ANLY-601
Advanced Pattern Recognition

Spring 2018

L14 – Principal Components

GEORGETOWN UNIVERSITY

Feature Extraction for Signal Representation Principal Components

- Principal component analysis is a classical statistical technique that eliminates correlation among variables and is can be used to reduce data dimensionality for
  - Visualization
  - Data compression (transform coding)
  - Dimension reduction prior to classification
- The method is covariance-based and doesn't "know" about higher order correlation in the data.
- We'll look at classic treatment, and a probabilistic interpretation.

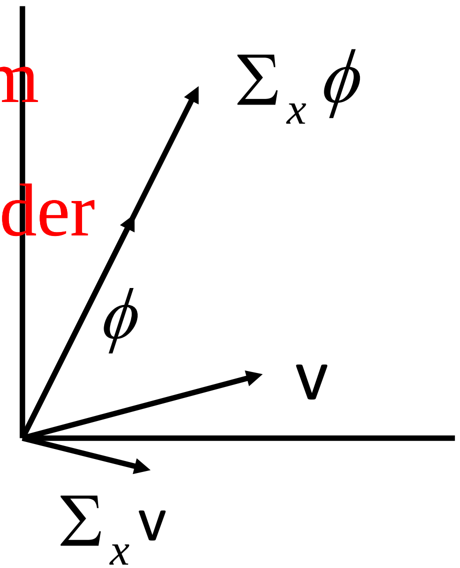
Eigenvectors and Eigenvalues of Covariance

The covariance matrix Σ_x of n-dimensional random variables *x* is a real, symmetric matrix. So there's a complete set of orthonormal eigenvectors ϕ that form a spanning set (complete basis) for R^n

$$\Sigma_x \;\phi_i \;=\; \lambda_i \;\phi_i, \quad i = 1 \ldots n$$

$$\phi_i^T \,\phi_j \,=\, \delta_{ij} \qquad \textit{orthonormal}$$

$$\sum_{i=1}^{n} \phi_i \,\phi_i^T \,=\, I \qquad \textit{complete}$$



Positive Semi-Definite Covariance

The covariance is positive semi-definite

$$V^T \Sigma_x V / \|V\|^2 = E\left[V^T (x - E[x])(x - E[x])^T V \right] / \|V\|^2$$

$$= E\left[\left(V^T (x - E[x]) \right)^2 \right] / \|V\|^2 = \sigma_V^2 \geq 0$$

Since the ϕ form a complete set, we can expand V as

$$V = \sum_{i=1}^{n} c_i \, \phi_i, \quad with \quad c_i = \phi_i^T V$$

then the inner product above becomes

$$V^T \Sigma_x V = \sum_{i,j} c_i \, c_j \, \phi_i^T \Sigma_x \phi_j = \sum_{i,j} c_i \, c_j \, \lambda_j \, \phi_i^T \phi_j$$

$$= \sum_{i,j} c_i \, c_j \, \lambda_j \, \delta_{ij} = \sum_i c_i^2 \, \lambda_i \geq 0 .$$ Hence $\lambda_i \geq 0$.

Positive Semi-Definite

The two definitions of positive semi-definite

$$V^T \Sigma_x V / \|V\|^2 = \sigma_V^2 \geq 0$$

and

$$\lambda_i \geq 0$$

are equivalent.

Usually, unless there is a strict linear constraint in the system, the covariance is positive-definite $\lambda_i > 0$

Zero Eigenvalues, Singularities, and Correlations

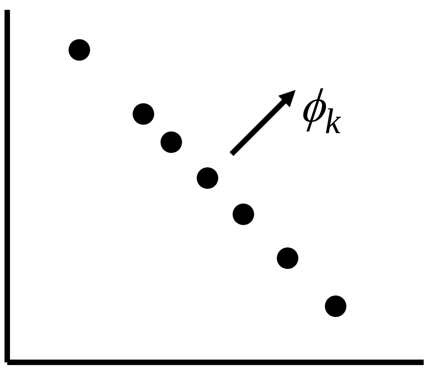
To get a zero eigenvalue, it must be true that the corresponding eigenvector is in the null space

$$\Sigma_x \ \phi_k = 0$$

and hence the covariance is <u>singular</u>. A reflection of this is that the determinant vanishes

$$|\Sigma_x| = \prod_{i=1}^{n} \lambda_i = 0$$

Geometrically, we get a singular covariance if the data has <u>zero spread</u> along ϕ_k



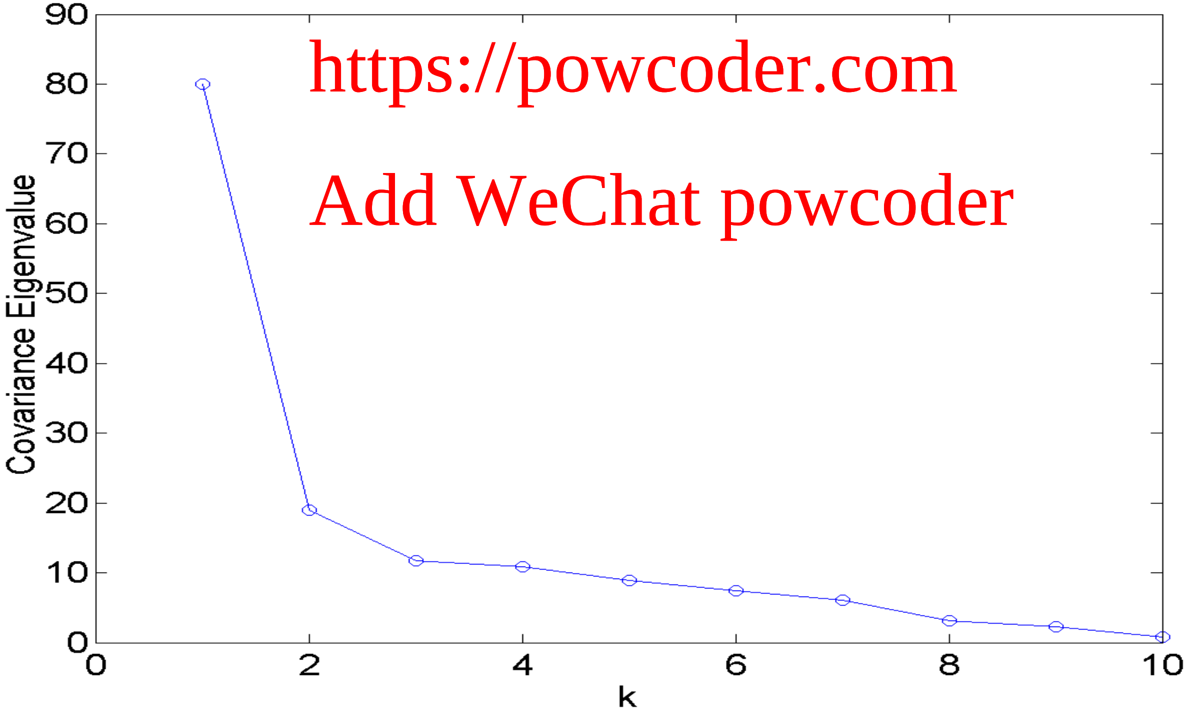
Consequently, some components of *x* are <u>perfectly</u> correlated, and the data are on an *n-1* dimensional (or smaller) sub-space.

