CPSC 340: Machine Learning and Data Mining

Sparse Matrix Factorization

Fall 2015

Admin

- Assignment 2 grades posted.
- Midterm back soon.
- Assignment 4 out tomorrow.
- Tomorrow at 6pm is DataSense’s Data Science Seminar Series:
  - IBM Watson Analytics and Panel Discussion.
  - https://www.facebook.com/events/975146559243561

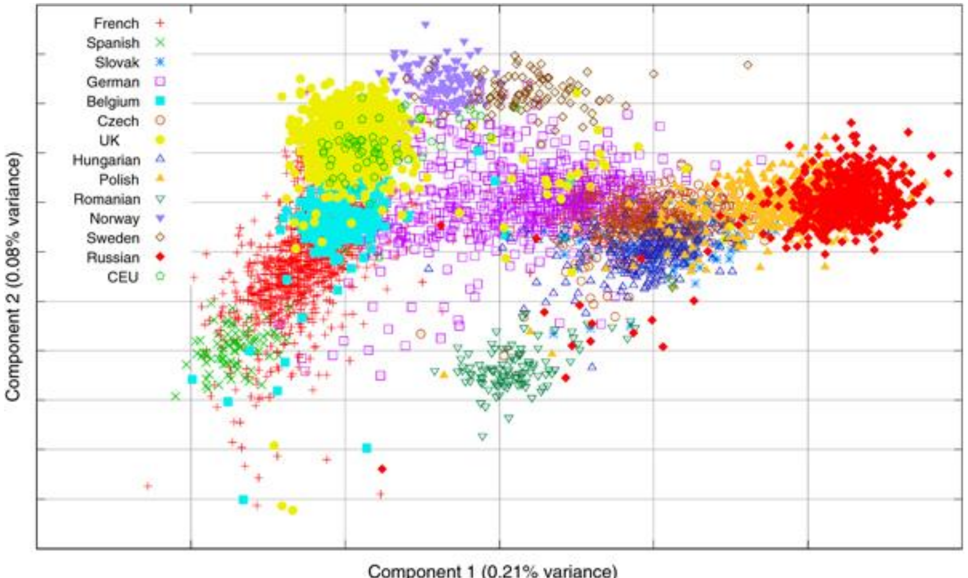
Last week: Principal Component Analysis

- PCA represents x_{ij} as linear combination of latent vectors:

$$f(W, Z) = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - w_j^T z_i)^2$$

- The w_c are 'latent factors', and z_i is low-dimensional representation.
- Why this model? Do we really all this math?

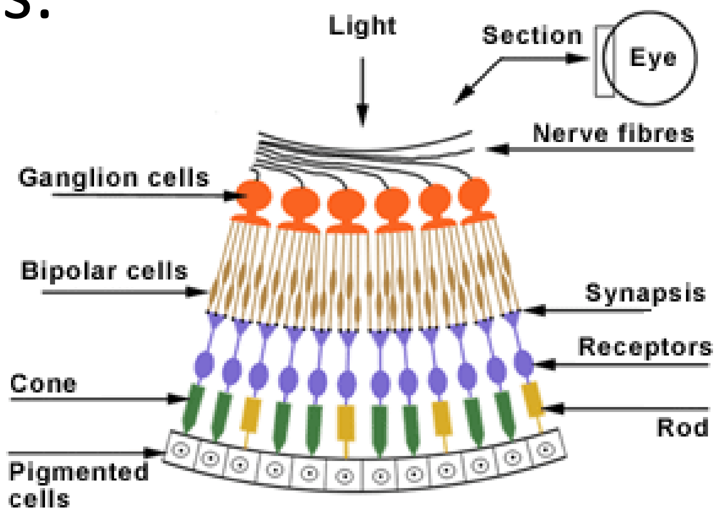
Trait	Description
Openness	Being curious, original, intellectual, creative, and open to new ideas.
Conscientiousness	Being organized, systematic, punctual, achievement-oriented, and dependable.
Extraversion	Being outgoing, talkative, sociable, and enjoying social situations.
Agreeableness	Being affable, tolerant, sensitive, trusting, kind, and warm.
Neuroticism	Being anxious, irritable, temperamental, and moody.

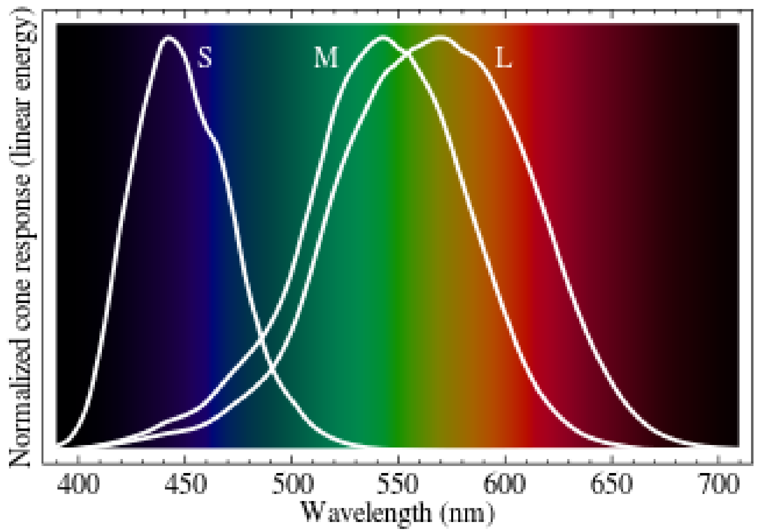


http://infoproc.blogspot.ca/2008/11/european-genetic-substructure.html
https://new.edu/resources/big-5-personality-traits
http://mikedusenberry.com/on-eigenfaces/

Colour Opponency in the Human Eye

- Classic model of the eye is with 4 photoreceptors:
  - L-Cones (most sensitive to red).
  - M-Cones (most sensitive to green).
  - S-Cones (most sensitive to blue).
  - Rods (more sensitive to brightness).
- Two problems with this system:
  - Correlation between receptors (not orthogonal).
    - Particularly between red/green.
  - We have 4 receptors for 3 colours.

