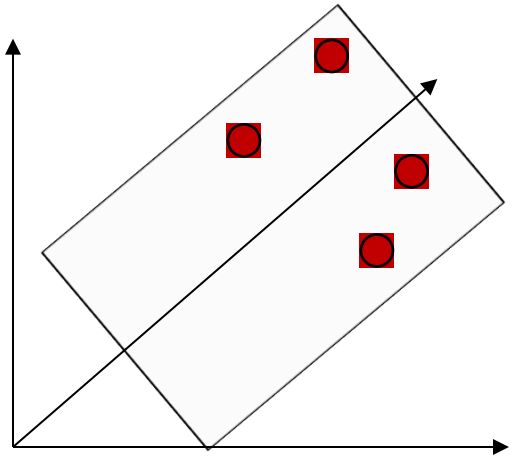
Dimension Reduction and Sparse representation (Week-11: Lectures 51 – 54)

Jayanta Mukhopadhyay
Dept. of Computer Science and Engg.

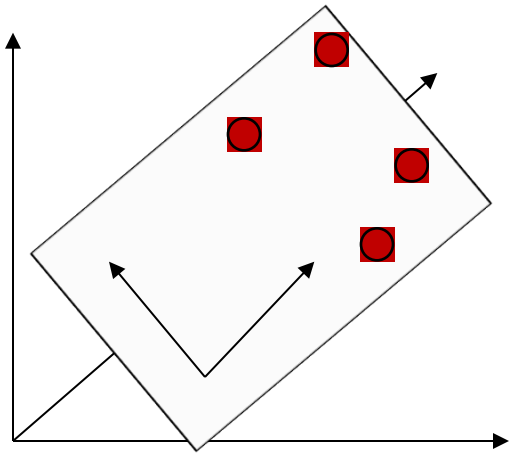
Data dimension

- Consider a set of data points $S=\{x_i | \ x_i \text{ in } R^n\}$.
  - The dimension of the space R^n is n.
  - Does it mean dimension of the set S also n?



Data dimension

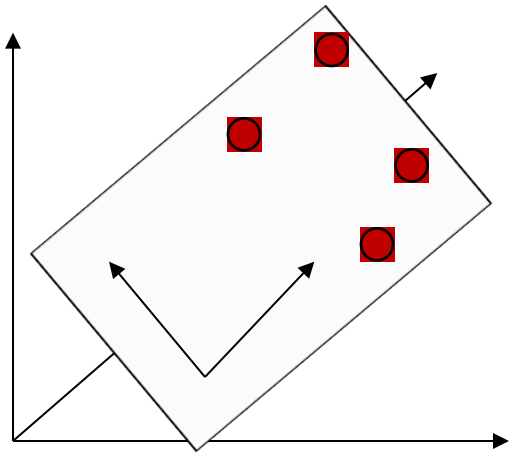
- Consider a set of data points $S=\{x_i | \ x_i \text{ in } R^n\}$.
  - The dimension of the space R^n is n.
  - Does it mean dimension of the set S also n?



S could be represented as a set of points on a 2D space (R^2).

Principal component analysis

- Consider a set of data points $S=\{x_i | \ x_i \text{ in } R^n\}$.
  - The dimension of the space R^n is n.
  - Does it mean dimension of the set S also n?

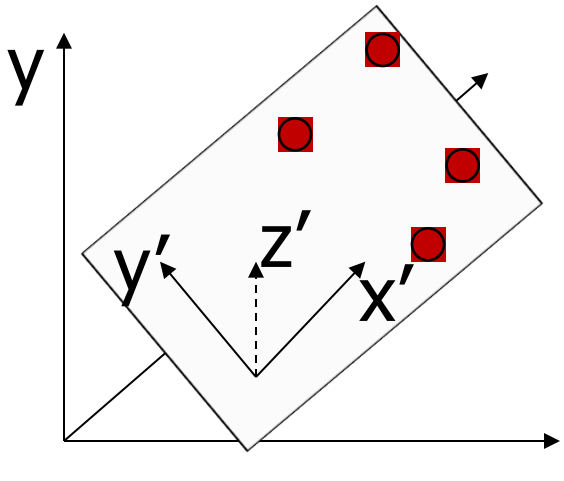


Principal component analysis (PCA) finds the minimum dimensional subspace for representing data.

S could be represented as a set of points on a 2D space (R^2).

Principal component analysis

- Consider a set of data points $S=\{x_i | \ x_i \text{ in } R^n\}$. The dimension of the space R^n is n.
  - Does it mean dimension of the set S also n?



Principal component analysis (PCA) finds the minimum dimensional subspace for representing data.

S could be represented as a set of points on a 2D space (R^2).

Computes a new set of orthogonal axes.

- Coordinate transformation

Maximizing variance of a component

- Consider a feature vector: $X=(x_1,x_2,\ldots x_n)$.
- Variance of x_i: $\mathrm{var}(x_i)=\frac{1}{N}\sum_{j=1}^{N}(x_{ij}-\bar{x}_i)^2$
- Dominant component:
  - Component with maximum variance.
- PCA maximizes variance of the dominant component.
  - Consider $W=(w_1,w_2,\ldots w_n)$ a unit vector.
  - Consider the mean of feature vectors: $\bar{S}$
  - For every X_j translated to the mean vector compute the component along W. $y_j=(X_j-\bar{S}).W$
  - Find W which maximizes variance of y_j's.

Maximizing variance of a component

- A set of data points: $S=\{X_j=(x_{1j},x_{2j},\ldots x_{nj})|\ X_j \text{ in } R^n\}$.
- Mean vector of S: $\bar{S} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} X_1^T - \bar{S} \\ X_2^T - \bar{S} \\ \vdots \\ X_N^T - \bar{S} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \tilde{X}^T W$$

- Compute W which maximizes: $\frac{1}{N} Y^T Y \Rightarrow \frac{1}{N} (\tilde{X}^T W)^T \tilde{X}^T W$

$$\Downarrow$$

$$W^T \frac{\tilde{X}\tilde{X}^T}{N} W$$

Such that $||W^T W||=1$

Maximizing variance of a component

- Compute W which maximizes:

$$W^T \frac{\tilde{X}\tilde{X}^T}{N} W$$

Such that $||W^T W||=1$

Covariance matrix (C): $C_{kl} = \frac{1}{N}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$

To maximize: Lagrange multiplier

$$L(W) = W^T C W - \lambda(W^T W - 1)$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow W^T W = 1$$

$$\frac{\partial L}{\partial W} = 0 \Rightarrow 2CW - 2\lambda W = 0 \Rightarrow CW = \lambda W$$

Eigen vector of C

Maximum eigen value

Principal components

Covariance matrix (C): $$C_{kl} = \frac{1}{N}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$$

- Dominant principal component.
  - Eigen vector corresponding to maximum eigen value of C.
- The set of eigen vectors corresponding to decreasing eigen values provide the principal components.
  - Ev=$\{\mathbf{e_1}, \mathbf{e_2}, \ldots \mathbf{e_n}\}$ with corresponding eigen values in increasing order.

$$\{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\}$$

  - All vectors are normalized.

