Lecture 7: Seeing patterns in high-dimensional data

R. Ranganathan
Green Center for Systems Biology, ND11.120E

The principles of principal components analysis (PCA) and independent components analysis (ICA)

Karl Pearson
1857 - 1936

Harold Hotelling
1895 - 1973

Pierre Comon
CNRS

Terry Sejnowski
Salk

The large variable limit….the idea of **apparent complexity** and **dimension reduction**

	n = 1	n = 2 or 3	n >> 1	continuum
Linear	exponential growth and decay single step conformational change fluorescence emission pseudo first order kinetics	second order reaction kinetics linear harmonic oscillators simple feedback control sequences of conformational change	electrical circuits molecular dynamics systems of coupled harmonic oscillators equilibrium thermodynamics diffraction, Fourier transforms	Diffusion Wave propagation quantum mechanics viscoelastic systems
Nonlinear	fixed points bifurcations, multi stability irreversible hysteresis overdamped oscillators	anharmonic oscillators relaxation oscillations predator-prey models van der Pol systems Chaotic systems	systems of non-linear oscillators non-equilibrium thermodynamics protein structure/ function neural networks the cell ecosystems	Nonlinear wave propagation Reaction-diffusion in dissipative systems Turbulent/chaotic flows

adapted from S. Strogatz

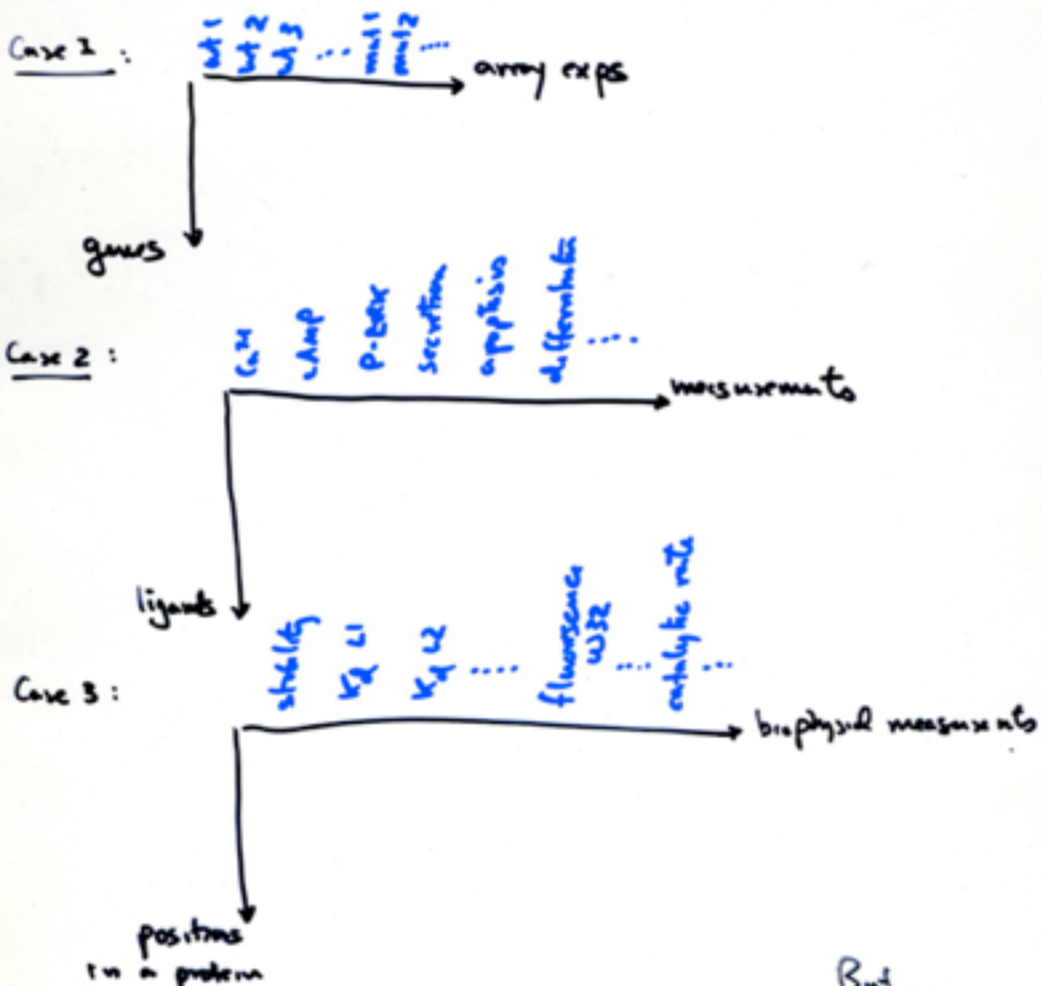
So... low dimensional representations of high-dimensional data

PCA…principal components analysis (also called Karhunen-Loeve transform (KLT), Hotelling transform, eigenvalue decomposition, factor analysis, spectral decomposition).

ICA...independent components analysis (also sometimes called blind-source separation)

What is the problem, fundamentally?

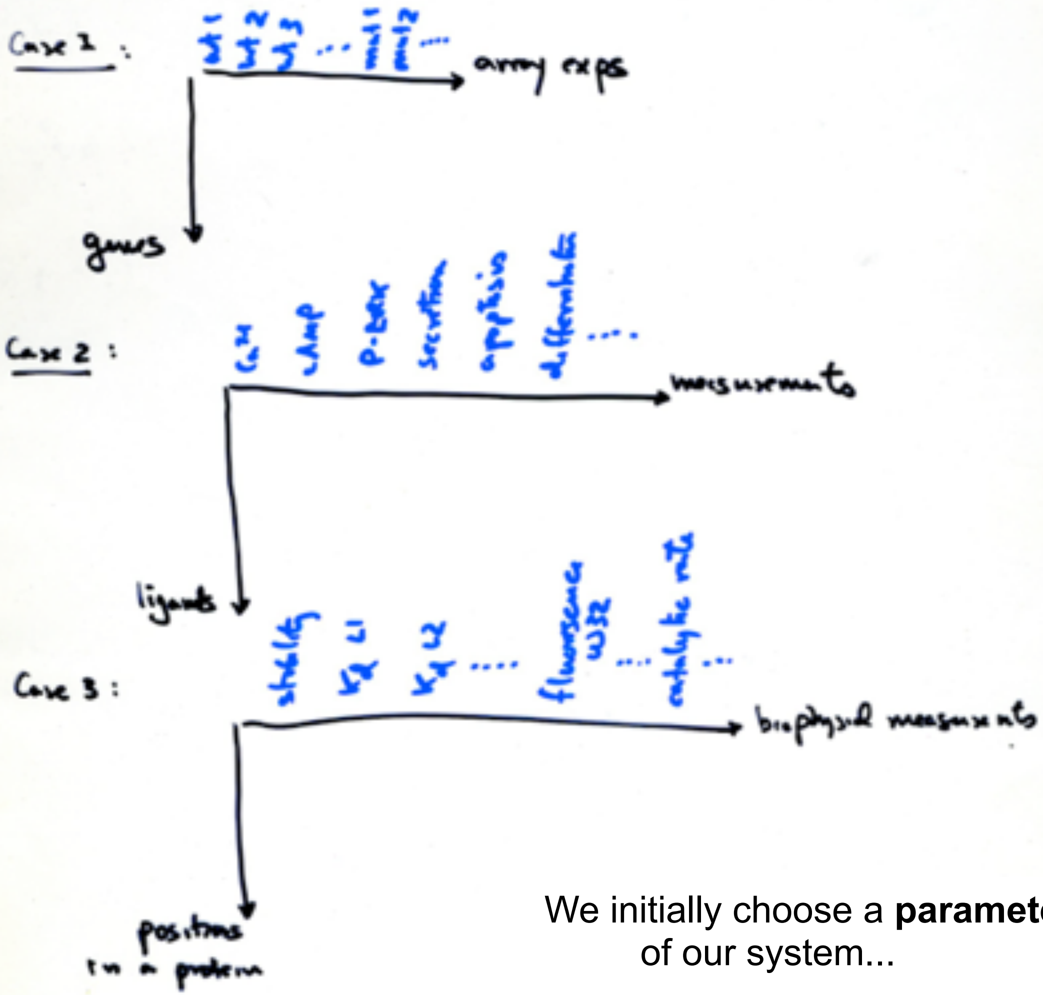
Often, we collect data about biological systems in many individual parameters that represent our experimental analysis ...



But....

What is the problem, fundamentally?

Often, we collect data about biological systems in many individual parameters that represent our experimental analysis...



We initially choose a **parameterization** of our system...

genes in a genome

ligands applied to a cell

positions in a protein

What is the problem, fundamentally?

We initially choose a **parameterization** of our system...

genes in a genome

ligands applied to a cell

positions in a protein

But are these the "units" of relevance? Might the data help tell us about the way these parameters are behaving that might lead to a new parameterization that is potentially simpler?

So... a "cluster of genes" → a shared biochemical cascade?

a "cluster of ligands" → similar transduction pathway

a "cluster of aa positions" → enzyme active site what is a good approach for achieving this **re-parameterization**?

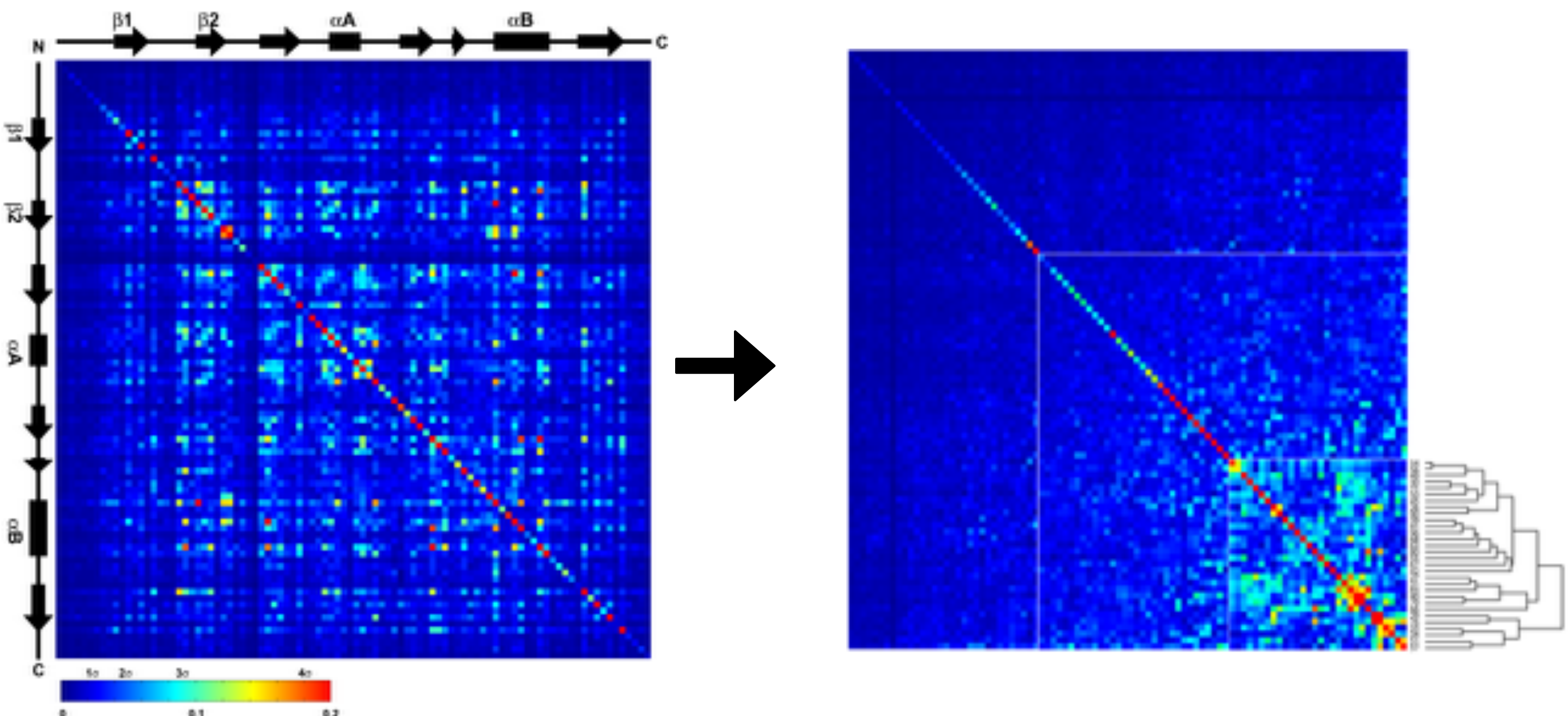
What is the problem, fundamentally?

what is a good approach for achieving this **re-parameterization**?

