MIT Department of Brain and Cognitive Sciences
9.641J, Spring 2005 - Introduction to Neural Networks
Instructor: Professor Sebastian Seung

Principal component analysis

Hypothesis: Hebbian synaptic plasticity enables perceptrons to perform principal component analysis

Outline

- Variance and covariance
- Principal components
- Maximizing parallel variance
- Minimizing perpendicular variance
- Neural implementation
  - covariance rule

Principal component

- direction of maximum variance in the input space
- principal eigenvector of the covariance matrix
- goal: relate these two definitions

Variance

- A random variable fluctuating about its mean value.

$$\delta x = x - \langle x \rangle$$

$$\left\langle (\delta x)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2$$

- Average of the square of the fluctuations.

Covariance

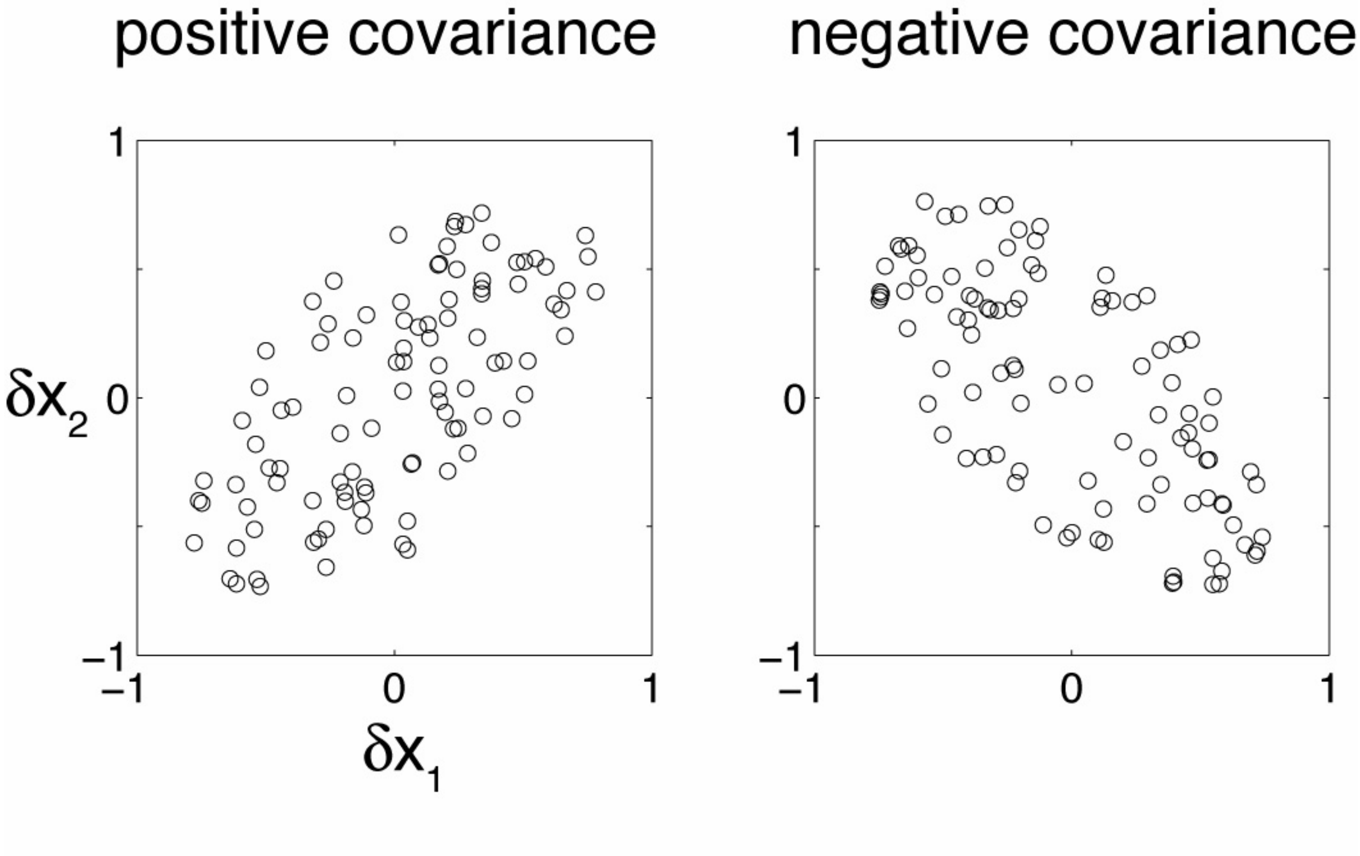
- Pair of random variables, each fluctuating about its mean value.

$$\delta x_1 = x_1 - \langle x_1 \rangle$$
$$\delta x_2 = x_2 - \langle x_2 \rangle$$

$$\langle \delta x_1 \delta x_2 \rangle = \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle$$

- Average of product of fluctuations.

