Algorithms in Nature

Non-negative matrix factorization

Slides adapted from Marshall Tappen and Bryan Russell

Dimensionality Reduction

The curse of dimensionality:

Too many features makes it difficult to visualize and interpret data

Harder to efficiently learn robust statistical models

Problem statement: Given a set of images..

1. Create basis images that can be linearly combined to reconstruct the original (or new) images

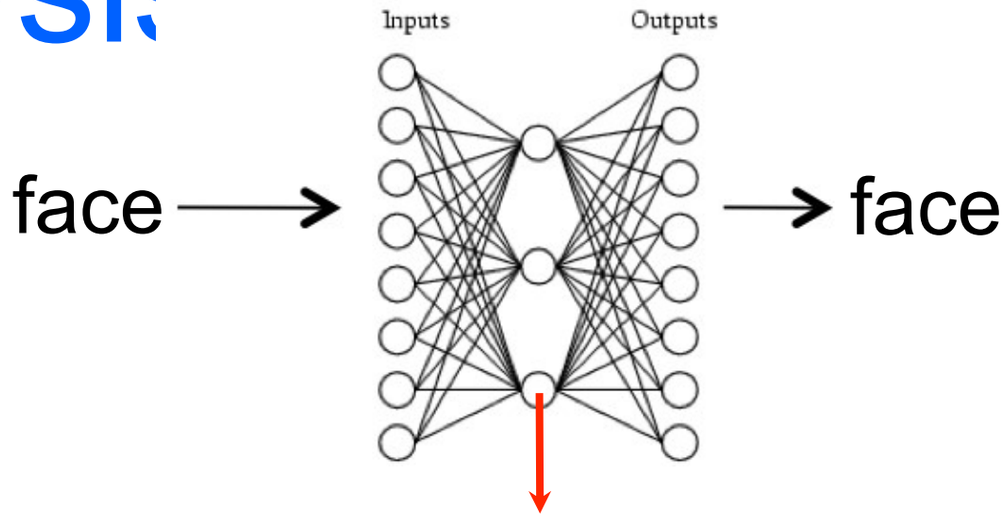
2. Find weights to reproduce every input image from the basis images

One set of weights for each input image

Principal Components Analysis

A low-dimensionality representation that minimizes reconstruction error

$$\text{face}_i = \sum_k c_{ik}\ \text{eigenface}_k$$



“eigenfaces”

PCA weaknesses

- Only allows *linear* projections
- Co-variance matrix is of size dxd. If $d=10^4$, then $|\Sigma| = 10^8$
  - *Solution*: singular value decomposition (SVD)
- PCA restricts to *orthogonal* vectors in feature space that minimize reconstruction error
  - *Solution*: independent component analysis (ICA) seeks directions that are *statistically independent,* often measured using information theory
- Assumes points are multivariate Gaussian
  - *Solution*: Kernel PCA that transforms input data to other spaces

PCA vs. Neural Networks

PCA	Neural Networks
Unsupervised dimensionality reduction	Supervised dimensionality reduction
Linear representation that gives best squared error fit	Non-linear representation that gives best squared error fit
No local minima (exact)	Possible local minima (gradient descent)
Non-iterative	Iterative
Orthogonal vectors (“eigenfaces”)	Auto-encoding NN with linear units may not yield orthogonal vectors

Is this really how humans characterize and identify faces?

