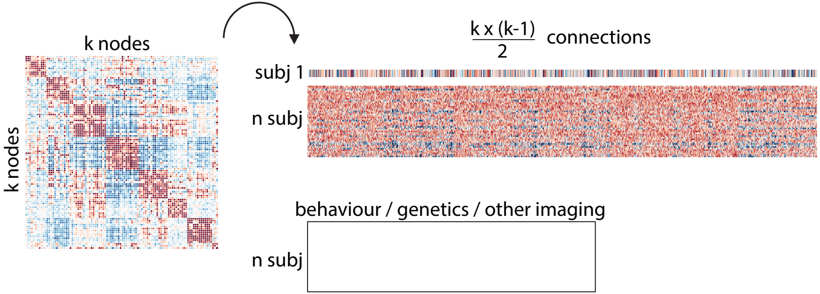
Week 3: dimensionality reduction

Bratislav Mišić

Montréal Neurological Institute

September 29^{th} 2017

Why multivariate statistics?



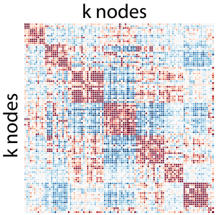
1. how to operationalize network property?
2. how to deal with with more variables than observations?
3. how to relate multiple data sets to one another?

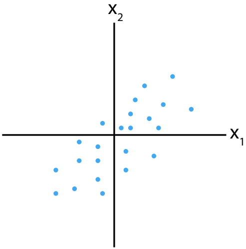
Dimensionality reduction

1. linear: PCA, FA, ICA, MDS
2. nonlinear: kernel PCA, LLE, diffusion maps, t-SNE, autoencoders

- how is similarity represented?
- is there an underlying generative model?
- inference? reliability?

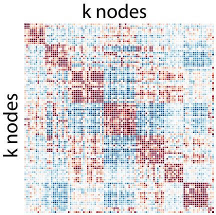
Principal component analysis (PCA)

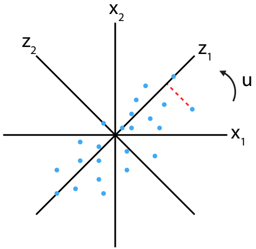




Hotelling (1933) *J Educ Psychol*

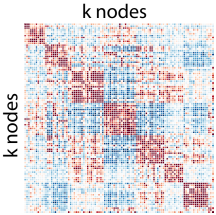
Principal component analysis (PCA)

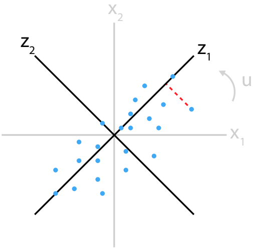




Hotelling (1933) *J Educ Psychol*

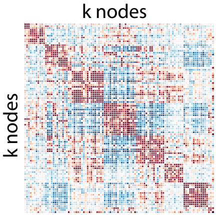
Principal component analysis (PCA)

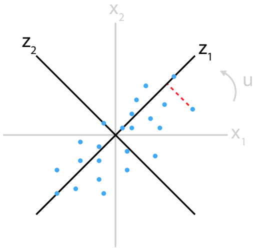


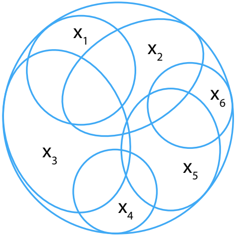


Hotelling (1933) *J Educ Psychol*

Principal component analysis (PCA)

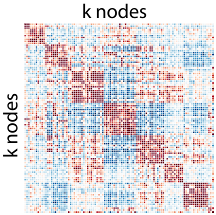


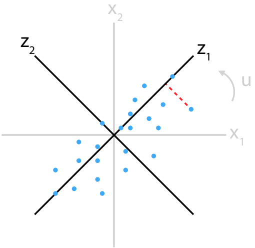


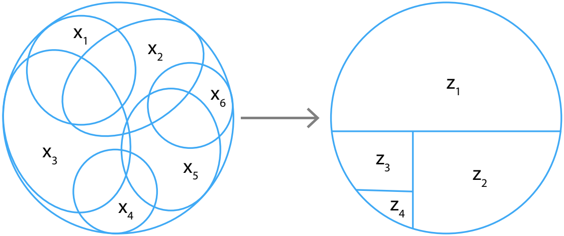


Hotelling (1933) *J Educ Psychol*

Principal component analysis (PCA)