First, what is the **target goal**, quantitatively?
We will see...

Hierarchical Clustering

what is a good approach for achieving this **re-parameterization**?

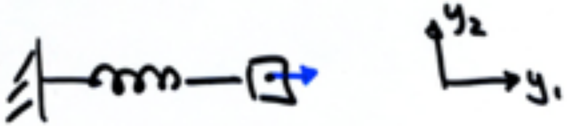


...but not a statistically rigorous method

Is there a better way?

principal components analysis, or **PCA**

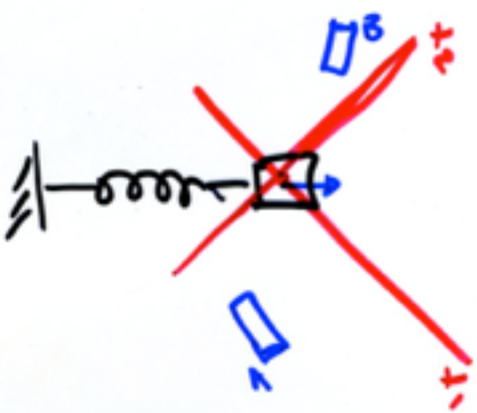
Start with an example:



So, for a frictionless, massless spring, we know that if we pull the mass along $y_{10} \to y_{10} + \delta y$ it will oscillate forever around y_{10}.

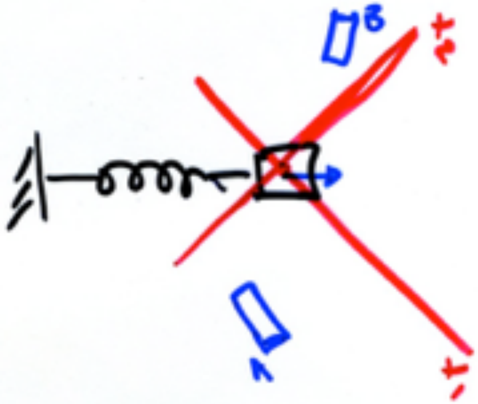
=> a 1-D motion ...that's it.

But say we are stupid experimentalists (as we are) and do something arbitrary since we don't know better:



we set up two cameras to record the action, A + B. And... we set up some laboratory frame of reference axes for recording the position of the mass x_1, and x_2.

principal components analysis, or **PCA**



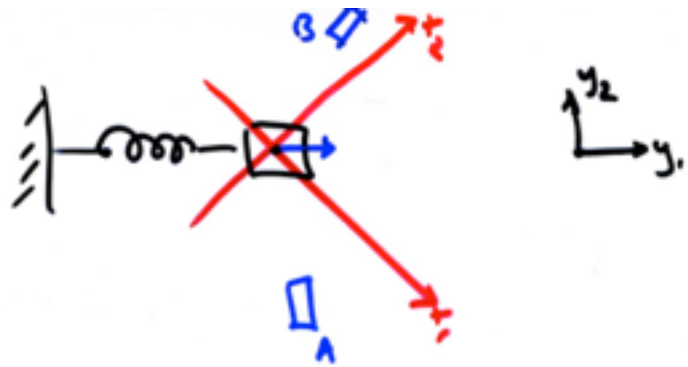
So ... some definitions. ① $B = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = I$ is called the initial "basis" set. These are the unit vectors along the x_1 and x_2 directions.

② our observations X are $4 \times N_{obs}$ and are with reference to basis B:

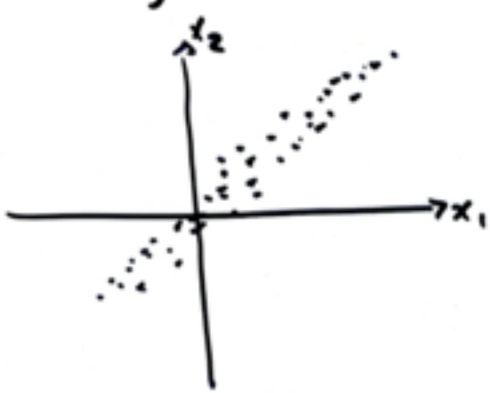
$$X = \begin{array}{c} 1 \longrightarrow N \\ \begin{bmatrix} x_{1A} & \cdots \\ x_{2A} & \cdots \\ x_{1B} & \\ x_{2B} & \end{bmatrix} \end{array}$$

2 cameras, each producing one (x_1, x_2) coordinate at each time t.

principal components analysis, or **PCA**



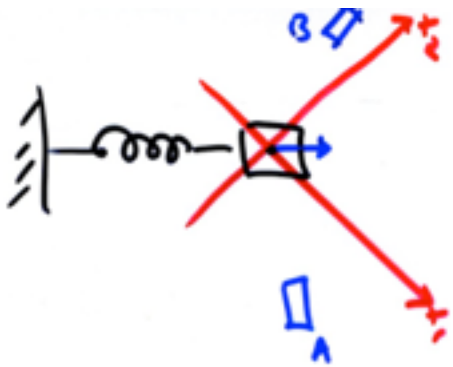
So... signal from camera A might be:

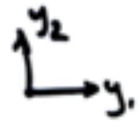


Now we ask ... is there another basis set comprised of a linear combination of original axes that is better? yes....

is it obvious?

principal components analysis, or **PCA**





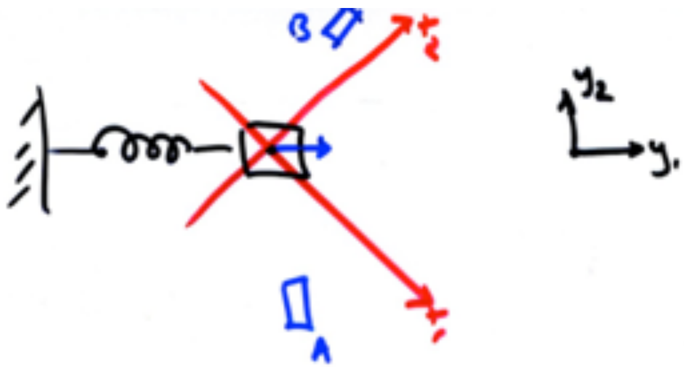
So... look for a transformation of X :

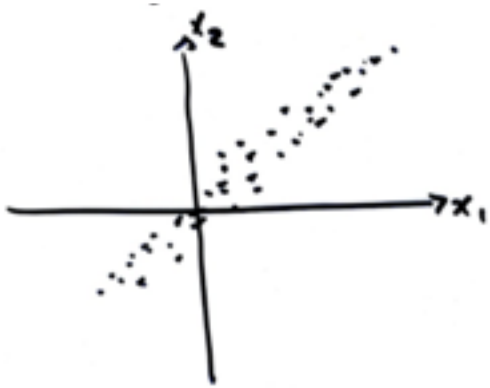
$$\boxed{Y = PX}$$

where..

1. P is the transformation matrix that takes $X \rightarrow Y$.
2. P is a rotation and stretching matrix
3. rows of P are the new basis set for representing the columns of $X \longrightarrow$ columns of Y.

principal components analysis, or **PCA**





So... look for a transformation of X:

$$\boxed{Y = PX}$$

where..

① P is the transformation matrix that takes $X \rightarrow Y$.

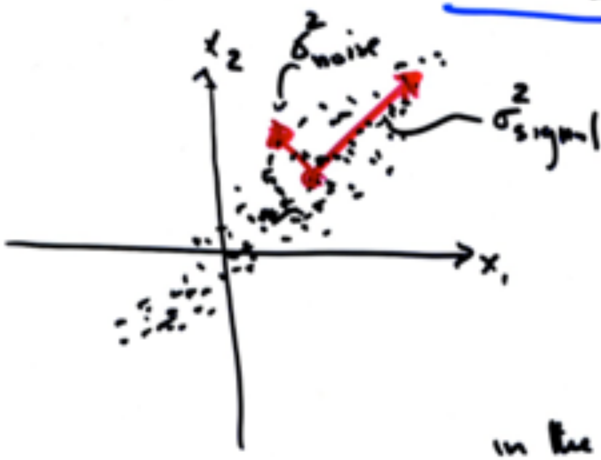
② P is a rotation and stretching matrix

③ rows of P are the new basis set for representing the columns of $X \longrightarrow$ columns of Y.

So.. what P do we want? what is a "better" basis set? well... one that captures the fact that this is a 1-D dynamic rather than a 2-D dynamic.

The two guiding principles of **PCA**...

① minimize noise ... i.e. maximize SNR



$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

PCA assumes directions with larger variance in the dataset are "interesting" directions.

Thus choose P such that the first "principal component" or first basis vector points in the direction of max variance. Then fix this, and search for the second orthogonal vector that next maximally accounts for variance and so on ...

The two guiding principles of **PCA**...

② eliminate redundancy

low redundancy ⟶ high redundancy.

Think about our two cameras...

So... choose P to minimize redundancy in dimensions of Y.

Variance and covariance….

SNR is captured in the variances of the variables

redundancy is captured in the co-variation of variables.

what is co-variation?

consider two variables with zero mean:

$$a: \{a_1, a_2, \ldots, a_n\}$$

$$b: \{b_1, b_2, \ldots, b_n\}$$

Variance and covariance….

consider two variables with zero mean :

$a : \{a_1, a_2, \dots, a_n\}$

$b : \{b_1, b_2, \dots, b_n\}$

the variance of $\vec{a}$ and $\vec{b}$ are defined as :

$\sigma_a^2 = \langle a_i a_i \rangle_i$ … the $\langle \, \rangle$ means expectation or average over n values.

$\sigma_b^2 = \langle b_i b_i \rangle_i$

the covariance of $\vec{a}$ & $\vec{b}$ is :

$\sigma_{ab}^2 = \langle a_i b_i \rangle_i$

Variance and covariance….

consider two variables with zero mean:

$a: \{a_1, a_2, \ldots, a_n\}$

$b: \{b_1, b_2, \ldots, b_n\}$

$$\sigma^2_{ab} = \langle a_i b_i \rangle_i$$

now.. obviously,

$\sigma^2_{ab} = 0$ iff a and b are not correlated

$\sigma^2_{ab} = \sigma^2_a = \sigma^2_b$ if $a = b$.

In vector notation,

$\sigma^2_{ab} = \frac{1}{n-1} a b^T$ ... remember

$$[a_1 a_2 \ldots a_n] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = [a_1 b_1, a_2 b_2, \ldots a_n b_n]$$

Now, this is for **two vectors**...

Variance and covariance….

Now, remember we have a matrix of observations: 12

$$X = \begin{bmatrix} x_1 & \cdots \\ \vdots & \\ x_m & \end{bmatrix} \quad \cdots \text{ an } m \times n \text{ matrix.}$$

$1 \longrightarrow N$ obs.

Each row is a measurement variable, and columns are observations of that variable.

So, the covariance matrix is an $m \times m$ matrix:

$$\boxed{S_X = \frac{1}{n-1} X X^T}$$

This is the **covariance matrix** for a set of variables
(here, for rows of our data matrix)

Variance and covariance….

Now, remember we have a matrix of observations:

$$X = \begin{bmatrix} x_1 & \cdots & \cdots \\ \vdots & & \\ x_m & & \end{bmatrix}$$

$1 \longrightarrow N$ obs.