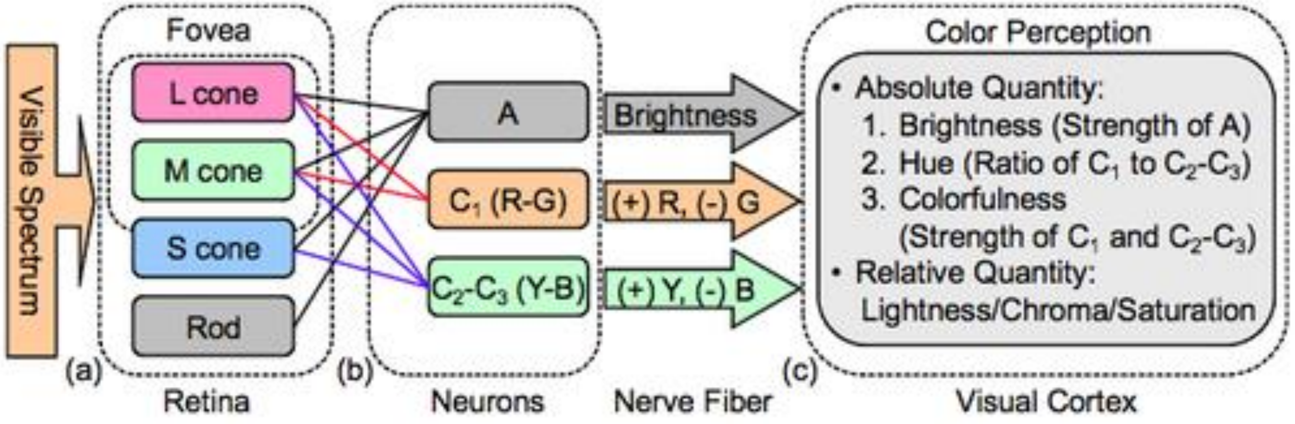


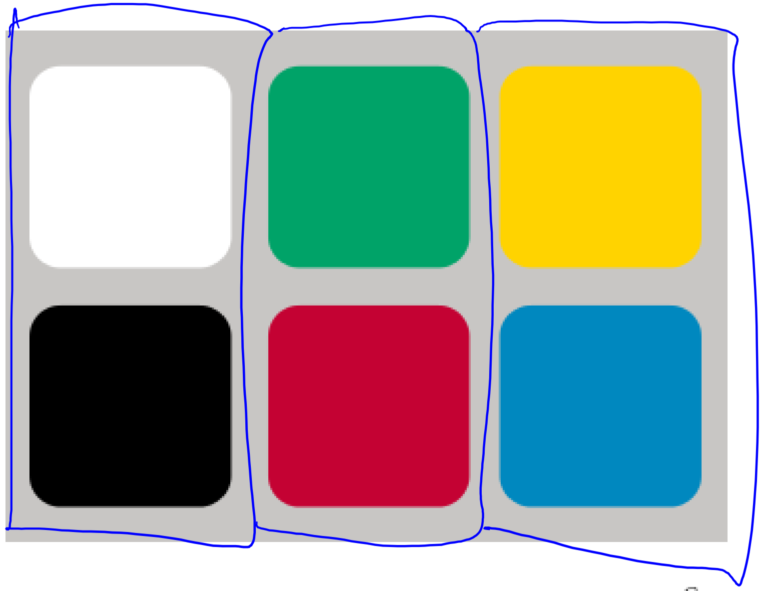


http://oneminuteastronomer.com/astro-course-day-5/
https://en.wikipedia.org/wiki/Color_visio

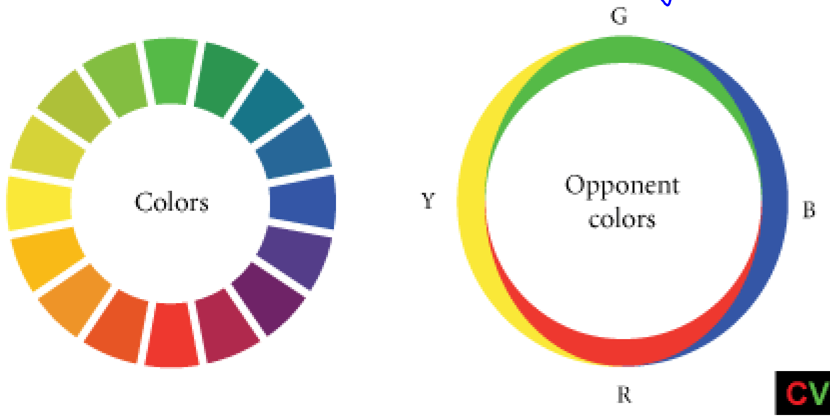
Colour Opponency in the Human Eye

- Bipolar and ganglion cells seem to code using ‘opponent colors’:
  - 3-variable orthogonal basis:





- This operation is similar to PCA.



http://oneminuteastronomer.com/astro-course-day-5/
https://en.wikipedia.org/wiki/Color_visio
http://5sensesnews.blogspot.ca/

Colour Opponency Representation

$= W_1$

Brightness

$+ W_2$

Red / Green

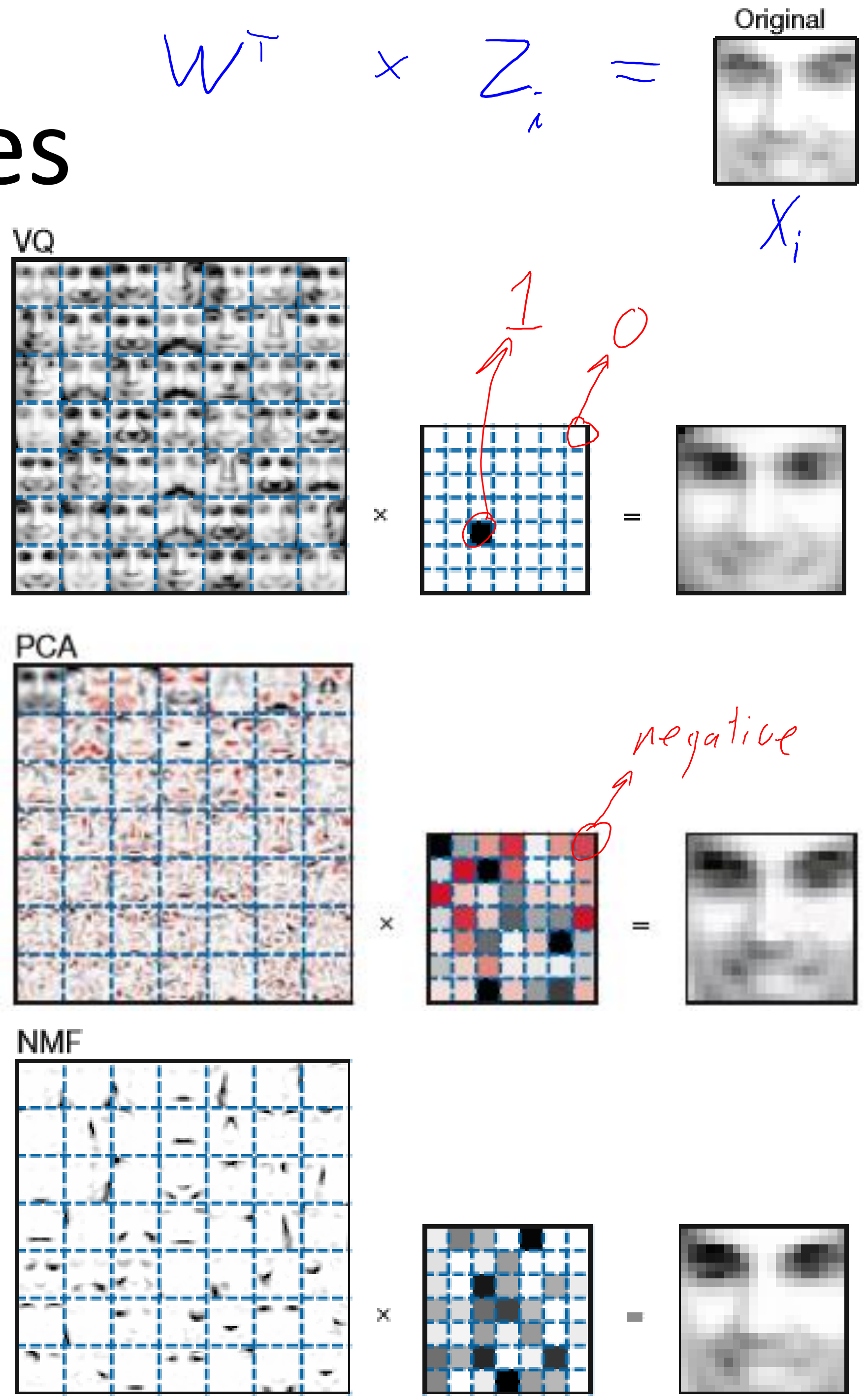
$+ W_3$

Blue / yellow

Representing Faces

But how should we represent faces?

