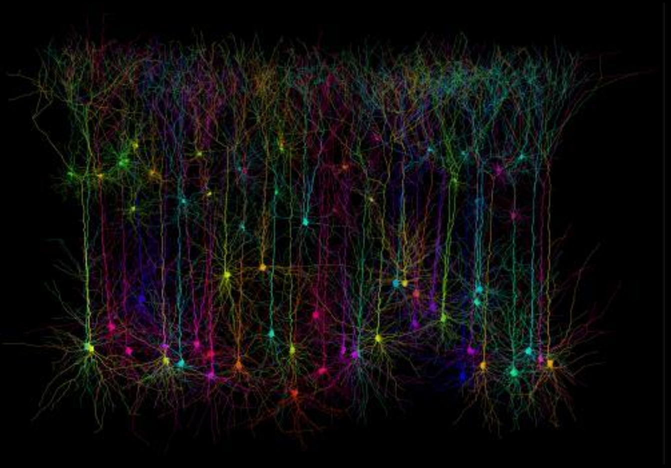
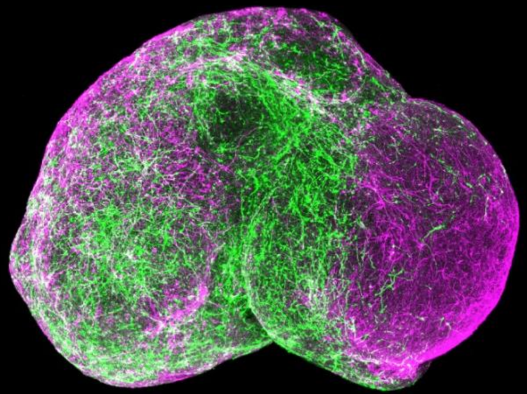
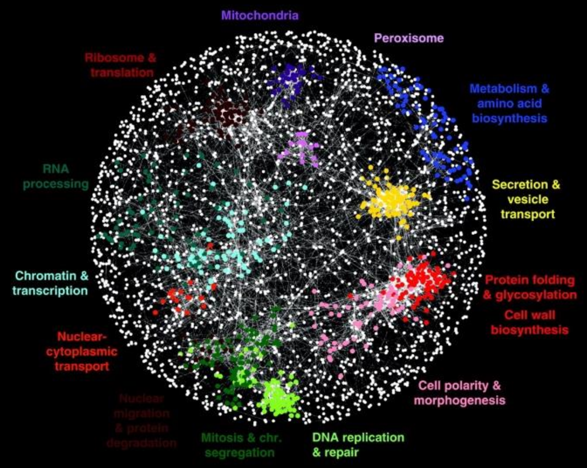
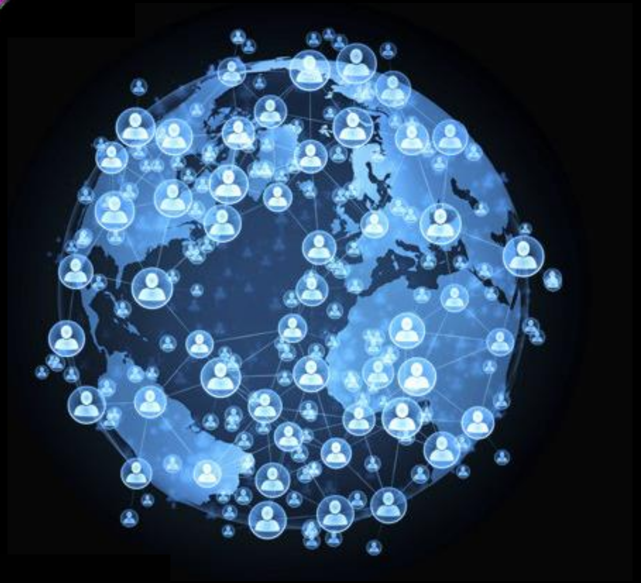
NEUR 608: Dimensionality Reduction

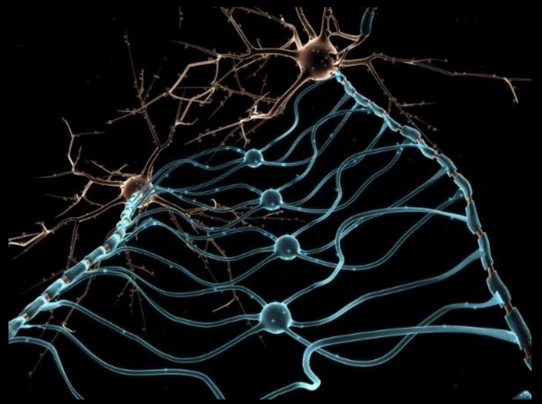
Bratislav Misic
McConnell Brain Imaging Centre
Montréal Neurological Institute
McGill University

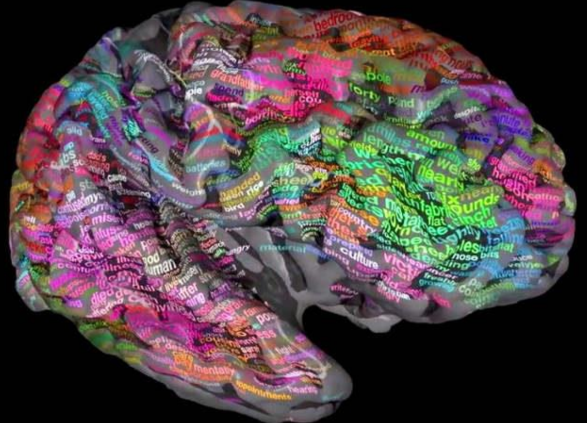





Mitochondria
Peroxisome
Ribosome & translation
Metabolism & amino acid biosynthesis
RNA processing
Secretion & vesicle transport
Chromatin & transcription
Protein folding & glycosylation
Cell wall biosynthesis
Nuclear-cytoplasmic transport
Cell polarity & morphogenesis
Nuclear migration & protein degradation
Mitosis & chr. segregation
DNA replication & repair

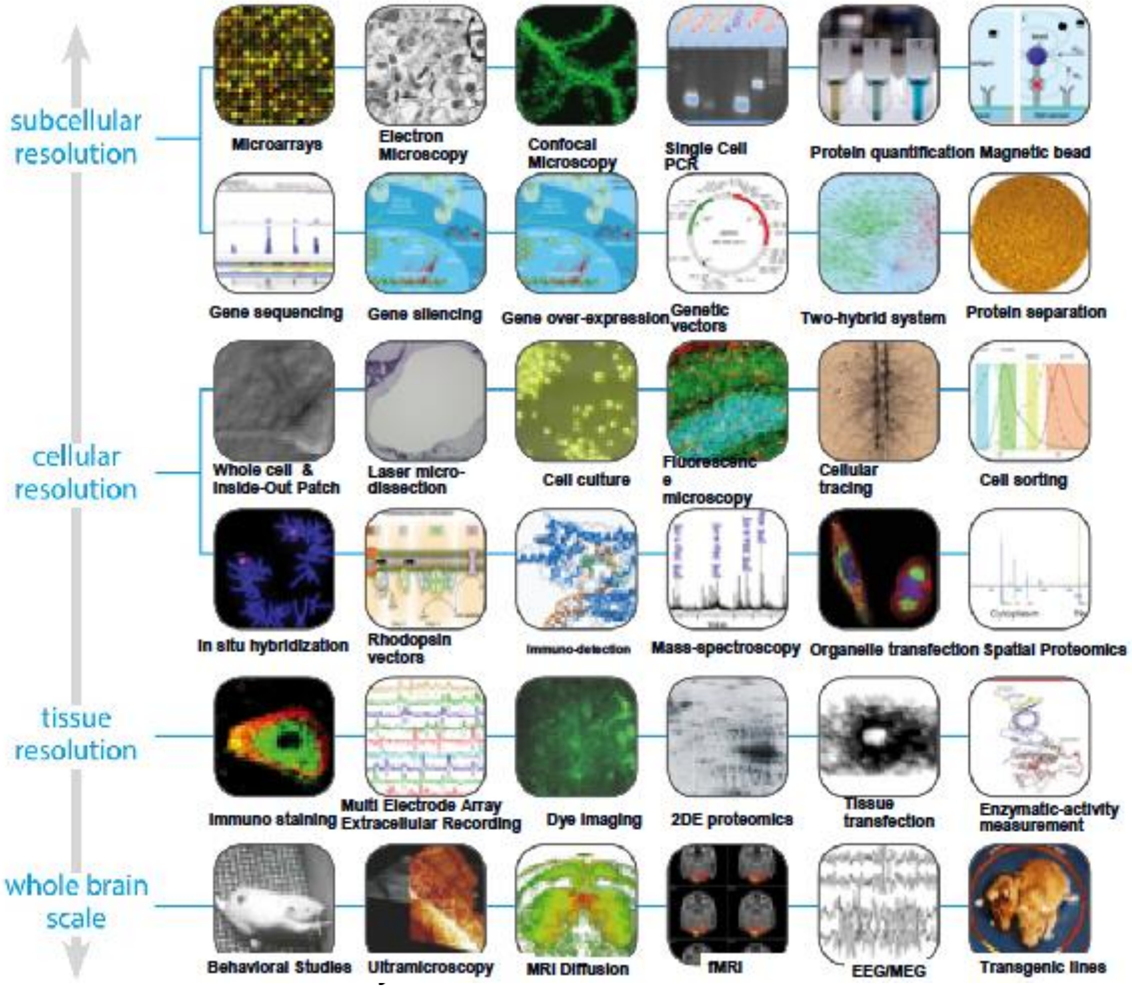






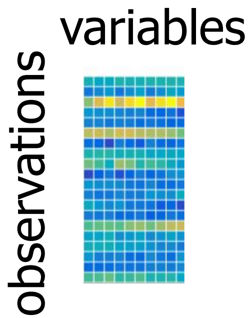
Towards multivariate analysis

- activation: mapping individual elements to individual external variables
- connectivity: mapping individual structural or functional connections to individual external variables
- networks: mapping network attributes to individual external variables
- multivariate systems: mapping patterns of elements or connections to patterns of external variables

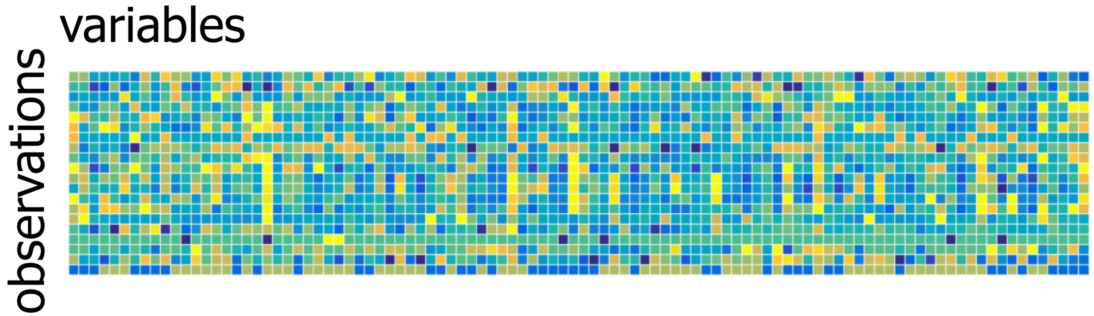


Misic & Sporns (2016) Curr Opin Neurobiol

Integrating scales and modalities



+



- How to deal with more variables than observations?
- How to operationalize the network property?
- How to relate multiple data sets to each other?