Hotelling (1933) *J Educ Psychol*

Maximizing variance

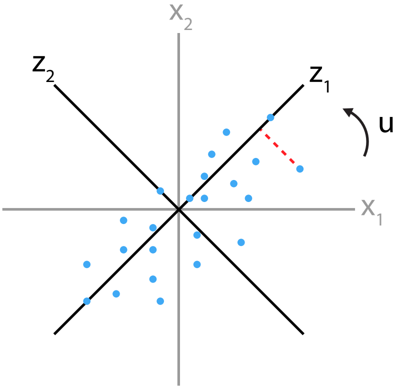
find new variable $\mathbf{z} = \mathbf{X}\mathbf{u}$

choose $\mathbf{u}$ to maximize $var(\mathbf{z})$

under the constraint $\mathbf{u}'\mathbf{u} = 1$

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

since $\mathbf{R} = \frac{1}{n-1}\mathbf{X}'\mathbf{X}$



Maximizing variance

$L = \mathbf{u}'\mathbf{R}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1)$

$\frac{\partial L}{\partial \mathbf{u}} = 2\mathbf{R}\mathbf{u} - 2\mathbf{u}\lambda = 0$

$\mathbf{R}\mathbf{u} = \mathbf{u}\lambda$

$(\mathbf{R} - \lambda\mathbf{I})\mathbf{u} = 0$

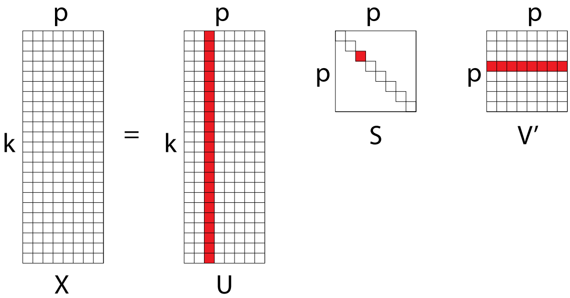
eigenvalue λ (variance) & eigenvector $\mathbf{u}$ (weights)

$var(\mathbf{z}) = \mathbf{u}'\mathbf{R}\mathbf{u} = \mathbf{u}'\mathbf{u}\lambda = \lambda$

Singular value decomposition

Spectral decomposition:
$\mathrm{EIG}(\mathbf{X}'\mathbf{X}) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$
$\mathrm{EIG}(\mathbf{X}\mathbf{X}') = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$

Singular value decomposition:
$\mathrm{SVD}(\mathbf{X}) = \mathbf{U}\mathbf{S}\mathbf{V}'$

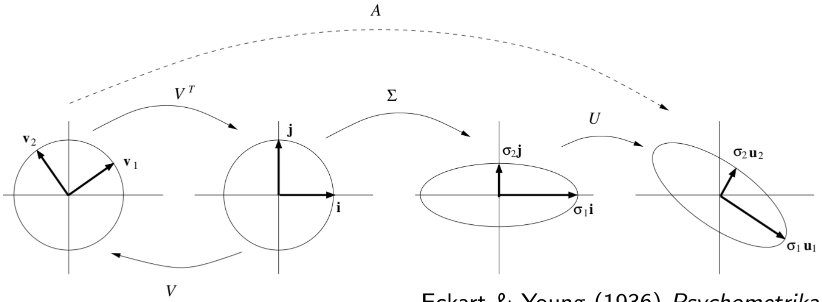


Eckart & Young (1936) *Psychometrika*

Singular value decomposition

Spectral decomposition:
$\mathrm{EIG}(\mathbf{X}'\mathbf{X}) = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$
$\mathrm{EIG}(\mathbf{X}\mathbf{X}') = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$

Singular value decomposition:
$\mathrm{SVD}(\mathbf{X}) = \mathbf{U}\mathbf{S}\mathbf{V}'$