… an $m \times n$ matrix.

Each row is a measurement variable, and columns are observations of that variable.

So, the covariance matrix is an $m \times m$ matrix:

$$\boxed{S_X = \frac{1}{n-1} X X^T}$$

properties:

1. The ij^{th} value of S_X is the co-variance of x_i and x_j
2. S_X is a square, symmetric $m \times m$ matrix
3. The diagonal terms are the variances of each x_i
4. The off-diagonal terms are the co-variances of x_i, x_j.

PCA

So... back to the problem. What do we want for P such that

$$y = Px \; ?$$

well, what do we want for S_y?

1. If we want to minimize redundancy, we want each transformed variable in y to co-vary as little as possible. Thus we want a transformation where all the off-diagonal terms in S_y are zero. $\Rightarrow$ want to diagonalize S_y. $\begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ & & & \lambda_m \end{bmatrix}$

2. If we want to maximize SNR, then choose the new basis vectors in an order where maximal variance is in the first basis vector .... and so on.

But, how do we **solve for P**?

PCA

So ... find some matrix P where:

$$Y = PX,$$

such that

$$\boxed{S_Y \equiv \frac{1}{n-1} YY^T}$$

$$\longrightarrow \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ \vdots & & & \lambda_m \end{bmatrix}$$

is diagonalized. Then the rows of P are the weights for each principal component of Y in X.

The path to finding P is **eigenvalue decomposition** of the covariation matrix of X, the initial variables...

PCA

To do this... begin by re-writing S_Y:

$$S_Y = \frac{1}{n-1} YY^T$$

$$= \frac{1}{n-1} (PX)(PX)^T$$

$$= \frac{1}{n-1} PXX^TP^T$$

$$= \frac{1}{n-1} P(XX^T)P^T$$

$$\boxed{S_Y = \frac{1}{n-1} PAP^T}$$

, where we define $\boxed{A \equiv XX^T.}$

Recognize the **matrix A**? What is it?

PCA

To do this... begin by re-writing S_Y:

$$S_Y = \frac{1}{n-1} YY^T$$

$$= \frac{1}{n-1} (PX)(PX)^T$$

$$= \frac{1}{n-1} PXX^TP^T$$

$$= \frac{1}{n-1} P(XX^T)P^T$$

$$\boxed{S_Y = \frac{1}{n-1} PAP^T}$$, where we define $\boxed{A \equiv XX^T.}$

A is a symmetric m×m matrix. Now, it turns out that a symmetric matrix is diagonalized by a matrix of its so-called eigenvectors:

$$\boxed{A = EDE^T}$$, where

D is a diagonal matrix, and E is the matrix of eigenvectors of A

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & \\ 0 & \lambda_2 & & & \\ \vdots & & \ddots & & \\ \vdots & & & & \\ \cdot & & & & \lambda_m \end{bmatrix}$$

The **eigenvalue decomposition**…

$$A = V \cdot \lambda \cdot V^T$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

eigenvector 1

eigenvector 2

In this process, a matrix is decomposed into its **eigenvalues** and associated **eigenvectors**….let's understand this more closely…

Basic concepts from linear algebra…

A **matrix** w/ m rows an n cols is:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{m1} & & & a_{mn} \end{bmatrix}$$

, and a_{ij} are elements of A.

A **square matrix** is a matrix with equal number of rows and cols.

The **order** of a square matrix is the no. of rows (or columns).

Basic concepts from linear algebra…

A matrix w/ m rows an n cols is:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{m1} & & & a_{mn} \end{bmatrix}$$

, and a_{ij} are elements of A.

A square matrix is a matrix with equal number of rows and cols.

~~The~~ The order of a square matrix is the no. of rows (or columns).

Every square matrix has a scalar value associated with it … the determinant. For a second order matrix …

$$\boxed{D = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc}$$

Higher order determinants are a pain in the ass …. use a computer.

Basic concepts from linear algebra…

There are some interesting properties of matrix determinants. Of most use to us now is that if the $\det(A) \neq 0$, then A is said to be non-singular. If so, then there exists an inverse matrix A^{-1} such that:

$$\boxed{A^{-1}A = I}$$

I is the identity matrix:

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & & \\ \vdots & & \ddots & \\ \vdots & & & 1 \end{bmatrix}$$

So… determinants are valuable since they tell us about the invertibility of square matrices.

So... now for eigenvalues and eigenvectors

Let A be a square matrix $(n \times n)$. A number is an eigenvalue of A if there exists a non-zero vector $\vec{v}$ such that:

$$A\vec{v} = \lambda\vec{v}$$

If so, then $\vec{v}$ is called an eigenvector of A corresponding to λ.

Computing eigenvalues and eigenvectors:

So.. $A\vec{v} = \lambda\vec{v}$

rewrite as

$$\boxed{(A - \lambda I)\vec{v} = 0}$$

Now... we want a non-zero $\vec{v}$ that makes this equation true. In this case the matrix $(A - \lambda I)$ must not be invertible.

So.. $A\vec{v} = \lambda\vec{v}$

rewrite as

$$\boxed{(A-\lambda I)\vec{v} = 0}$$

Now... we want a non-zero $\vec{v}$ that makes this equation true. In this case the matrix $(A-\lambda I)$ must not be invertible.

Why? If it was, then...

$$(A-\lambda I)^{-1}(A-\lambda I)\vec{v} = (A-\lambda I)^{-1}0 \quad , \text{ or}\ldots$$

$$\vec{v} = 0$$

So for $(A-\lambda I)$ to be non-invertible, it must be singular. That is,

$$\boxed{\det(A-\lambda I) = 0}$$

So.. $A\vec{v} = \lambda\vec{v}$

rewrite as

$$\boxed{(A-\lambda I)\vec{v} = 0}$$

Now... we want a non-zero $\vec{v}$ that makes this equation true. In this case the matrix $(A-\lambda I)$ must not be invertible.

Why? If it was, then...

$$(A-\lambda I)^{-1}(A-\lambda I)\vec{v} = (A-\lambda I)^{-1}0 \quad , \text{ or...}$$

$$\vec{v} = 0$$

So for $(A-\lambda I)$ to be non-invertible, it must be sigular. That is,

$$\boxed{\det(A-\lambda I) = 0}$$

The equation $p(\lambda) = \det(A-\lambda I)$ is called the "characteristic polynomial" of A, and the eigenvalues of A are just the roots of this polynomial ... values of λ where the polynomial goes to zero

Let’s do a **calculation**..

$$A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

First, we compute the **eigenvalues**...

Let's do a **calculation**..

$$A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

$$\det(A - \lambda I) = \det \begin{bmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{bmatrix}$$

$$= (2-\lambda)(-1-\lambda) - (-4)(-1)$$

$$= \lambda^2 - \lambda - 6$$

$$= (\lambda - 3)(\lambda + 2)$$

The roots are $\underline{\lambda = 3 \text{ and } \lambda = -2}$ ... these are the eigenvalues of A.

And now, for the **eigenvectors**...

Let's do a **calculation**..

To get associated eigenvectors .... solve the system of equations for each eigenvalue.

$$\boxed{A\vec{v} = \lambda\vec{v}} \quad \text{or} \ldots \quad (A-\lambda I)\vec{v} = 0$$

For $\lambda = 3$ ...

$$\begin{bmatrix} 2-3 & -4 \\ -1 & -1-3 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boxed{\begin{aligned} -v_1 - 4v_2 &= 0 \\ -v_1 - 4v_2 &= 0 \end{aligned}}$$

2 eqns, 2 unknowns. So.. $\vec{v} = \begin{bmatrix} 0.97 \\ -0.24 \end{bmatrix}$

For $\lambda = -2$ ...

$$\begin{bmatrix} 2-(-2) & -4 \\ -1 & -1-(-2) \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So... $\vec{v} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$

Now for the key point....

$A\vec{v} = \lambda\vec{v}$, so..

$$\begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 0.97 & 0.707 \\ -0.24 & 0.707 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} 0.97 & 0.707 \\ -0.24 & 0.707 \end{bmatrix}$$

Do you see that by multiplying by the eigenvectors we have diagonalized A?

Back to PCA...

we want to find P such that $Y = PX$ where:

$$S_Y = \frac{1}{n-1} YY^T$$

is diagonalized. we concluded that...

$$\boxed{\begin{aligned} S_Y &= \frac{1}{n-1} P(XX^T)P^T \\ &= \frac{1}{n-1} PAP^T \quad \text{where} \quad A = XX^T \end{aligned}}$$

Except for the scaling term, A is the co-variation matrix of the original data variables.

Back to PCA...

we want to find P such that $Y = PX$ where:

$$S_Y = \frac{1}{n-1} YY^T$$

is diagonalized. we concluded that...

$$\boxed{\begin{aligned} S_Y &= \frac{1}{n-1} P(XX^T)P^T \\ &= \frac{1}{n-1} PAP^T \quad \text{where} \quad A = XX^T \end{aligned}}$$

Except for the scaling term, A is the co-variation matrix of the original data variables.

Since A is a symmetric $m \times m$ matrix, we can diagonalize it by finding its eigenvectors...

$\boxed{A = EDE^T}$. where the matrix D is a diagonal matrix of eigenvalues of A, and E contains the corresponding eigenvectors.

Back to PCA...

we want to find P such that $Y = PX$ where:

$$S_Y = \frac{1}{n-1} Y Y^T$$

is diagonalized. we concluded that...

$$\boxed{\begin{aligned} S_Y &= \frac{1}{n-1} P(XX^T)P^T \\ &= \frac{1}{n-1} P A P^T \quad \text{where} \quad A = XX^T \end{aligned}}$$

$$\boxed{A = EDE^T}.$$

So, **what should we choose for P**, so that Sy is diagonalized?

Back to PCA...

$$S_y = \frac{1}{n-1} P A P^T$$

Now the trick is to select $P = E^T$; that is let's propose to choose P to contain (as rows) the eigenvectors of the covariance matrix of original variables.

$A = XX^T$ (the **covariance matrix** of initial variables)

$A = E D E^T$ (the **eigenvalue decomposition** of A)

$$\downarrow$$

P^T and so,

$$A = P^T D P$$

Back to PCA...

$$S_y = \frac{1}{n-1} P A P^T$$

Now the trick is to select $P = E^T$; that is let's propose to choose P to contain (as rows) the eigenvectors of the covariance matrix of original variables.

Then...

$$S_y = \frac{1}{n-1} P(P^T D P) P^T \quad , \quad \text{since } A = P^T D P$$

$$= \frac{1}{n-1} (PP^T) D (PP^T)$$

$$= \frac{1}{n-1} (PP^{-1}) D (PP^{-1})$$

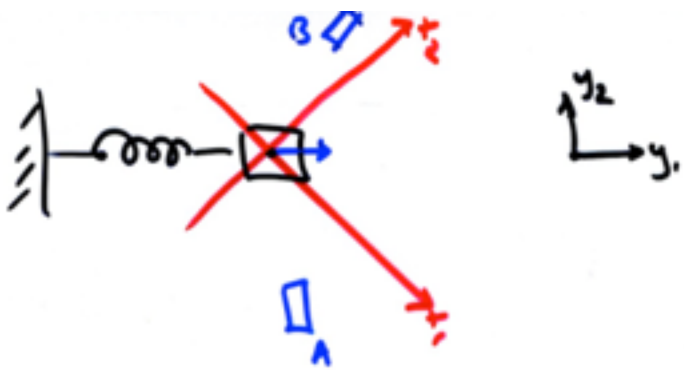
$$\boxed{S_y = \frac{1}{n-1} D}$$

So that's it ... our choice of $P = E^T$ diagonalizes S_y just as we wanted to.

So, to **summarize PCA**...

Summary of PCA...

Given an initial (non-optimal) parameterization of our system....