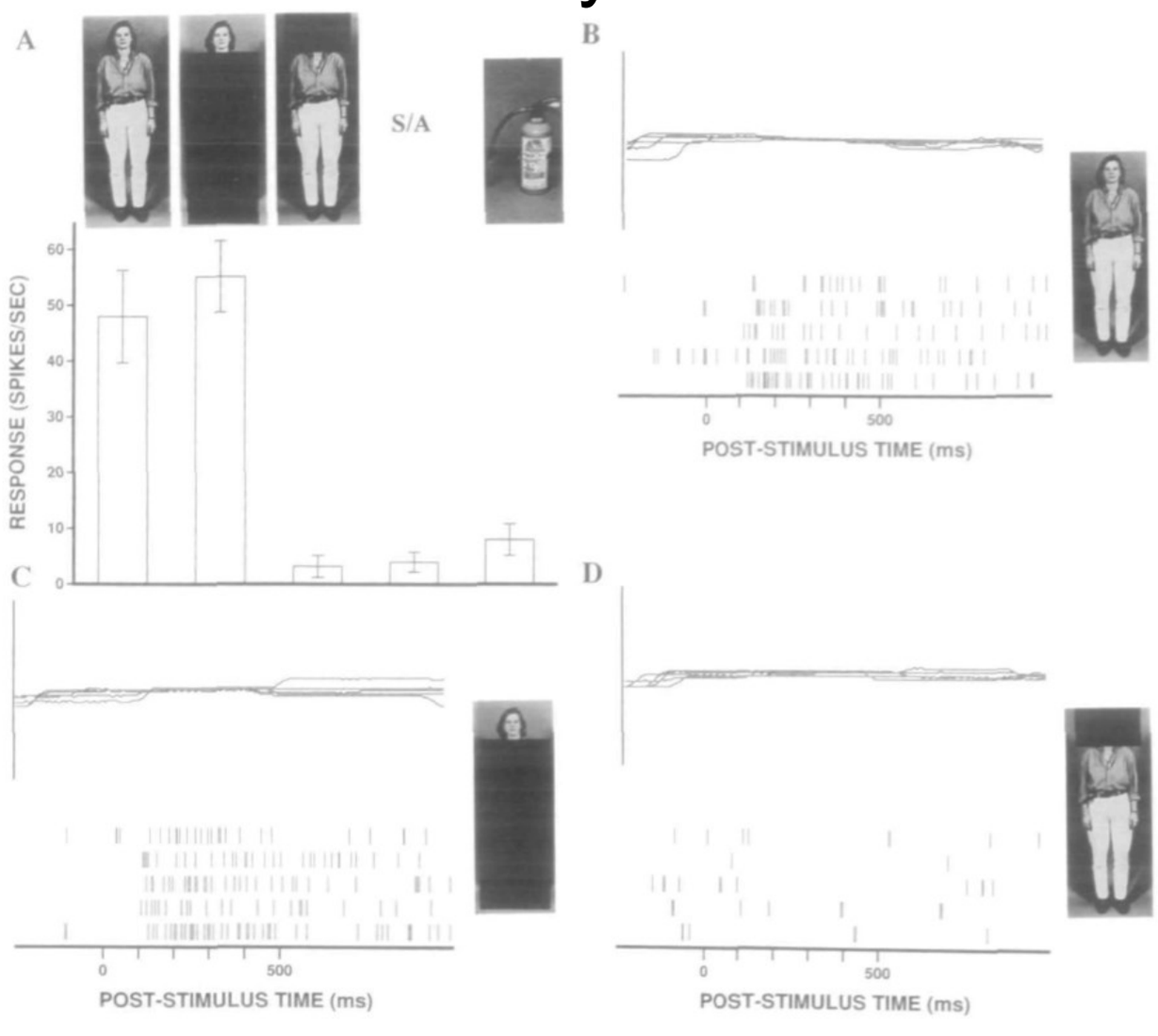
What don't we like about PCA?

- Basis images aren't physically intuitive
  - Humans can *explain* why a face is a face
- PCA involves adding up some basis images and subtracting others which may not make sense in some applications:
  - What does it mean to subtract a face? A document?

Going from the whole to parts..