Covariance examples



Covariance matrix

- N random variables
- NxN symmetric matrix

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

- Diagonal elements are variances

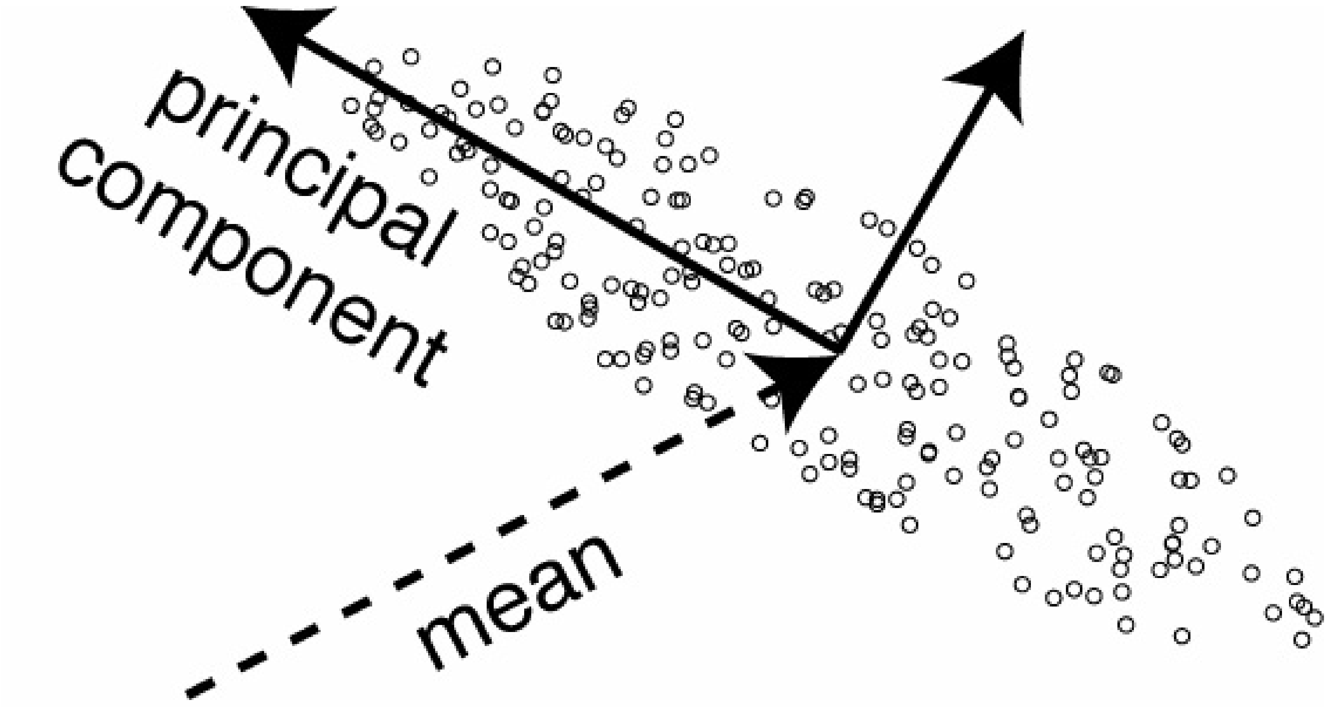
Principal components

$$Cv_1 = \lambda_1 v_1$$
$$Cv_2 = \lambda_2 v_2$$
$$\vdots$$
$$Cv_k = \lambda_k v_k$$
$$\vdots$$
$$Cv_N = \lambda_N v_N$$