Lecture plan

- Goal #1: why dimensionality reduction?
  - linear: PCA, FA, ICA, MDS
  - nonlinear: kernel PCA, LLE, diffusion maps, t-SNE, autoencoders
- Goal #2: how is similarity expressed?
- Goal #3: is there an underlying generative model?
- Goal #4: statistical inference and reliability (and lack thereof)
- Goal #5: examples + demo

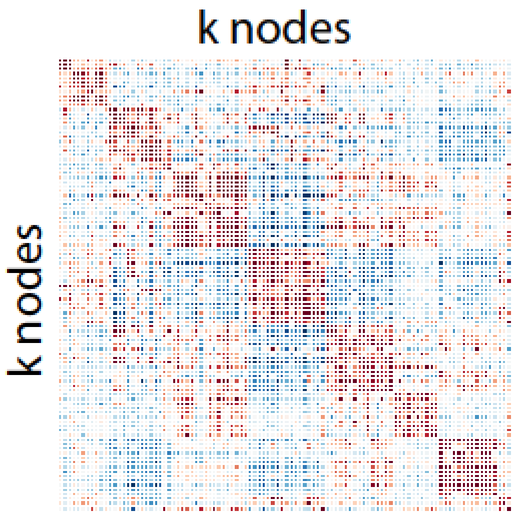
Principal component analysis (PCA)

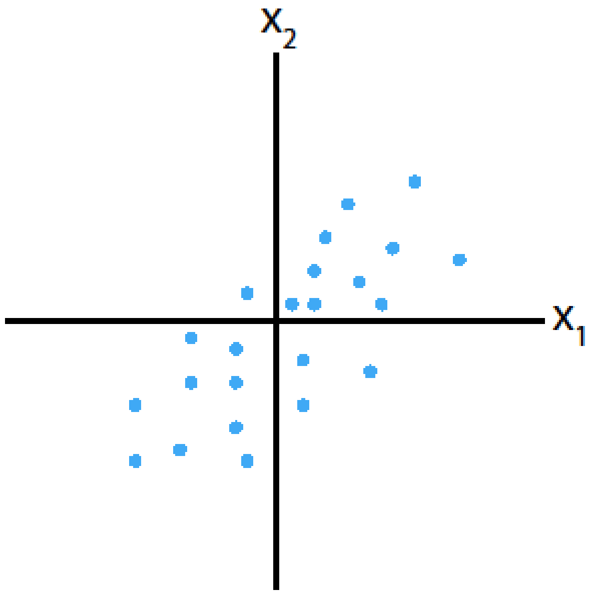
- original variables $\mathbf{x_1}$ and $\mathbf{x_2}$ are correlated but neither captures the dominant pattern of variance

- want a new variables $\mathbf{z_1}$ and $\mathbf{z_2}$ that
  (a) capture as much variance
  (b) are mutually uncorrelated

- need to find a rotation **u** to re-align the original axes (variables)

- **u** is chosen to maximize the variance of the new variables **z**





Hotelling (1933) J Educ Psychol

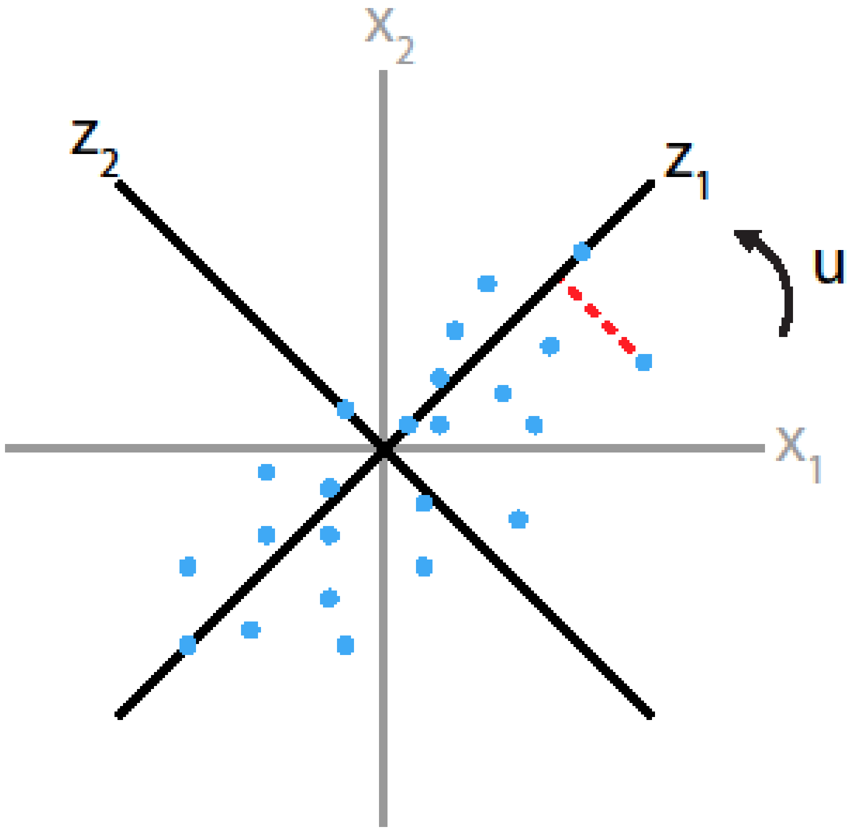
Maximzing variance

- want to find a new variable $\mathbf{z} = \mathbf{X}\mathbf{u}$
- choose $\mathbf{u}$ to maximize $var(\mathbf{z})$

  under the constraint that $\mathbf{u}'\mathbf{u} = \mathbf{1}$
- $var(\mathbf{z}) = \frac{1}{n-1}\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} = \mathbf{u}'\mathbf{R}\mathbf{u}$

  where $\mathbf{R} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$
- define Lagrangian: $L = \mathbf{u}'\mathbf{R}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1)$
- find maximum: $\frac{\partial L}{\partial \mathbf{u}} = 2\mathbf{R}\mathbf{u} \; - 2\mathbf{u}\lambda = 0$

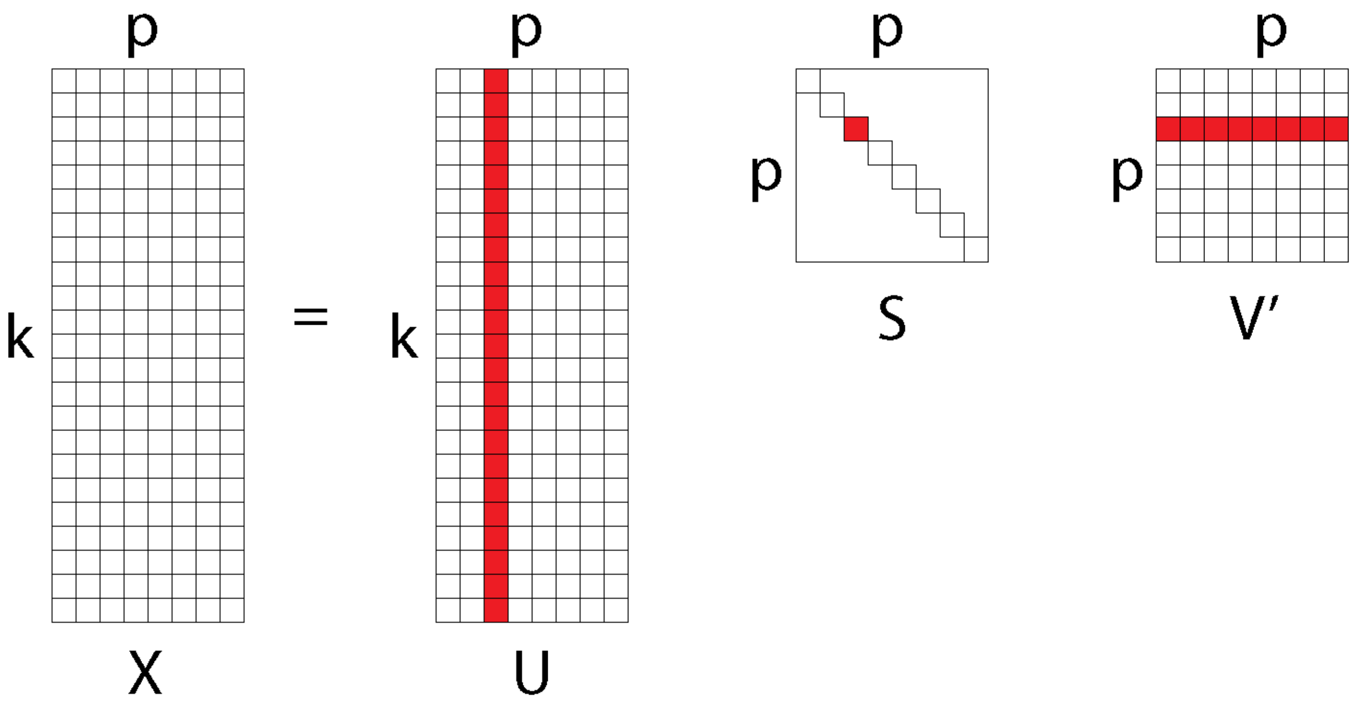
  $$\mathbf{R}\mathbf{u} = \lambda\mathbf{u} \qquad (\mathbf{R} - \lambda\mathbf{I})\mathbf{u} = 0$$
- $\lambda =$ eigenvalue, $\mathbf{u} =$ eigenvector
- $var(\mathbf{z}) = \mathbf{u}'\mathbf{R}\mathbf{u} = \mathbf{u}'\mathbf{u}\lambda = \lambda$