GEORGETOWN UNIVERSITY

Correlations in Real Data

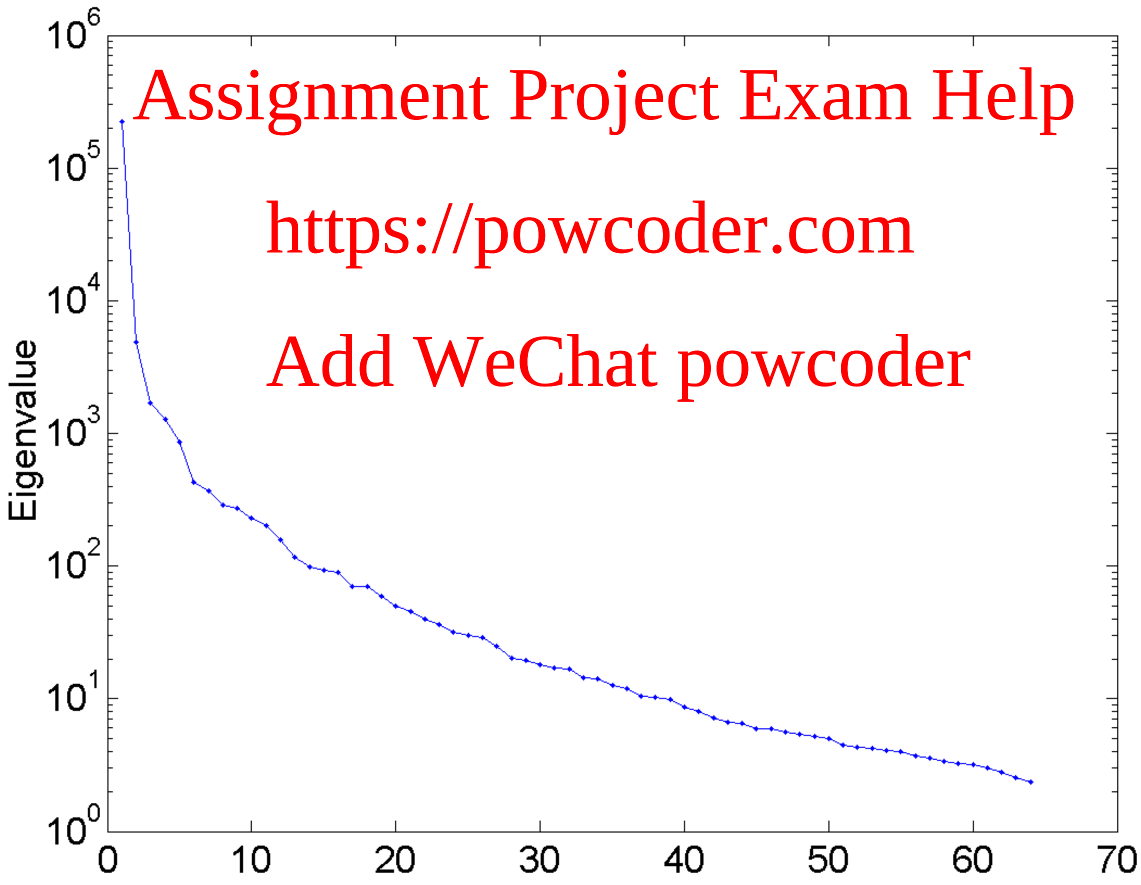
In real data, components are almost never perfectly correlated. However, high dimensional data often has many strongly correlated components. This is reflected in the eigenvalue *spectrum* of the covariance having some very small (but non-zero) eigenvalues.

Here's the eigenvalue spectrum of the 'm' phoneme data.



Correlations in Real Data

Here's the covariance eigenspectrum for 8x8 blocks (64-dim vectors) of pixel intensities from a grayscale image (4096 vectors) – note log scale!



GEORGETOWN UNIVERSITY

PCA

Transform to basis of eigenvectors:

$$y_i = \phi_i^T \ x \qquad or \qquad y = \Phi \ x \qquad where \quad \Phi = \begin{pmatrix} (& \phi_1 &) \\ (& \phi_2 &) \\ & \vdots & \end{pmatrix}$$

The y's are the principal components. Note that $\Phi \ \Phi^T = I$ (the identity matrix) and

$$\Sigma_x \quad \Phi^T \ = \ \Phi^T \ \Lambda \ , \qquad where \quad \Lambda \equiv \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \end{pmatrix}$$

Then the covariance of y is

$$\Sigma_y = E\left[(y - E[y])(y - E[y])^T\right]$$

$$= \Phi \ \Sigma_x \ \Phi^T = \Phi \Phi^T \Lambda = I \ \Lambda$$

Since cov(y) is diagonal, the components of y are uncorrelated.