So... look for a transformation of X :

$$\boxed{Y = PX}$$

P should be chosen to :

① maximize $SNR = \dfrac{\sigma^2_{sig}}{\sigma^2_{noise}}$

② eliminate redundancy ... that is minimize all off-diagonal elements of the $cov(X) \equiv S_X$

Summary of PCA...

Solution to PCA problem amounts to eigenvalue decomposition of S_x.

$$S_x = E^T D E \quad , \text{ where } \quad D = S_y$$

Remember...

$$\underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \\ \vdots & & \ddots \end{bmatrix}}_{S_x} = \underbrace{\begin{bmatrix} v_{11} & v_{21} & \cdots \\ v_{12} & v_{22} & \\ \vdots & & \ddots \end{bmatrix}}_{E^T} \underbrace{\begin{bmatrix} D_1 & 0 & 0 & \cdots \\ 0 & D_2 & & \\ 0 & & D_3 & \\ \vdots & & & \ddots \end{bmatrix}}_{S_y} \begin{bmatrix} v_{11} & v_{12} & \cdots \\ v_{21} & v_{22} & \\ \vdots & & \ddots \end{bmatrix}$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \downarrow \ \ \downarrow$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vec{v}_1 \ \ \vec{v}_2$$

The eigenvalues ($D_1, D_2 \ldots$) represent the quantity of information (variance) in S_x captured, and each associated eigenvector ($\vec{v}_1, \vec{v}_2 \ldots$) gives the weights for combining the original variables ($x_1, x_2 \ldots$) to form the new variables ($y_1, y_2 \ldots$).

Summary of PCA...

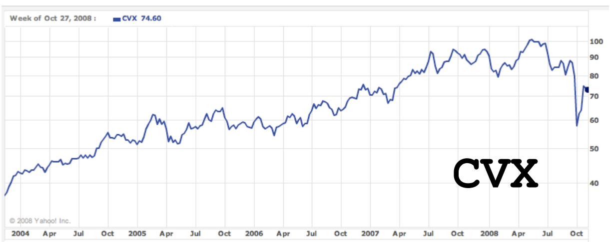
Solution to PCA problem amounts to eigenvalue decomposition of S_X.

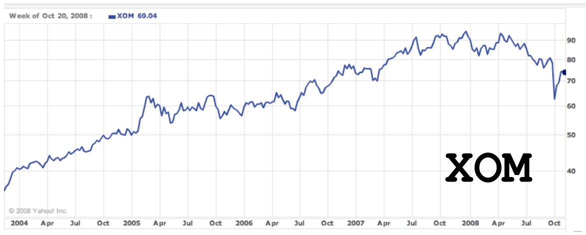
So... if we calculate the eigenvalues and eigenvectors of the S_X matrix and we calculate $Y = PX$ such that $P = E^T$, then S_Y is diagonalized. => PCA

Then to maximize SNR, we rank the eigenvalues (the diagonal elements of S_Y) by magnitude. The eigenvectors corresponding to the dominant eigenvalues are the principal components... directions along which the original data set is maximally decorrelated and which contain most of the signal.

Example from "econophysics"...a rational investment strategy for optimizing return

The idea is to understand the natural breakdown of the economy by looking at how stocks are correlated in their market performance....





Pilereu et al. (2002), Physical Review E 65, 066126
Laloux et al. (1999), PRL 83, p. 1467
Pilereu et al. (1999), PRL 83, p. 1471

An analogy from economics

To understand the natural breakdown of the economy by the statistics of stock market performance....

(1) Make a covariance matrix for the performance of a bunch of stocks over a time window (here, 7 years from the S&P 500).

Now, this matrix is contaminated with two kinds of noise.....(1) sampling noise (limited time series), and (2) global correlations of stocks due to overall market performance

Pilereu et al. (2002), Physical Review E 65, 066126
Laloux et al. (1999), PRL 83, p. 1467
Pilereu et al. (1999), PRL 83, p. 1471

An analogy from economics

To understand the natural breakdown of the economy by the statistics of stock market performance....

(1) Make a covariance matrix for a bunch of stocks (here, from the S&P 500).

(2) Compute the so-called eigenvalues of the covariance matrix. Each eigenvalue represents a collection of stocks that move together in the market.

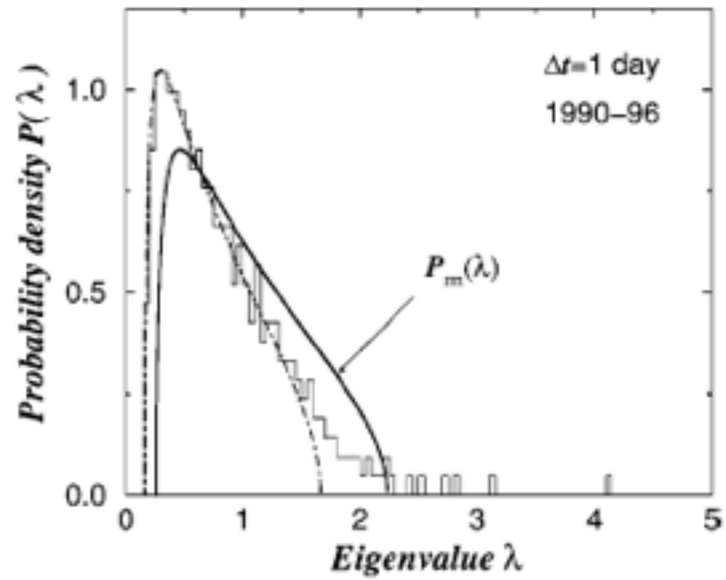


FIG. 4. $P(\lambda)$ for C constructed from daily returns of 422 stocks for the 7-yr period 1990–1996. The solid curve shows the RMT result $P_{rm}(\lambda)$ of Eq. (6]) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{rm}(\lambda)$ with λ_+ and λ_- as free parameters. We find similar results as found in Fig. 3(a) for 30-min returns. The largest eigenvalue (not shown) has the value $\lambda_{422}=46.3$.

Pilereu et al. (2002), Physical Review E 65, 066126
Laloux et al. (1999), PRL 83, p. 1467
Pilereu et al. (1999), PRL 83, p. 1471

An analogy from economics

To understand the natural breakdown of the economy by the statistics of stock market performance....

(1) Make a covariance matrix for a bunch of stocks (here, from the S&P 500).

(2) Compute the so-called eigenvalues of the covariance matrix. Each eigenvalue represents a collection of stocks that move together in the market.

(3) Find the “significant” eigenvalues by making a random correlation matrix.

(4) Analyze the remaining eigenvalues....

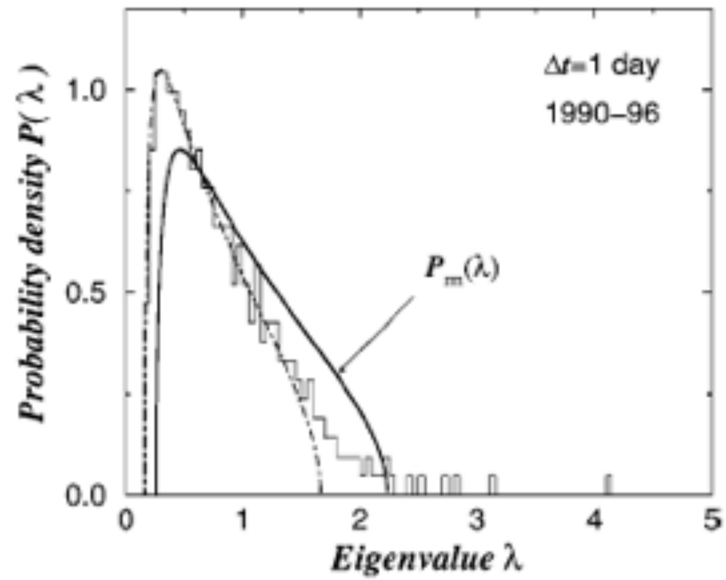


FIG. 4. $P(\lambda)$ for C constructed from daily returns of 422 stocks for the 7-yr period 1990–1996. The solid curve shows the RMT result $P_{rm}(\lambda)$ of Eq. (6]) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{rm}(\lambda)$ with λ_+ and λ_- as free parameters. We find similar results as found in Fig. 3(a) for 30-min returns. The largest eigenvalue (not shown) has the value $\lambda_{422}=46.3$.

Pilereu et al. (2002), Physical Review E 65, 066126
Laloux et al. (1999), PRL 83, p. 1467
Pilereu et al. (1999), PRL 83, p. 1471

Breakdown of the economy into "sectors"

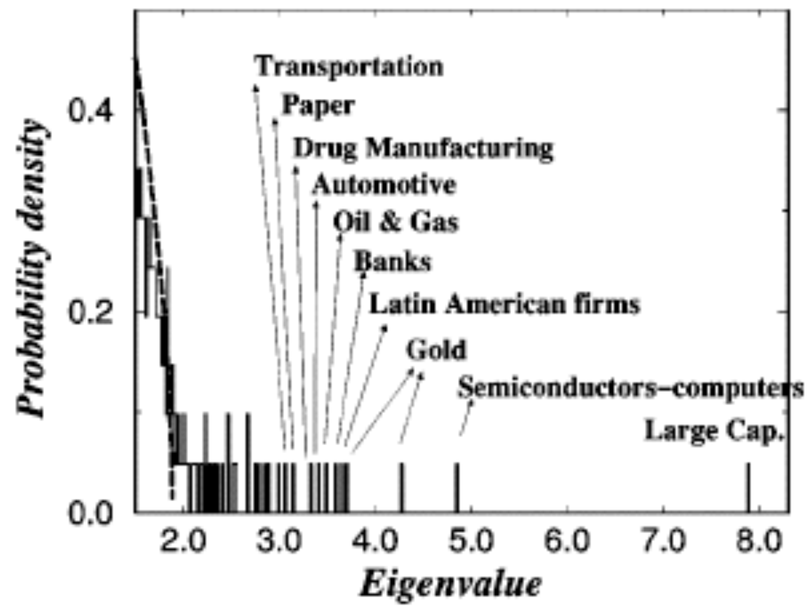


FIG. 13. Schematic illustration of the interpretation of the eigenvectors corresponding to the eigenvalues that deviate from the RMT upper bound. The dashed curve shows the RMT result of Eq. (6).

Pilereu et al. (2002), Physical Review E 65, 066126

In biology....

A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis

Kevin A. Janes,[1,2*] John G. Albeck,[2,3*] Suzanne Gaudet,[2,3]
Peter K. Sorger,[1,2,3] Douglas A. Lauffenburger,[1,2,3]
Michael B. Yaffe[1,2,3]†

Signal transduction pathways control cellular responses to stimuli, but it is unclear how molecular information is processed as a network. We constructed a systems model of 7980 intracellular signaling events that directly links measurements to 1440 response outputs associated with apoptosis. The model accurately predicted multiple time-dependent apoptotic responses induced by a combination of the death-inducing cytokine tumor necrosis factor with the prosurvival factors epidermal growth factor and insulin. By capturing the role of unsuspected autocrine circuits activated by transforming growth factor–α and interleukin-1α, the model revealed new molecular mechanisms connecting signaling to apoptosis. The model derived two groupings of intracellular signals that constitute fundamental dimensions (molecular "basis axes") within the apoptotic signaling network. Projection along these axes captures the entire measured apoptotic network, suggesting that cell survival is determined by signaling through this canonical basis set.

Many measurements in a cellular apoptotic signaling network...but **a small number** of reactions suffice to predict the probability of apoptosis...

Janes et al (2005), Science 310, p. 1646

Limitations and extensions of PCA...

Now... what are the assumptions and limitations of PCA? This will lead us to ICA.

① Linearity. Assumes that original variables contribute additively to the information. That is, assumes no higher order correlations in the data that influence the pattern of pairwise correlations.