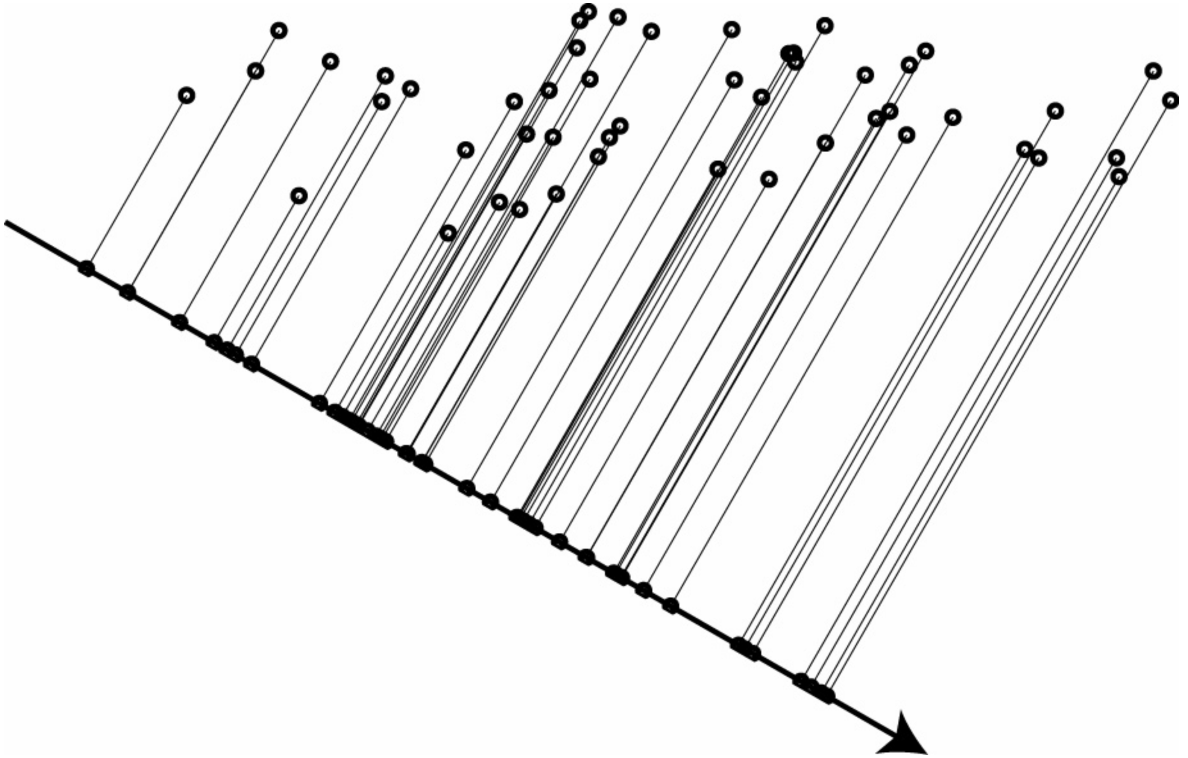
- eigenvectors with *k* largest eigenvalues
- Now you can calculate them, but what do they mean?

Handwritten digits

Principal component in 2d



One-dimensional projection



Covariance to variance

- From the covariance, the variance of any projection can be calculated.
- Let *w* be a unit vector

$$\left\langle \left(w^T x\right)^2 \right\rangle - \left\langle w^T x \right\rangle^2 = w^T C w$$

$$= \sum_{ij} w_i C_{ij} w_j$$

Maximizing parallel variance

- Principal eigenvector of C
  - the one with the largest eigenvalue.