Eckart & Young (1936) *Psychometrika*

EIG vs SVD

SVD: $\mathbf{X} = \mathbf{USV}'$

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

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

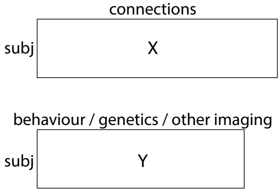
1. eigenvector of $(\mathbf{XX}')$ = left singular vector $(\mathbf{U})$
2. eigenvector of $(\mathbf{X}'\mathbf{X})$ = right singular vector $(\mathbf{V})$
3. eigenvalue = squared singular value

A family of techniques

PCA: SVD($\mathbf{X}$)

PLS: SVD($\mathbf{X}'\mathbf{Y}$)

CCA: SVD($(\mathbf{X}'\mathbf{X}')^{-\mathbf{1/2}}(\mathbf{X}'\mathbf{Y})(\mathbf{Y}'\mathbf{Y})^{-\mathbf{1/2}}$)

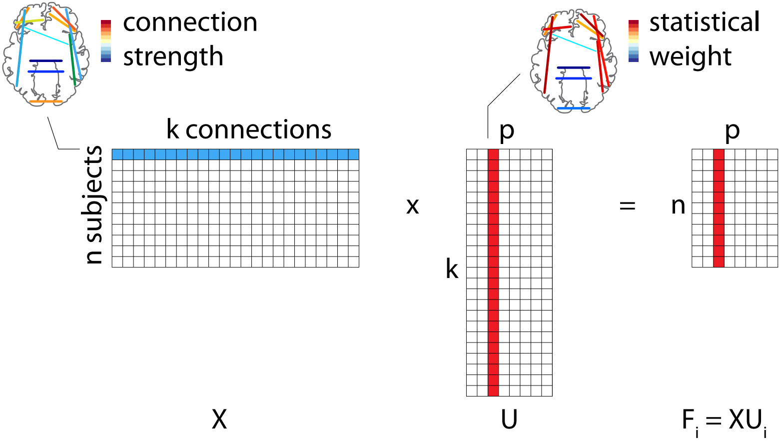


Worsley et al. (1997) *NeuroImage*

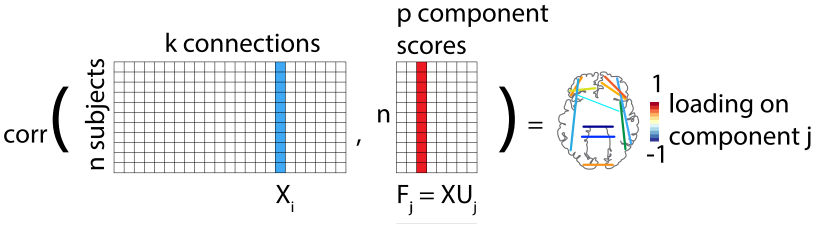
Tijl et al. (2005) *Handbook of Geometric Computing - Springer*

McIntosh & Mišić (2013) *Annu Rev Psychol*

Individual participants

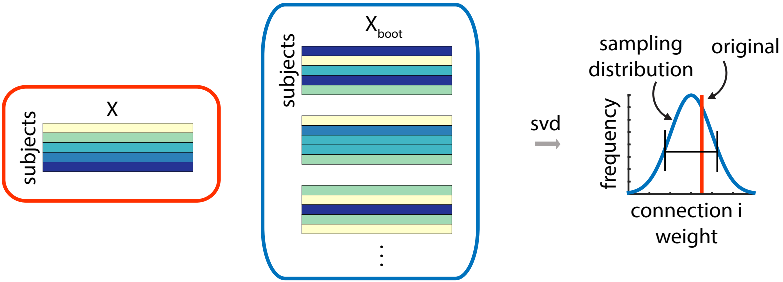


Which variables are important?



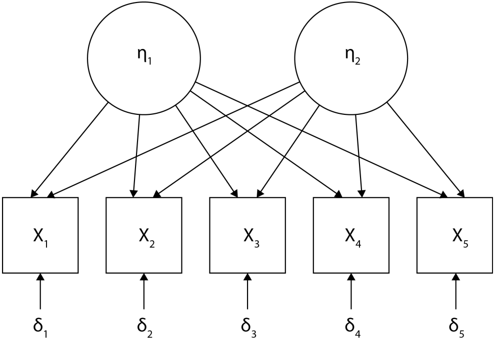
1. eigenvector weights
2. loadings
3. bootstrap resampling

Which variables are important?