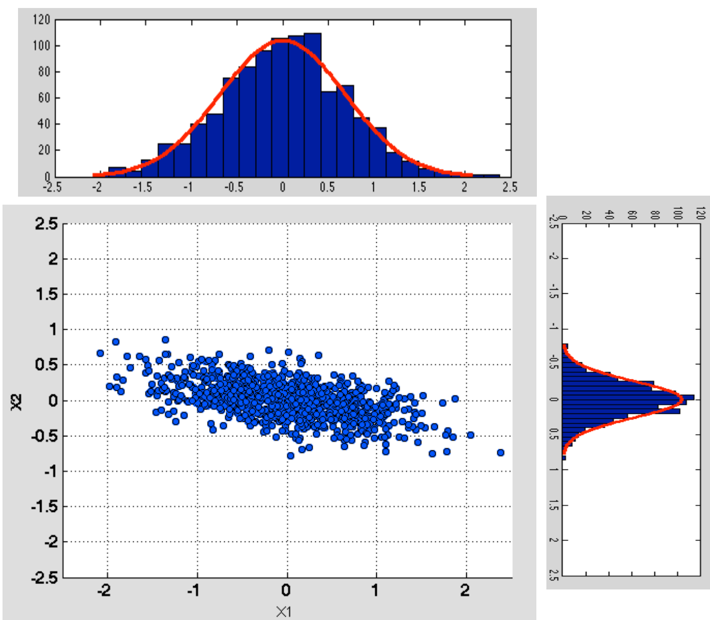
② Assumes that the mean and variance of data variables sufficiently captures their distribution.

When are these things true? When are they not?

Limitations and extensions of PCA...

well... the only distribution that is fully characterized by its variance and mean is the normal or Gaussian distribution... the bell shaped curve.

Thus... for non-Gaussian distributed variables, just decorrelation (i.e. PCA) may not identify the statistically independent components.
Example to follow.



Limitations and extensions of PCA...

what if variables are <u>not Gaussian distributed</u>? Then decorrelation does not mean statistical independence.

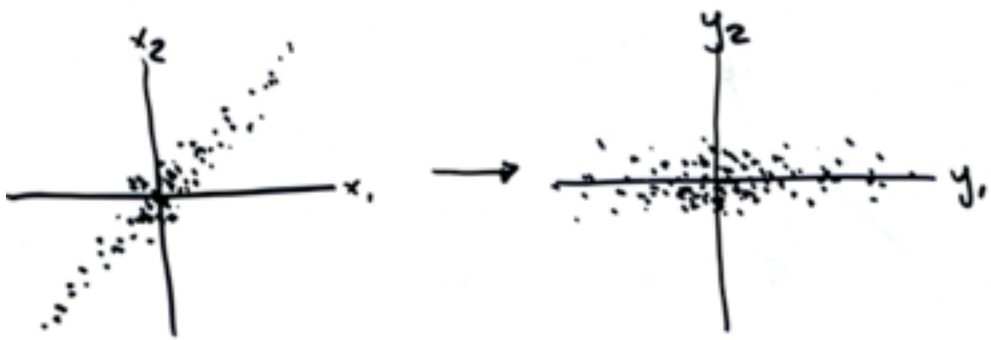
Limitations and extensions of PCA...

What if variables are not Gaussian distributed? Then decorrelation does not mean statistical independence.

So... to summarize. If variables are Gaussian distributed, then all the statistical information can be represented by the mean and variance.

Then... for many such variable dimensions, the covariation matrix contains all the information about interactions between variables.

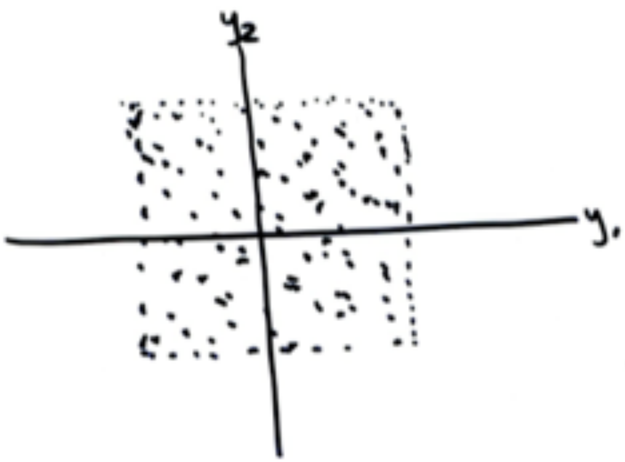
Then... PCA finds the independent components. ~~[illegible]~~ That is... the linear combination of variables that represent the statistically independent "new" variables.



Limitations and extensions of PCA...

Now consider this example...

Let's take 2 variables on orthogonal dimensions, each drawn from the identical uniform random distribution...



$$p(s) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{if } |s_i| < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$

These are truly uncorrelated and statistically independent.

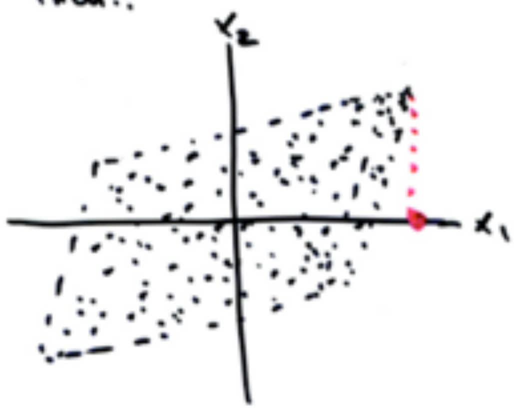
Limitations and extensions of PCA...

Now lets mix y_1 and y_2 by applying the following transform:

$$x = Ay \,,$$

where $A = \begin{pmatrix} 5 & 10 \\ 10 & 2 \end{pmatrix}$

Then..



Now, x_1 and x_2 are still uncorrelated right? Still random values drawn from a uniform distribution.

But... they are clearly not independent! Consider the value of x_1 marked in red. It fully determines the x_2 value.

There is information about x2 in x1....despite decorrelation!! This is due to the non-Gaussian nature of the distributions of variables…more than just mean and variance required to represent the statistics...

Limitations and extensions of PCA...

Statistical independence is a more rigorous concept than decorrelation. Independence implies decorrelation, but decorrelation does not demonstrate independence except for the case of Gaussian distributed variables.

Put simply... statistical independence means that

$$\boxed{p(a,b) = p(a) \cdot p(b)}$$

=> the joint distribution of two variables is unconditionally factorizable into the marginal distributions.

This only happens if not only the covariance but all higher order dependencies are zero.

Limitations and extensions of PCA...

Statistical independence is a more rigorous concept than decorrelation. Independence implies decorrelation, but decorrelation does not demonstrate independence except for the case of Gaussian distributed variables.

If the data contain higher order correlations, then variables will likely not be Gaussian distributed, and PCA is expected to fail.

Limitations and extensions of PCA...

Statistical independence is a more rigorous concept than decorrelation. Independence implies decorrelation, but decorrelation does not demonstrate independence except for the case of Gaussian distributed variables.

If the data contain higher order correlations, then variables will likely not be Gaussian distributed, and PCA is expected to fail.

So, **Independent Components Analysis (ICA)** is an extension of PCA to find new variables that are not just decorrelated, but truly statistically independent.

The idea of ICA...

As we said, if two variables are statistically independent, then they are uncorrelated. If so, then any function operating on either or both variables will result in new variables that are still uncorrelated.

For example, if y_1 and y_2 are independent, then

$$\boxed{p(y_1, y_2) = p(y_1) \cdot p(y_2)}$$

But also,

$$\boxed{p(h(y_1), g(y_2)) = p(h(y_1)) \cdot p(g(y_2))}$$

This is true for any h or g as long as they are well-behaved (i.e. continuously differentiable) functions.

The idea of ICA...

For example, if y_1 and y_2 are independent, then

$$\boxed{p(y_1, y_2) = p(y_1) \cdot p(y_2)}$$

But also,

$$\boxed{p(h(y_1), g(y_2)) = p(h(y_1)) \cdot p(g(y_2))}$$

- PCA finds new variables y given data x for the special case in which h & g are linear functions.
- ICA looks for new variables y given data x for the general case where h & g are some non-linear functions. That is, it looks for a transformation where variables y remain uncorrelated despite non-linear transformation.

Example… “**the cocktail party problem**”

Given audio recordings from some number of microphones placed randomly in the room, how can we extract the individual conversations without knowledge of the number of conversations or the information content of the conversations?

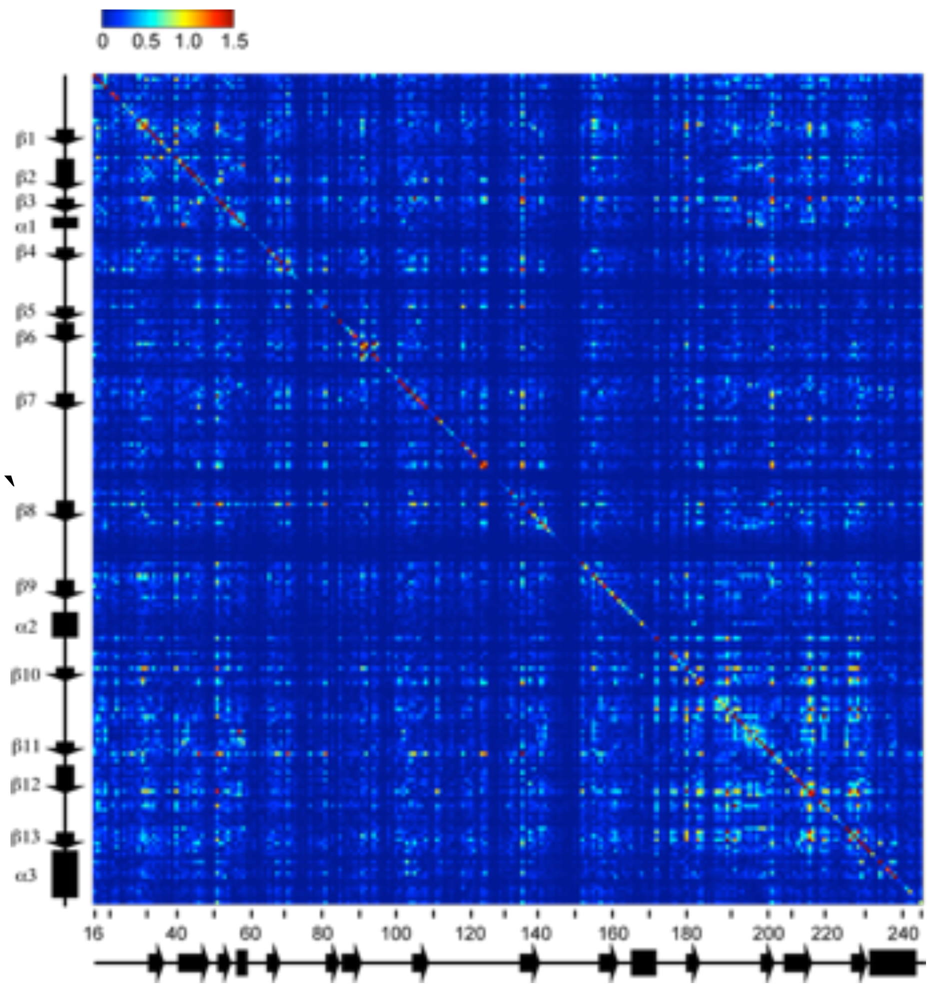
How can we quantitatively extract the information content in this matrix? The signal processing method of Independent Component Analysis (ICA)

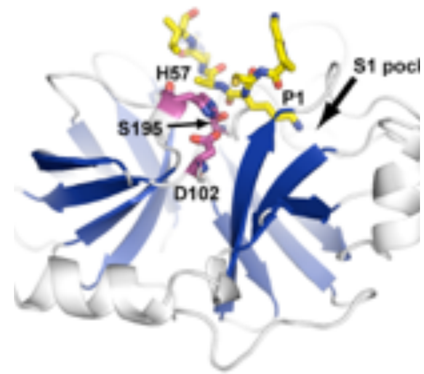
The method works by using the principle that source signals are statistically independent of each other.