- K-means (vector quantization):
  - 'Grandmother cell': one neuron = one face.
  - Almost certainly not true: too few neurons.
- Principal components analysis (PCA):
  - 'Distributed representation'.
    - We'll cover artificial neural networks next week.
  - Coded by pattern of group of neurons.
  - PCA uses all variables to make cancelling parts.
- Non-negative matrix factorization (NMF):
  - 'Sparse coding'.
  - Coded by activation of small set of neurons.
  - NMF makes object out small number of 'parts'.



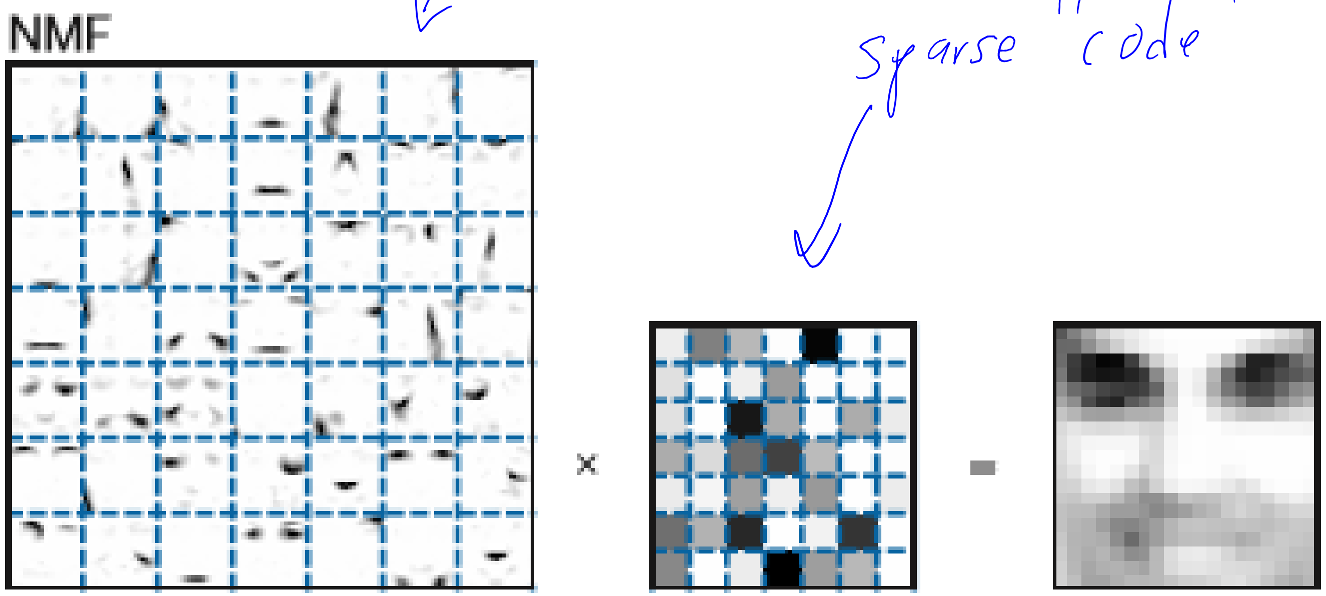
http://www.columbia.edu/~jwp2128/Teaching/W4721/papers/nmf_nature.pdf

Representing Faces

- Why sparse coding?
  - ‘Parts’ are intuitive, and brains seem to use sparse representation.
  - Energy efficiency if using sparse code.
  - Increase number of concepts you can memorize?
    - Some evidence in fruit fly olfactory system.

Sparse basis or "dictionary"

sparse "code"



http://www.columbia.edu/~jwp2128/Teaching/W4721/papers/nmf_nature.pdf

Warm-up to NMF: Non-Negative Least Squares

- Consider our usual least squares problem:

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

- Assume that y_i and elements of x_i are non-negative:
  - Could be sizes (‘height’, ‘milk’, ‘km’) or counts (‘vicodin’, ‘likes’, ‘retweets’).
- We may want elements of w to be non-negative, too:
  - No physical interpretation to negative weights.
  - If x_{ij} is amount of product you produce, what does $w_j < 0$ mean?
- Non-negativity constraint has interesting property:
  - Solution w tends to be sparse.

Non-Negative Least Squares

- The non-negative least squares formulation:

$$\underset{w \in \mathbb{R}^d}{\text{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 \quad \text{subject to} \quad w_j \geq 0 \text{ for all } j.$$

- This can be solved with projected-gradient iteration:

$$w^{t+1} = \underbrace{P_{w \geq 0}}_{\text{"projection"}}[\underbrace{w^t - \alpha_t \nabla f(w^t)}_{\text{usual gradient descent step}}] \quad \text{where } P_{w\geq 0} \text{ sets negative elements to } 0.$$

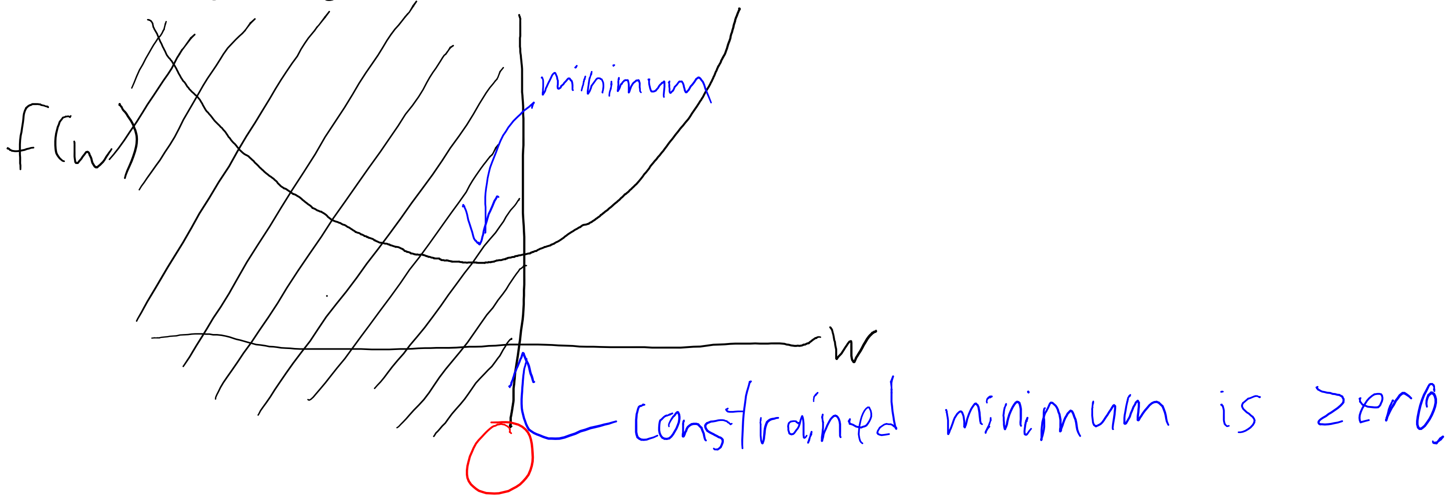
- Projected-gradient has similar properties to gradient descent.
  - Guaranteed to decrease objective for small enough α_t.
  - Guaranteed to find constrained local minimum.
  - Can also add projection to stochastic gradient.