1. eigenvector weights
2. loadings
3. bootstrap resampling

Factor analysis (FA)



- assume that variance in variable X_i comes from a set of latent factors η_i and measurement error δ_i, i.e.

$$X_1 = \lambda_{1,1}\eta_1 + \lambda_{1,2}\eta_2 + \delta_1$$
$$X_2 = \lambda_{2,1}\eta_1 + \lambda_{2,2}\eta_2 + \delta_2$$
$$\ldots$$

Factor analysis (FA)

- unique factors (δ_i) are uncorrelated, so they only contribute to diagonals of the covariance matrix

$$\mathrm{var}(X_i) = \mathrm{var}(\lambda_{i,1}\eta_1 + \lambda_{i,2}\eta_2 + \delta_i) \tag{1}$$
$$\mathrm{var}(X_i) = \lambda_{i,1}^2 + \lambda_{i,2}^2 + \theta_i^2, \text{where } \theta_i^2 = \mathrm{var}(\delta_i) \tag{2}$$

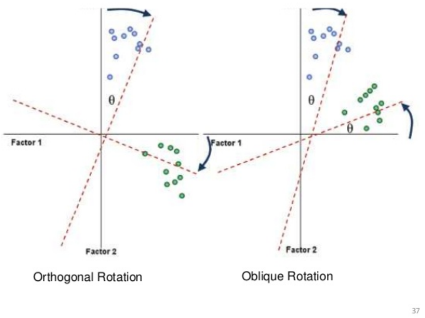
- so, estimate θ_i^2, subtract them and decompose the matrix
- PCA: decompose correlation matrix **R**, with 1's on diagonal
- FA: decompose matrix with diagonal elements $1 - \theta_i^2$

Factor analysis (FA)

- how to estimate $1 - \theta_i^2$ (communalities)?
- option #1: guess and re-adjust
- option #2: estimate squared multiple correlation (SMC)
  e.g. regress X_i on all $X_{j \neq i}$

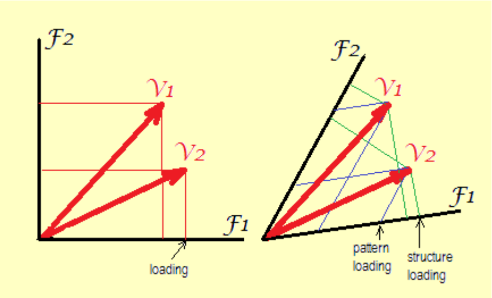
1. $\mathbf{R}_{adj}$
2. EIG($\mathbf{R}_{adj}$)
3. communalities stable?
   no = recalculate $\mathbf{R}_{adj}$
   yes = stop

Rotating factors



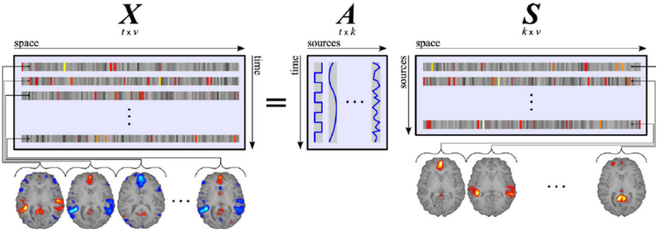
- unlike PCA, FA has no rotational indeterminacy
- we want to improve interpretability: each variable has high loadings only on a few factors, and near zero on all others
- orthogonal: factors remain uncorrelated
- oblique: factors may correlate

Rotating factors



- unlike PCA, FA has no rotational indeterminacy
- we want to improve interpretability: each variable has high loadings only on a few factors, and near zero on all others
- orthogonal: factors remain uncorrelated
- oblique: factors may correlate

Independent component analysis (ICA)



- assumption: there exist a finite number of sources (independent components) that are mixed to give the observed variables
- any linear mixture of independent variables (e.g. voxels) will be more Gaussian than the original variables
- solution: create new axes with maximally non-Gaussian projections

Ylipaavalniemi & Vigário (2008) *NeuroImage*

Independent component analysis (ICA)