Hotelling (1933) J Educ Psychol

Singular value decomposition (SVD)

- how can we factorize a data matrix?
  - $SVD(\mathbf{X}) = \mathbf{USV}'$
- this is a generalization of the spectral decomposition:
  - $EIG(\mathbf{X}'\mathbf{X}) = \mathbf{U\Lambda U}'$
  - $EIG(\mathbf{XX}') = \mathbf{V\Lambda V}'$
- **U** and **V** are orthonormal singular vectors; represent how you should weigh the original variables in **X**
- **S** is a diagonal matrix of singular values; represent how strongly paired the **U** and **V** are



Eckart & Young (1936) Psychometrika

Singular value decomposition (SVD)

  - $SVD(\mathbf{X}) = \mathbf{USV'}$

  - $EIG(\mathbf{X'X}) = \mathbf{U\Lambda U'}$
  - $EIG(\mathbf{XX'}) = \mathbf{V\Lambda V'}$

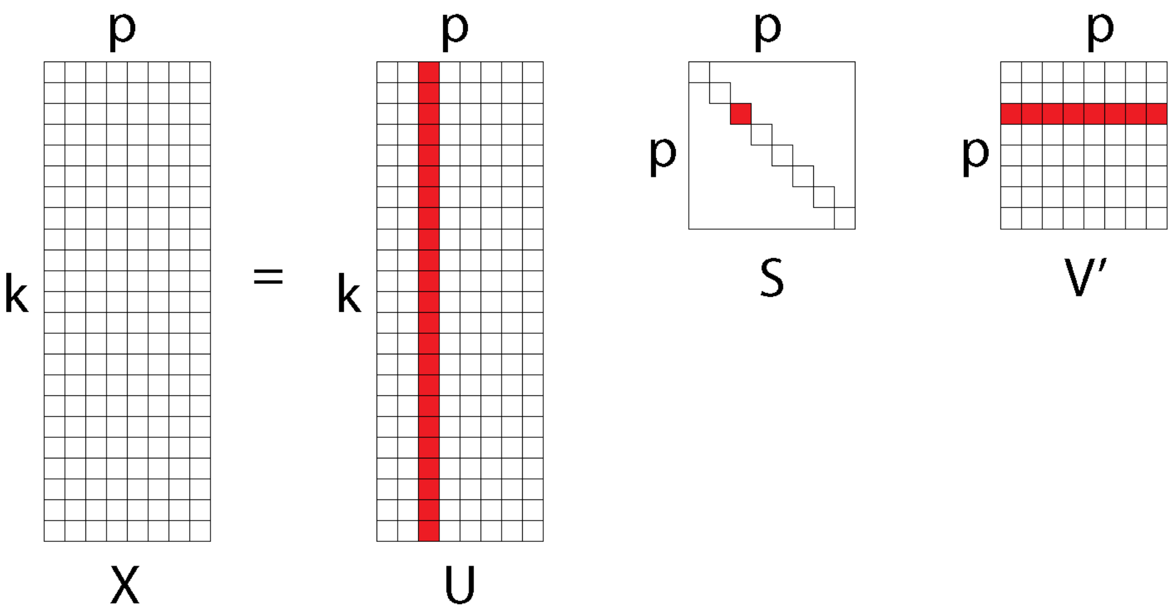
- $$\begin{aligned}\mathbf{X'X} &= (\mathbf{VS'U'})(\mathbf{USV'}) = \mathbf{VS'(U'U)SV'} \\ &= \mathbf{V(S'S)V'} = \mathbf{V\Lambda V}\end{aligned}$$

- $$\begin{aligned}\mathbf{XX'} &= (\mathbf{USV'})(\mathbf{VS'U'}) = \mathbf{US(V'V)S'U'} \\ &= \mathbf{U(SS')U'} = \mathbf{U\Lambda U'}\end{aligned}$$

- eigenvector of $\mathbf{XX'}$ = left singular vector of $\mathbf{X}$

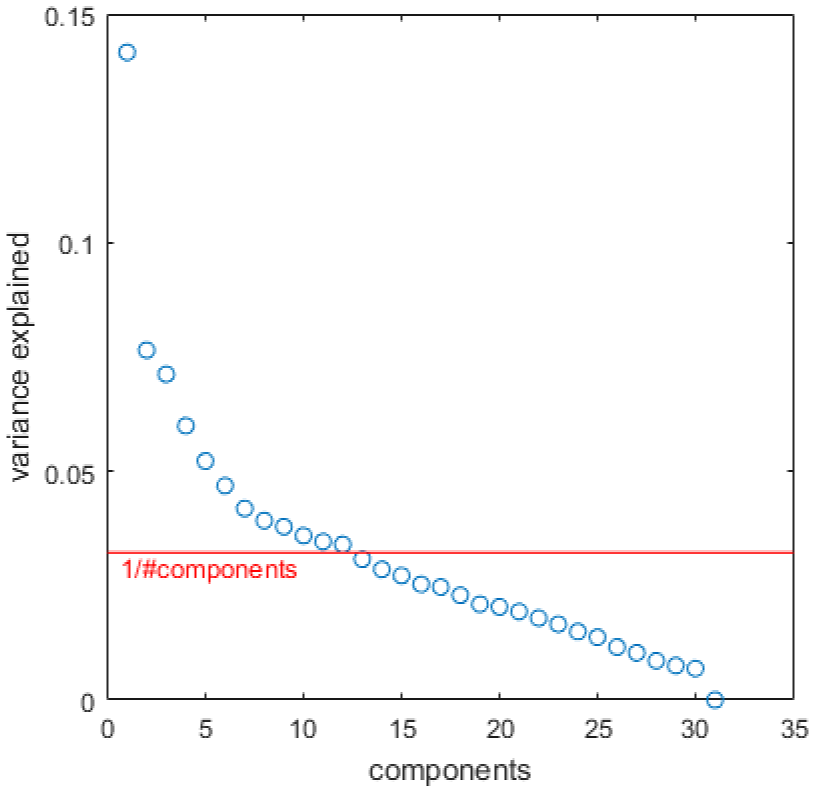
- eigenvector of $\mathbf{X'X}$ = right singular vector of $\mathbf{X}$

- eigenvalue = squared singular value



Eckart & Young (1936) Psychometrika

How many components to retain?

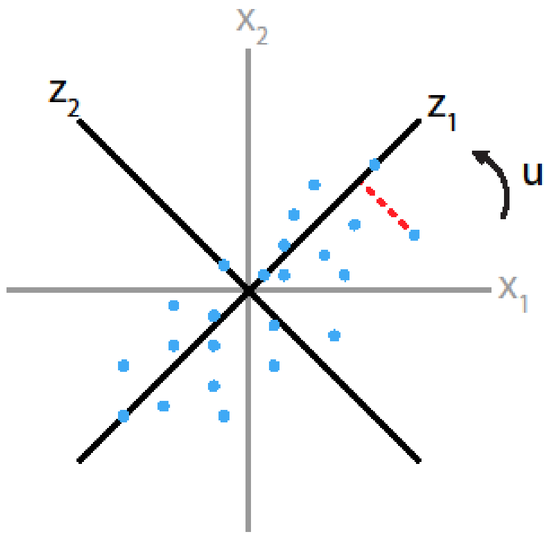


- Kaiser criterion: drop all eigenvalues lower than 1
- “elbow” rule: look for the biggest drop-off in the scree plot

Cattell (1966) Mult Behav Res

Thorndike (1953) Psychometrika

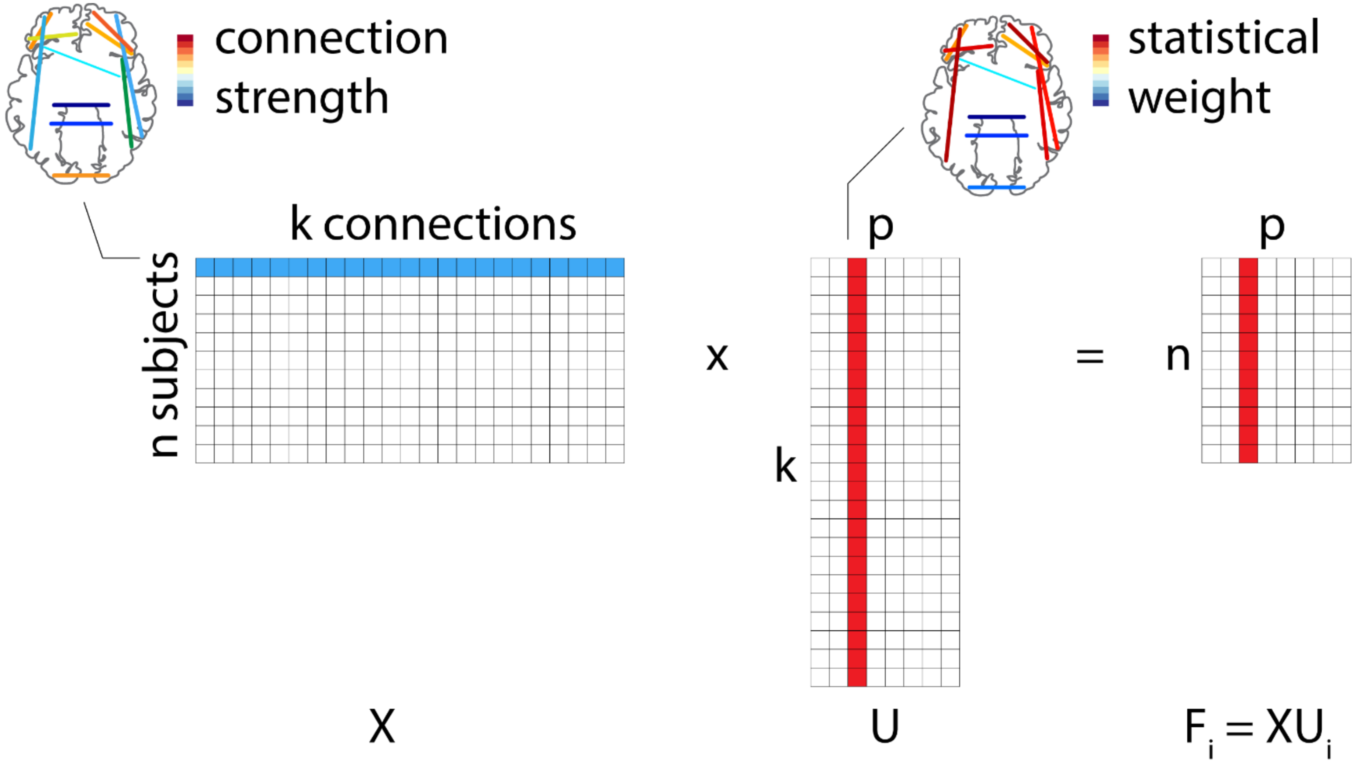
Individual participants



- weights that tell us how much each variable (connection, region, behaviour) contributes to the pattern

- but every participant is different – how do individual participants express the overall pattern?