Maximal Variance Directions

The variance of the data along an arbitrary direction *V* (with |V|=1) is

$$V^T \Sigma_x V = \sigma_V^2$$

Let's find *V* that maximizes this subject to the constraint that *V* is unit norm. Construct the Lagrangian

$$L = V^T \Sigma_x V - \gamma (V^T V - 1)$$

And find it's critical points by setting the derivative with respect to the components of *V* equal to 0

Maximal Variance Directions

Let's find *V* that maximizes this subject to the constraint that *V* is unit norm. Construct the Lagrangian

$$L = V^T \Sigma_x V - \gamma\,(V^T V - 1)$$

And find it's critical points by setting the derivative with respect to the components of *V* equal to 0

$$\frac{\partial L}{\partial V_i} = \frac{\partial}{\partial V_i}\left(\sum_{k,l} V_k \Sigma_{x_{kl}} V_l - \gamma \left(\sum_k V_k V_k - 1 \right) \right)$$

$$= \sum_l \Sigma_{x_{il}} V_l + \sum_k V_k \Sigma_{xki} - 2\gamma V_i$$

$$= 2\left(\Sigma_x V\right)_i - 2\,\gamma V_i = 0$$

$$So \quad \boxed{\Sigma_x V = \gamma V}$$

Maximal Variance

$$We \;\; have \quad \Sigma_x V = \gamma V$$

The data has maximal variance along an eigenvector of the covariance. Which one?

$$\sigma_{\phi_i}^2 = \phi_i^T \Sigma_x \phi_i = \lambda_i \phi_i^T \phi_i = \lambda_i$$

The variance is maximum along the eigenvector corresponding to the largest eigenvalue.

Typically we order the indices such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

Deflation

So the leading variance direction is ϕ_1. Let's project the data orthogonal to this eigenvector with a projection operator

$$x' = (1 - \phi_1 \phi_1^T)\, x \equiv \Pi_1 x$$

where the projection operator Π_1 satisfies $\Pi_1^T = \Pi_1, \quad \Pi_1 \Pi_1 = \Pi_1$

The x' have covariance

$$\Sigma_{x'} = \Pi_1 \Sigma_x \Pi_1$$

with eigenvalues and eigenvectors

$$\Sigma_{x'} \phi_i = \lambda_i \phi_i \quad i > 1$$

The highest variance direction of the new data *x'* is the old second eigenvector ϕ_2 … etc.

Dimension Reduction

For visualization and as a precursor to compression or other processing, PCA is useful as a dimension reduction technique.

$$x(m) = \sum_{i=1}^{m} \phi_i \ \phi_i^T \ x \quad \textit{while the original } x \textit{ is} \quad x = \sum_{i=1}^{n} \phi_i \ \phi_i^T \ x$$

Approximate the vector *x* by its expansion in terms of the leading m<n eigenvectors (translate the origin so that E[x]=0 for simplicity)

Dimension Reduction

The expected squared distance between *x* and *x(m)* is

$$E\left[\left\|x-x(m)\right\|^2\right]=E\left[\left\|\sum_{i=m+1}^{n}\phi_i\ \ \phi_i^T\ \ x\right\|^2\right]=E\left[\left(\sum_{i=m+1}\phi_i\ \ (\phi_i^T\ \ x)\right)^T\left(\sum_{j=m+1}\phi_j\ \ (\phi_j^T\ \ x)\right)\right]$$

$$= E\left[\sum_{i=m+1}^{n} y_i^2\right] = \sum_{i=m+1}^{n}\lambda_i$$

the sum of the discarded eigenvalues !

Example --- Eigenfaces

Developed for face characterization --- M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for characterization of human faces. IEEE Trans. Patt. Anal. Mach. Int., 12, 103-108, 1990.

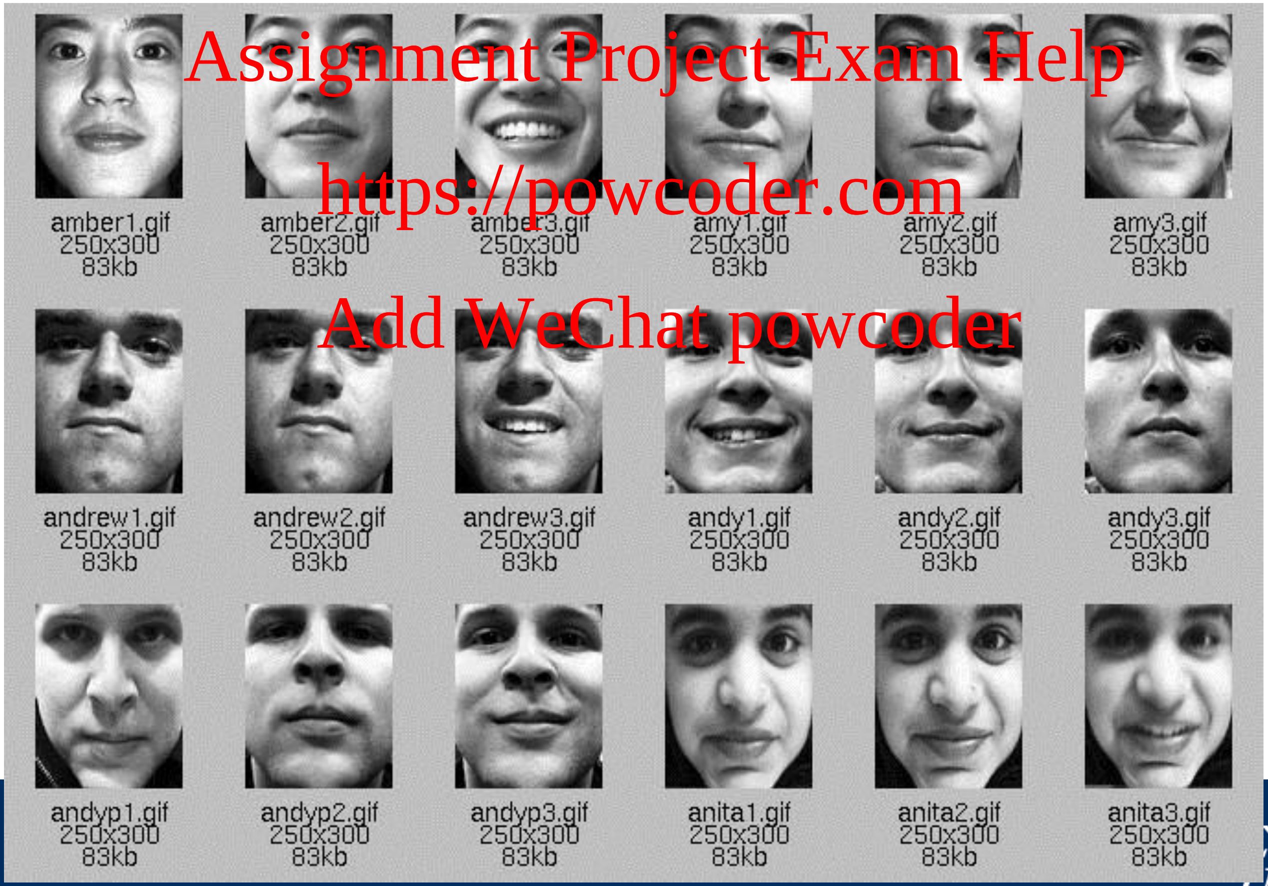
First applied to face recognition --- M. Turk and Alex Pentland. Eigenfaces for Recognition. J. Cog. Neurosci. 3, 71-86, 1991.