Sparsity and Non-Negative Least Squares

- Consider 1D non-negative least squares objective:

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 \quad \text{with } w \geq 0$$

- Plotting the (constrained) objective function:



- Instead of setting w negative, NNLS will set w to zero.
- In higher-dimensions, NNLS also implicitly regularizes non-zero values:
  - Positive w_j are smaller because no ‘cancellation’ with negative values.

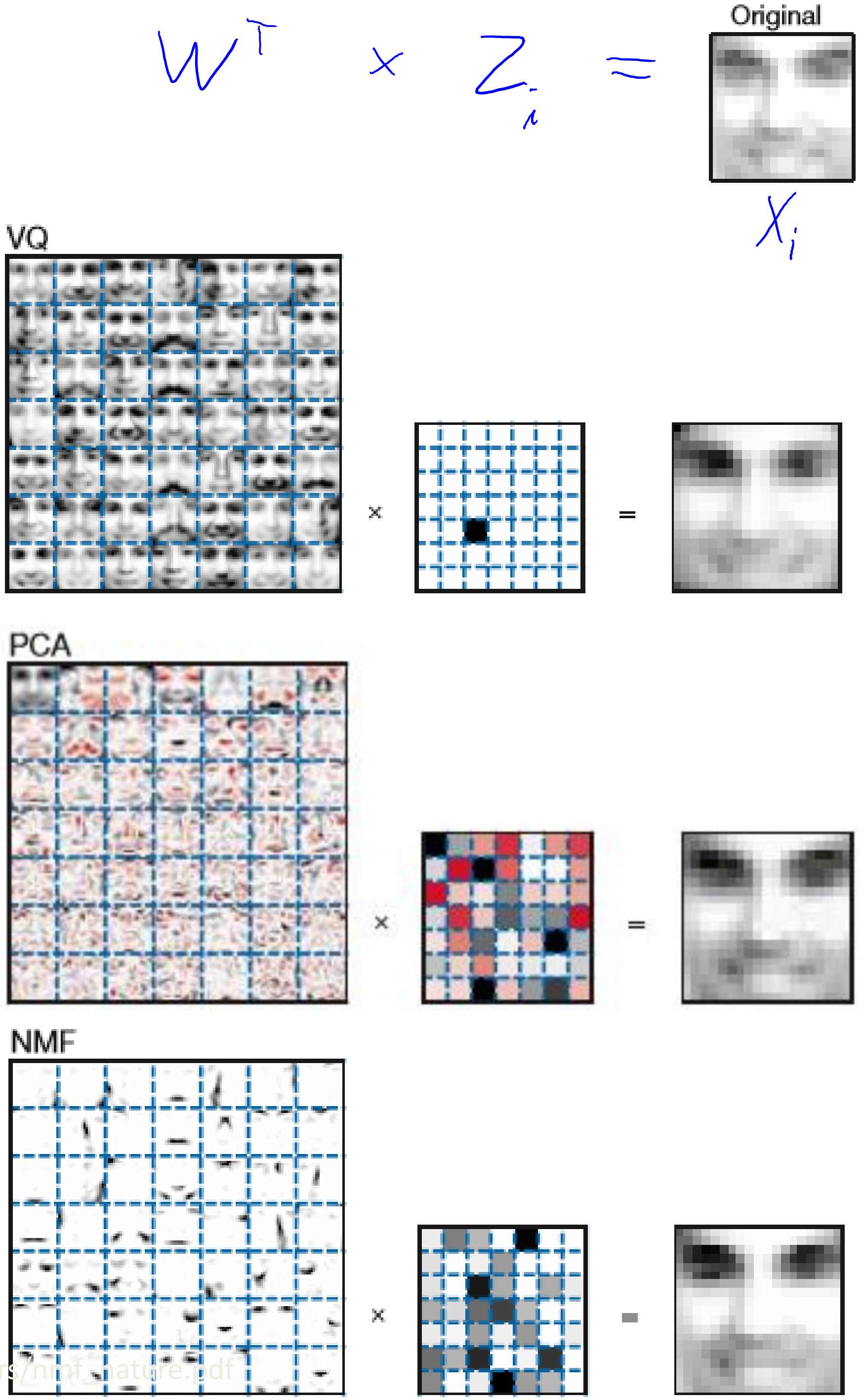
Non-Negative Matrix Factorization (NMF)

- Recall our objective for latent-factor models:

$$f(W,Z) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - w_j^T z_i)^2$$

- We get different models with different constraints:
  - K-means: each z_i has one '1' and the rest are zero.
  - Least squares: we only have one variable (d=1) and the z_i are fixed.
  - PCA: the w_c have a norm of 1 and have an inner product of zero.
  - NMF: all elements of W and Z are non-negative:
    - Latent-factors w_c are sparse (sparse 'dictionary').
    - Low-dimensional representation z_i is sparse (sparse 'code').

- We can also fit NMF with projected-gradient.
- Usually, alternate between updating 'W' and 'Z'.
- Not convex, initialization matters:.
  - Usually, random initial values.
- You can't initialize w_c the same:
  - They would stay the same.
  - Use different random values.



http://www.columbia.edu/~jwp2128/Teaching/W4721/papers/nmf_nature.pdf

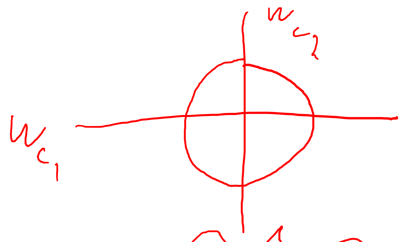
Other Latent-Factor Models

- Recall our objective for latent-factor models (LFM):

$$f(W,Z) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - w_j^T z_i)^2$$

- We can use our linear regression tricks in this framework:
  - Use robust loss function like absolute error (robust LFM).
  - Use logistic loss for binary x_{ij} (binary LFM).
  - Add regularization of W and/or Z to improve test error (regularized LFM).
  - Instead of non-negativity, use L1-regularization to encourage sparsity.

Sparse Coding and Sparse PCA



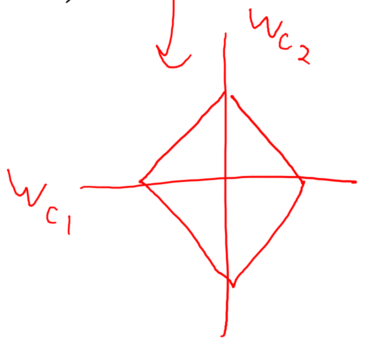
- Sparse coding:

$$\underset{W,Z}{\operatorname{argmin}} \overbrace{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\|x_{ij} - w_j^T z_i\|^2}^{\text{usual LFM}} + \lambda \sum_{i=1}^{n}\|z_i\|_1 \quad \text{subject to } \|w_c\| \leq 1$$

- Sparse PCA:

$$\underset{W,Z}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\|x_{ij} - w_j^T z_i\|^2 \quad \text{subject to } \|w_c\|_1 \leq 1$$

(some enforce orthogonality, too.)