- participant scores – project original data onto the latent variables

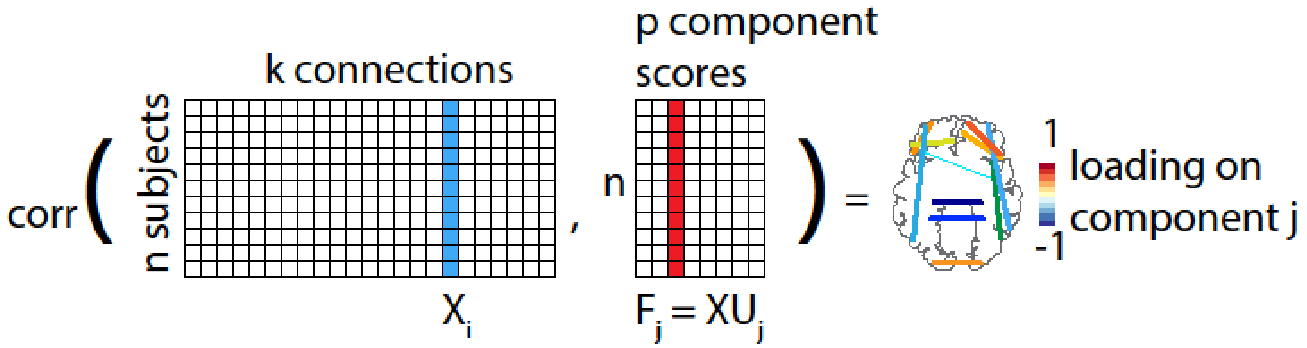


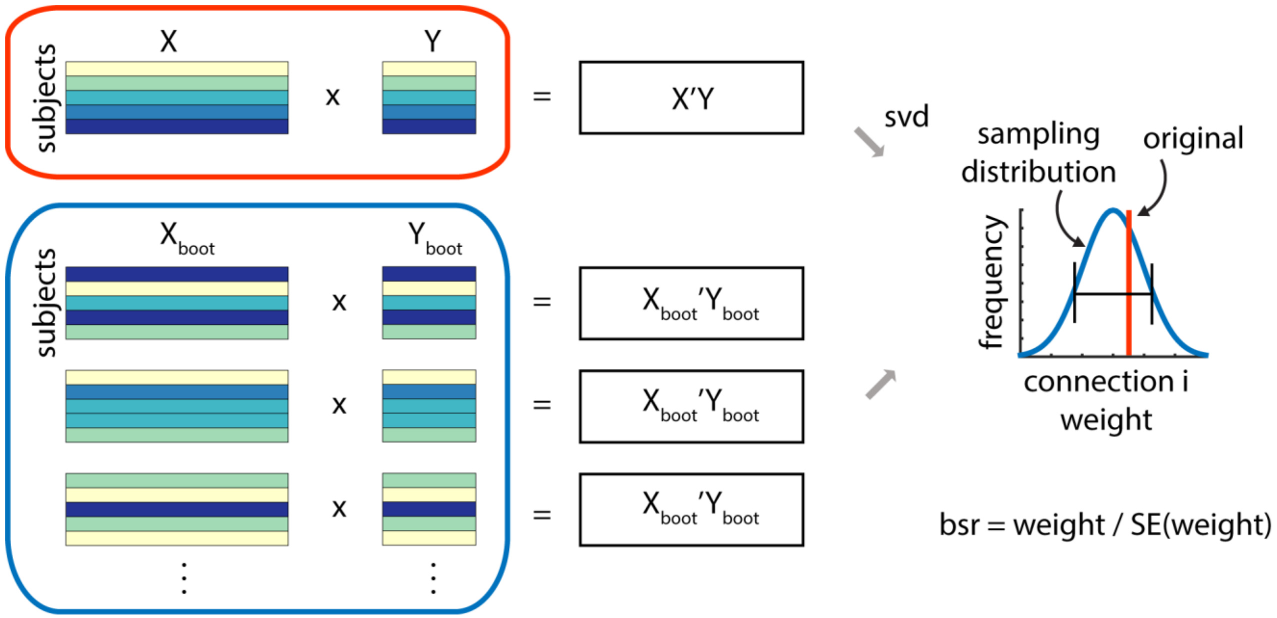
Which variables are important?

- singular vector weight: how much should I weigh the variables?

- loading: does expression of the original variable correlate with expression of the latent variable?

- “bootstrap ratio”: which original variables are weighed highly but also stable across participants?



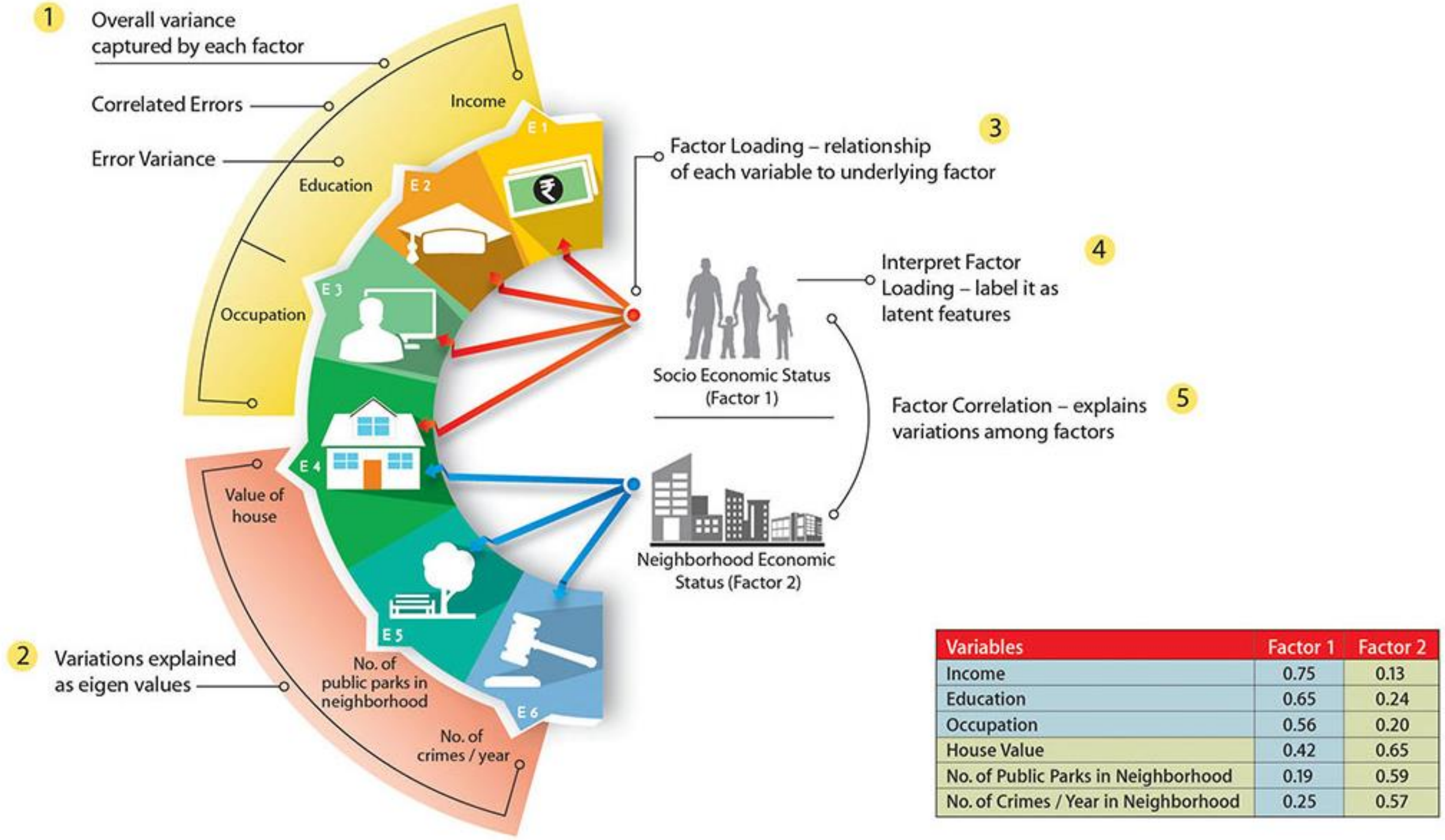


bsr = weight / SE(weight)

Efron & Tibshirani (1986) Stat Sci

McIntosh et al. (1996) NeuroImage

Factor Analysis (FA)