Still a standard method for faces

Eigenfaces

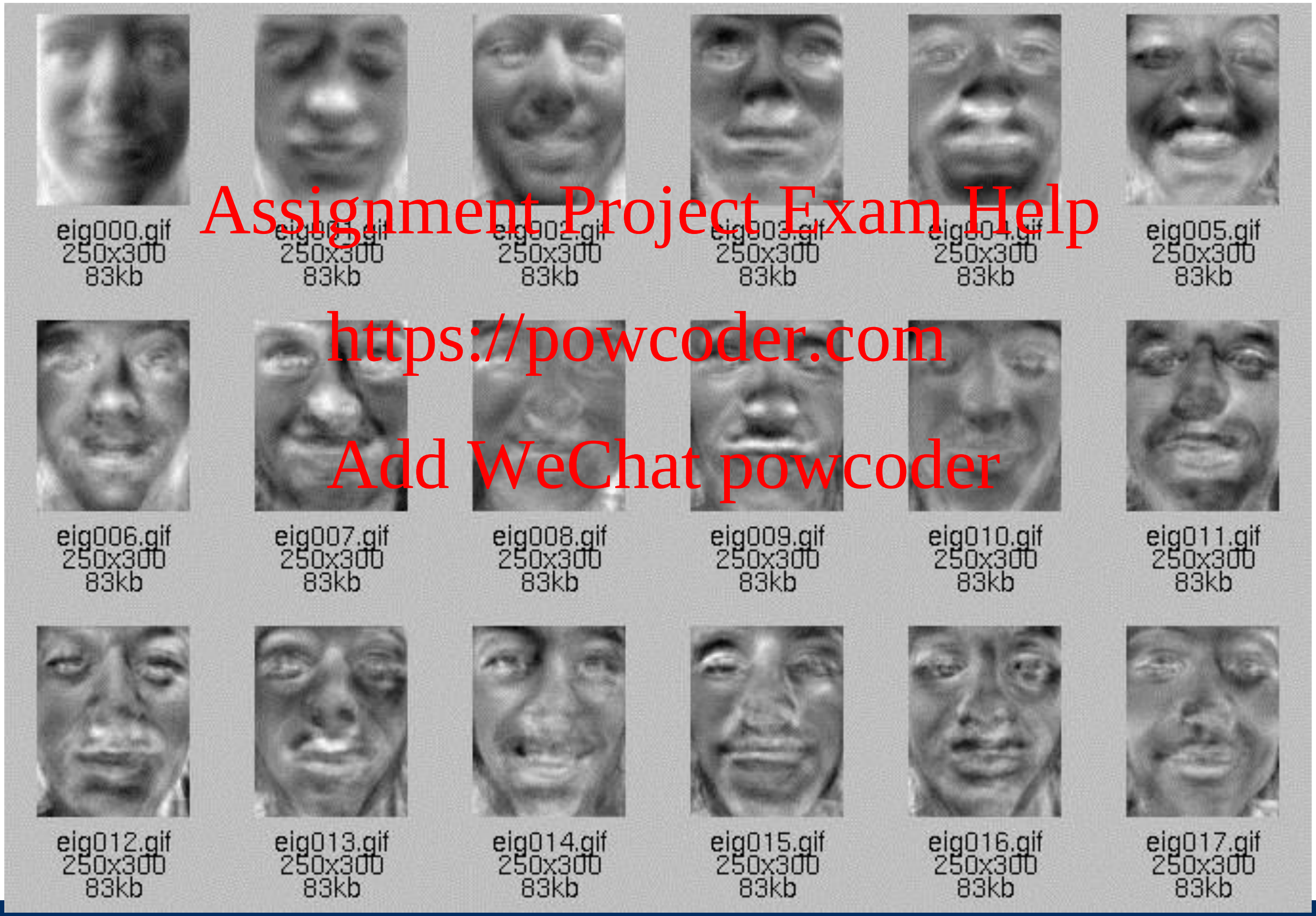
Examples from
http://www.owlnet.rice.edu/~elec301/Projects99/faces/results.html

Original Images: greyscale 250x300=75,000 pixel images, 31 subjects, 86 images (only 86 non-zero eigenvalues --- why?)



Leading Eigenvectors (Eigenfaces)

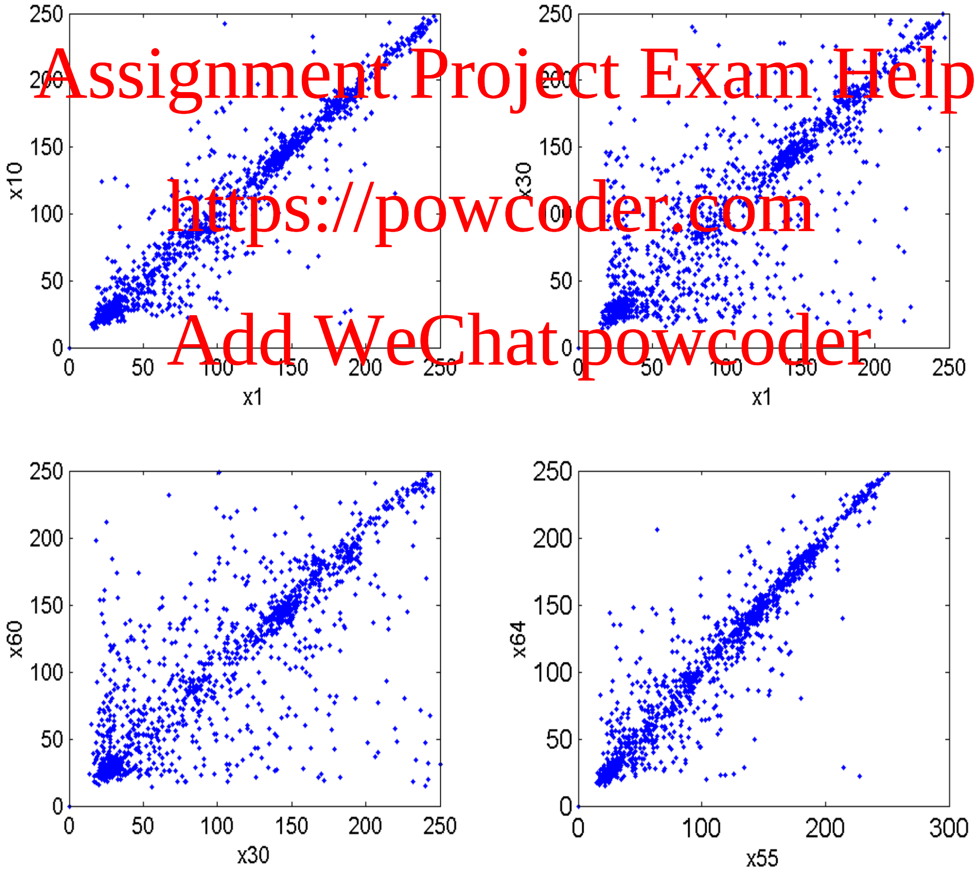
18 eigenvectors corresponding to largest eigenvalues