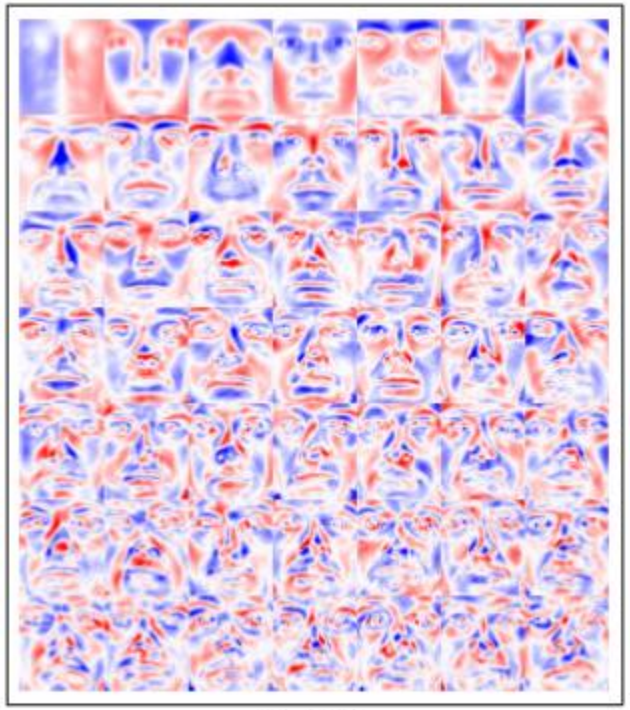


- K-SVD: constrain L0-norm of z_i.
- Literature is messy: can mix/match regularizers/constraints.

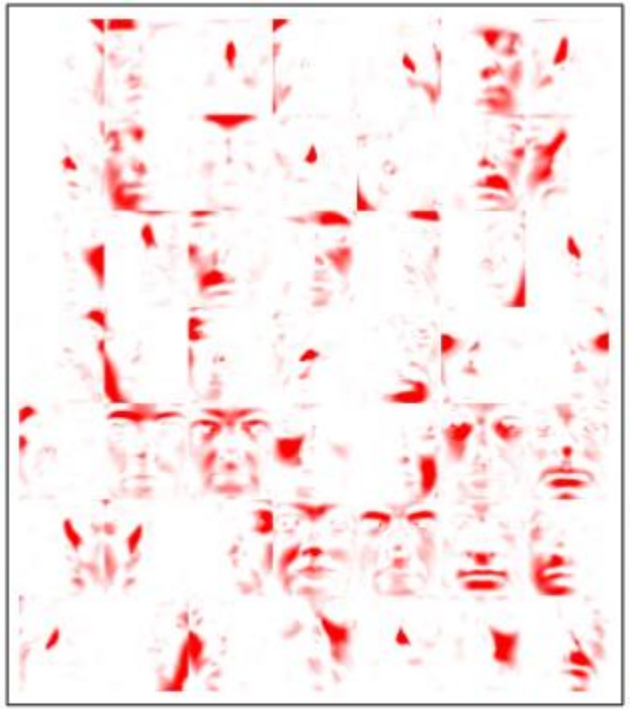
Latent-Factor Models for Face Representations

Each x_i is a black and white face image.