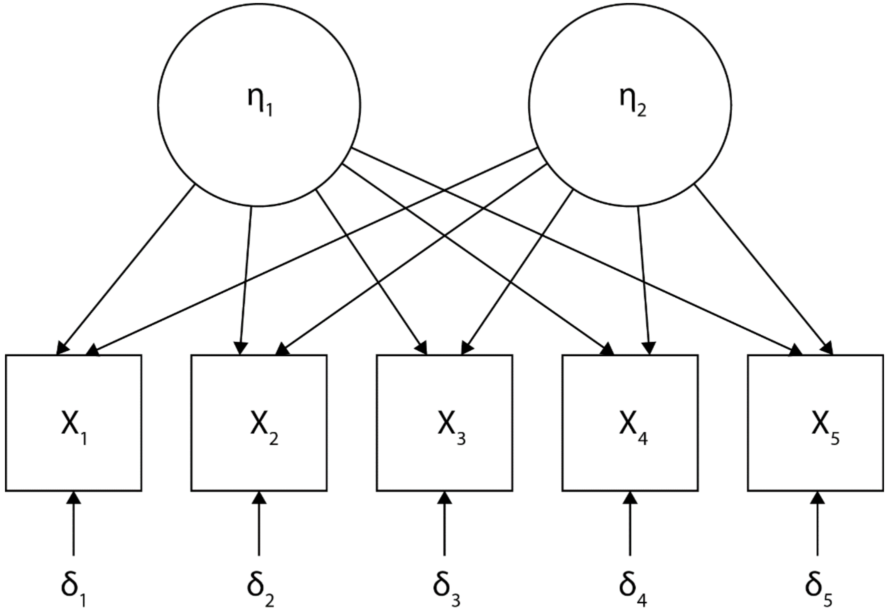

1
Overall variance captured by each factor
Correlated Errors
Error Variance
Income
E 1
Education
E 2
E 3
Occupation
E 4
Value of house
E 5
No. of public parks in neighborhood
E 6
No. of crimes / year
2
Variations explained as eigen values
3
Factor Loading – relationship of each variable to underlying factor
4
Interpret Factor Loading – label it as latent features
Socio Economic Status (Factor 1)
5
Factor Correlation – explains variations among factors
Neighborhood Economic Status (Factor 2)

Variables	Factor 1	Factor 2
Income	0.75	0.13
Education	0.65	0.24
Occupation	0.56	0.20
House Value	0.42	0.65
No. of Public Parks in Neighborhood	0.19	0.59
No. of Crimes / Year in Neighborhood	0.25	0.57

Factor Analysis (FA)

- Assume that variance in variable X_i comes from a set of latent factors η_i and measurement error δ_i:
  - $X_1 = \lambda_{1,1}\eta_1 + \lambda_{1,2}\eta_2 + \delta_1$
  - $X_2 = \lambda_{2,1}\eta_1 + \lambda_{2,2}\eta_2 + \delta_2$

  …
- Unique factors δ_i are uncorrelated, so they only contribute to the diagonals of the covariance matrix
  - $var(X_i) = var(\lambda_{i,1}\eta_1 + \lambda_{i,2}\eta_2 + \delta_i)$
  - $var(X_i) = \lambda_{i,1}^2 + \lambda_{i,2}^2 + \theta_i^2$, where $\theta_i^2 = var(\delta_i)$
- The idea:
  - Estimate θ_i^2, subtract them, and decompose the matrix
  - PCA: decompose correlation matrix **R**, with 1's on diagonal
  - FA: decompose matrix with diagonal elements $1 - \theta_i^2$



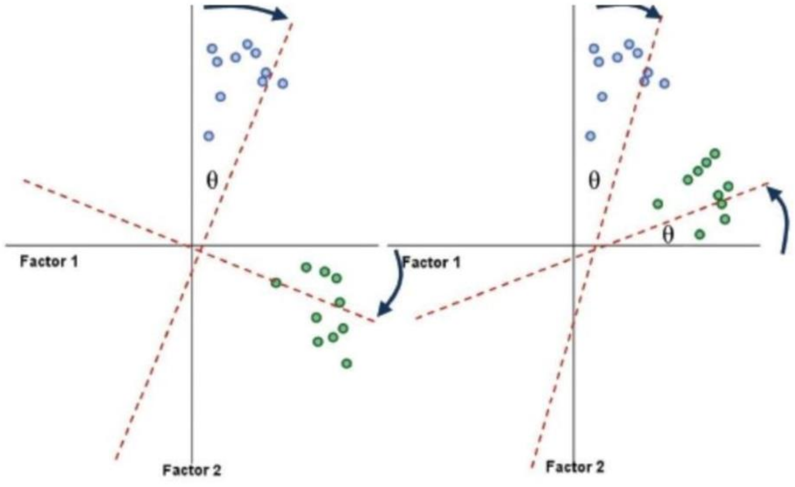
Factor Analysis (FA)

- How do we estimate $1-\theta_i^2$ ("communalities")?
  - option #1: guess and re-adjust
  - option #2: estimate squared multiple correlation (SMC), i.e. regress X_i on all $X_{j\neq i}$
- Procedure:
  1. calculate $\boldsymbol{R}$
  2. calculate $\boldsymbol{R}_{adj}$
  3. $EIG(\boldsymbol{R}_{adj})$
  4. are communalities stable?
     - no: recalculate $\boldsymbol{R}_{adj}$
     - yes: stop

Factor Analysis (FA): rotation

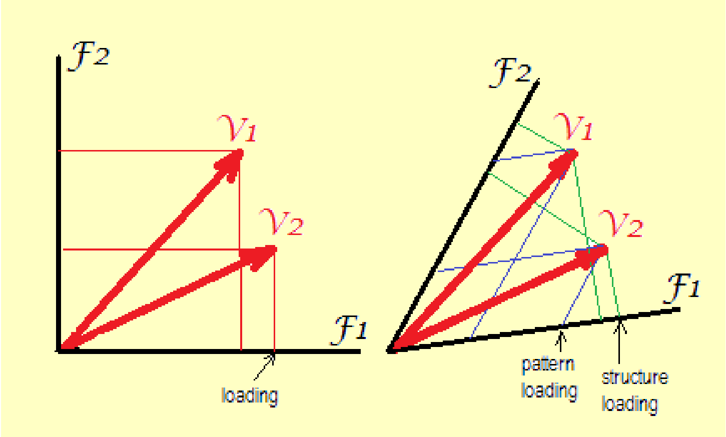
- We want to improve interpretability: each factor has high loadings on only a few variables, and near zero for all others
  - orthogonal: factors remain uncorrelated (e.g. *varimax*)
  - oblique : factors may correlate (e.g. *promax*, *oblimin*)

- If we perform an oblique rotation, it is no longer obvious how to interpret factors:
  - structure loading
  - pattern loading



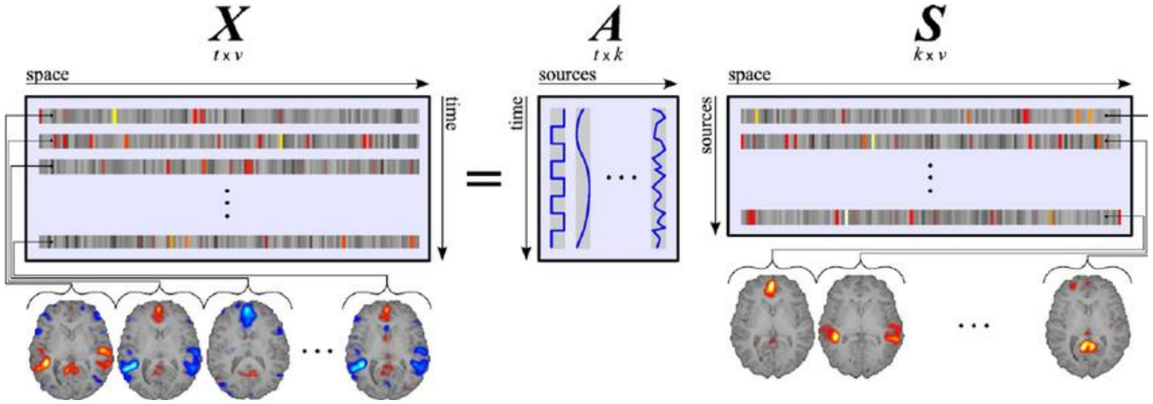
Orthogonal Rotation Oblique Rotation

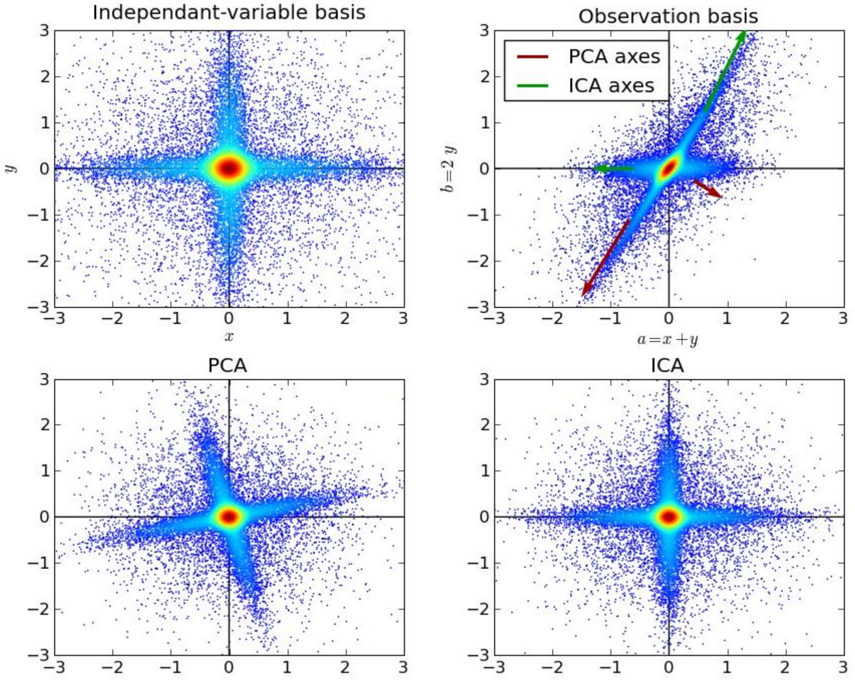
PSYC 4310/6310 Experimental Methods and Statistics © 2014, Michael Kalsher



Independent Components Analysis (ICA)

- Blind source separation
- Assumption: there exist a finite number of sources (independent components) that are mixed to produce observed neural activity
- Solution:
  - Any linear mixture of independent variables (e.g. voxels) will be more Gaussian than the original variables
  - Create new axes with maximally non-Gaussian projections
  - Many algorithms + objective functions