[Wachsmuth et al. 1994]

Recording from neurons in the temporal lobe in the macaque monkey



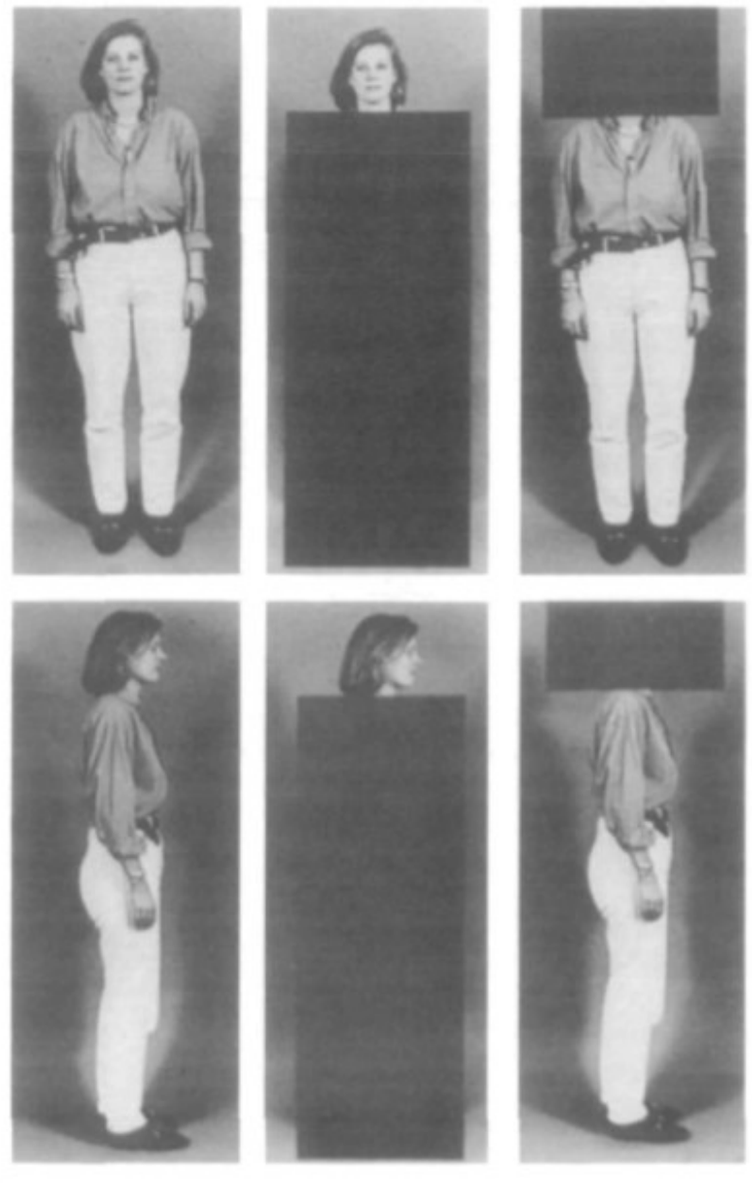
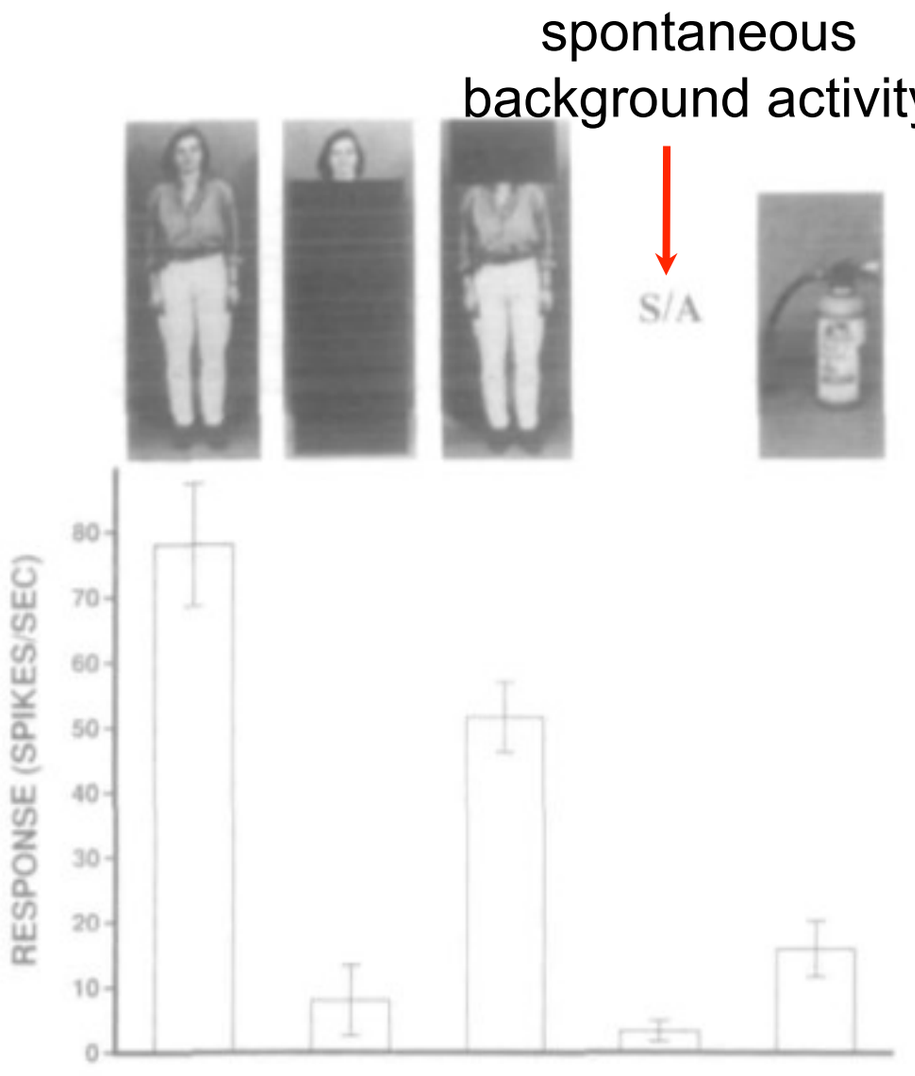