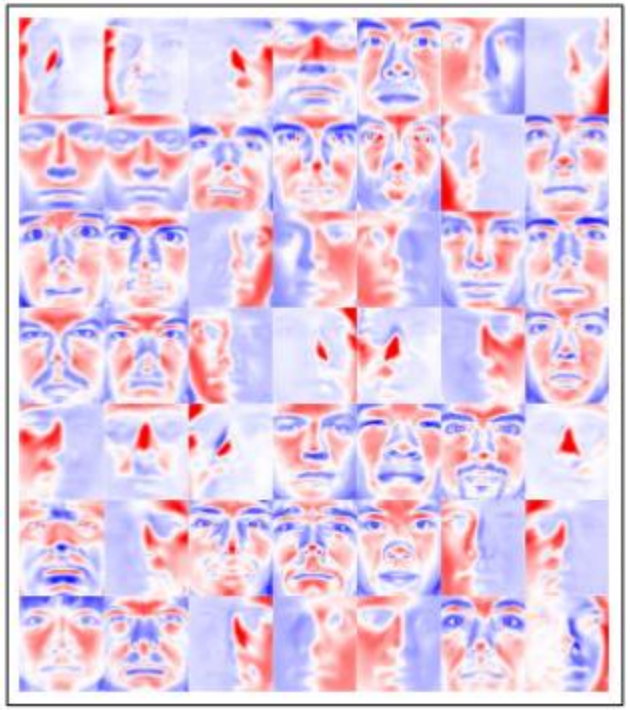
red: positive   blue: negative



(a) PCA

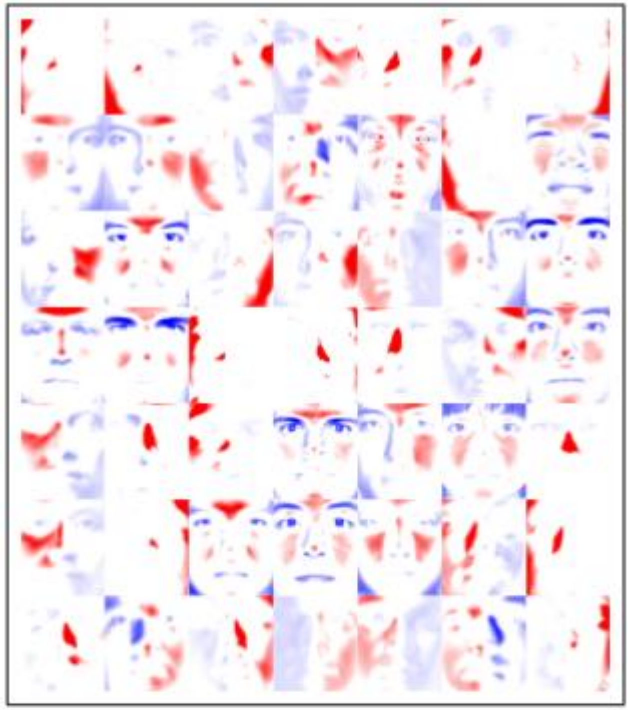


(c) NMF



(e) Dictionary Learning

no orthogonality constraint



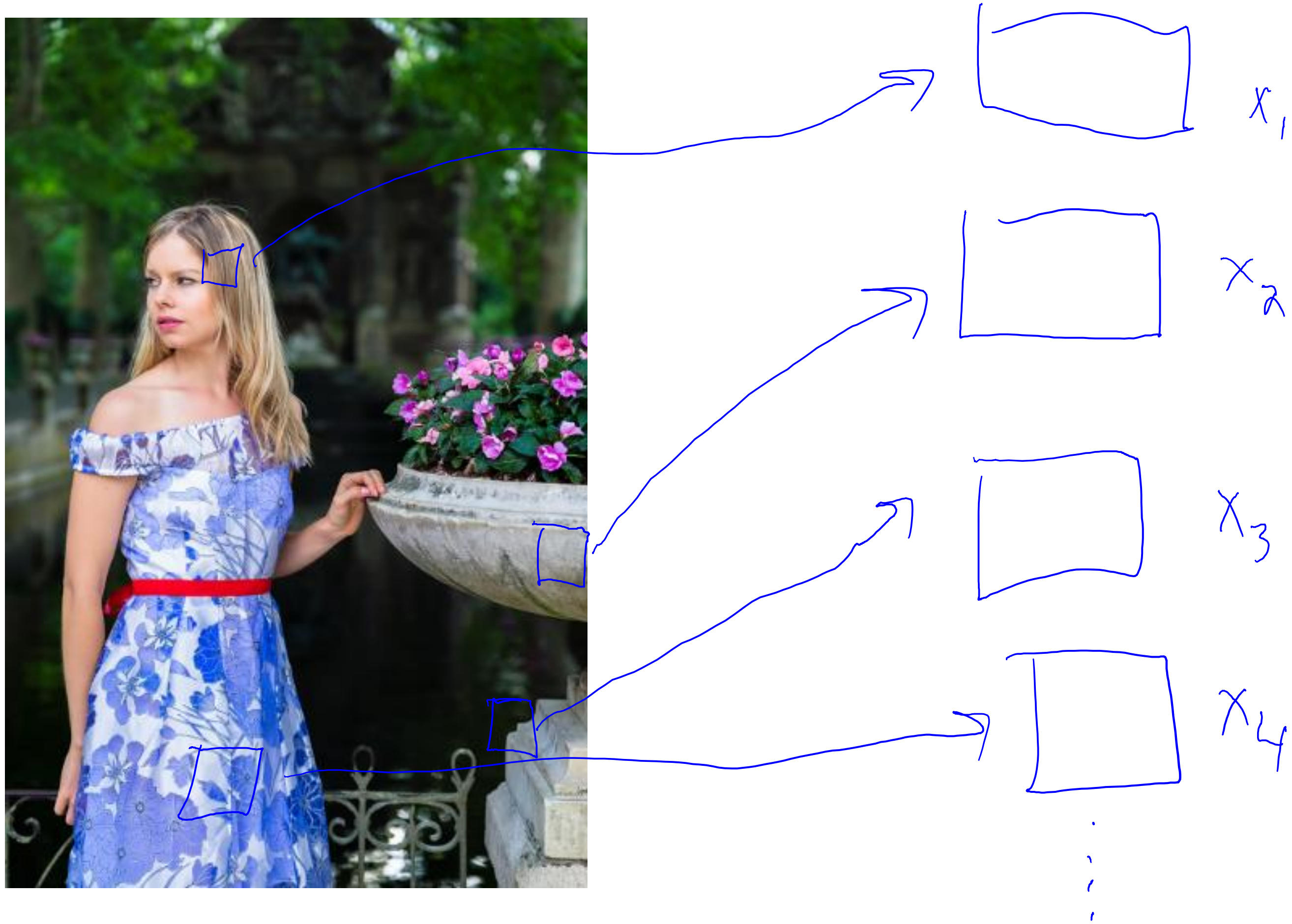
(d) SPCA, $\tau = 30\%$

http://www.jmlr.org/papers/volume11/mairal10a/mairal10a.pdf

Latent-Factor Models for Image Patches

- Consider building latent-factors for general image patches:

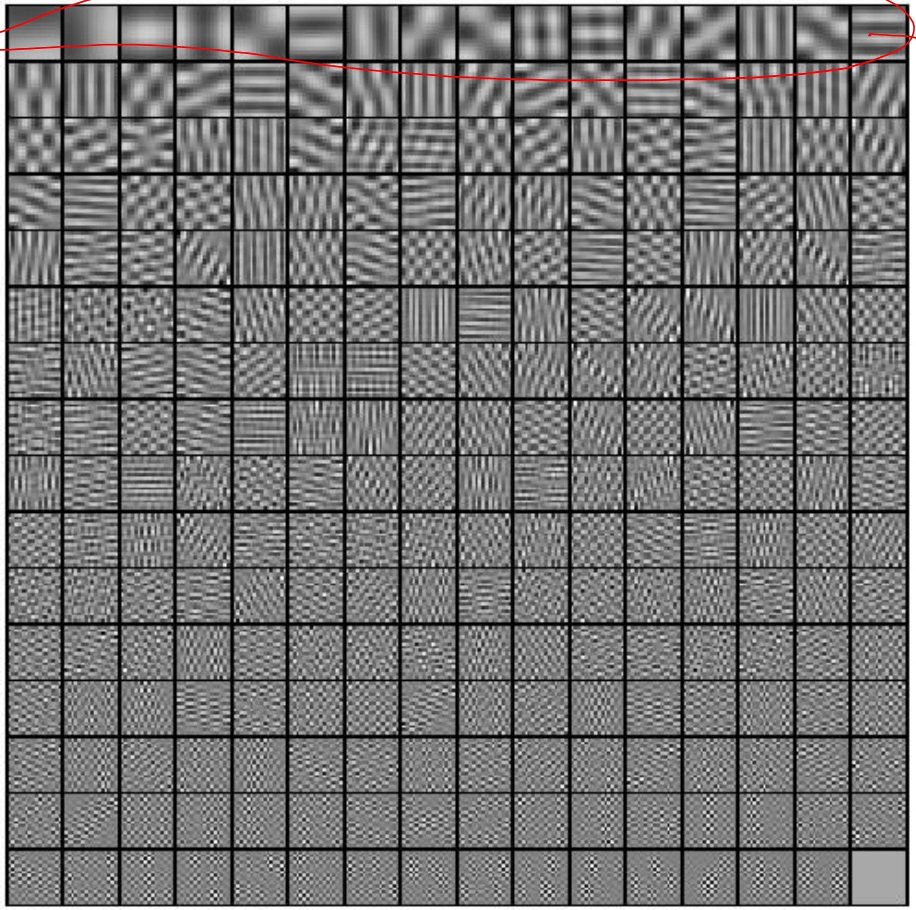
What are images made of?



Typical pre-processing:

center and ‘whiten’ patches.

Latent-Factor Models for Image Patches

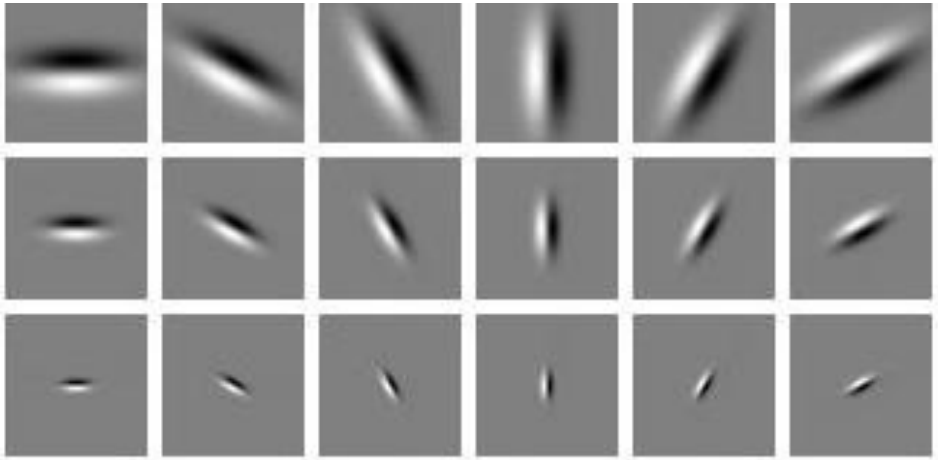


(b) Principal components.