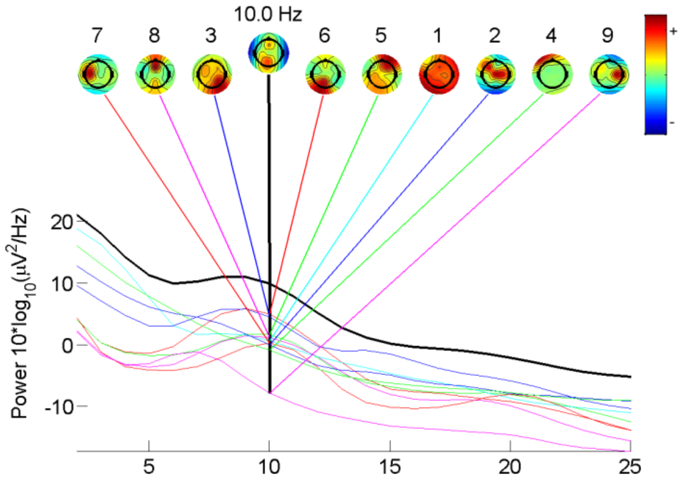


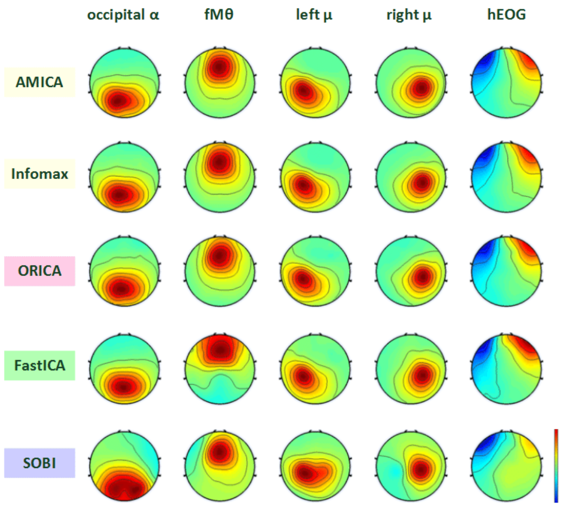


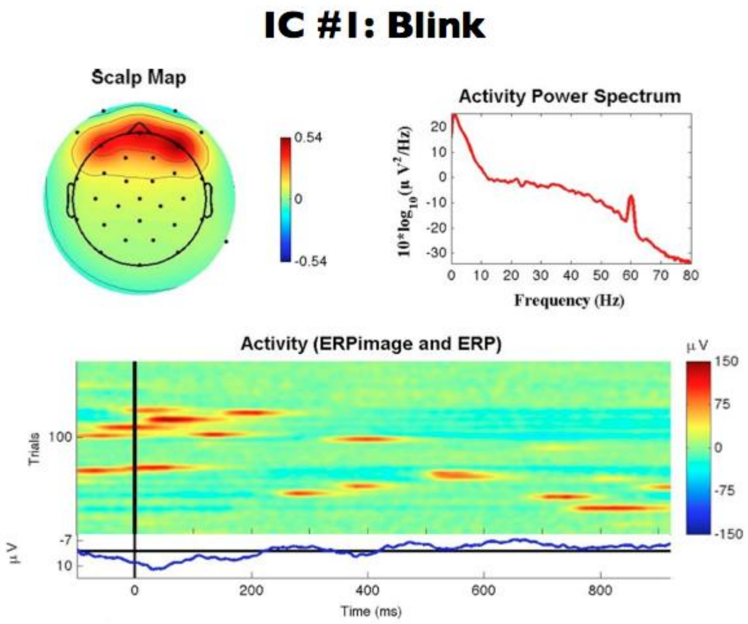
Ylipaavalniemi & Vigario (2008) NeuroImage

image: Gael Varoqaux

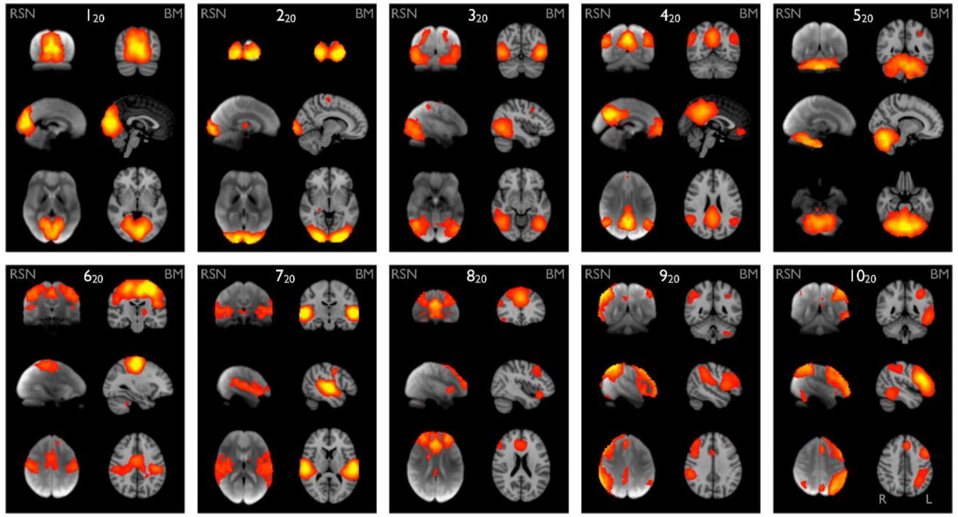
ICA in practice



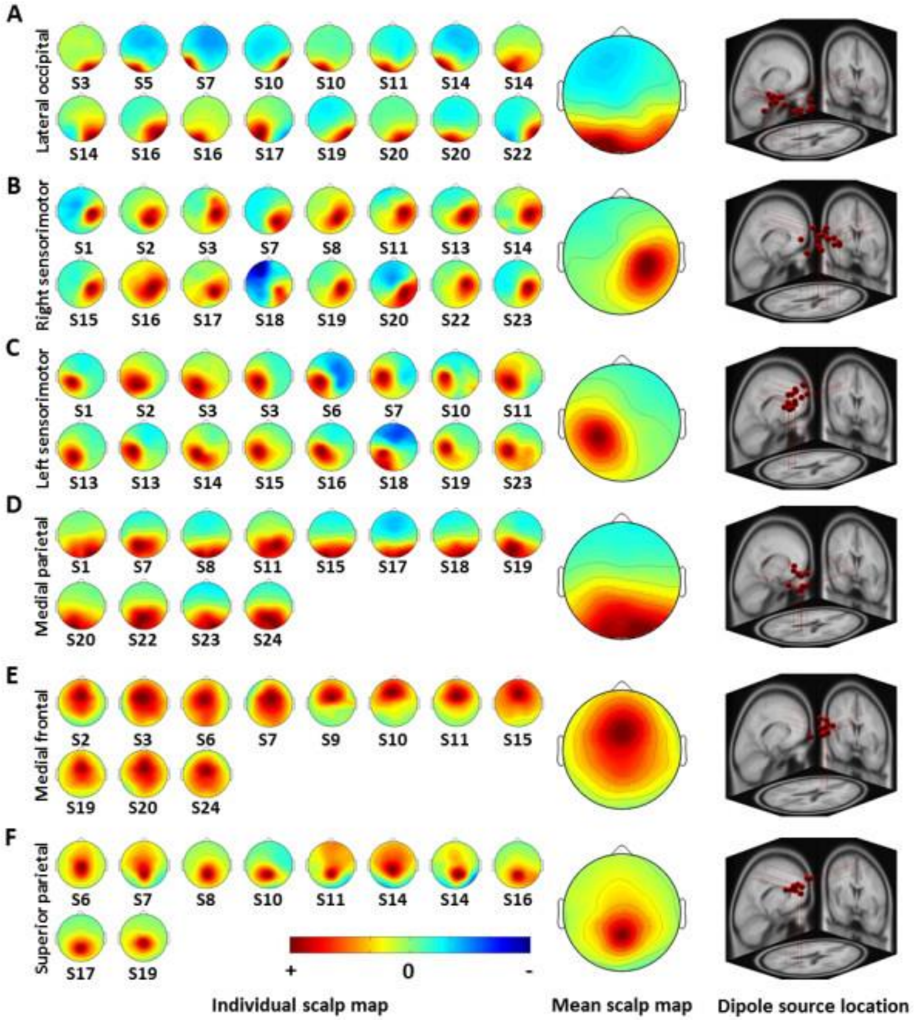


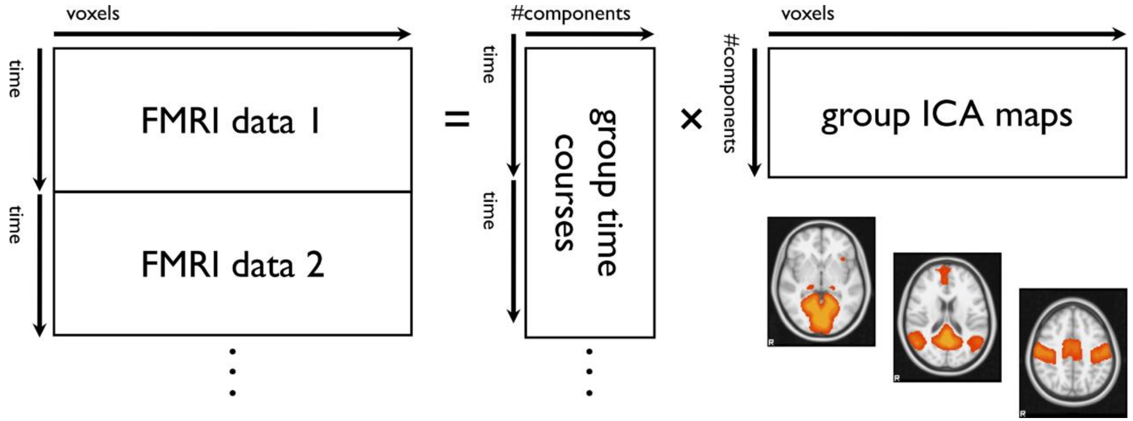


- Considerations:
  - Model selection
  - Starting point
  - Whitening
  - Algorithm
  - Subject- vs group-level
  - Spatial vs temporal independence
- Used for both preprocessing and for discovery