Principal components

- i th principal component:

Mean of data points

$$y_i = (X - \bar{S}).\mathbf{e_i}$$

- Dimension Reduction: Ignore eigenvectors of small eigen values.
  - Suppose all the eigen vectors till k th eigen value retained for representing data.
  - $Y=(y_1,y_2, \ldots, y_k)$ is k-dimensional representation of data.

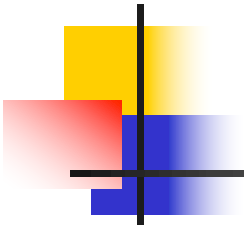
Dimension Reduction

$$Y = ((X - \bar{S}).\mathbf{e_1}, \ (X - \bar{S}).\mathbf{e_2}, \dots, (X - \bar{S}).\mathbf{e_k})$$

- k dimensional vector
- k <n

- Total Variance of data: $$V = \sum_{j=1}^{n} \left(\frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 \right)$$
  - Variance is the sum of eigen values. $$V = \sum_{j=1}^{n} \lambda_j$$
  - Ratio of sum of k eigen values to total sum (variance of data): fraction of variance accounted for.

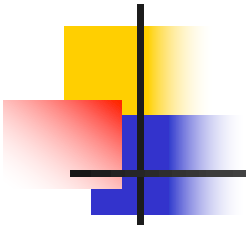
$$R^2 = \frac{\sum_{j=k+1}^{n} \lambda_j}{V}$$



PCA-Algorithm

- Input: A set of data points: $S=\{X_j=(x_{1j},x_{2j},\ldots x_{nj})|\ X_j \text{ in } R^n\}$.
- Output: A set of k eigen vectors: $E_v=\{e_1,e_2,\ldots,e_k\}$

1. Compute mean of data points.
2. Translate all data points to their mean.
3. Compute covariance matrix of the set.
4. Compute eigen vetcors and eigen values (in increasing order).
5. Choose k such that the fraction of variance accounted for is more than a threshold.
6. Use those k-components for representing any data point.



Example

- Data : {(5, 3, 2), (4, 6, 0), (3, -7, 14), (2, 5, 3), (3, 13, -6)}
- Perform PCA and if applicable, reduce the dimension of data.

Example (contd.)

$$X = \begin{bmatrix} 5 & 4 & 3 & 2 & 3 \\ 3 & 6 & -7 & 5 & 13 \\ 2 & 0 & 14 & 3 & -6 \end{bmatrix} \qquad \bar{S} = \begin{bmatrix} 3.4 \\ 4 \\ 2.6 \end{bmatrix}$$

$$\tilde{X} = X - \bar{S} = \begin{bmatrix} 1.6 & .6 & -.4 & -1.4 & -.4 \\ -1 & 2 & -11 & 1 & 9 \\ -.6 & -2.6 & 11.4 & .4 & -8.6 \end{bmatrix}$$

$$C = \frac{1}{5}\tilde{X}\tilde{X}^T \qquad C = \begin{bmatrix} 1.04 & -.2 & -.84 \\ -.2 & 41.6 & -41.4 \\ -.84 & -41.4 & 42.24 \end{bmatrix}$$

Example (contd.)

$$C = \begin{bmatrix} \mathbf{1.04} & -.2 & -.84 \\ -.2 & \mathbf{41.6} & -41.4 \\ -.84 & -41.4 & \mathbf{42.24} \end{bmatrix}$$

Total variance: Trace(C)=1.04+41.6+42.24=84.88

Eigen values of C: (83.3238, 1.5562, 0)

Sum of eigen values

Respective eigen vectors:

$$\mathbf{e_1} = \begin{bmatrix} -.0055 \\ -.7043 \\ .7099 \end{bmatrix} \quad \mathbf{e_2} = \begin{bmatrix} -.8165 \\ .413 \\ .4034 \end{bmatrix} \quad \mathbf{e_3} = \begin{bmatrix} -.5774 \\ -.5774 \\ -.5774 \end{bmatrix}$$

Example (contd.)

Respective eigen vectors:

$$\mathbf{e_1} = \begin{bmatrix} -.0055 \\ -.7043 \\ .7099 \end{bmatrix} \quad \mathbf{e_2} = \begin{bmatrix} -.8165 \\ .413 \\ .4034 \end{bmatrix} \quad \mathbf{e_3} = \begin{bmatrix} -.5774 \\ -.5774 \\ -.5774 \end{bmatrix}$$