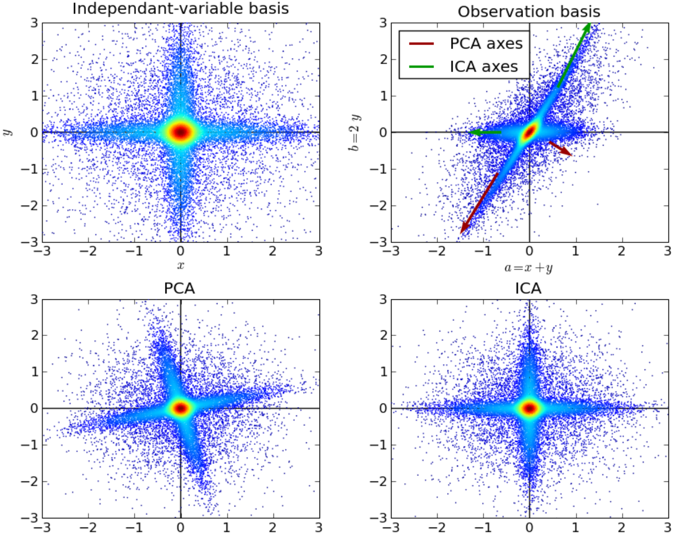
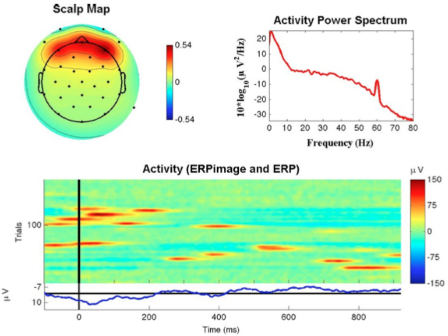


image: Gaël Varoqaux

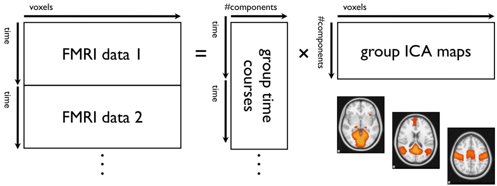
ICA in practice

IC #1: Blink



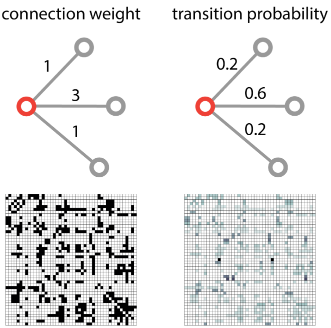
- model selection
- starting point
- whitening
- algorithm
- subject vs group

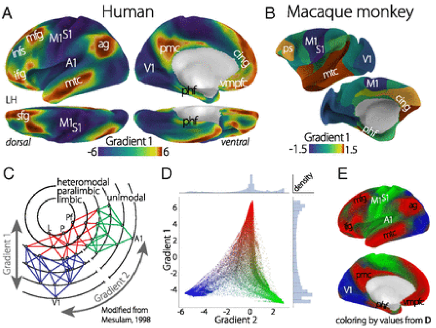
ICA in practice



- model selection
- starting point
- whitening
- algorithm
- subject vs group

Diffusion maps





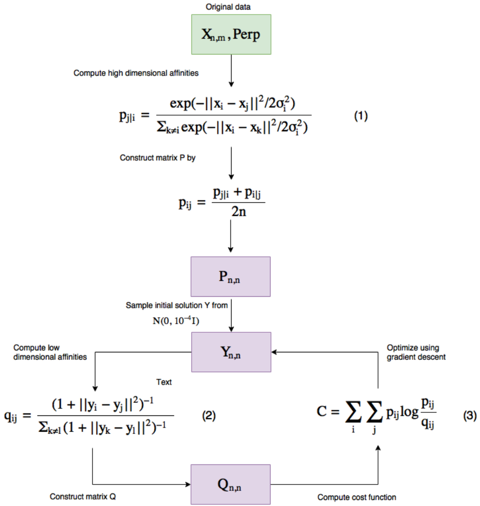
1. convert correlation matrix to a transition probability matrix
2. compute diffusion operator (Laplacian matrix), i.e.