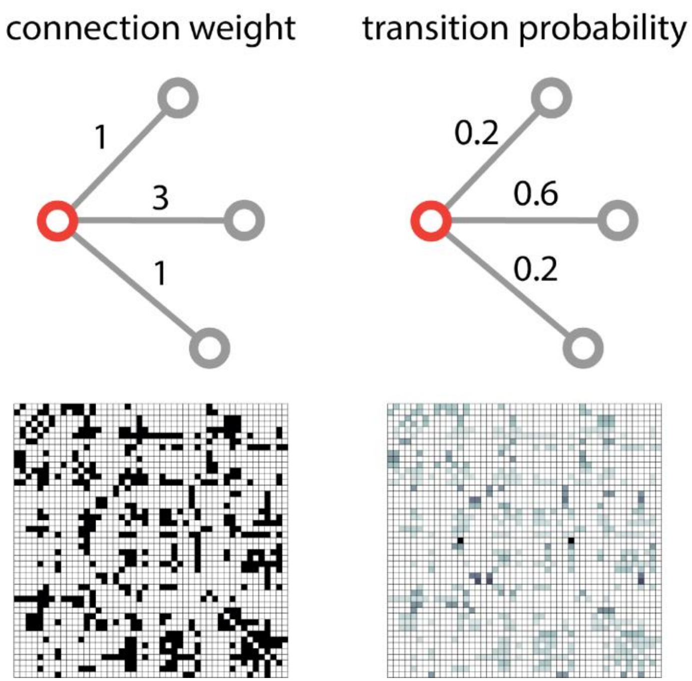


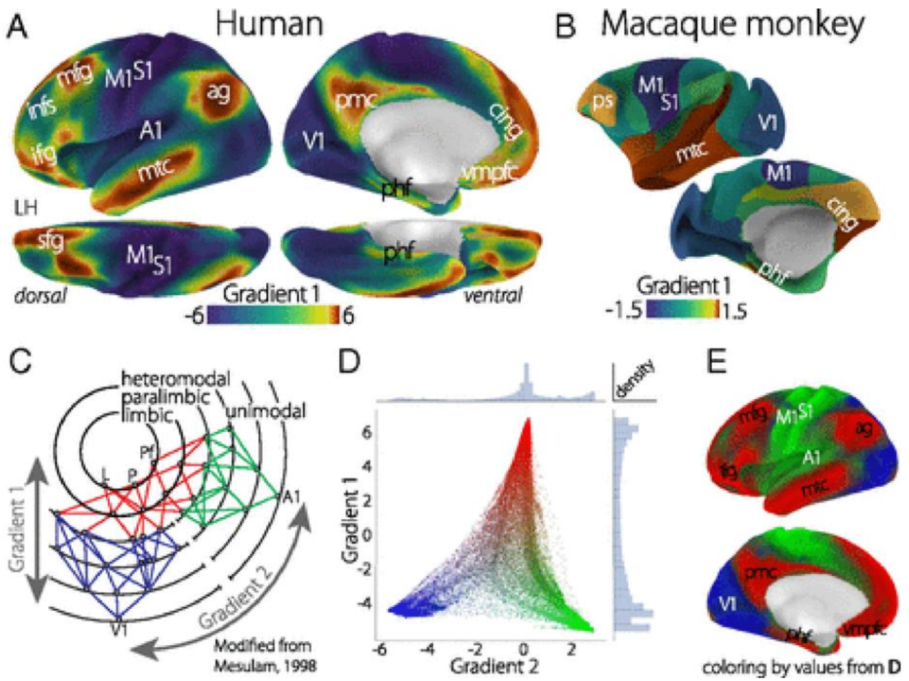
ICA: subject- vs group-level


A
Lateral occipital
S3 S5 S7 S10 S10 S11 S14 S14
S14 S16 S16 S17 S19 S20 S20 S22
B
Right sensorimotor
S1 S2 S3 S7 S8 S11 S13 S14
S15 S16 S17 S18 S19 S20 S22 S23
C
Left sensorimotor
S1 S2 S3 S3 S6 S7 S10 S11
S13 S13 S14 S15 S16 S18 S19 S23
D
Medial parietal
S1 S7 S8 S11 S15 S17 S18 S19
S20 S22 S23 S24
E
Medial frontal
S2 S3 S6 S7 S9 S10 S11 S15
S19 S20 S24
F
Superior parietal
S6 S7 S8 S10 S11 S14 S14 S16
S17 S19
+
0
-
Individual scalp map
Mean scalp map
Dipole source location



Diffusion embedding



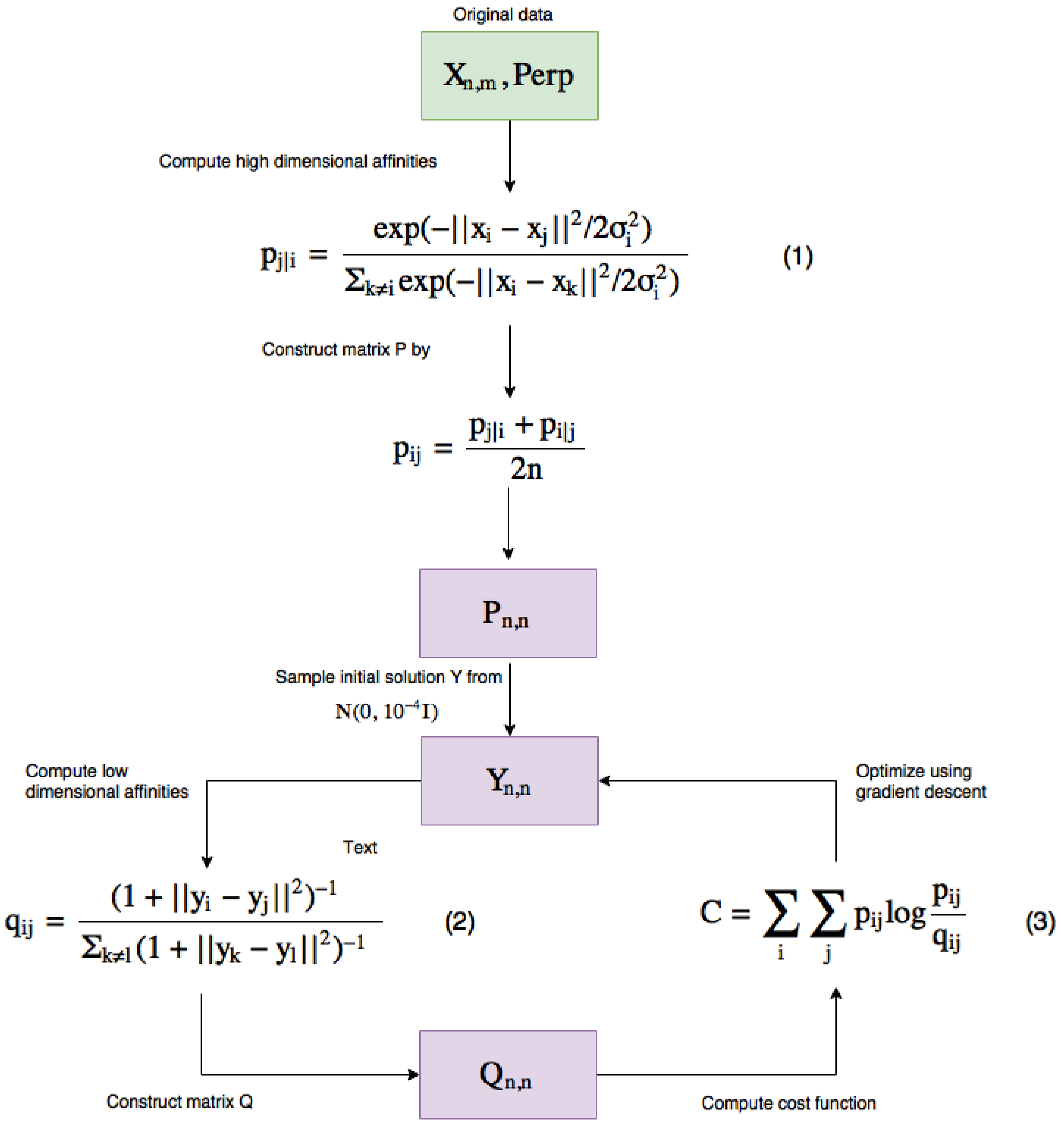


1. Convert correlation matrix to a transition probability matrix
2. Compute diffusion operator (Laplacian):
3. Get eigenvectors and eigenvalues of Laplacian

- The resulting components reflect dominant modes of diffusion
- Points (e.g. voxels) are grouped by how close or accessible they are to each other by a random walker

Margulies et al. (2016) PNAS

t-SNE



Original data

$X_{n,m}$, Perp

Compute high dimensional affinities

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\Sigma_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)} \qquad (1)$$

Construct matrix P by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

$P_{n,n}$

Sample initial solution Y from $N(0, 10^{-4}I)$

$Y_{n,n}$

Compute low dimensional affinities

Text

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\Sigma_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} \qquad (2)$$

Construct matrix Q

$Q_{n,n}$

Compute cost function

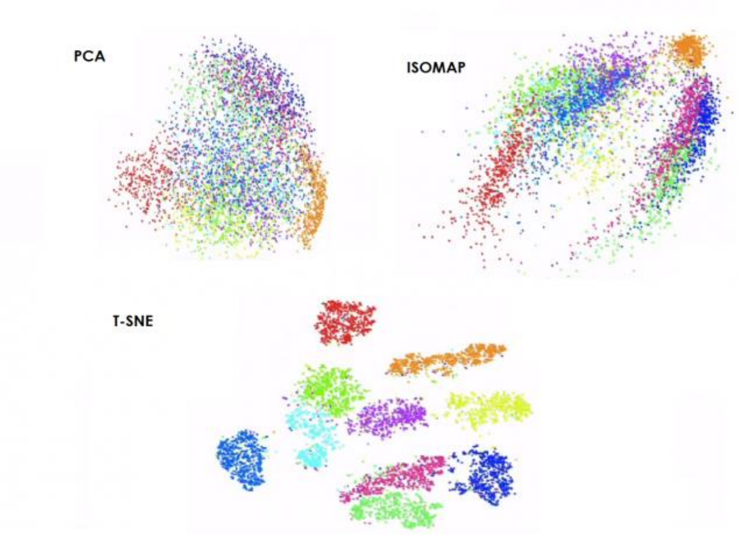
$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (3)$$