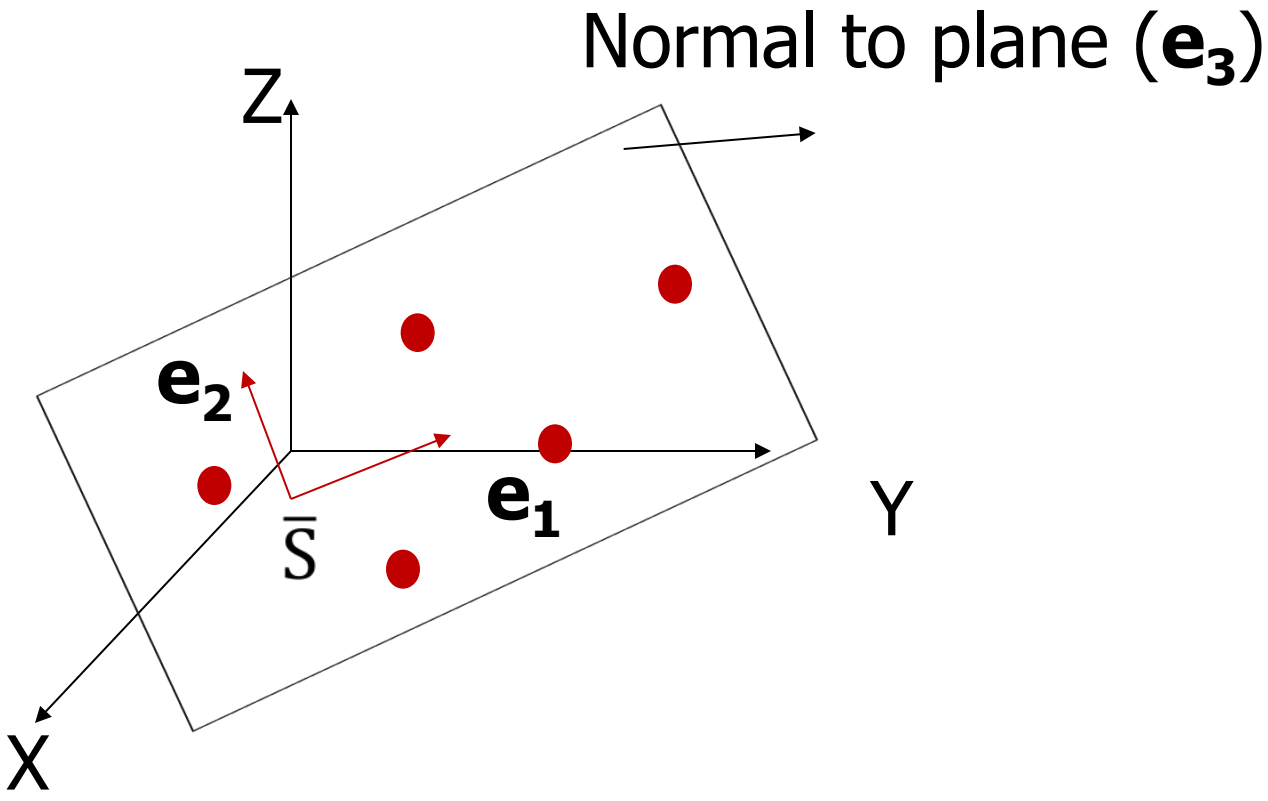
$$\mathrm{B} = [\mathbf{e_1} \quad \mathbf{e_2} \quad \mathbf{e_3}] = \begin{bmatrix} -.0055 & -.8165 & -.5774 \\ -.7043 & .413 & -.5774 \\ .7099 & .4034 & -.5774 \end{bmatrix}$$

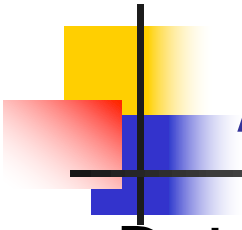
$$\tilde{X}^T . \mathrm{B} = \begin{bmatrix} .2696 & -1.9615 & 0 \\ -3.2576 & -0.7128 & 0 \\ 15.8421 & 0.3825 & 0 \\ -.4126 & 1.7175 & 0 \\ -12.4415 & 0.5742 & 0 \end{bmatrix}$$

Points lying in the plane:
X+Y+Z=10

Redundant dimension

Coordinate transformation





Application of PCA

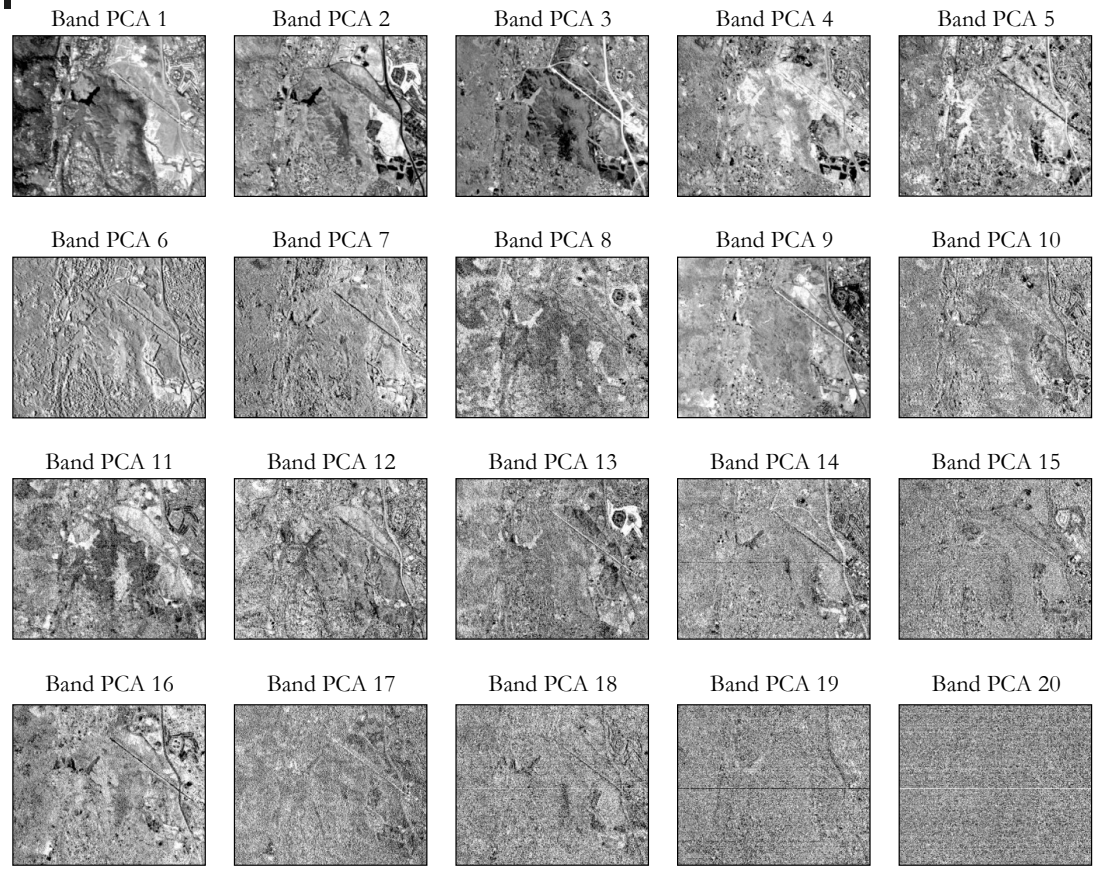
- Data compression
  - Provides optimum set of orthonormal basis vectors for a set of data points.
    - Data dependent.
    - Basis vectors also called as 'Karhunen-Loeve' basis, and the transform called 'Karhunen-Loeve Transform' (KLT).
    - Type-2 DCT basis vectors are approximately the eigen vectors of a 2-D matrix with (j,k) the entries as $r^{|j-k|}$.
      - Covariance matrix for a useful class of signals, where r is the measure of correlation between adjacent samples and a value near to 1.

Application of PCA

- Decorrelating components
  - Color images in RGB space highly correlated.
    - By performing PCA with different blocks of color images a color transformation matrix obtained, useful for segmentation.
      - (R+G+B)/3, R-B, (2G-R-B)/2
  - Multispectral, hyperspectral and ultraspectral remote sensing images.
    - Multispectral – 10's of bands
    - Hyperspectral – 100's of bands
    - Ultraspectral - 1000's of bands
    - PCA required to highlight decorrelated information.