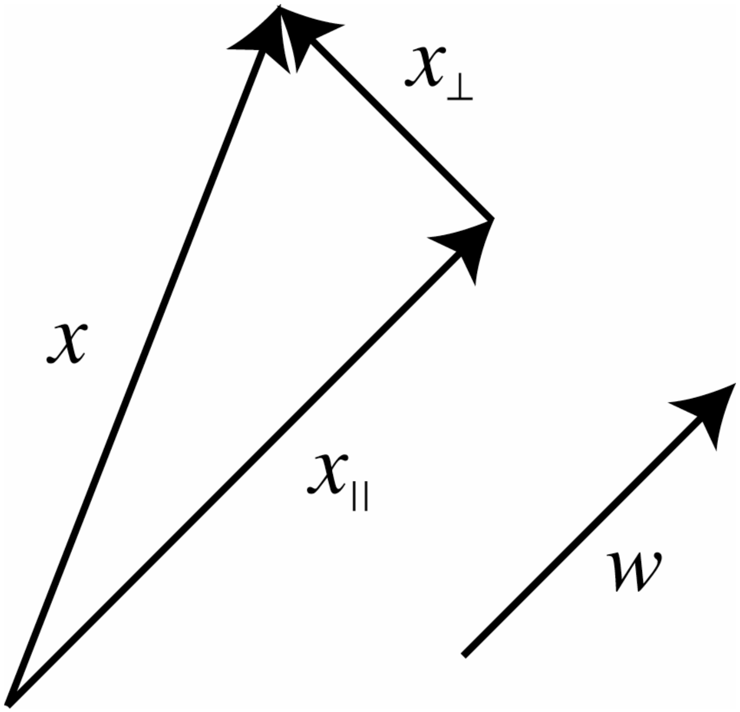
$$w^* = \arg\max_{w:|w|=1} w^T C w$$

$$\lambda_{\max}(C) = \max_{w:|w|=1} w^T C w$$

$$= w^{*T} C w^*$$

Orthogonal decomposition

$$|x|^2 = |x_{||}|^2 + |x_\perp|^2$$



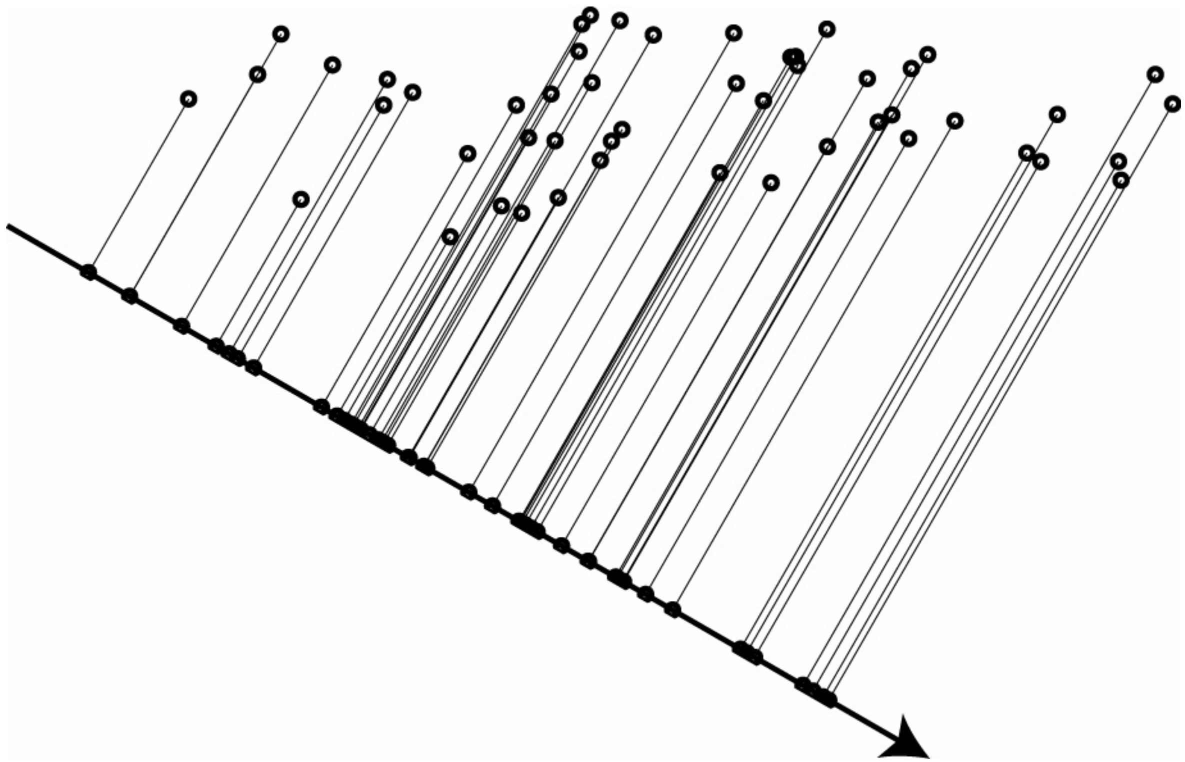
Total variance is conserved

$$\langle |x|^2 \rangle = \langle |x_{||}|^2 \rangle + \langle |x_\perp|^2 \rangle$$

- Maximizing parallel variance = Minimizing perpendicular variance

$$\underset{w:|w|=1}{\operatorname{argmax}} \langle |x_{||}|^2 \rangle = \underset{w:|w|=1}{\operatorname{argmin}} \langle |x_\perp|^2 \rangle$$

Rubber band computer



Correlation/covariance rule

- presynaptic activity x
- postsynaptic activity y
- Hebbian

$$\Delta w = \eta y x$$

$$\Delta w = \eta (y - \langle y \rangle)(x - \langle x \rangle)$$

Stochastic gradient ascent

- Assume data has zero mean
- Linear perceptron

$$y = w^T x \qquad \Delta w = \eta x x^T w$$

$$\langle \Delta w \rangle = \eta \frac{\partial E}{\partial w} \qquad E = \frac{1}{2} w^T C w$$

$$C = \langle x x^T \rangle$$

Preventing divergence

- Bound constraints

Oja's rule

- Converges to principal component
- Normalized to unit vector

$$\Delta w = \eta\left(yx - y^2 w\right)$$

Multiple principal components

- Project out principal component
- Find principal component of remaining variance

clustering vs. PCA

- Hebb: output x input
- binary output
  - first-order statistics
- linear output
  - second-order statistics

Data vectors

- x_a means *a*th data vector
  - *a*th column of the matrix X.
- X_{ia} means matrix element X_{ia}
  - *i*th component of x_a
- x is a generic data vector
- x_i means *i*th component of x

Correlation matrix

- Generalization of second moment

$$\left\langle x_i x_j \right\rangle = \frac{1}{m} \sum_{a=1}^{m} X_{ia} X_{ja}$$

$$\left\langle x x^T \right\rangle = \frac{1}{m} X X^T$$