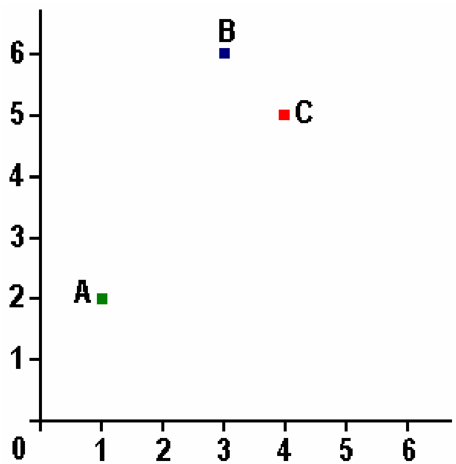


so think what you are doing!

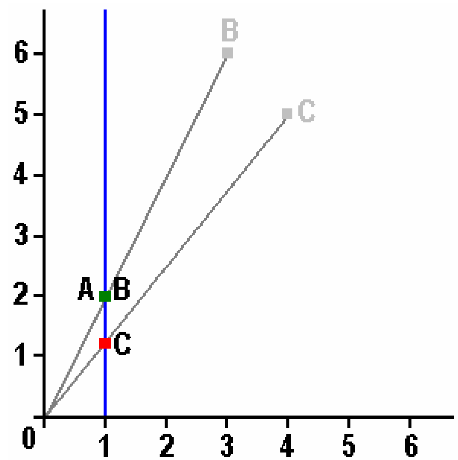
just like other types of transformations:



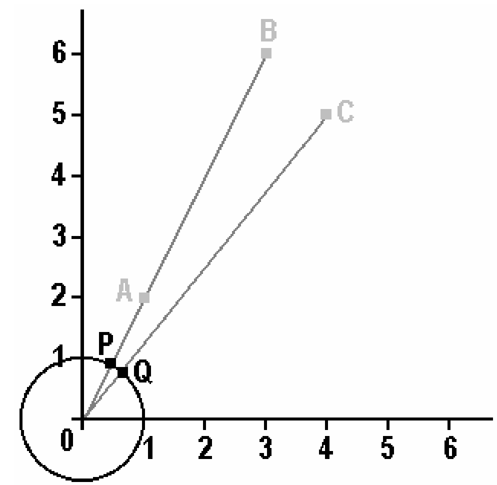
	t_1	t_2
gene A	1	2
gene B	3	6
gene C	4	5



raw signals:
C = {{A}, {B,C}}



ratio s_{t2}/s_{t1}:
C = {{A,B}, {C}}

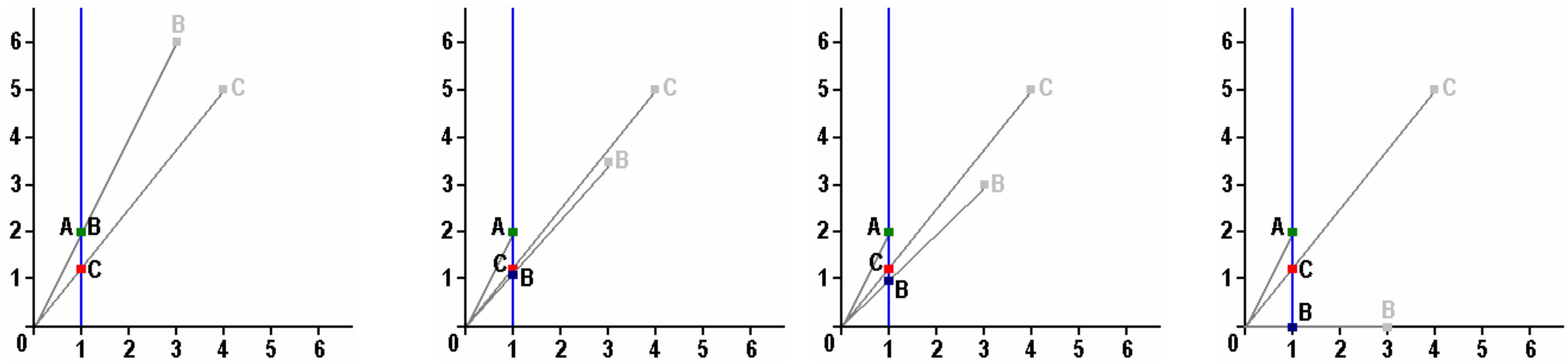


normalized s_{t2}/s_{t1}:
C = {{A,B}, {C}}

do not guess missing values:

	t_1	t_2
gene A	1	2
gene B	3	6
gene C	4	5

	t_1	t_2
gene A	1	2
gene B	3	?
gene C	4	5



ratio s_{t2}/s_{t1}:
C = {{A,B}, {C}}

$B_2=\mu_{t2}=3.5$ $B_2=\mu_B=3$
C = {{A}, {B,C}}

$B_2=0$:
C = {{A,C}, {B}}

Each clustering result needs further evaluation (visualisation, etc).

The only quality measure of a clustering is its usefulness in practice.