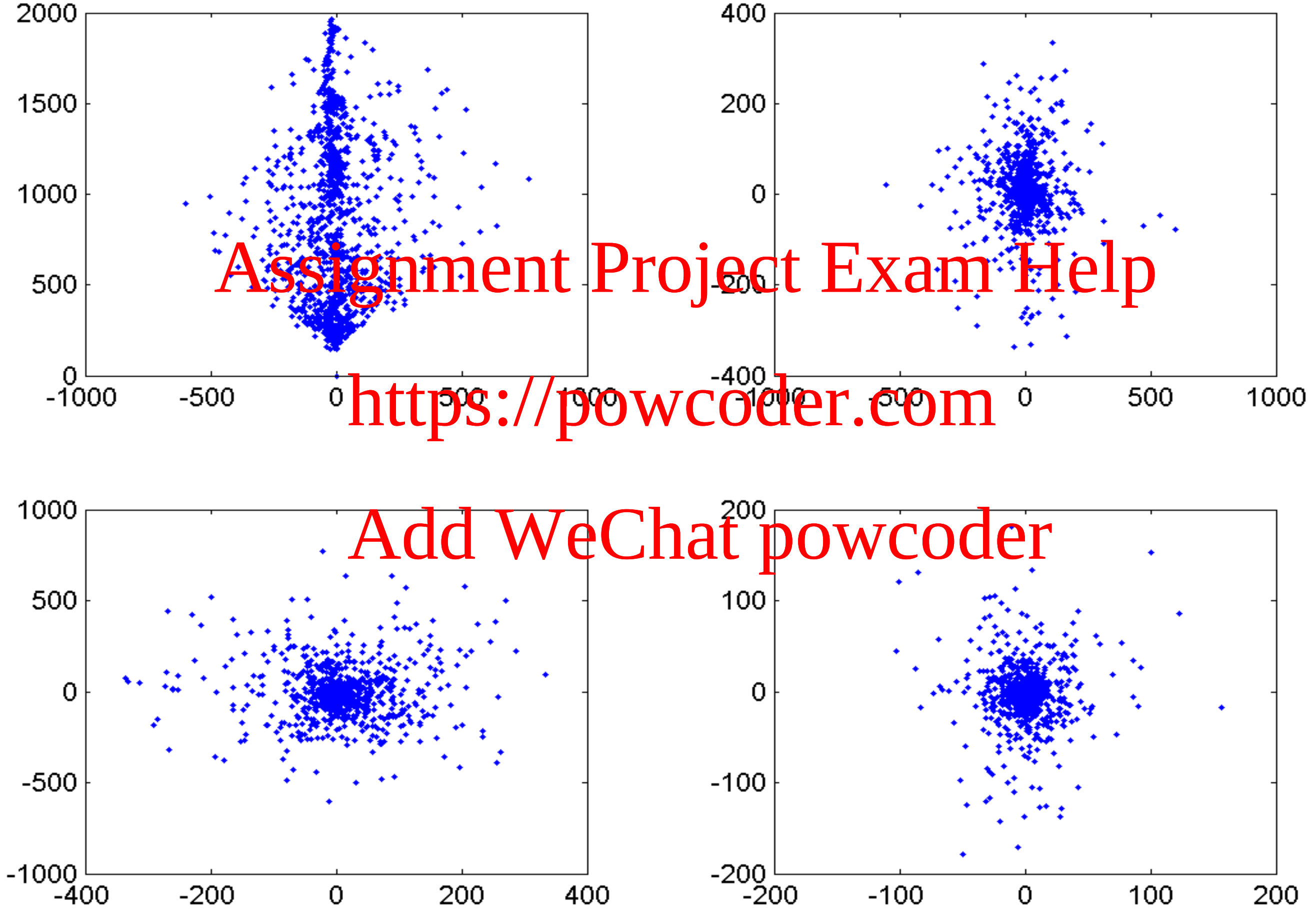
GEORGETOWN UNIVERSITY

Visualization

Since the principal components – y_i are uncorrelated, they can provide a better basis for visualization than the old coordinates which can be strongly correlated.



Visualization PC



GEORGETOWN UNIVERSITY

Estimation from Data

Estimating PCs from data is straightforward. Find the sample covariance, find its eigenvectors and eigenvalues, and project the data onto the eigenvectors.

What's the bias and variance of the estimates of the eigenvectors and eigenvalues?

Fukunaga gives results for Gaussian data – the restriction arises because the calculation requires fourth moments of the data. For Gaussian data, these are simply related to the covariance.

$$E_D\left[\hat{\phi}_k\right] = \phi_k + O(1/N) \qquad \mathrm{var}_D(\hat{\lambda}_k) \cong \tfrac{2}{N}\lambda_k^2$$

$$E_D\left[\hat{\lambda}_k\right] = \lambda_k + O(1/N) \qquad E_D\left[\left\|\hat{\phi}_i - \phi_i\right\|^2\right] \cong \tfrac{1}{N}\sum_{j\neq i}\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}$$

GEORGETOWN UNIVERSITY

Probabilities and Principal Components

Principal components rely on computation of the data's mean and covariance.

Since they only depend on second moments, we might suppose that PCA has something to do with Gaussian distributions.