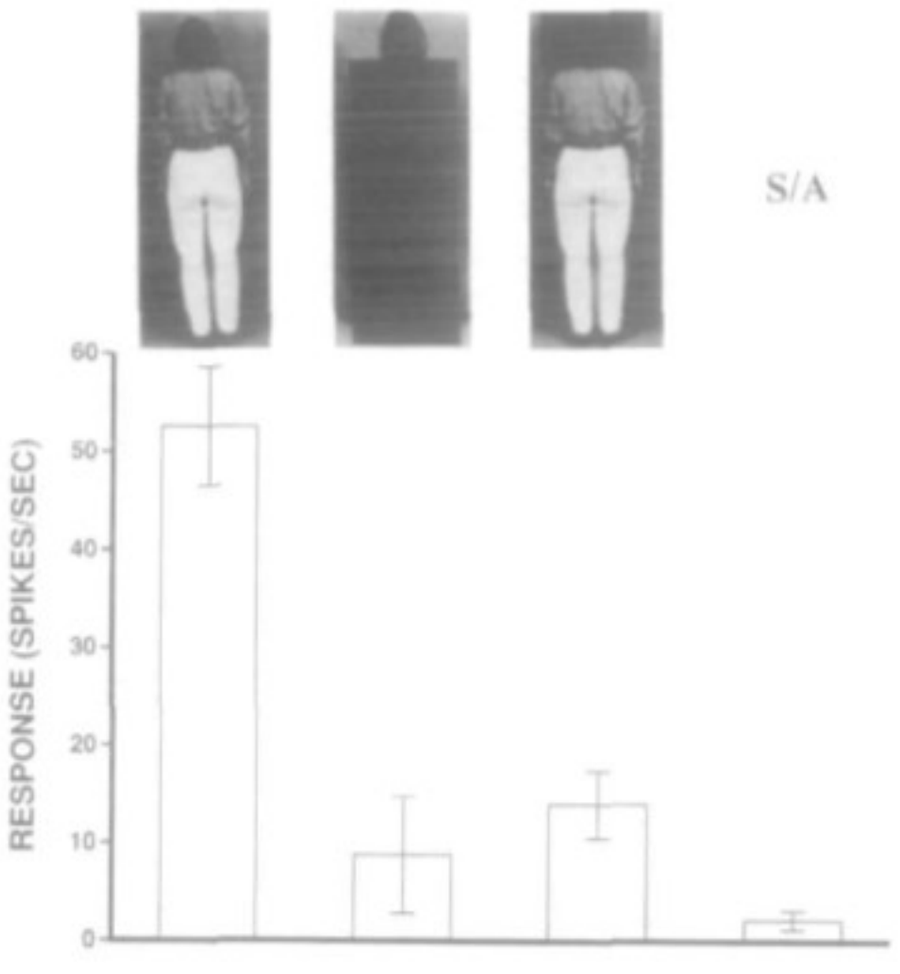
Figure 1. Examples of stimuli used for testing: whole-body, head-only, and body-only stimuli in different views.