Y.I. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation", Computer Graphics and Image Processing, 13, 222-241,

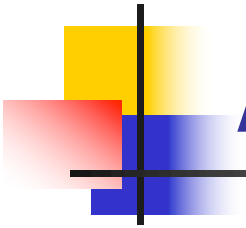
PCA components of a hyperspectral image



After component 20, not much details are available.

Removal of data redundancy.

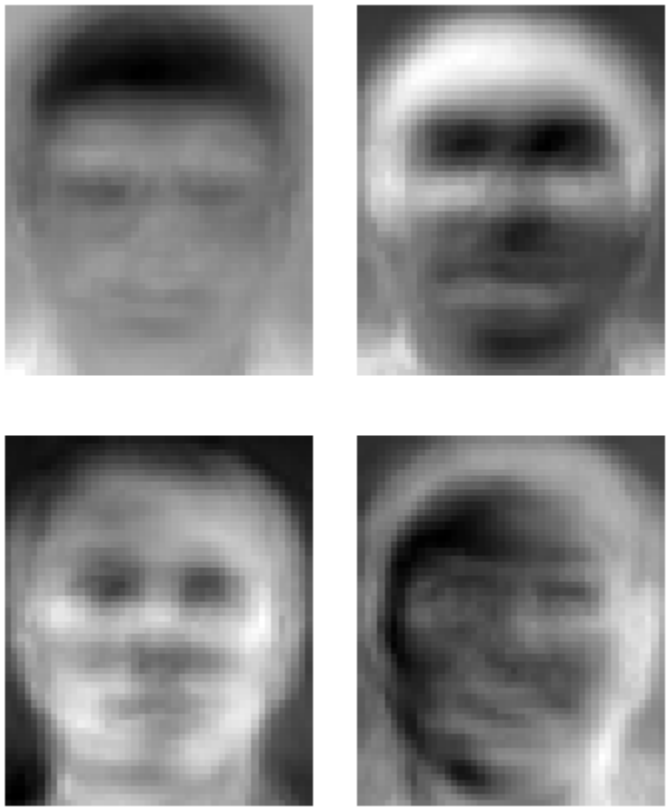
Courtesy: Li et al, "A New Subspace Approach for Supervised Hyperspectral Image Classification", 2011 IEEE International Geoscience and Remote Sensing Symposium.



Application of PCA

- Factor analysis.
  - Highlights decorrelated factors.
    - Useful for classification.
  - For example, eigen faces for representing human faces.
    - Performs PCA on a large set of images of human faces cropped to the same size.
    - Any arbitrary face expressed as linear combination of them.
    - Coefficients of linear combination represent an arbitrary face.

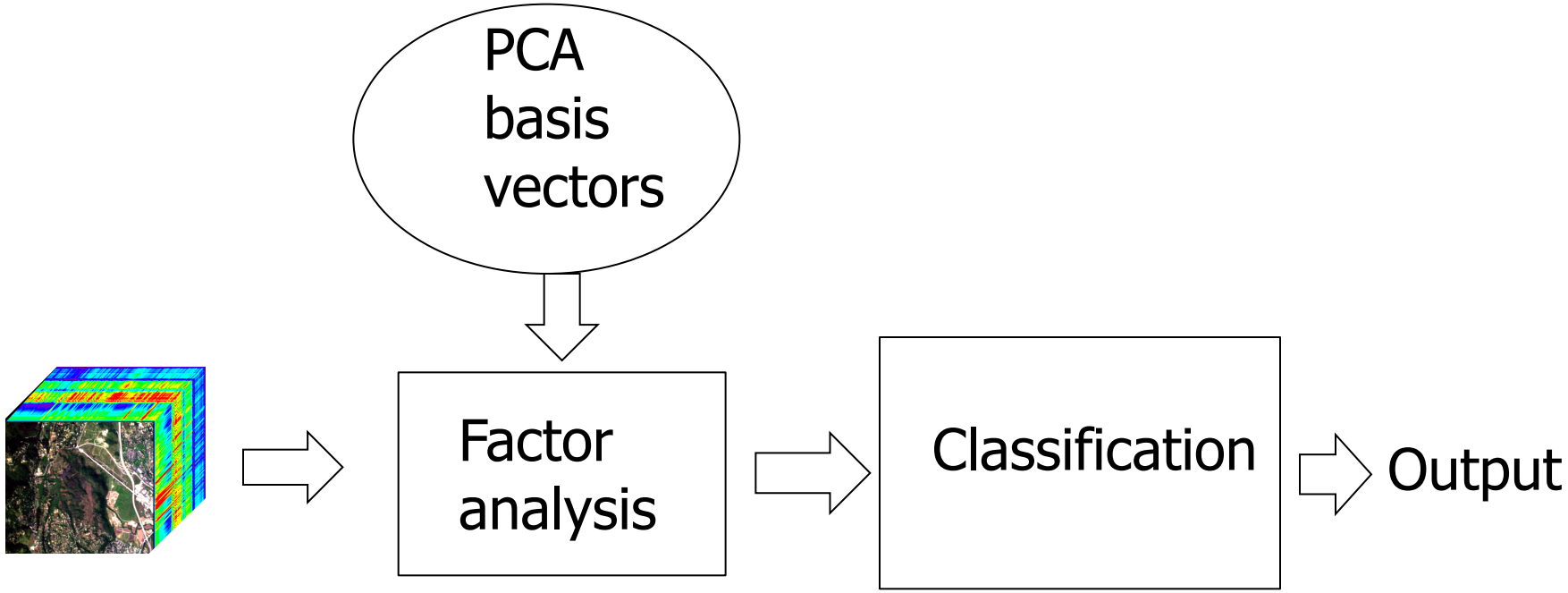
PCA: Eigen faces



http://en.wikipedia.org/wiki/Image:Eigenfaces.png

Application of PCA

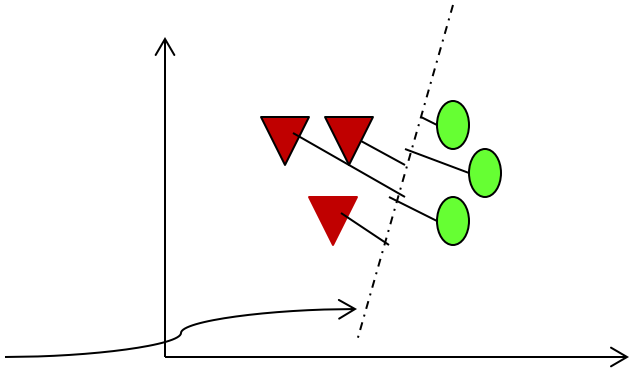
- Classification / High level processing
  - Using the representation derived by factor analysis or component analysis.