We don't think this is the right representation:

- Few PCs do almost everything.
- Most PCs do almost nothing.

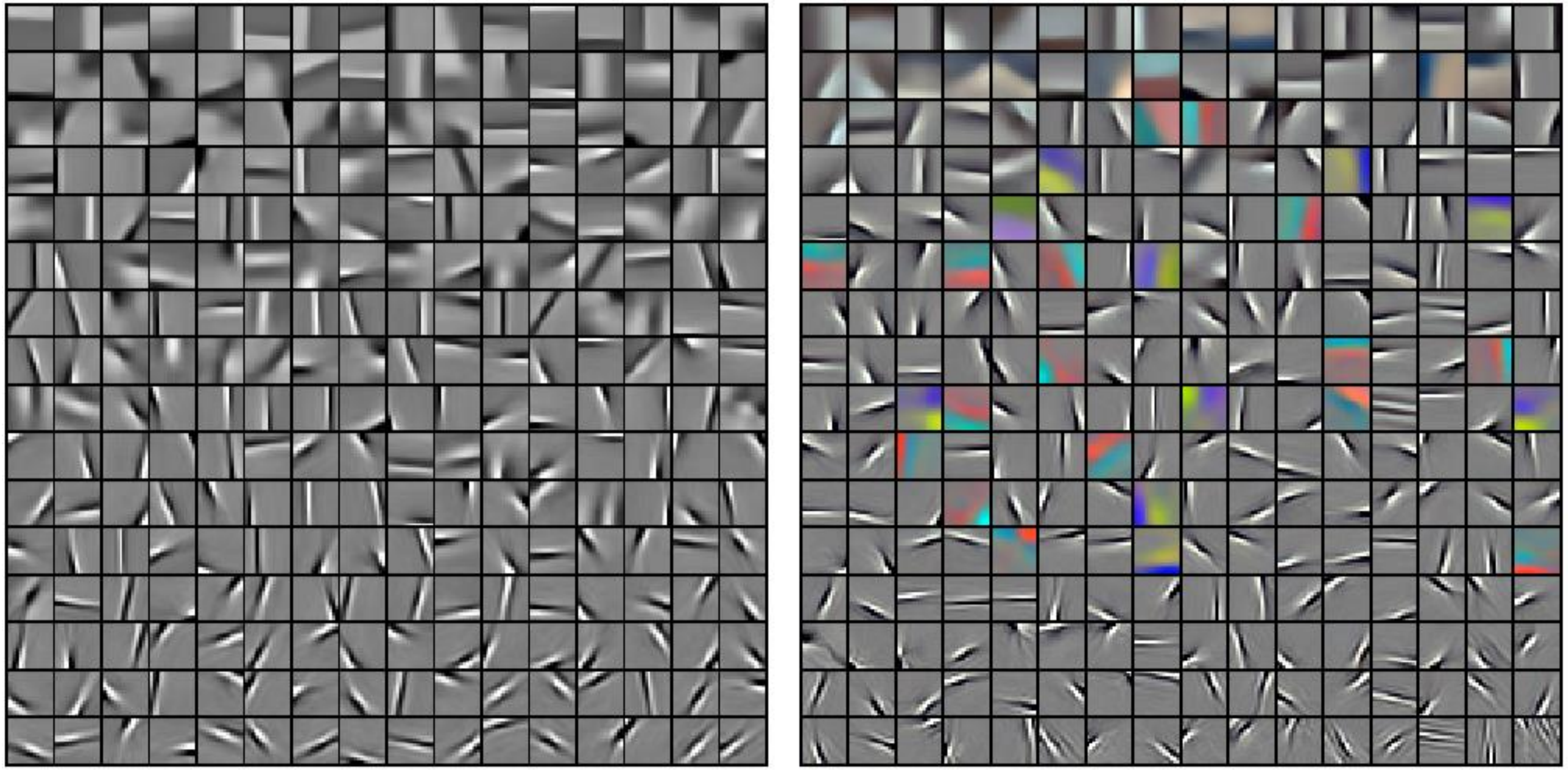
We believe 'simple cells' in visual cortex look like:



'Gabor' filters

Latent-Factor Models for Image Patches

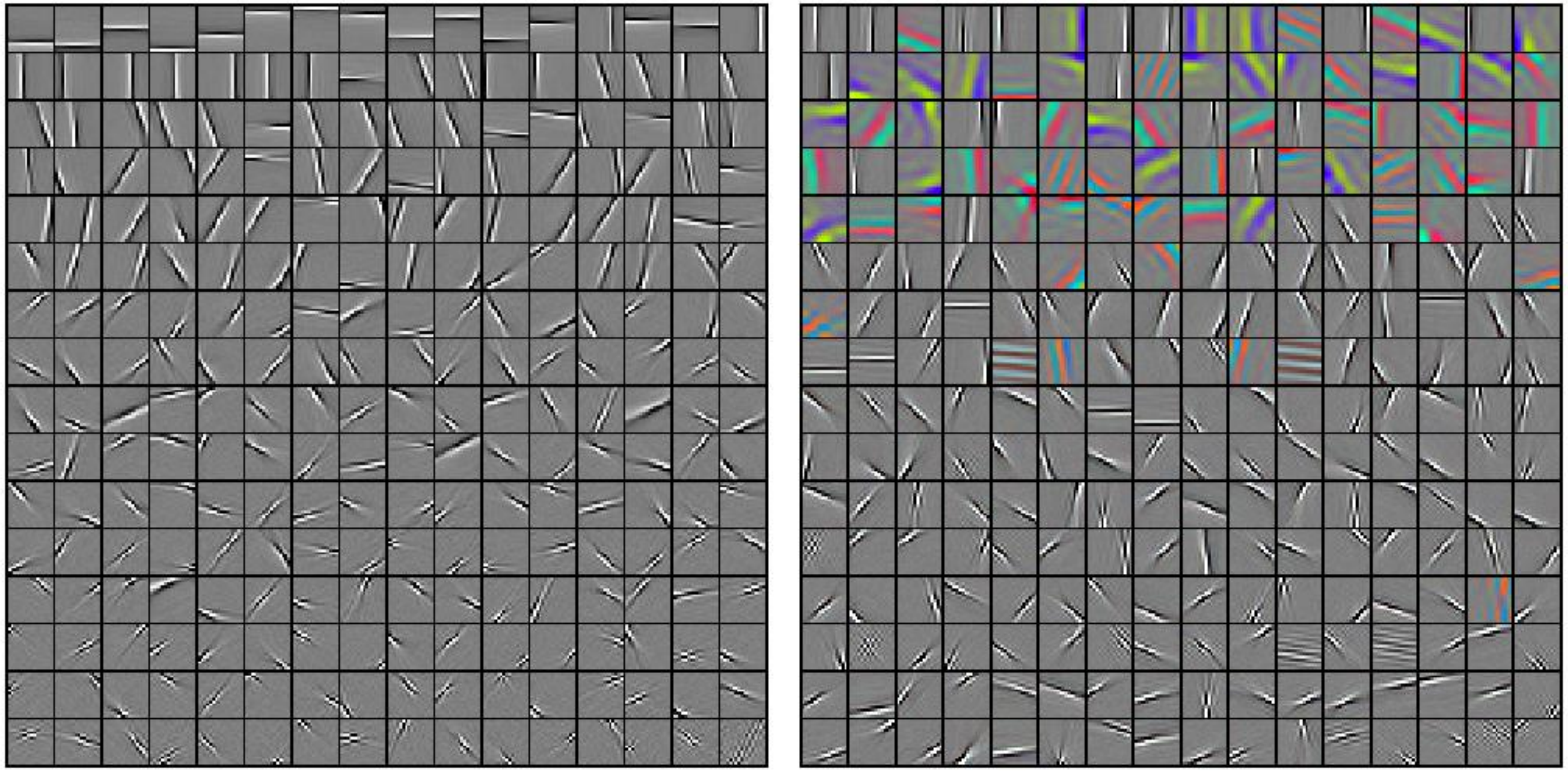
- Latent factors from sparse coding on B+W and colour patches:



(a) With centering - gray. (b) With centering - RGB.