Example…discovering patterns of coevolution in protein sequence alignments

We can make a matrix of the correlated conservation (or coevolution) of pairs of sequence positions….

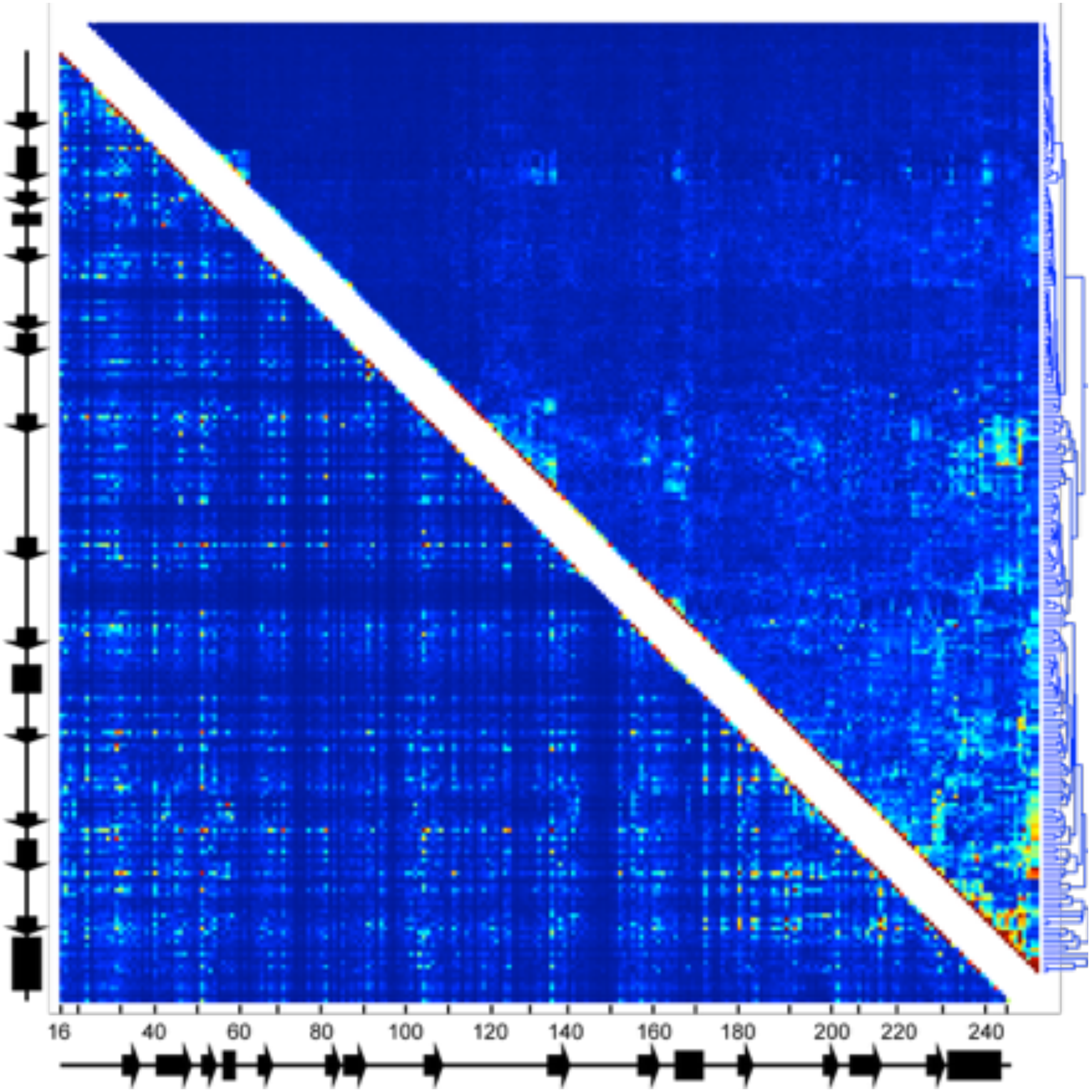
For example, in the S1A serine proteases (1470 sequences from diverse eukaryotic organisms)

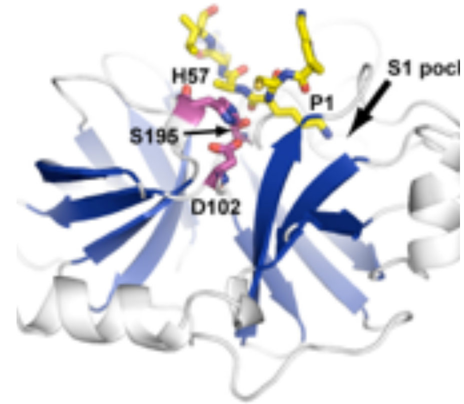




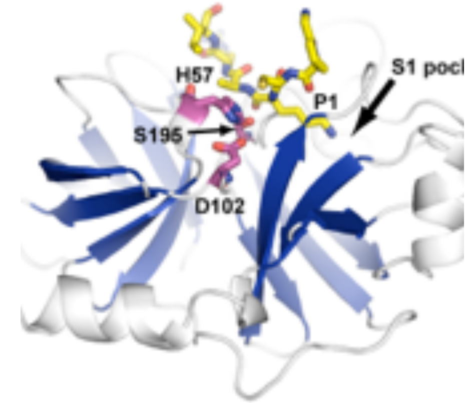
Halabi et al., Cell (2009) 138: 774-86..

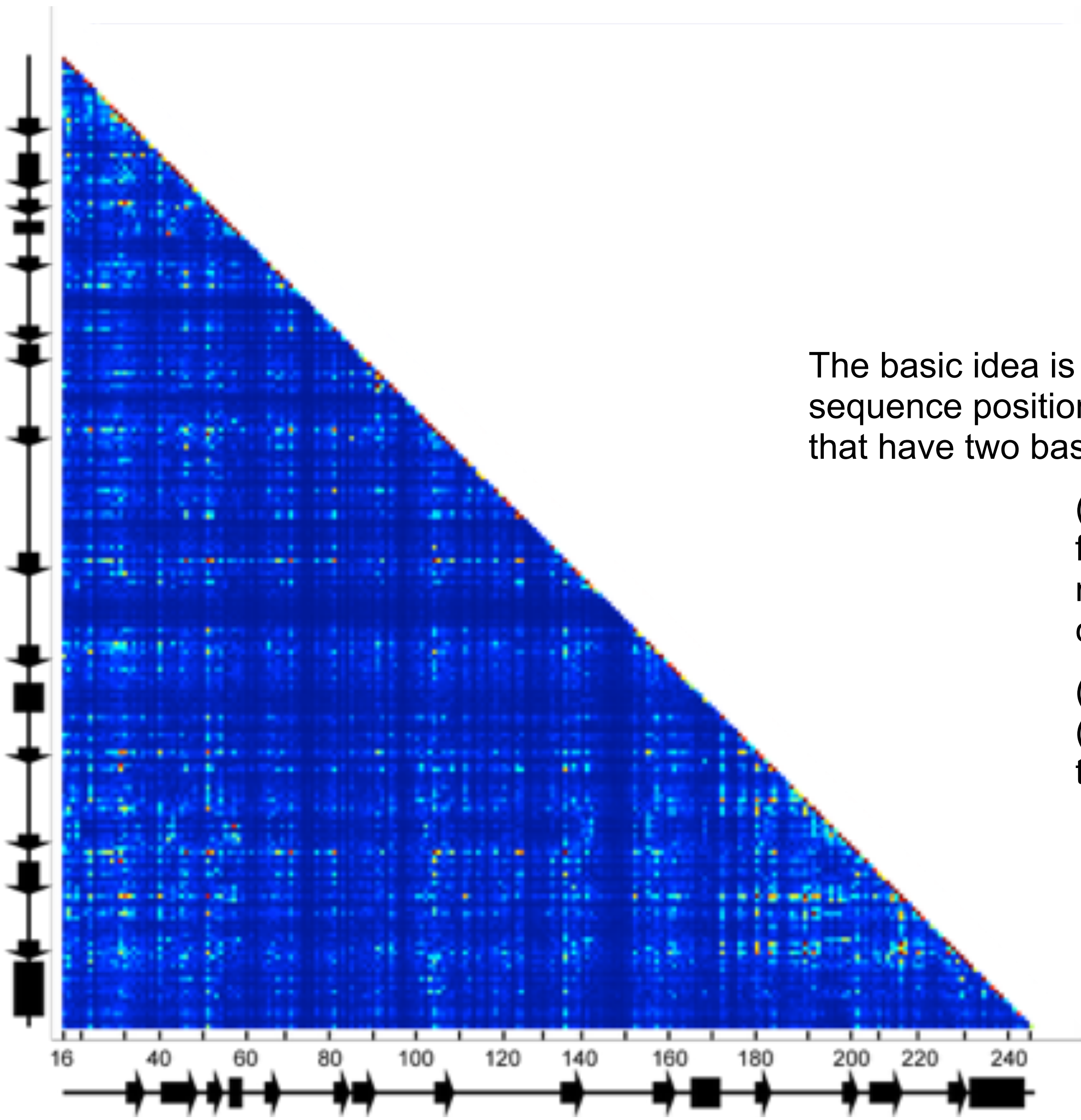
Clustering….





PCA…..





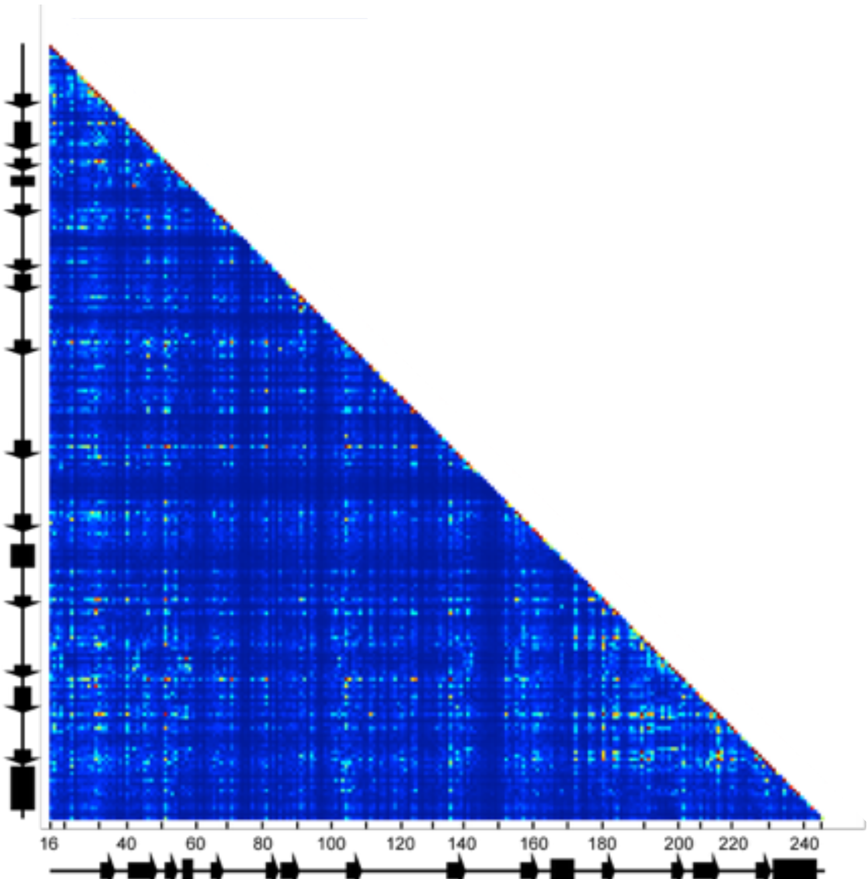
The basic idea is to transform the current variables (the sequence positions) into new variables (eigenmodes) that have two basic properties:

(1) they capture the information in a few new dimensions as possible (i.e. maximize variance per principal component).