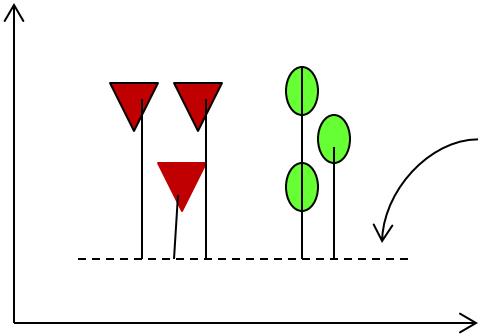


Fisher linear discriminant

- For the purpose of classification, dimensional reduction using PCA may not work.
  - It captures the direction of maximum variance for a data set.
  - For labelled data sets, it does not capture the direction of maximum separation between the groups of data points of differing labels.

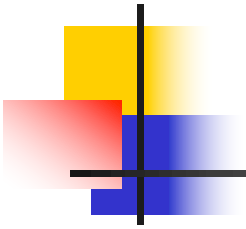
Direction for principal component.





Well separated but not along the direction of principal component.

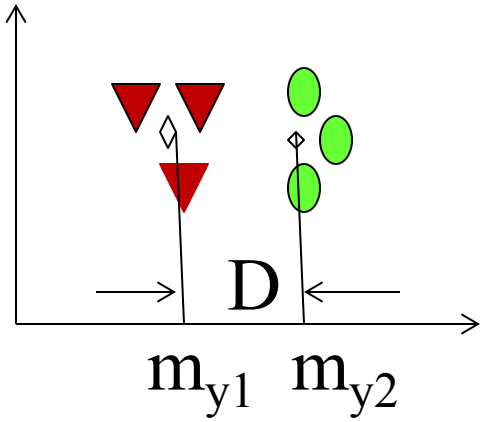
Fisher linear discriminant



- Consider a set of data points $S=\{x_i | \ x_i \text{ in } R^n\}$.
  - N_1 points in class w_1.
  - N_2 points in class w_2.
  - Say, $N_1 + N_2 = N$ (total data points).
- Consider a line with direction u.
- Projection of data x_i on u: $y_i = x_i^T u$
  - One dimensional subspace representing data.

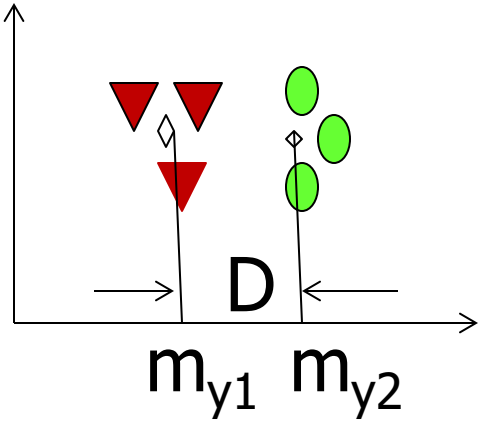
Separation between projected data of different classes

- m_1= mean of data points in w_1.
- m_2= mean of data points in w_2.
- Projection of means:
  - $m_{y1}=m_1^T u$
  - $m_{y2}=m_2^T u$
- A measure of separation:
  - $D=|m_{y1} - m_{y2}|$
  - Does not consider variance of data.



A better measure of separation

- Normalized by a factor proportional to class variances.
- Scatter of data belonging to class C:



$$s^2 = \sum_{y \in C} (y - m_c)^2$$

Class Variance x Number of samples

Mean

- Measure of separation: $J(u) = \frac{D^2}{(s_1^2 + s_2^2)}$

Scatter of class w1

Scatter of class w2

- To obtain u maximizing $J(u)$.

Scatter of projected samples should be small.

Scatter matrix

- Scatter matrix for samples of class C in original space :

$$S_C = \sum_{x \epsilon C} (x - m_C)(x - m_C)^T$$

Within the class Scatter matrix

Scatter matrixes for class w1 and w2.

Within the class scatter matrix: $S_w = S_1 + S_2$

$$s_1^2 = \sum_{y \in W1} (y - m_{y1})^2 \Rightarrow \sum_{x \in W_1} (u^T x - u^T m_1)(u^T x - u^T m_1)^T$$

$$\sum_{x \in W_1} u^T (x - m_1)(x - m_1)^T u$$

$$u^T S_1 u$$

$$u^T \left(\sum_{x \in W_1} (x - m_1)(x - m_1)^T \right) u$$

$$\Rightarrow \quad s_1^2 + s_2^2 = u^T S_w u$$

Between the class scatter matrix

Between the class scatter matrix:

Means of w1 and w2

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$

$$D^2 = (m_{y1} - m_{y2})^2 \Rightarrow (u^T m_1 - u^T m_2)(u^T m_1 - u^T m_2)^T$$

$$\Downarrow$$

$$u^T S_B u \Leftarrow u^T (m_1 - m_2)(m_1 - m_2)^T u$$

Rewriting optimization function

To maximize $J(u) = \dfrac{D^2}{(s_1^2 + s_2^2)}$ $\Rightarrow$ $J(u) = \dfrac{u^T S_B u}{u^T S_W u}$

Solution

To maximize $J(u) = \dfrac{D^2}{(s_1^2 + s_2^2)}$ ⇨ $J(u) = \dfrac{u^T S_B u}{u^T S_W u}$

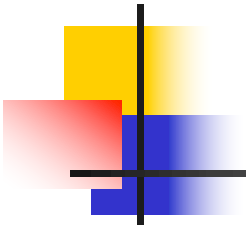
u should be such that $S_W^{-1} S_B u = \lambda u$

Should be invertible — Eigen value problem.

$S_B u$ has the eigen vector along $(m_1 - m_2)$

$$(m_1 - m_2)\underbrace{(m_1 - m_2)^T u}_{k} = k\,(m_1 - m_2)$$

⇨ $u = S_W^{-1}\,(m_1 - m_2)$



Example

- Data points:
  - X_1={(5, 3, 2), (4, 6, 0), (3, -7, 14)}
  - X2={(-2 -5 17), (3 -13 10), (-4 -2 16)}
- Perform LDA and get the optimum direction. Check separability in the line of projection.
- Perform PCA on the whole data set ignoring class information and get the dominant principal direction. Check the separability of projected points on it.

Example (contd.)