Optimize using gradient descent

- t-distributed stochastic neighbour embedding: visualizing data in 2 or 3 dimensions

- Input: data X and perplexity σ

1. Estimate high-dimensional distance P
2. Estimate low-dimensional distance Q
3. Estimate distance between distances



van der Maaten & Hinton (2008) J Mach Learn Res

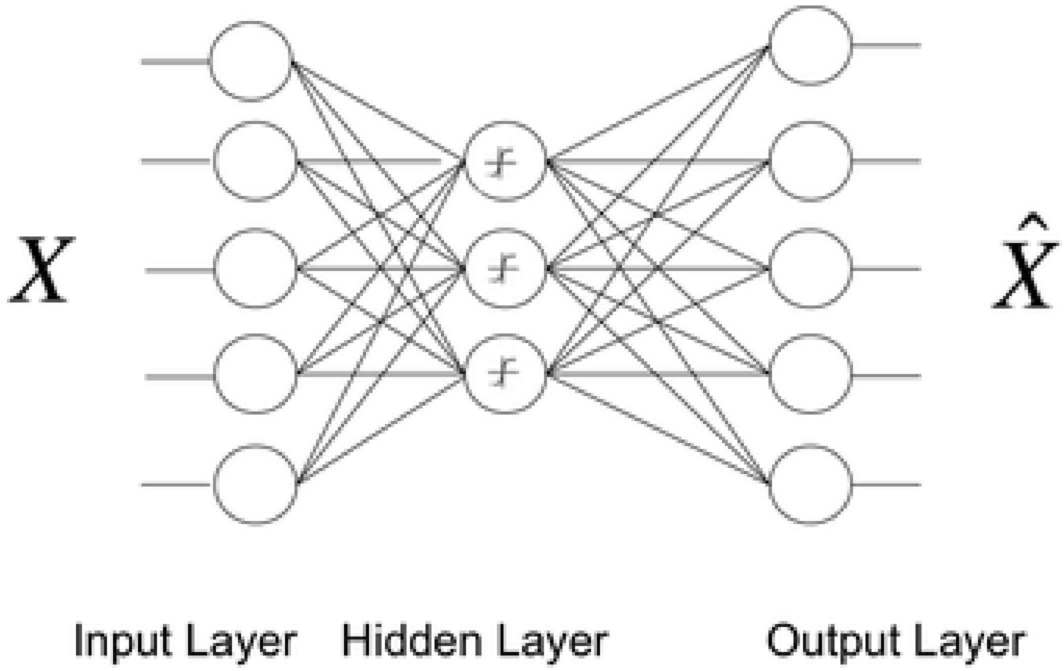
Autoencoders

- Idea: input and output layers have the same number of nodes (i.e. equal to the number of variables), but hidden layers with fewer nodes

- The neural network has to efficiently encode and then reconstruct the input

- Typically feedforward, non-recurrent

- If activations are linear, autoencoders will approximate SVD (Bourlard & Kemp, 1988)

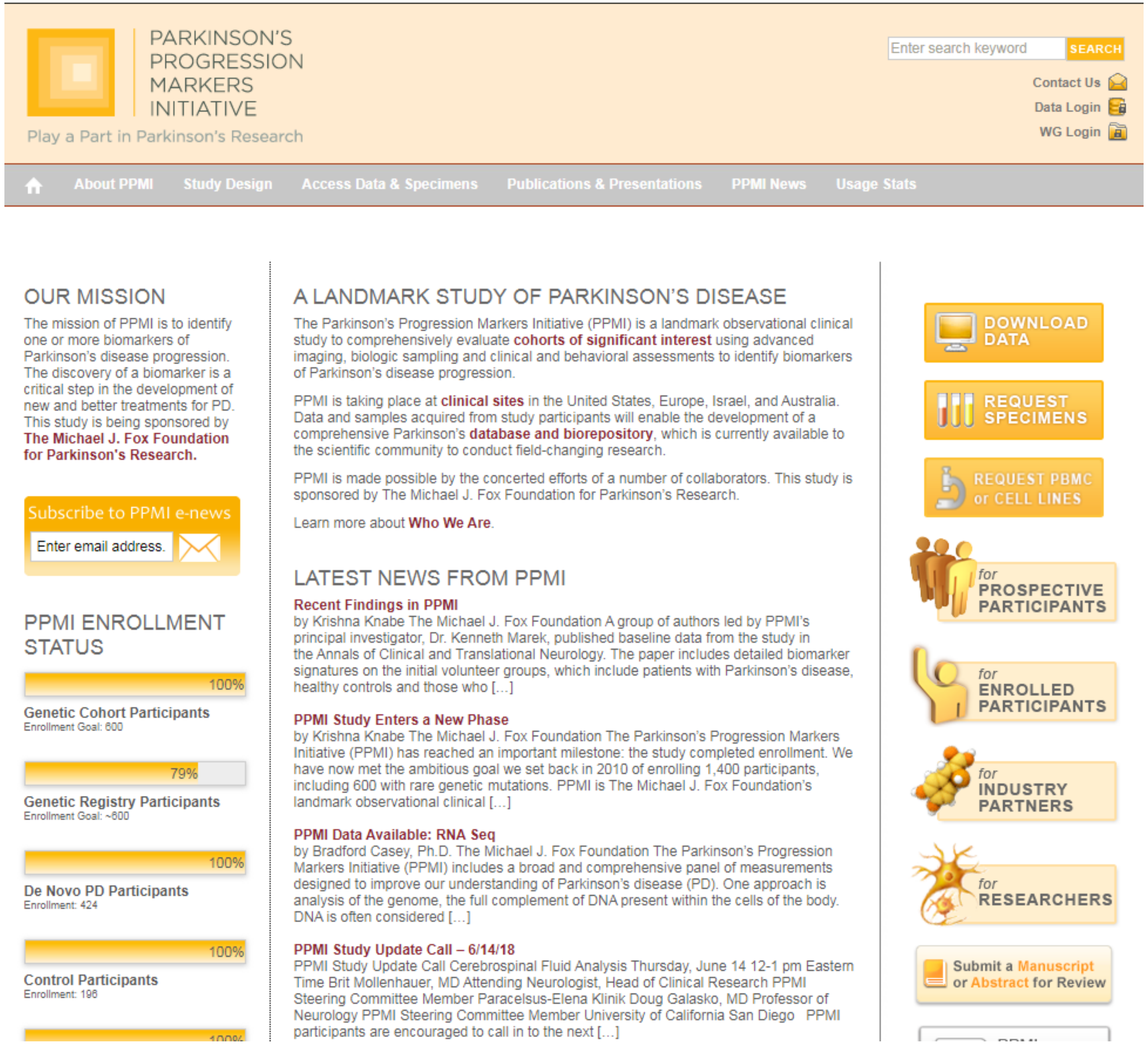


Limitations and Considerations

- Overfitting
- Linear or nonlinear?
- Identifiable?
- Unique partitioning of variance
- Inference on individual variables

Demo: PPMI

- Parkinson’s Progression Markers Initiative (http://www.ppmi-info.org)
- Deep phenotyping in patients with Parkinson’s disease
- n = 232 *de novo* patients
- p = 31 clinical, demographic and physiological variables
- Question: can we feasibly reduce this 31-dimensional dataset? Can we find dominant latent patterns?



OUR MISSION

The mission of PPMI is to identify one or more biomarkers of Parkinson’s disease progression. The discovery of a biomarker is a critical step in the development of new and better treatments for PD. This study is being sponsored by **The Michael J. Fox Foundation for Parkinson's Research.**

Subscribe to PPMI e-news

Enter email address.

PPMI ENROLLMENT STATUS

100%
Genetic Cohort Participants
Enrollment Goal: 600

79%
Genetic Registry Participants
Enrollment Goal: ~600

100%
De Novo PD Participants
Enrollment: 424

100%
Control Participants
Enrollment: 196

A LANDMARK STUDY OF PARKINSON’S DISEASE

The Parkinson’s Progression Markers Initiative (PPMI) is a landmark observational clinical study to comprehensively evaluate **cohorts of significant interest** using advanced imaging, biologic sampling and clinical and behavioral assessments to identify biomarkers of Parkinson’s disease progression.

PPMI is taking place at **clinical sites** in the United States, Europe, Israel, and Australia. Data and samples acquired from study participants will enable the development of a comprehensive Parkinson’s **database and biorepository**, which is currently available to the scientific community to conduct field-changing research.

PPMI is made possible by the concerted efforts of a number of collaborators. This study is sponsored by The Michael J. Fox Foundation for Parkinson’s Research.

Learn more about **Who We Are**.

LATEST NEWS FROM PPMI

Recent Findings in PPMI
by Krishna Knabe The Michael J. Fox Foundation A group of authors led by PPMI’s principal investigator, Dr. Kenneth Marek, published baseline data from the study in the Annals of Clinical and Translational Neurology. The paper includes detailed biomarker signatures on the initial volunteer groups, which include patients with Parkinson’s disease, healthy controls and those who […]

PPMI Study Enters a New Phase
by Krishna Knabe The Michael J. Fox Foundation The Parkinson’s Progression Markers Initiative (PPMI) has reached an important milestone: the study completed enrollment. We have now met the ambitious goal we set back in 2010 of enrolling 1,400 participants, including 600 with rare genetic mutations. PPMI is The Michael J. Fox Foundation’s landmark observational clinical […]

PPMI Data Available: RNA Seq
by Bradford Casey, Ph.D. The Michael J. Fox Foundation The Parkinson’s Progression Markers Initiative (PPMI) includes a broad and comprehensive panel of measurements designed to improve our understanding of Parkinson’s disease (PD). One approach is analysis of the genome, the full complement of DNA present within the cells of the body. DNA is often considered […]

PPMI Study Update Call – 6/14/18
PPMI Study Update Call Cerebrospinal Fluid Analysis Thursday, June 14 12-1 pm Eastern Time Brit Mollenhauer, MD Attending Neurologist, Head of Clinical Research PPMI Steering Committee Member Paracelsus-Elena Klinik Doug Galasko, MD Professor of Neurology PPMI Steering Committee Member University of California San Diego PPMI participants are encouraged to call in to the next […]

Marek et al. (2011) Prog Neurobiol