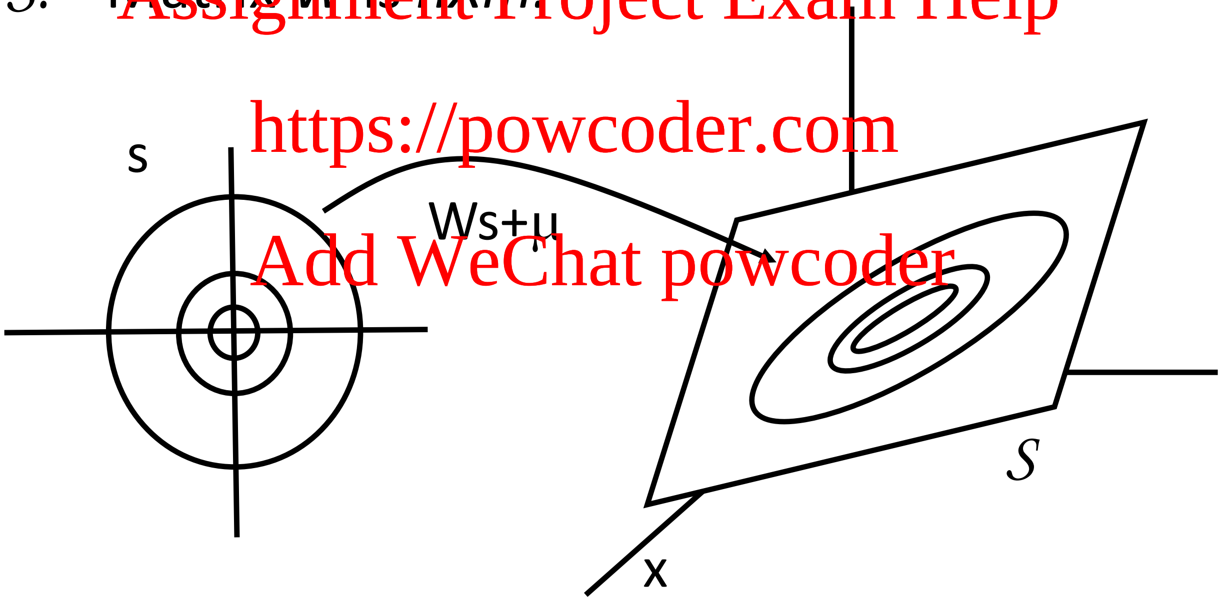
Here's a sense in which it does.

PCA Probability Model

Latent data s – *dim m,* Gaussian, $\mathcal{N}(0, I)$

Observed data space x – *dim n>m.*

Map from S into x *:* W s + μ has an image **in** the m-dimensional hyperplane $\mathcal{S}$. Matrix W is *nXm.*



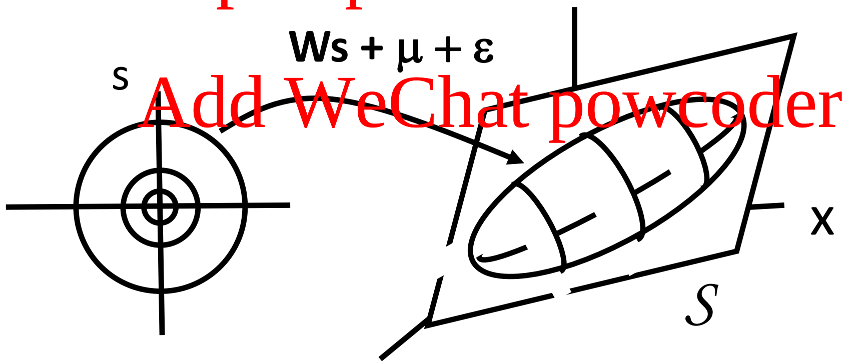
A. Basilevsky. *Statistical Factor Analysis and Related Methods,* Wiley 1994.

Probabilistic PCA

Next add spherical Gaussian noise ε with density $N(0, \sigma^2 I)$. The resulting

$$x = W s + \mu + \varepsilon$$

also has Gaussian density, and occupies R^n rather than the subspace $\mathcal{S}$.



Probabilistic PCA

To see that *x* is Gaussian, note that

$$p(x) = \int p(x \mid s)\ p(s)\ ds$$

where $p(x|s)$ is normal with mean $Ws + \mu$ and covariance σ_ε^2,
and $p(s)$ is normal with mean zero and covariance I

where *p(x | s)* is normal with mean $Ws + \mu$ and covariance $\sigma^2{}_\varepsilon$
and *p(s)* is normal with mean zero and covariance *I*

It's easy to show that *x* has mean μ, and covariance

$$\Sigma_x = W W^T + \sigma_\varepsilon^2\ I$$

Fitting Model Parameters to Data
Maximum Likelihood Estimates

- MLE of μ is the data mean.
- MLE of W is

$$W = U\,\Gamma^{1/2}$$

where U are the leading m eigenvectors of the data covariance, Γ is a diagonal matrix with entries

$$\Gamma_i = \lambda_i - \sigma_\varepsilon^2, \quad i = 1 \ldots m$$

where λ_i are the leading m eigenvalues of the data covariance.

- MLE estimate of the noise variance is the average of the trailing eigenvalues of the data covariance.

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-m} \sum_{i=m+1}^{n} \lambda_i$$

Probabilistic Interpretation of Projection onto the Hyperplane $\mathcal{S}$

- Projection of *x* back onto $\mathcal{S}$ corresponds to a two-step process
  - Given *x*, find the most likely *s* that generated it

$$\hat{s}(x) = \arg\max_{s} \; p(x \mid s)$$

  - Then map this s onto the hyperplane by

$$W\,\hat{s}(x) + \mu$$

This corresponds *exactly* to the projection of *x* onto the space spanned by the leading *m* eigenvectors of the data covariance.

PCA and SVD

PCA and singular value decomposition (SVD) are often spoken of as if they are identical. Here's the connection.

The SVD theorem says that any matrix X_{Nxn} can be decomposed as the product

$$X = U\ S\ V^T$$

where U_{Nn} is column – orthogonal

S_{nn} is diagonal –– the 'singular values' of X

V_{nn} is orthogonal

PCA and SVD

Suppose we have a collection of N vectors, each of dimension n, with zero mean (not absolutely necessary).