- LDA:

$$X1 = \begin{bmatrix} 5 & 4 & 3 \\ 3 & 6 & -7 \\ 2 & 0 & 14 \end{bmatrix} \quad X2 = \begin{bmatrix} -2 & 3 & -4 \\ -5 & -13 & -2 \\ 17 & 10 & 16 \end{bmatrix}$$

$$\text{mean1} = \begin{bmatrix} 4 \\ .67 \\ 5.33 \end{bmatrix} \quad \text{mean2} = \begin{bmatrix} -1 \\ -6.67 \\ 14.33 \end{bmatrix} \quad S1 = \begin{bmatrix} 2 & 10 & -12 \\ 10 & 92.66 & -102.67 \\ -12 & -102.67 & 114.67 \end{bmatrix}$$

$$S1 = (X1 - \text{mean1})(X1 - \text{mean1})^T \quad S2 = \begin{bmatrix} 26 & -41 & -25 \\ -41 & 64.67 & 39.67 \\ -25 & 39.67 & 28.66 \end{bmatrix}$$

SW=S1+S2

$$Sw = \begin{bmatrix} 28 & -31 & -37 \\ -31 & 157.33 & -63 \\ -37 & -63 & 143.33 \end{bmatrix}$$

$$u = SW^{-1}(\text{mean1} - \text{mean2}) \quad u = \begin{bmatrix} 3.2070 \\ -1.1952 \\ 1.2904 \end{bmatrix}$$

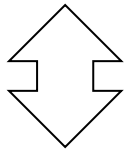
Example (contd.)

- LDA: Separability

$$u = \begin{bmatrix} 3.2070 \\ -1.1952 \\ 1.2904 \end{bmatrix}$$

$$Y1 = X1^T u \qquad Y1 = \begin{bmatrix} 22.2 \\ 19.99 \\ 19.31 \end{bmatrix}$$

$$Y2 = X2^T u \qquad Y2 = \begin{bmatrix} 9.55 \\ 6.99 \\ 5.43 \end{bmatrix}$$



Well separated.

Example (contd.)

- PCA: $X = \begin{bmatrix} 5 & 4 & 3 & -2 & 3 & -4 \\ 3 & 6 & -7 & -5 & -13 & -2 \\ 2 & 0 & 14 & 17 & 10 & 16 \end{bmatrix}$ $\bar{S} = \begin{bmatrix} 1.5 \\ -3 \\ 9.83 \end{bmatrix}$

$$C = \begin{bmatrix} 10.92 & 4 & -17.42 \\ 4 & 39.67 & -27 \\ -17.42 & -27 & 44.14 \end{bmatrix}$$

Eigen values: 72.96, 20.29, 1.47

Eigen vectors:

$$\mathbf{e_1} = \begin{bmatrix} -0.25 \\ -0.63 \\ 0.74 \end{bmatrix} \quad \mathbf{e_2} = \begin{bmatrix} -.52 \\ .73 \\ .44 \end{bmatrix} \quad \mathbf{e_3} = \begin{bmatrix} -0.82 \\ -0.27 \\ -.51 \end{bmatrix}$$

Example (contd.)

- PCA: Separability

$$\mathbf{e_1} = \begin{bmatrix} -0.25 \\ -0.63 \\ 0.74 \end{bmatrix}$$

$$Z1 = X1^T \mathbf{e_1} \qquad Z1 = \begin{bmatrix} -1.65 \\ -4.76 \\ 13.98 \end{bmatrix}$$

Overlapping.

$$Z2 = X2^T \mathbf{e_1} \qquad Z2 = \begin{bmatrix} 16.18 \\ 14.8 \\ 14.05 \end{bmatrix}$$

Sparse Representation: Problem Statement

- Consider a dictionary of N elementary n-D vectors known as atoms.
  - $D=\{d_i|\ i=1,2,..,\mathrm{N}\},\ N>n$
- Consider any arbitrary vector n-D vector X.
- Compute the best linear approximation using a subset of D as basis vectors.
  - The number of atoms should be minimum.
  - Reconstruction should be as close as possible.

$$X \approx \sum_{d_j \epsilon\, S \subset D} a_j d_j \qquad |S| \leq n$$

Exact / Approximate Representation

- Exact reconstruction.

$$X = \sum_{d_j \in S \subset D} a_j d_j \qquad |S| \leq n$$

- Keeping the number of atoms fixed (say, m).

$$X \approx \sum_{d_j \in S \subset D} a_j d_j \qquad |S| \leq m$$

Sparse Approximation

- The problem of approximating a signal with the best linear combination of elements from a redundant dictionary.
  - Optimal / Near optimal representation
  - Fast computation
  - Optimal dictionary (joint optimization problem)

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

Sparse approximation

- Minimize the approximation error using L_2 norm using *m* terms.

Dictionary $D=\{d_i|\ i=1,2,..,\mathrm{N}\},\ N>n$

Optimization task

Linear combination

$$\min_{|S|=m} \min_{\{a_k\}} \left\| X - \sum_{d_{i_k} \epsilon S} a_k d_{i_k} \right\|_2$$

Fixed no. of atoms.

Data vector

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

Reconstruction given S

$$S = \{d_{i_1}, d_{i_2}, \dots, d_{i_m}\} \subset D$$

Construct matrix B from S with the columns as elements of S.

$$B = [d_{i_1} | d_{i_2} | \dots | d_{i_m}]$$

$$\sum_{k=1}^{m} a_k \, d_{i_k}$$

Dimension: $n x m$

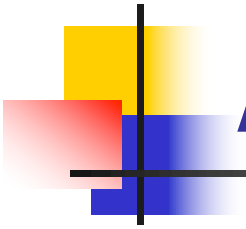
$$X = [d_{i_1} | d_{i_2} | \dots | d_{i_m}] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

$$X = BY$$

Y

How to get the best approximation for m elements ?

$$Y = (B^T B)^{-1} B^T X$$



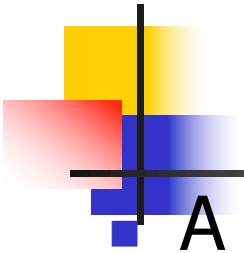
Approaches

- Two major approaches
  - Orthogonal Matching pursuit (OMP)
  - Basis pursuit (BP)

OMP

- An iterative greedy algorithm
  - selects at each step the dictionary element best correlated with the residual part of the input vector.
  - produces a new approximant by projecting the residual onto the dictionary elements that have already been selected.
  - extends the trivial greedy algorithm that succeeds for an orthonormal system.

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.



BP

- A more sophisticated approach that replaces the original sparse approximation problem by a linear programming problem.

