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**

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

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 experimental analysis...

Case 2: $\sum_{i=1}^{n} x_i$ array exprs

- genes

Case 2:

- Ca
- ATP
- pH
- specfic opps
- phosphates

... measurements

Case 3:

- ligands

- weakly
  - $Y_i$, $Y_i$...

- fluorescence

... weakly

- biophysical measurements

But...
What is the problem, fundamentally?

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

Case 2: \( \overbrace{\text{array exps}}^{\text{genes}} \)

Case 2: \( \overbrace{\text{Ca, phospho, etc.}}^{\text{ligands}} \rightarrow \overbrace{\text{fluorescence, etc.}}^{\text{measurements}} \rightarrow \text{biophysical measurements} \)

Case 3: \( \overbrace{\text{positions in a protein}}^{\text{genes in a genome}} \rightarrow \text{ligands applied to a cell} \rightarrow \text{positions in a protein} \)

We initially choose a parameterization of our system...
What is the problem, fundamentally?

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

But are those 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:

\[ \frac{\text{dy}}{\text{dy}} \rightarrow y_1 \]

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

\[ \Rightarrow \text{a 1-D motion...that's it.} \]

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 and B. And...we set up some laboratory frames of reference axes for recording the position of the mass \( x_1 \) and \( y_2 \).
principal components analysis, or **PCA**

So... some definitions.

1. $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ 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 $B$:

$$X = \begin{bmatrix} x_{1A} \\ x_{1B} \\ \vdots \\ x_{2A} \\ x_{2B} \end{bmatrix}$$

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 $X$:

$$y = PX$$

where...

1. $P$ is the transform 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 $X \rightarrow$ columns of $y$. 
principal components analysis, or **PCA**

So... look for a transformation of $X$:

$$Y = PX$$

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 $X \rightarrow$ 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...

- **minimize noise**... i.e., maximize SNR

\[
\text{SNR} = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}}
\]

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**

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:

\[ a = \{a_1, a_2, \ldots, a_n\} \]
\[ b = \{b_1, b_2, \ldots, b_n\} \]

The variance of \( a \) and \( b \) are defined as:

\[ \sigma_a^2 = \langle a_i a_i \rangle \quad \text{and} \quad \sigma_b^2 = \langle b_i b_i \rangle \]

The covariance of \( a \) and \( b \) is:

\[ \sigma_{ab}^2 = \langle a_i b_i \rangle \]
Variance and covariance....

consider two variables with zero mean:

\[ a = \{a_1, a_2, \ldots, a_n\} \]
\[ b = \{b_1, b_2, \ldots, b_n\} \]

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

now... obviously,

\[ \sigma_{ab}^2 = 0 \text{ iff } a \text{ and } b \text{ are not correlated} \]

\[ \sigma_{ab}^2 = \sigma_a^2 = \sigma_b^2 \text{ if } a = b. \]

In vector notation,

\[ \sigma_{ab}^2 = \frac{1}{n-1} a b^T \ldots \text{ remember} \]

\[ [a_1, a_2, \ldots, a_n] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = [a_1 b_1, a_2 b_2, \ldots, a_n b_n] \]
Variance and covariance....

Now, remember we have a matrix of observations:

\[
\begin{pmatrix}
    x_1 \\
    \vdots \\
    x_n \\
\end{pmatrix}
\]

... 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:

\[
S_x = \frac{1}{n-1} XX^T
\]

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:

\[ X = \begin{bmatrix} x_1 & \cdots & x_n \\ \vdots & \ddots & \vdots \\ x_m & \cdots & x_N \end{bmatrix} \quad \text{... an } m \times n \text{ matrix.} \]

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

So, the covariance matrix is an \( m \times m \) matrix:

\[ S_x = \frac{1}{n-1} XX^T \]

**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 \( x_i \) and \( x_j \).
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 $S_y$.

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

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 $P$?
The path to finding $P$ is **eigenvalue decomposition** of the covariation matrix of $X$, the initial variables...
To do this... begin by re-writing $S_y$:

$$S_y = \frac{1}{n-1} y y^T$$
$$= \frac{1}{n-1} (p x) (p x)^T$$
$$= \frac{1}{n-1} p x x^T p^T$$
$$= \frac{1}{n-1} p (x x^T) p^T$$

$$S_y = \frac{1}{n-1} P A P^T$$, where we define $A = x x^T$.

Recognize the matrix $A$? What is it?
To do this... begin by re-writing $S_y$:

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

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

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

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

$$S_y = \frac{1}{n-1} P A P^T$$, where we define $A = X X^T$.