Going from the whole to parts..

[Wachsmuth et al. 1994]

Neurons that respond primarily to the body





Going from the whole to parts..

[Wachsmuth et al. 1994]

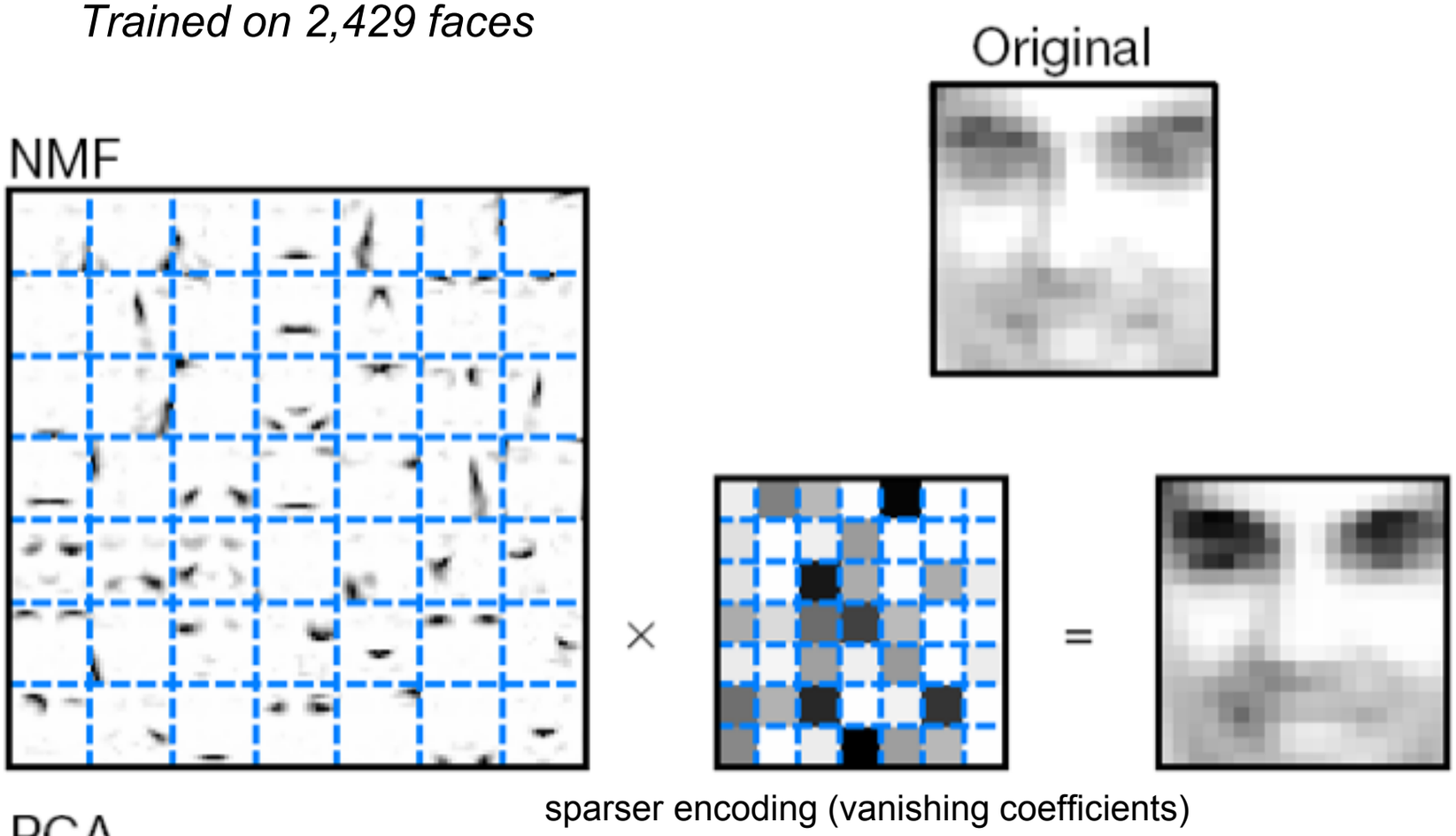
Overall, recorded from 53 neurons:

17 (32%) responded to the head only
5 (9%) responded to the body only
22 (41%) responded to both the head and the body in isolation
9 (17%) responded to the whole body *only (*neither part in isolation)

Suggestive of a *parts-based* (Today) representation with possible *hierarchy*

Non-negative matrix factorization

Trained on 2,429 faces



Like PCA, except the coefficients in the linear combination must be *non-negative*

Forcing positive coefficients implies an additive combination of basis parts to reconstruct whole

Several versions of mouths, noses, etc.

Better physical analogue in neurons

Formal definition of NMF

$$V_{iu} \approx (WH)_{iu} = \sum_{a=1}^{r} W_{ia} H_{au}$$

V_{iu}: n×m matrix of image database. n= # pixels/face; m= # faces

W: n×r matrix; r columns are the basis images, each of size n; "eigenfaces"

H: r×m matrix; r coefficients to represent each of the m faces

$\text{subject to } W, H \geq 0$: non-negativity constraints

WH is a compressed version of V

How to choose the rank r? Want (n+m)r < nm

A similar neural network view

$$V_{iu} \approx (WH)_{iu} = \sum_{a=1}^{r} W_{ia} H_{au}$$

n×m matrix; input image database. n=# of pixels/face; m = # of faces

n×r matrix; r columns are the basis images, each of size n

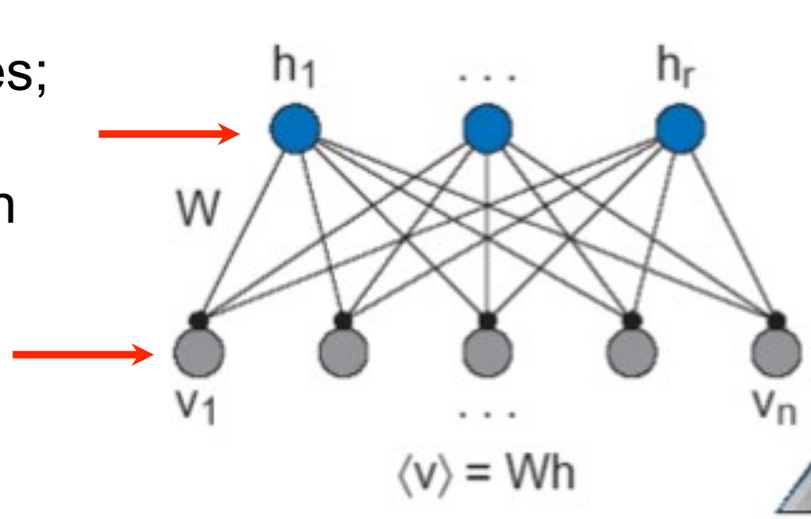
r×m matrix; r coefficients to represent each of the m faces