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

Matching pursuit

$$D=\{d_i|\ i=1,2,..,\mathrm{N}\},\ N>n$$

$$\min_{|S|=m}\ \min_{\{a_k\}} \left\| X - \sum_{d_{i_k}\epsilon S} a_k d_{i_k} \right\|_2$$

- Minimize the approximation error using L_2 norm using *m* terms.

Residue (r_k) Approximate representation (a_k)

Initialization $r_0=X$ $a_0=0$

At *k* th step:

$$i^* = \operatorname*{argmax}_j \langle r_{k-1}, d_j \rangle$$

$$a_k = a_{k-1} + \langle r_{k-1}, d_{i^*} \rangle d_{i^*}$$

$$r_k = X - a_k \quad \Longleftrightarrow \quad r_k = r_{k-1} - \langle r_{k-1}, d_{i^*} \rangle d_{i^*}$$

MP may select the same atom multiple times.

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

$D=\{d_i|\ i=1,2,..,N\},\ N>n$

$$\min_{|S|=m}\min_{\{a_k\}}\left\|X-\sum_{d_{i_k}\epsilon S}a_k d_{i_k}\right\|_2$$

OMP

- Minimize the approximation error using L_2 norm using m terms.

Initialization $r_0=X$ $a_0=0$ $S_0=\{\ \}$

At k th step:

$$i^*=\underset{j}{\operatorname{argmax}}\langle r_{k-1},d_j\rangle$$

$$S_k=S_{k-1}\cup\{d_{i^*}\}$$

$$\{a_k^*\}=\min_{\{a_k\}}\left\|X-\sum_{d_{i_k}\epsilon S_k}a_k d_{i_k}\right\|_2$$

$$r_k=X-\sum_{d_{i_k}\epsilon S_k}a_k^* d_{i_k}$$

This minimization can be performed incrementally with standard least-squares techniques.

OMP selects an atom only once, as the residual is always orthogonal to selected set.

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

BP

$D=\{d_i|\ i=1,2,..,N\}, N>n$

- Minimize the approximation error using L_1 norm.
  - A convex function, hence can be minimized in polynomial time.

$$\min_{\{a_k\}} \sum_{k=1}^{N} |a_k| \quad \text{subject to} \quad X = \sum_{k=1}^{N} a_k\, d_k$$

There exists different approaches to solve this problem.

Joel A. Tropp, , Greed is Good: Algorithmic Results for Sparse Approximation, IEEE TRANSACTIONS ON INFORMATION THEORY, 50 (10), 2004, 2231-2242.

Ex.1

- Consider the following set of basis vectors.

$$\begin{bmatrix} \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \quad \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{bmatrix}$$

(a) Show that they form an orthonormal set of basis vectors.
(b) Decompose a vector $[1\ 2\ 3]^T$ as a linear combination of the above set.

Ans. 1(a)

$$\begin{bmatrix} \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \quad \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \quad \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{bmatrix}$$

- Take any pair of vectors and perform the dot product, it would be zero.
- Magnitude of these vectors is 1.
- Hence, a set of orthonormal basis vectors.

Ans. 1(b)

$$\begin{bmatrix}1\\2\\3\end{bmatrix} \cdot \begin{bmatrix}\frac{1}{\sqrt{3}}\\-\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{3}}\end{bmatrix} = \frac{2}{\sqrt{3}} \quad \begin{bmatrix}1\\2\\3\end{bmatrix} \cdot \begin{bmatrix}-\frac{1}{\sqrt{6}}\\\frac{1}{\sqrt{6}}\\\frac{2}{\sqrt{6}}\end{bmatrix} = \frac{7}{\sqrt{6}} \quad \begin{bmatrix}1\\2\\3\end{bmatrix} \cdot \begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\end{bmatrix} = \frac{3}{\sqrt{2}}$$

$$\begin{bmatrix}1\\2\\3\end{bmatrix} = \frac{2}{\sqrt{3}}\begin{bmatrix}\frac{1}{\sqrt{3}}\\-\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{3}}\end{bmatrix} + \frac{3}{\sqrt{2}}\begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\end{bmatrix} + \frac{7}{\sqrt{6}}\begin{bmatrix}-\frac{1}{\sqrt{6}}\\\frac{1}{\sqrt{6}}\\\frac{2}{\sqrt{6}}\end{bmatrix}$$

Ex. 2

- Consider a dictionary in a 3-D space consisting of following atoms:

$$\{[\,1\ 1\ 1\,]^T, [1\ -1\ \ 1]^T, [-1\ -1\ \ 1]^T, [-1\ \ 1\ \ 1]^T\}$$

Derive the best representation of the vector [1 2 3] using 2 atoms of the above dictionary following orthogonal matching pursuit (OMP).

Ans.

- 1st selection of an atom:

  $<[1\ 2\ 3]^T, [1\ 1\ 1]^T>$= 6 and maximum.

- Therefore, r_1=$[1\ 2\ 3]^T$-6$[1\ 1\ 1]^T$

  =$[-5\ -4\ -3]^T$

- 2nd selection: $<[-5\ -4\ -3]^T, [-1\ -1\ 1]^T>$=6 and maximum.

- Therefore, a_2= x.$[1\ 1\ 1]^T$+ y.$[-1\ -1\ 1]^T$ LSE solution approximating $[1\ 2\ 3]^T$.

$$\begin{bmatrix}1\\2\\3\end{bmatrix} \approx \overbrace{\begin{bmatrix}1 & -1\\1 & -1\\1 & 1\end{bmatrix}}^{A} \begin{bmatrix}x\\y\end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix}x\\y\end{bmatrix} = (A^TA)^{-1}A^T \begin{bmatrix}1\\2\\3\end{bmatrix}$$

Ans. (contd.)

$$A^T A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

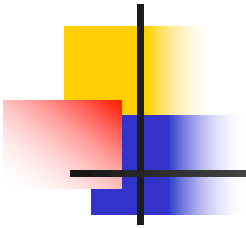
$$(A^T A)^{-1} = \frac{1}{10}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{18}{10} \\ \frac{6}{10} \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \approx \frac{18}{10}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \frac{6}{10}\begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

Learning a dictionary

- Given a set of data points $X=\{x_i | i=1,2,\ldots,N,\ x_i \text{ in } R^n\}$, what should be a dictionary D of K atoms so that it would provide best possible sparse representation for each member of the set.



Motivation

- Use of dictionaries adaptive to specific classes of signals or data of interest.
  - Application specific.
- Dictionaries learned from exemplars with sparse representation property ensured.