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

$$A = E D E^T$$, where

$D$ is a diagonal matrix, and $E$ is the matrix of eigenVectors of $A$.

$$D = \begin{bmatrix}
\lambda_1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \cdots \\
0 & 0 & \ddots & \ddots \\
0 & \cdots & \cdots & \lambda_m
\end{bmatrix}$$
The eigenvalue decomposition...

\[
\mathbf{A} = \mathbf{V} \cdot \mathbf{\Lambda} \cdot \mathbf{V}^T
\]

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

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 with $m$ rows and $n$ columns is:

\[
A = \begin{bmatrix}
    a_{11} & a_{12} & \cdots & a_{1n} \\
    a_{21} & a_{22} & \cdots & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}, \quad \text{and } a_{ij} \text{ are elements of } A.
\]

A square matrix is a matrix with equal number of rows and columns. The order of a square matrix is the no. of rows (or columns).
Basic concepts from linear algebra...

A matrix with \( m \) rows and \( n \) cols is:

\[
A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}, \text{ and } a_{ij} \text{ are elements of } A.
\]

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).

Every square matrix has a scalar value associated with it ... the determinant \( \det \) For a second order matrix...

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

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

\[
A^{-1} A = I
\]

\( I \) is the identity matrix:

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

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 $\mathbf{v}$ such that:

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

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

Computing eigenvalues and eigenvectors:
So... \[ A \mathbf{v} \propto \lambda \mathbf{v} \]

rewrite as

\[ (A - \lambda I) \mathbf{v} = 0 \]

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

Rewrite as:

\[
(A - \lambda I)\mathbf{v} = 0
\]

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

Why? If it was, then...

\[
(A - \lambda I)^r (A - \lambda I)\mathbf{v} = (A - \lambda I)^0 \mathbf{v} = 0
\]

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

\[
\text{det}(A - \lambda I) = 0
\]
So \( A^2 = 2A \)

 rewrite as

\[
(A - \lambda I)v = 0
\]

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

Why? If it was, then...

\[
(A - \lambda I)^{-1}(A - \lambda I)v = (A - \lambda I)^{-1}0 \quad \text{or} \quad v = 0
\]

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

\[
\text{det}(A - \lambda I) = 0
\]

The equation \( p(\lambda) = \text{det}(A - \lambda I) \) is called the "characteristic polynomial" of \( A \) and the eigenvalues of \( A \) are just the roots of this polynomial... values of \( \lambda \) 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 $\lambda = 3$ 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.

\[ A \vec{v} = \lambda \vec{v} \text{ or } (A - \lambda I) \vec{v} = 0 \]

For \( \lambda = 3 \) ...

\[
\begin{bmatrix}
2 & -3 & -4 \\
-1 & -1 & -3
\end{bmatrix}
\begin{bmatrix}
\vec{v}_1 \\
\vec{v}_2
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

\[-\vec{v}_1 - 4\vec{v}_2 = 0 \\
-\vec{v}_1 - 4\vec{v}_2 = 0
\]

2 equations, 2 unknowns. So ...

\[ \vec{v} = \begin{bmatrix} 0.97 \\ -0.24 \end{bmatrix} \]

For \( \lambda = -2 \) ...

\[
\begin{bmatrix}
2 & -(-2) & -4 \\
-1 & -1 & -(-2)
\end{bmatrix}
\begin{bmatrix}
\vec{v}_1 \\
\vec{v}_2
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

So ...

\[ \vec{v} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \]
Now for the key point...

\[ A^T v = \lambda v, \quad \infty \ldots \]

\[
\begin{bmatrix}
2 & -4 \\
-1 & -1
\end{bmatrix}
\begin{bmatrix}
0.97 & 0.797 \\
-0.24 & 0.797
\end{bmatrix}
= 
\begin{bmatrix}
3 & 0 \\
0 & -2
\end{bmatrix}
\begin{bmatrix}
0.97 & 0.797 \\
-0.24 & 0.797
\end{bmatrix}
\]

Do you see that by multiplying by the eigenvalues we have diagonalized \( A \)?
we want to find $P$ such that $y = PX$ where:

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

is diagonalized. we concluded that...

$$S_y = \frac{1}{n-1} P(xx^T)P^T = \frac{1}{n-1} PAP^T \text{ where } A = xx^T$$

Except for the scaling term, $A$ is the covariance matrix of the original data variables.
we want to find $P$ such that $y = Px$ where:

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

is diagonalized. we concluded that...

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

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

where $A = x x^T$

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

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

$$A = E \Lambda E^T$$

where the matrix $\Lambda$ 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:

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

is diagonalized. We concluded that...

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

$$= \frac{1}{n-1} PA P^T \quad \text{where} \quad A = \nu x^T$$

So, what should we choose for $P$, so that $S_y$ is diagonalized?
Back to PCA...

\[ S_y = \frac{1}{n-1} \mathbf{P} \mathbf{A} \mathbf{P}^T \]

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

\[ \mathbf{A} = \mathbf{xx}^T \]

\[ \mathbf{A} = \mathbf{ED} \mathbf{E}^T \]

\[ \downarrow \]

\[ \mathbf{AP} \]

\[ \mathbf{A} = \mathbf{P}^T \mathbf{D} \mathbf{P} \]

(the **covariance matrix** of initial variables)

(the **eigenvalue decomposition** of \( \mathbf{A} \))

and so,
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...

\[ S_Y = \frac{1}{n-1} P (P^T D P) P^T \quad \text{since} \quad A = P^T D P \]

\[ = \frac{1}{n-1} (P P^T) D (P P^T) \]

\[ = \frac{1}{n-1} (P P^T) D (P P^T) \]

\[ S_Y = \frac{1}{n-1} D \]

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 $X$:

$$y = Px$$

$P$ should be chosen to:

1. maximize $\text{SNR} = \frac{\sigma^2}{\sigma^2_{\text{noise}}}$

2. eliminate redundancy ... that is minimize all off-diagonal elements of the $\text{cov}(X) = S_x$
Solution to PCA problem amounts to eigenvalue decomposition of $S_x$.