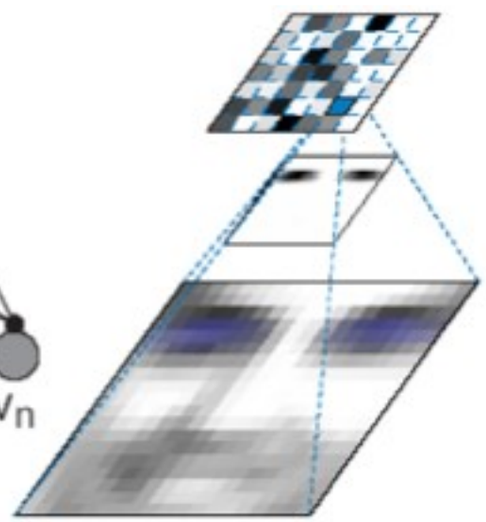
$$\text{subject to } W, H \geq 0$$

non-negativity constraints

hidden variables; parts-based representation

original image pixels





$$V_{iu} \approx (WH)_{iu} = \sum_{a=1}^{r} W_{ia} H_{au}$$

One possible objective function

Reconstruction error:

$$\arg\min_{W,H} E_r = \|V - WH\|^2, \text{s.t. } W, H \geq 0$$

Update rule:

$$H_{au} \leftarrow H_{au} \sum_i W_{ia} \frac{V_{iu}}{(WH)_{iu}}$$

- H_{au}: update a^{th} coefficient for the u^{th} face
- W_{ia}: a^{th} basis
- i: sum over all pixels
- V_{iu}: projection for i^{th} pixel
- $(WH)_{iu}$: ratio of actual to reconstructed pixel value for the u^{th} face

$$W_{ia} \leftarrow W_{ia} \sum_u \frac{V_{iu}}{(WH)_{iu}} H_{au}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

Normalize

$$\arg\min_{W,H} E_r = \|V - WH\|^2, \text{s.t. } W, H \geq 0$$

$$V_{iu} \approx (WH)_{iu} = \sum_{a=1}^{r} W_{ia} H_{au}$$

One possible objective function

Update rule:

$$H_{au} \leftarrow H_{au} \sum_i W_{ia} \frac{V_{iu}}{(WH)_{iu}}$$

- H_{au}: update a^{th} coefficient for the u^{th} face
- W_{ia}: a^{th} basis projection for i^{th} pixel
- $\sum_i$: sum over all pixels
- $\frac{V_{iu}}{(WH)_{iu}}$: ratio of actual to reconstructed pixel value for the u^{th} face

$$W_{ia} \leftarrow W_{ia} \sum_u \frac{V_{iu}}{(WH)_{iu}} H_{au}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

Normalize

Basic idea: multiply current value by a factor depending on the quality of the approximation.

If ratio > 1, then we need to increase denominator.

If ratio < 1, then we need to decrease denominator.

If ratio = 1, do nothing.

What is significant about this?

- The update rule is *multiplicative* instead of additive
  - In the initial values for W and H are non-negative, then W and H can never become negative
- This guarantees a non-negative factorization
- Will it converge?
  - Yes, to a local optima: see [Lee and Seung, NIPS 2000] for proof

PCA vs. NMF

PCA	NMF
Unsupervised dimensionality reduction	Unsupervised dimensionality reduction
Orthogonal vectors with positive and negative coefficients	Non-negative coefficients
“Holistic”; difficult to interpret	“Parts-based”; easier to interpret
Non-iterative	Iterative (the presented algorithm)
CS developed	Biologically-“inspired” (alas, there are inhibitory neurons in the brain)

The ‘Jennifer Aniston’ neuron

[Quiroga et al., Nature 2005]

- UCLA neurosurgeon Itzhak Fried and researcher Quian Quiroga operating on patients with epileptic seizures
- Procedure requires implanting a probe in the brain, but doctor first needs to map surgical area (fyi, open brains do not hurt)
- “Mind if I try some exploratory science?”
  - Flashed one-second snapshots of celebrities, animals, objects, and landmark buildings. Each person shown ~2,000 pictures.
- When Aniston was shown, one neuron in the medial temporal lobe always flashed
  - Invariant to: different poses, hair styles, smiling, not smiling, etc.
  - Never flashed for: Julia Roberts, Kobe Bryant, other celebrities, places, animals, etc.

Hierarchical models of object recognition

<u>Stirred a controversy</u>:

Are there ‘grandmother cells’ in the brain? [Lettvin, 1969]

Or are there populations of cells that respond to a stimuli?

Are the cells organized into a hierarchy? (Riesenhuber and Poggio model; see website)