Latent-Factor Models for Image Patches

- Latent factors from sparse coding on B+W and colour patches:



(c) With whitening - gray. (d) With whitening - RGB.

colour opponency

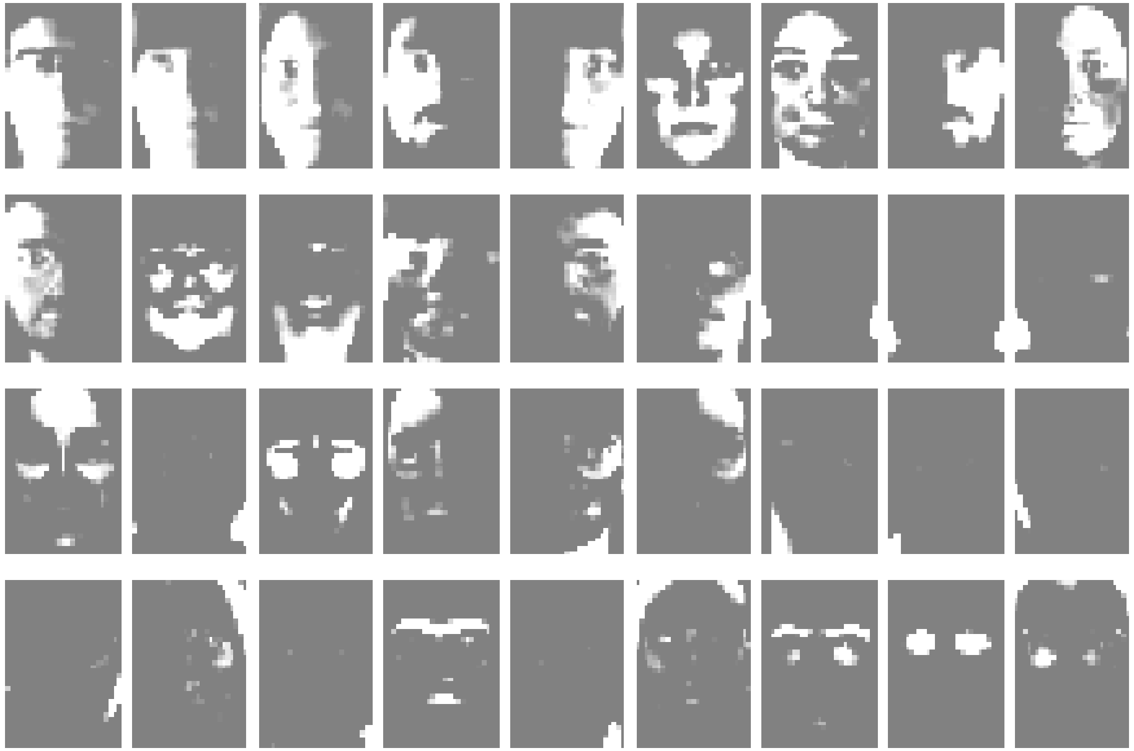
http://lear.inrialpes.fr/people/mairal/resources/pdf/review_sparse_arxiv.pdf

Application: Image Inpainting

http://www.jmlr.org/papers/volume11/mairal10a/mairal10a.pdf

Recent Work: Structured Sparsity

- ‘Structured sparsity’ considers dependencies in sparsity patterns.



NMF

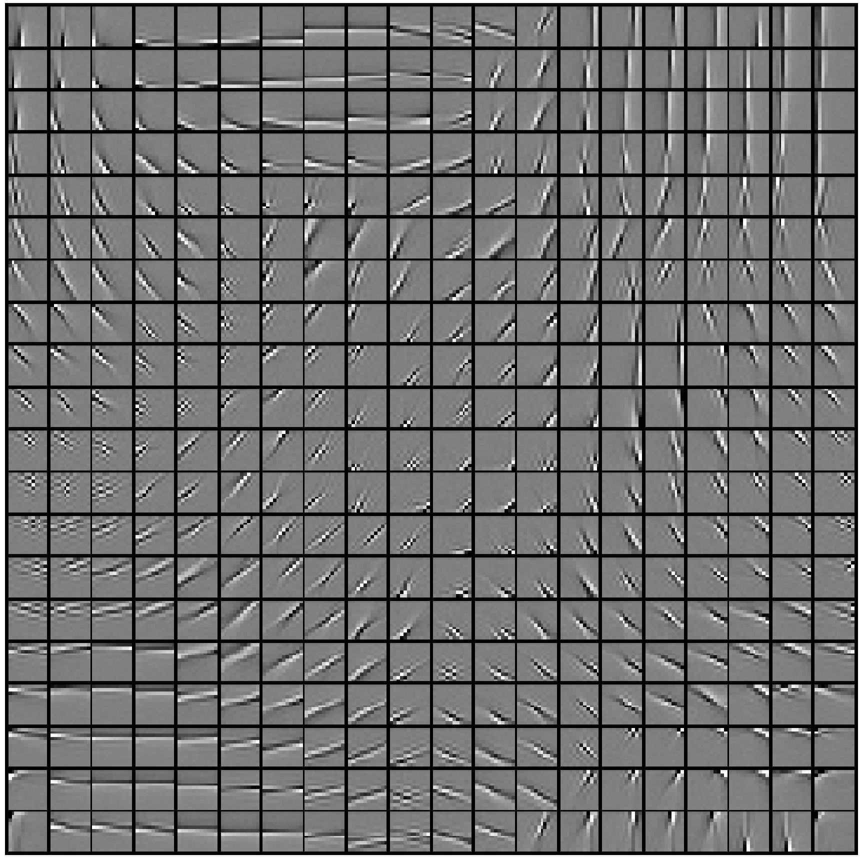
"Structured" sparse PCA

http://jmlr.org/proceedings/papers/v9/jenatton10a/jenatton10a.pdf

Recent Work: Structured Sparsity

- 'Structured sparsity' considers dependencies in sparsity patterns.

Factors with "structured" sparse coding



(b) With 4 × 4 neighborhood.

This is similar to 'cortical columns' theory in visual cortex.

Summary

- Biological motivation for orthogonal and sparse latent factors.
- Non-negativity leads to a form of sparsity.
- Non-negative matrix factorization leads to sparse LFM.
- L1-regularization leads to other sparse LFMs.

- Next time: predicting which movies you are going to like.