$$L = S - W = \begin{cases} s_i & i = j \\ -w_{ij} & i \neq j \end{cases}$$

3. get eigenvalues and eigenvectors of L

Margulies et al. (2016) *PNAS*

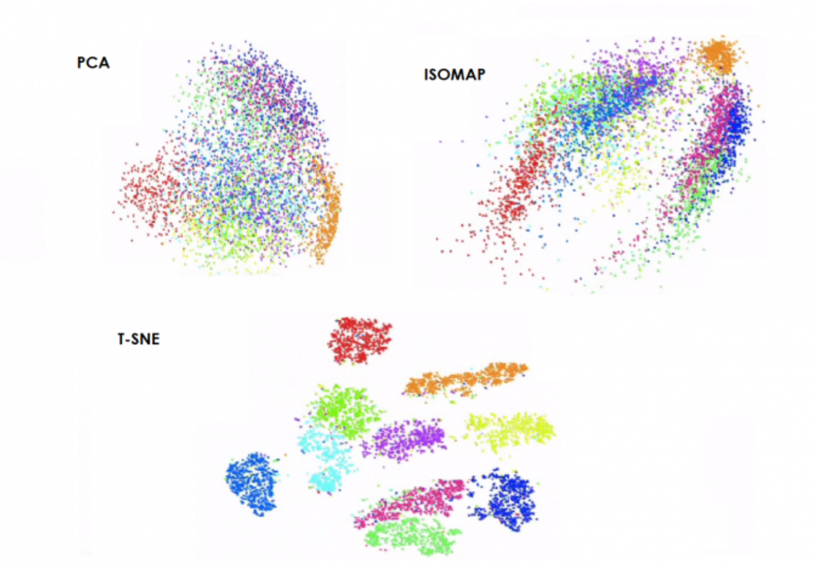
t-SNE



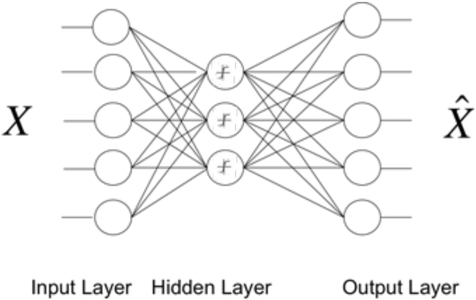
- t-distributed stochastic neighbor embedding
- visualizing high-dimensional data in 2 or 3 dimensions
- input: data X and perplexity σ_i
- (1) high-dimensional distance P
- (2) low-dimensional distance Q
- (3) distance between distances

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

t-SNE



Autoencoders



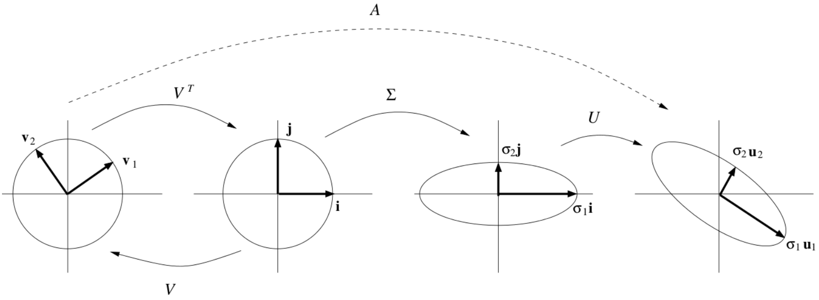
- idea: input and output layers have same no. of nodes, but hidden layer has fewer nodes
- so you have to efficiently encode and reconstruct the input
- typically feedforward, non-recurrent
- if activations are linear, approximates SVD (Bourlard & Kamp, 1988; *Biol Cybern*)

Limitations and considerations

- overfitting
- linear or nonlinear?
- identifiable?
- unique partitioning of variance/covariance
- inference on individual variables

Matching randomized components

$$\mathbf{X}'\mathbf{Y} = \mathbf{USV}' \longleftrightarrow \mathbf{X}'_{\mathrm{boot}}\mathbf{Y} = \mathbf{U}_{\mathrm{boot}}\mathbf{S}_{\mathrm{boot}}\mathbf{V}'_{\mathrm{boot}}$$



Milan & Whittaker (1995) *J Roy Stat Soc C*