(2) they are maximally non-redundant (i.e minimize co-variation in the transformed variables

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & & \\ a_{31} & & \ddots & \\ \vdots & & & \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & v_{13} & \cdots \\ v_{21} & v_{22} & & \\ v_{31} & & \ddots & \\ \vdots & & & \end{bmatrix} = \overset{\text{eigenvalues}}{\begin{bmatrix} D_{11} & 0 & 0 & \cdots \\ 0 & D_{22} & & \\ 0 & & \ddots & \\ \vdots & & & \end{bmatrix}} \overset{\text{eigenvectors}}{\begin{bmatrix} v_{11} & v_{12} & v_{13} & \cdots \\ v_{21} & v_{22} & & \\ v_{31} & & \ddots & \\ \vdots & & & \end{bmatrix}}$$

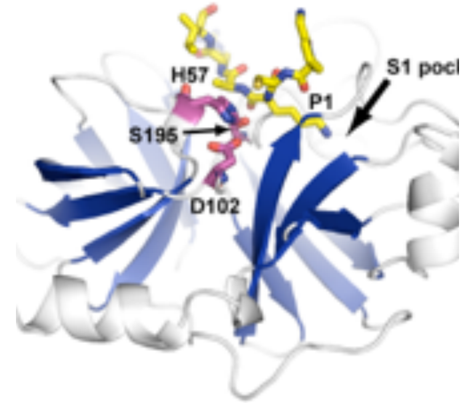
$\vec{v}_1 \quad \vec{v}_2$



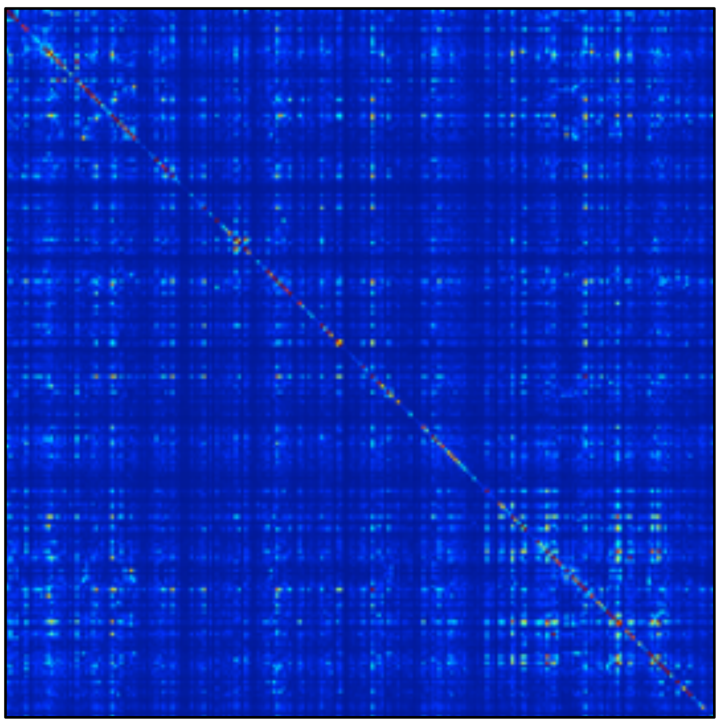
Mathematically, this amounts to computing the eigenvalues and eigenvectors of the SCA matrix...

$$A\vec{v} = \lambda\vec{v}$$

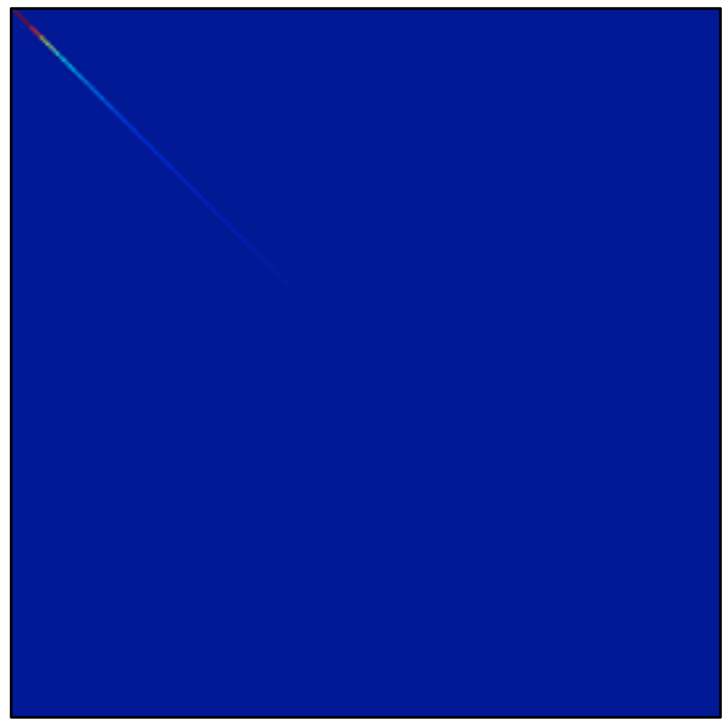
Here, the eigenvalues represent the quantity of variance captured in each new dimension, and each associated eigenvector contains the weights of each of the original sequence positions.