Load these vectors *row-wise* into the matrix X

$$X = \begin{bmatrix} (& x^{(1)} &) \\ (& x^{(2)} &) \\ & \vdots & \\ (& x^{(N)} &) \end{bmatrix}$$

PCA and SVD

- Next, SHOW THAT

$$X^T X = N \hat{\Sigma}_x$$

- Substitute the SVD of X into the above

$$X^T X = V S U^T U S V^T = V S^2 V^T = N \hat{\Sigma}_x$$

rearrange to give $V \frac{1}{N} S^2 = \hat{\Sigma}_x V$

Compare with the equation in the middle of p3 to conclude that

The columns of V are the eigenvectors of $\hat{\Sigma}_x$

The eigenvalues of $\hat{\Sigma}_x$ are $\frac{1}{N} S^2$

PCA and SVD: Numerical Considerations

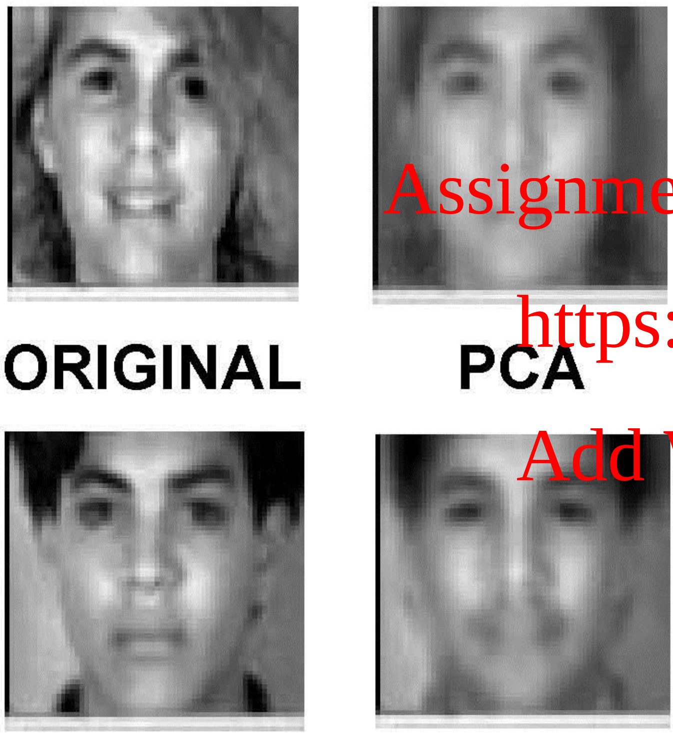
- Whenever you do algebraic operations on a computer you have to worry about truncation error and its propagation.

  If the ratio of the smallest eigenvalue to the largest eigenvalue (or singular values) of a matrix approaches the machine precision, you're sure to be in trouble. This ratio is called the *condition number* of the matrix.

- Notice that the eigenvalues of the covariance Σ are S^2. So the condition number of the covariance matrix is the square of the condition number of *X*.

  Numerically, you're better off using SVD than computing the covariance and diagonalizing it directly.

Application: Dimensionality Reduction of Faces

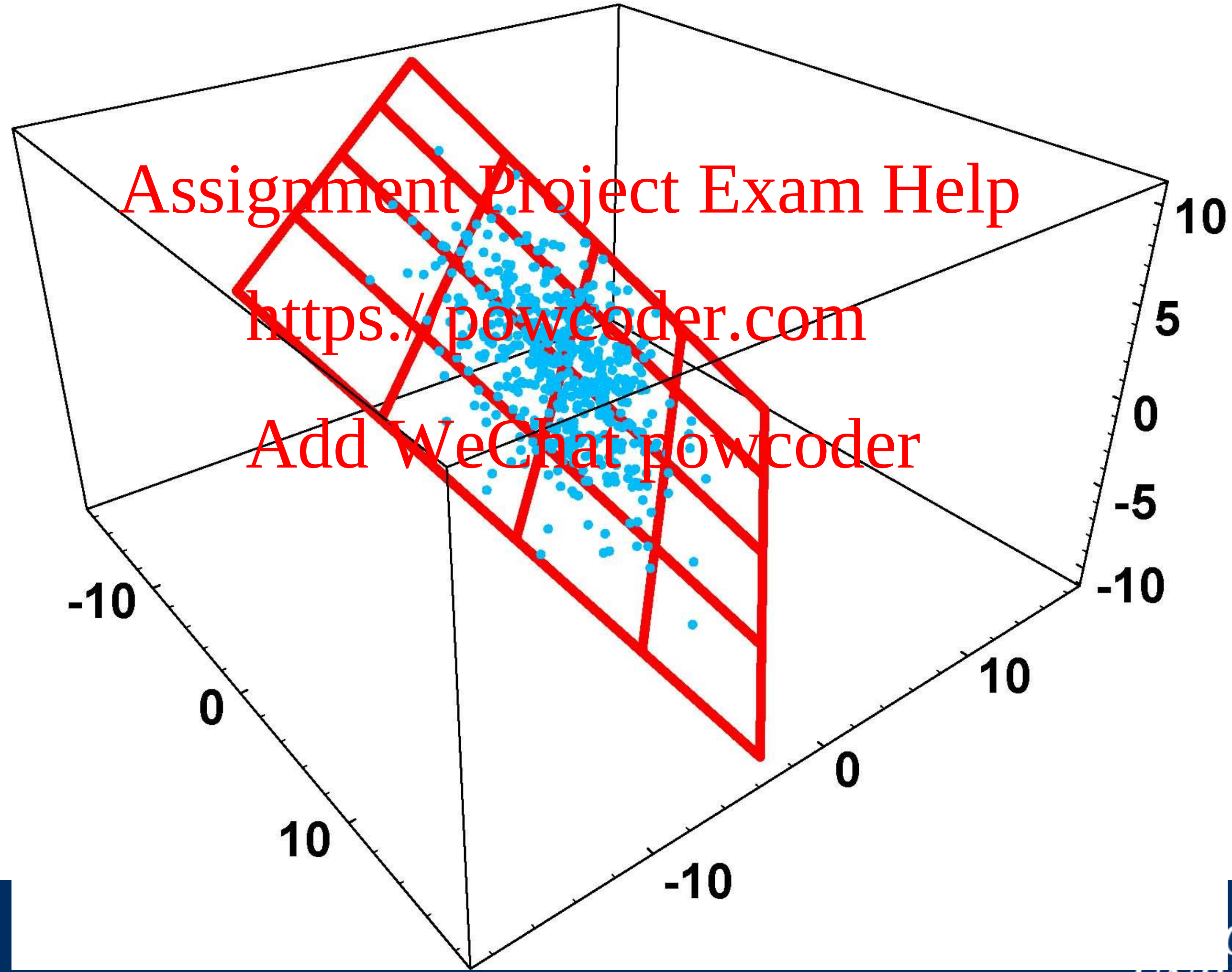


64x64 pixel 8-bit grayscale images (64x64=4096 dimensional points).