Problem statement

$$n \text{ x } N \longrightarrow X = [x_1\ x_2\ \ldots\ x_N]\ \ x_i \epsilon R^n$$

$$n \text{ x } K \longrightarrow D = [d_1\ d_2\ \ldots\ d_K]\ d_i \epsilon R^n$$

$$K \text{ x } N \longrightarrow Y = [y_1\ y_2\ \ldots\ y_N]\ \ y_i \epsilon R^K$$

- To obtain a sparse Y in R^K such that
  - $X=DY$, or $X \sim DY$

Various Sparsity constraints:

$$\min_y \|y\|_0 \ subject\ to\ x = Dy$$

$$\min_y \|y\|_0 \ subject\ to\ \|x - Dy\|_2 \leq \epsilon$$

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

K-SVD: Forming dictionary for sparse representation

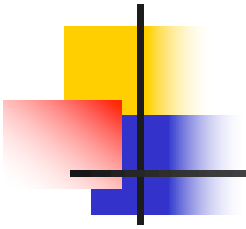
- Given a set of training signals $\{x_i\}_{i=1}^{N}$, to obtain the dictionary of K elements that leads to the best possible representations for each member in this set with strict sparsity constraints.

Various Sparsity constraints:

$$\min_y \|y\|_0 \; subject\ to\ x = Dy$$

$$\min_y \|y\|_0 \; subject\ to\ \|x - Dy\|_2 \leq \epsilon$$

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322



K-SVD

- Generalizes K-means clustering problem.
  1. Choose a dictionary of K atoms.
  2. Obtain sparse representation.
  3. Update dictionary atoms.
  4. Repeat steps 2 and 3 till convergence.
- K-means clustering: Extreme sparse representation of a signal by a single atom only.
- K-SVD: A sparse linear combination of K atoms.

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

K-means clustering

- Given a set of atoms $D=\{d_i\}_1^K$
  - Assign the training examples $\{x_i\}_{i=1}^N$ to their nearest neighbor in *D.*
    - *Usually L_2 norm used.*
  - Given the assignment update *D* to better fit the examples.
    - *Update mean of each partition of assignment.*
- *Start with any initial set of distinct atoms.*

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

K-means clustering: A code book with extreme sparse representation

- The code book: $D=\{d_i\}_1^K=[d_1\ d_2\ \dots\ d_K\ ]_{n\text{x}K}$
- The training examples: $[X\ ]_{n\text{x}N}=\{x_i\}_{i=1}^N$
- Extreme sparse vector: e_j= [0 0 …0 1 0 ... 0]^T
  - Only *j* th term is 1 of K-dim. vector.
- Sparse representation: $Y=[y_1\ y_2\ \dots\ y_N]_{K\text{x}N}$
  - where y_i is one of e_j's .
- Optimization problem: Minimize $||X-DY||^2_F$ Frobenius norm
  - $y_i=e_r$ if $||x_i-d_r||_2$ is minimum among all atoms.
- Update atoms: d_j= Mean ($\{x_i \mid y_i=e_j\}$), for all j.

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

KSVD: Generalization of K-means clustering

- The code book: $D=\{d_i\}_1^K=[d_1\ d_2\ \dots\ d_K\ ]_{n\text{x}K}$
- The training examples: $[X\ ]_{n\text{x}N}=\{x_i\}_{i=1}^N$
- Sparse representation: $Y=[y_1\ y_2\ \dots\ y_N]_{K\text{x}N}$
  - where y_i provides linear combination of maximum T_0 *nonzero* terms.
- Optimization problem:
  - Minimize $||X-DY||^2_F$ subject to $||y_i||_0 \leq T_0$, for all i.

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

Minimize $||X - DY||_F^2$ subject to $||y_i||_0 \leq T_0$, for all i.

Rewriting optimization function

- $y^j{}_T$: j th row of Y.

$$\|X - DY\|_F^2 = \left\|X - \sum_{j=1}^{K} d_j y_T^j\right\|_F^2$$

$$\left\|\left(X - \sum_{j \neq k} d_j y_T^j\right) - d_k y_T^k\right\|_F^2$$

E_k

Consider effect of minimizing w.r.t. k th row of Y associated with code vector d_k keeping other terms fixed.

Perform SVD: $E_k = UDV^T$ and take columns of U and V for max singular value (say $D(1,1)$).

1st Column of U: d_k

$D(1,1)$ x 1st column of V: y_T^k

But the column vector may not be sparse.

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

K-SVD: Enforcing sparsity

$$\left\| \left(X - \sum_{j \neq k} d_j y_T^j \right) - d_k y_T^k \right\|_F^2$$

$y^j{}_T$: j th row of Y.

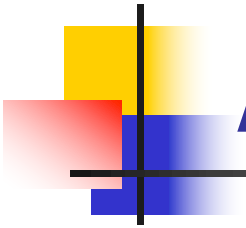
Performing SVD K times for K atoms in each iteration.

- Choose only samples from X which have a nonzero component along d_k.
- Form reduced E_k (denoted E_{kR}) and $y^k{}_T$ by $\mathrm{y^k{}_R}$.
- Perform SVD of E_{kR} to get d_k and $y^k{}_R$.
- Update d_k and $y^k{}_T$.
- Repeat for all d_j's and obtain updated D and Y.
- Repeat till convergence

Michal Aharon, Michael Elad, and Alfred Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006, 4311-4322

The algorithm

- Input: $X=\{x_i \mid i=1,2,\ldots,N\}$, x_i in R^n.
- Output: $D=\{d_i|i=1,2,..,K\}$, d_i in R^n. $Y=\{y_i|i=1,2,..,N\}$, y_i in R^K.
- Form an initial dictionary of K atoms.
  - K-means clustering.
- Obtain an initial sparse representation Y using any pursuit algorithm.
  - OMP
- Iterate for updating j th atom and sparse representation associated with this atom (j th row of Y).

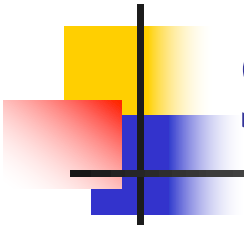


Applications

- Compression.
- Denoising
- Deblurring
- Super-resolution
  - Mapping of learned dictionaries
- Inpainting

Summary

- Dimension reduction techniques
  - Principal Component Analysis
    - Data represented in minimal subspace.
    - Involves coordinate transformation.
    - Chooses a direction maximizing variance of dominant component.
    - Decorrelating data across different dimensions.
  - Fisher's Linear Discriminant
    - Data projected on an 1-D subspace.
    - Appropriate for classification using a linear discriminant function.



Summary

- Sparse representation
  - Pursuit algorithms
    - Matching pursuit
    - Orthogonal matching pursuit.
    - Basis pursuit
- Dictionary learning and sparse representation.
  - K- SVD