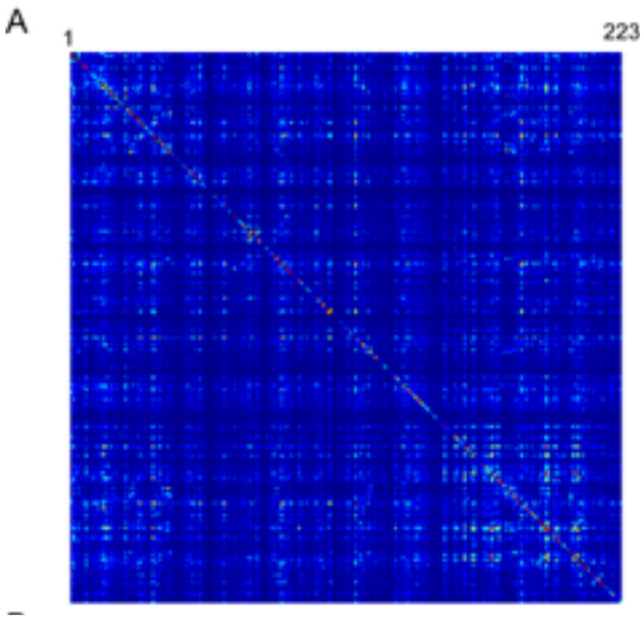
SCA matrix

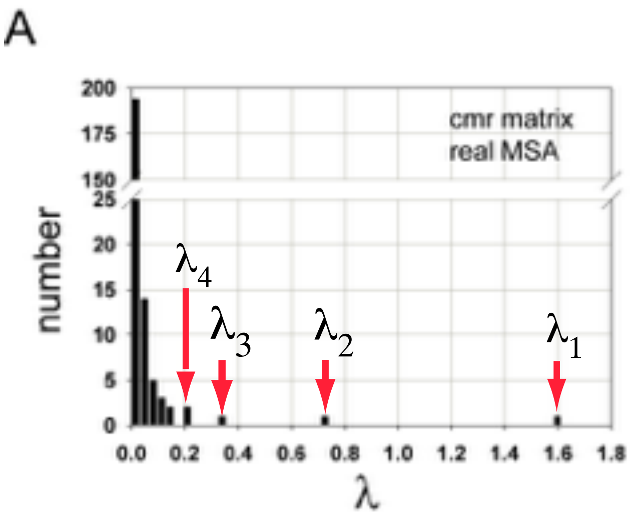


eigenvalues



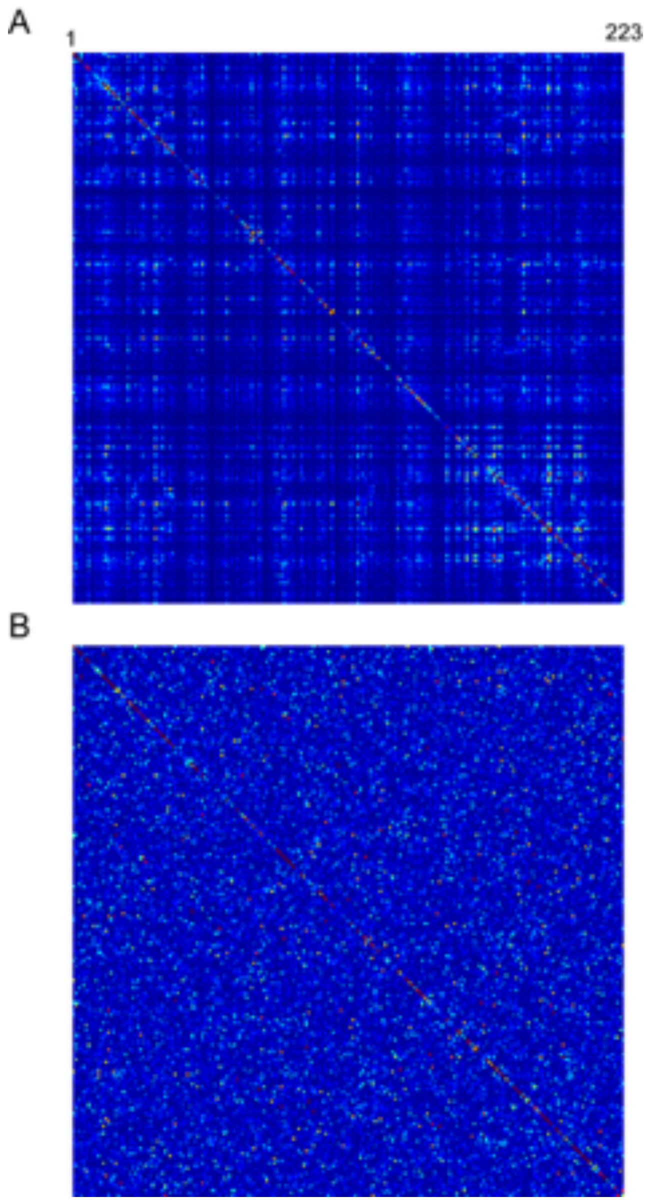
The eigenvalue spectrum

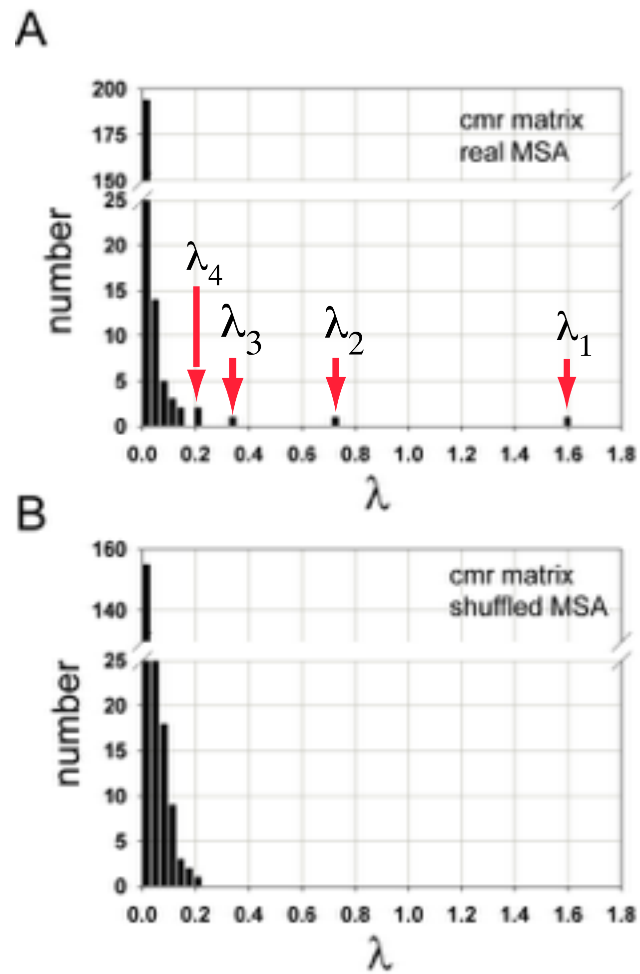




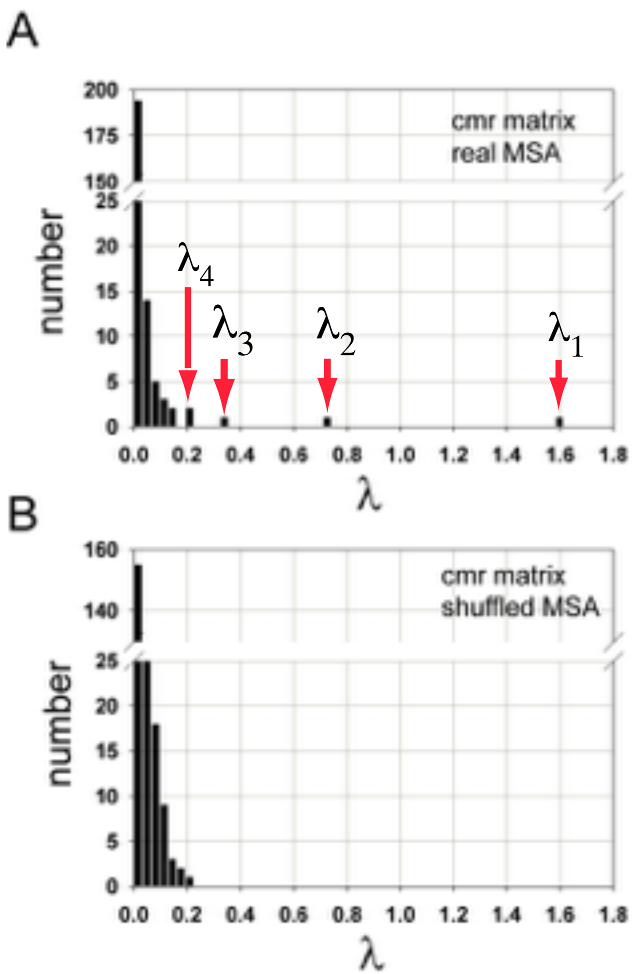
how many dimensions to keep?

The eigenvalue spectrum...and its random matrix counterpart





An analogy from economics



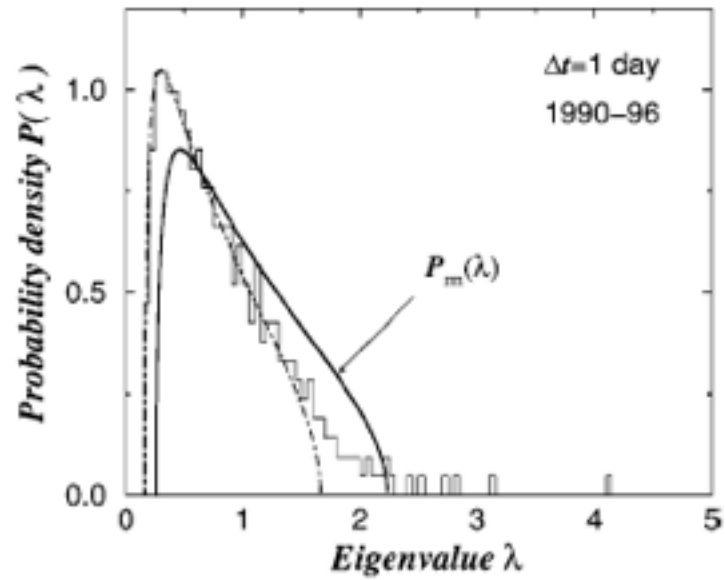
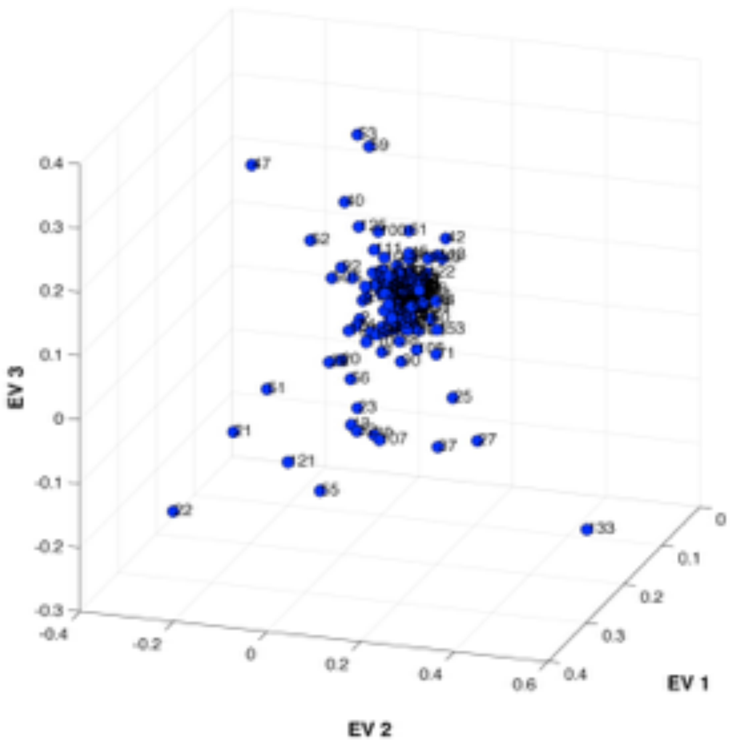


FIG. 4. $P(\lambda)$ for C constructed from daily returns of 422 stocks for the 7-yr period 1990–1996. The solid curve shows the RMT result $P_{rm}(\lambda)$ of Eq. (6]) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{rm}(\lambda)$ with λ_+ and λ_- as free parameters. We find similar results as found in Fig. 3(a) for 30-min returns. The largest eigenvalue (not shown) has the value $\lambda_{422}=46.3$.

Pilereu et al. (2002), Physical Review E 65, 066126
Laloux et al. (1999), PRL 83, p. 1467
Pilereu et al. (1999), PRL 83, p. 1471

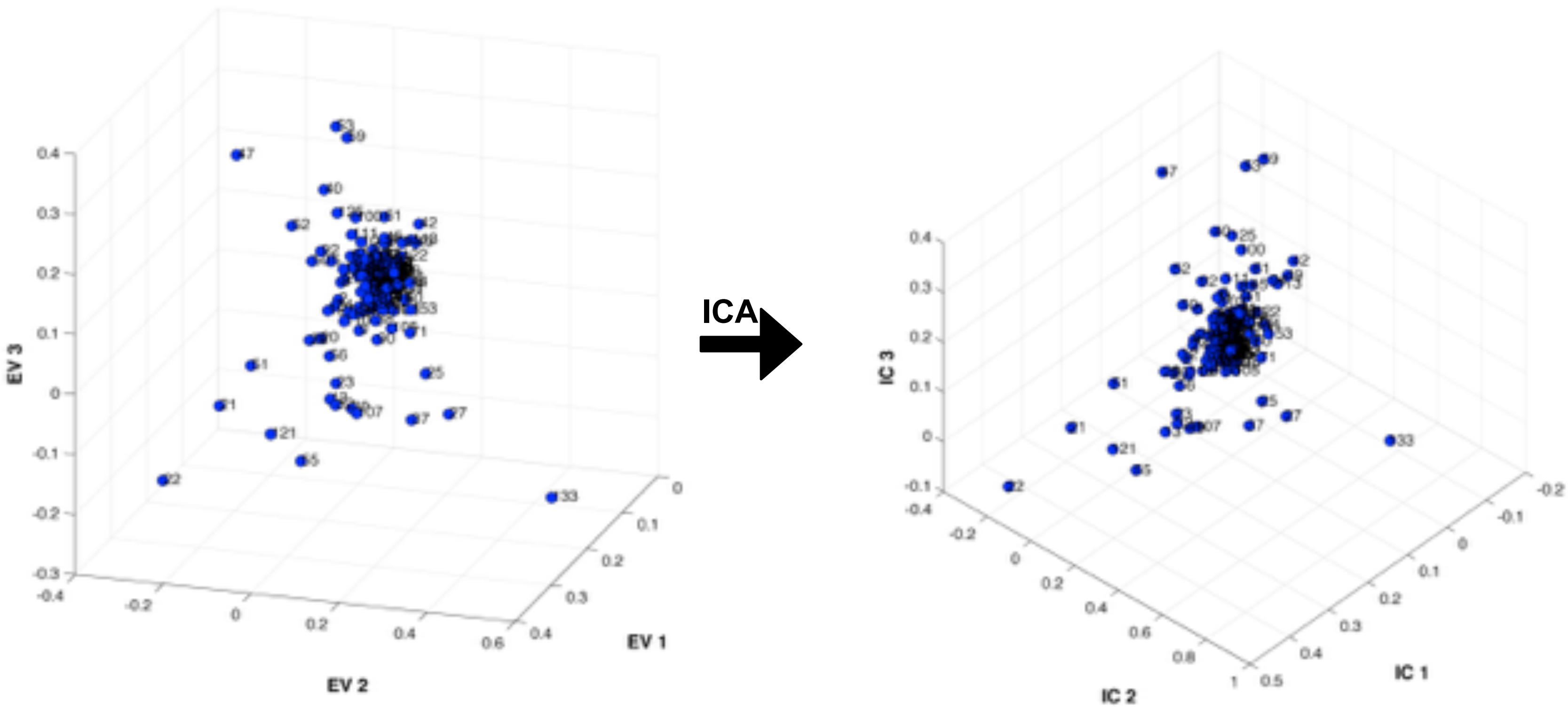
The top three eigenvectors….



As we know ad nauseum, eigenvectors need not represent maximally independent directions....

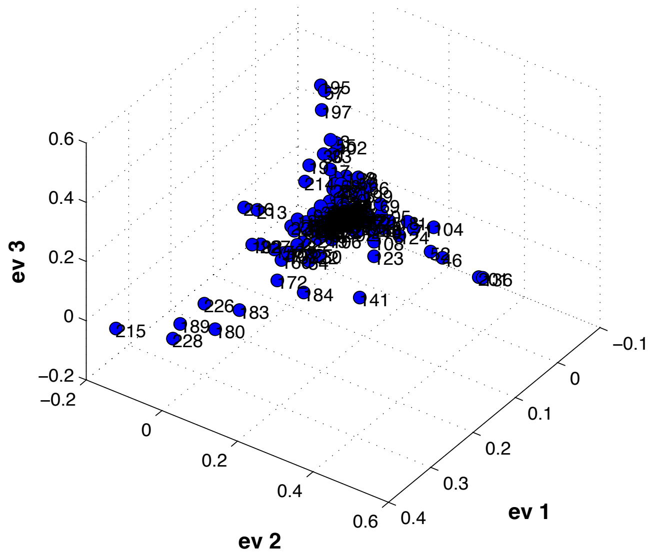
ICA…

```
kmax=3;
[SA.Vp,W]=rot_ica(SA.V,kmax);
```



ICA provides a “better” representation of quasi-independent modes.

Another example….



ICA provides a “better” representation of quasi-independent modes. We will come back to this later….

The large variable limit….linear decomposition methods

	n = 1	n = 2 or 3	n >> 1	continuum
Linear	exponential growth and decay single step conformational change fluorescence emission pseudo first order kinetics	second order reaction kinetics linear harmonic oscillators simple feedback control sequences of conformational change	electrical circuits molecular dynamics systems of coupled harmonic oscillators equilibrium thermodynamics diffraction, Fourier transforms	Diffusion Wave propagation quantum mechanics viscoelastic systems
Nonlinear	fixed points bifurcations, multi stability irreversible hysteresis overdamped oscillators	anharmonic oscillators relaxation oscillations predator-prey models van der Pol systems Chaotic systems	systems of non-linear oscillators non-equilibrium thermodynamics protein structure/ function neural networks the cell ecosystems	Nonlinear wave propagation Reaction-diffusion in dissipative systems Turbulent/chaotic flows

adapted from S. Strogatz