$S_x = E^T D E$, where $D = S_y$

Remember...

The eigenvalues ($D_1, D_2, \ldots$) represent the quantity of information (variance) in $S_x$ captured, and each associated eigenvector ($v_1, v_2, \ldots$) 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 eigenvector of the $S_x$ matrix and we calculate $y = Px$ such that $P = E^T$, then $S_y$ is diagonalized. $\Rightarrow$ 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, i.e., directions along which the original dataset 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.

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....

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).
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.** Assume that original variables contribute additively to the information. That is, assume no higher order correlations in the data that influence the pattern of pairwise correlations.

2. Assume 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 this 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 de-correlation (i.e. PCA) may not identify the statistically independent components.

Example to follow.
Limitations and extensions of PCA...

what if variables are not Gaussian distributed? Then
decomposition does not mean statistical independence.
Limitations and extensions of PCA...

What if variables are not Gaussian distributed? Then de-correlation 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. 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...

\[
p(z) = \begin{cases} 
\frac{1}{2\sqrt{3}}, & \text{if } |z| < \sqrt{3} \\
0, & \text{otherwise}
\end{cases}
\]

These are truly uncorrelated and statistically independent.
Limitations and extensions of PCA...

Now let's write \( y_1 \) and \( y_2 \) by applying the following transform:

\[
x = Ay,
\]

where

\[
A = \begin{pmatrix} 5 & 10 \\ 10 & 2 \end{pmatrix}
\]

Then...

Now, \( y_1 \) and \( y_2 \) 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 \( x_2 \) in \( x_1 \)... 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...
Statistical independence is a more rigorous concept than de-correlation. Independence implies de-correlation, but de-correlation does not demonstrate independence except for the case of Gaussian distribution variables.

Put simply... statistical independence means that

\[ p(a, b) = p(a) \cdot p(b) \]

\[ \Rightarrow \] 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.
Statistical independence is a more rigorous concept than de-correlation. Independence implies de-correlation, but de-correlation 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.
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

\[
p(y_1, y_2) = p(y_1) \cdot p(y_2).
\]

But also,

\[
p(h(y_1), g(y_2)) = p(h(y_1)) \cdot p(g(y_2))
\]

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

$$p(y_1, y_2) = p(y_1) \cdot p(y_2).$$

But also,

$$p(h(y_1), g(y_2)) = p(h(y_1)) \cdot p(g(y_2)).$$

- 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.
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) works by using the principle that source signals are statistically independent of each other.

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....
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).
Mathematically, this amounts to computing the eigenvalues and eigenvectors of the SCA matrix...\[
\hat{A} \vec{v} = \lambda \vec{v}
\]

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.
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 $P_{rs}(\lambda)$ of Eq. (6) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{ms}(\lambda)$ with $\lambda_{-}$ and $\lambda_{-}$ 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 provides a “better” representation of quasi-independent modes.

\[ \text{kmax}=3; \]
\[ [\text{SA.Vp},W]=\text{rot}\_ica(\text{SA.V},\text{kmax}); \]
ICA provides a “better” representation of quasi-independent modes. We will come back to this later....
The large variable limit….linear decomposition methods

<table>
<thead>
<tr>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 1</strong></td>
<td><strong>n = 2 or 3</strong></td>
</tr>
<tr>
<td>exponential growth and decay</td>
<td>second order reaction kinetics</td>
</tr>
<tr>
<td>single step conformational change</td>
<td>linear harmonic oscillators</td>
</tr>
<tr>
<td>fluorescence emission</td>
<td>simple feedback control</td>
</tr>
<tr>
<td>pseudo first order kinetics</td>
<td>sequences of conformational change</td>
</tr>
<tr>
<td><strong>Linear</strong></td>
<td><strong>Nonlinear</strong></td>
</tr>
<tr>
<td>fixed points</td>
<td>anharmonic oscillators</td>
</tr>
<tr>
<td>bifurcations, multi stability</td>
<td>relaxation oscillations</td>
</tr>
<tr>
<td>irreversible hysteresis</td>
<td>predator-prey models</td>
</tr>
<tr>
<td>overdamped oscillators</td>
<td>van der Pol systems</td>
</tr>
<tr>
<td>Chaotic systems</td>
<td>Chaotic systems</td>
</tr>
</tbody>
</table>

adapted from S. Strogatz