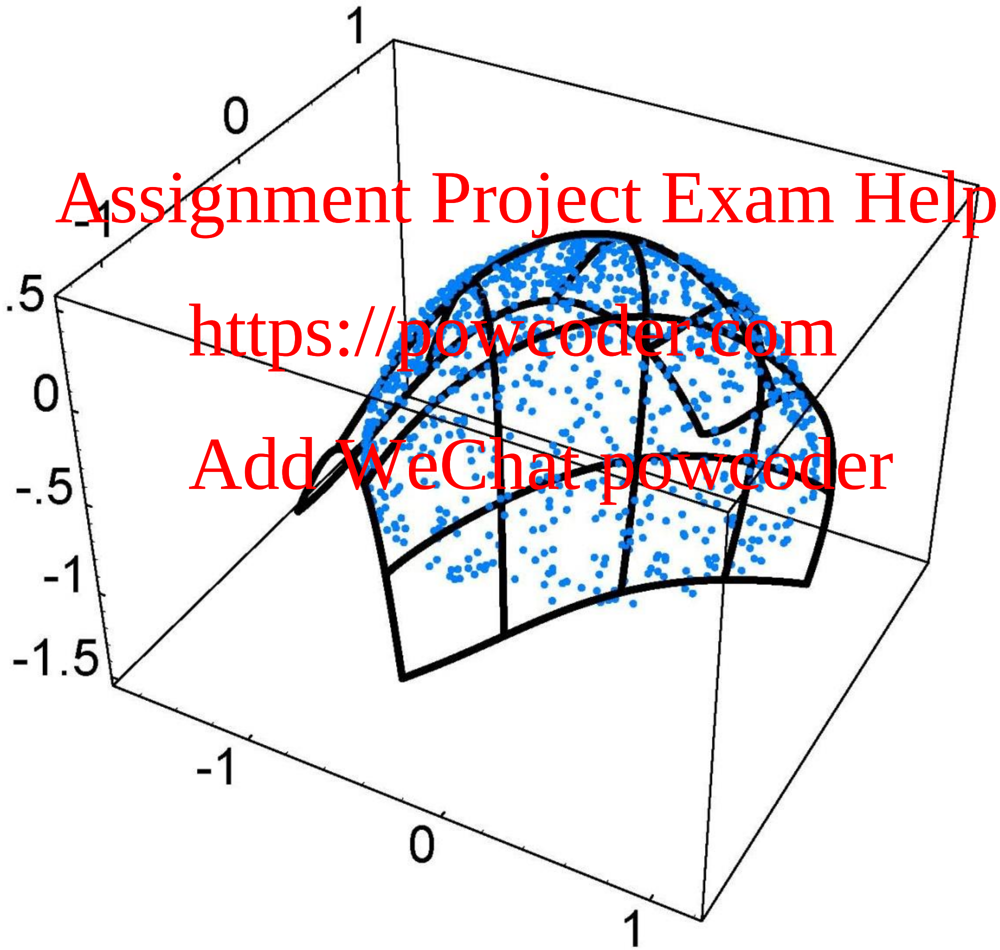
20 different subjects, 160 total images.

PCA reduction to <u>5 dimensions</u>.

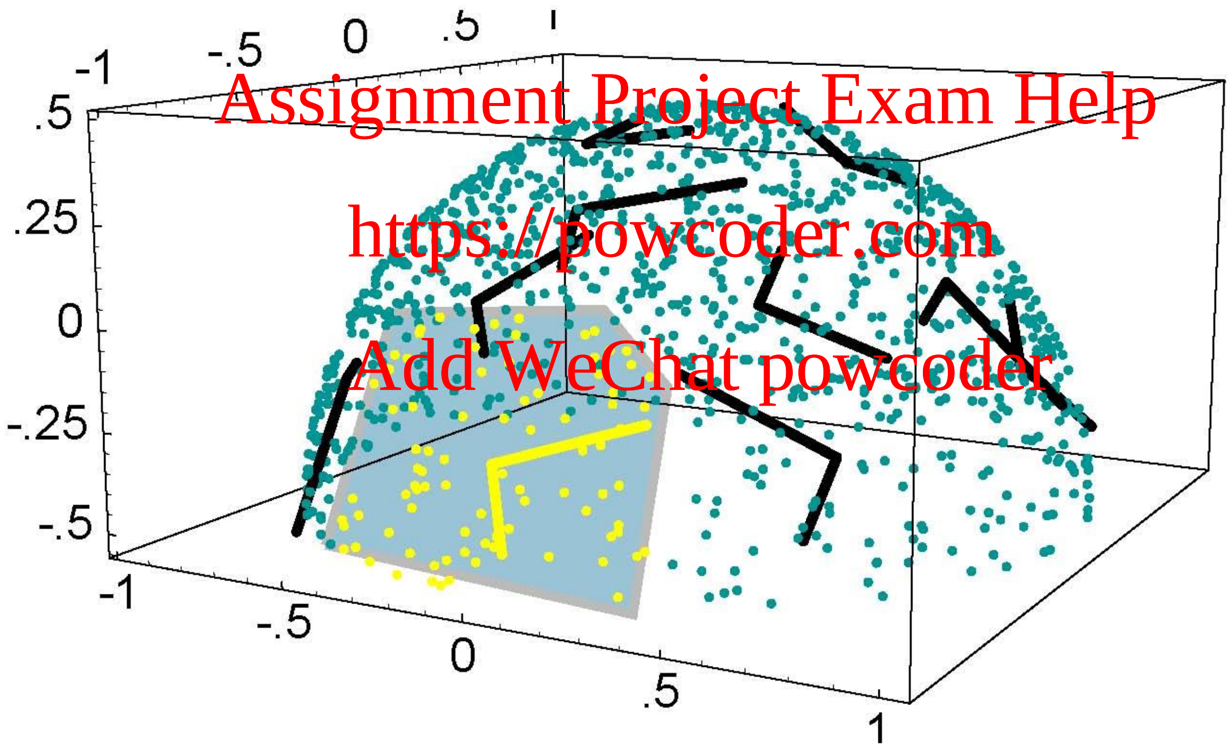
PCA – Linear Subspace Model



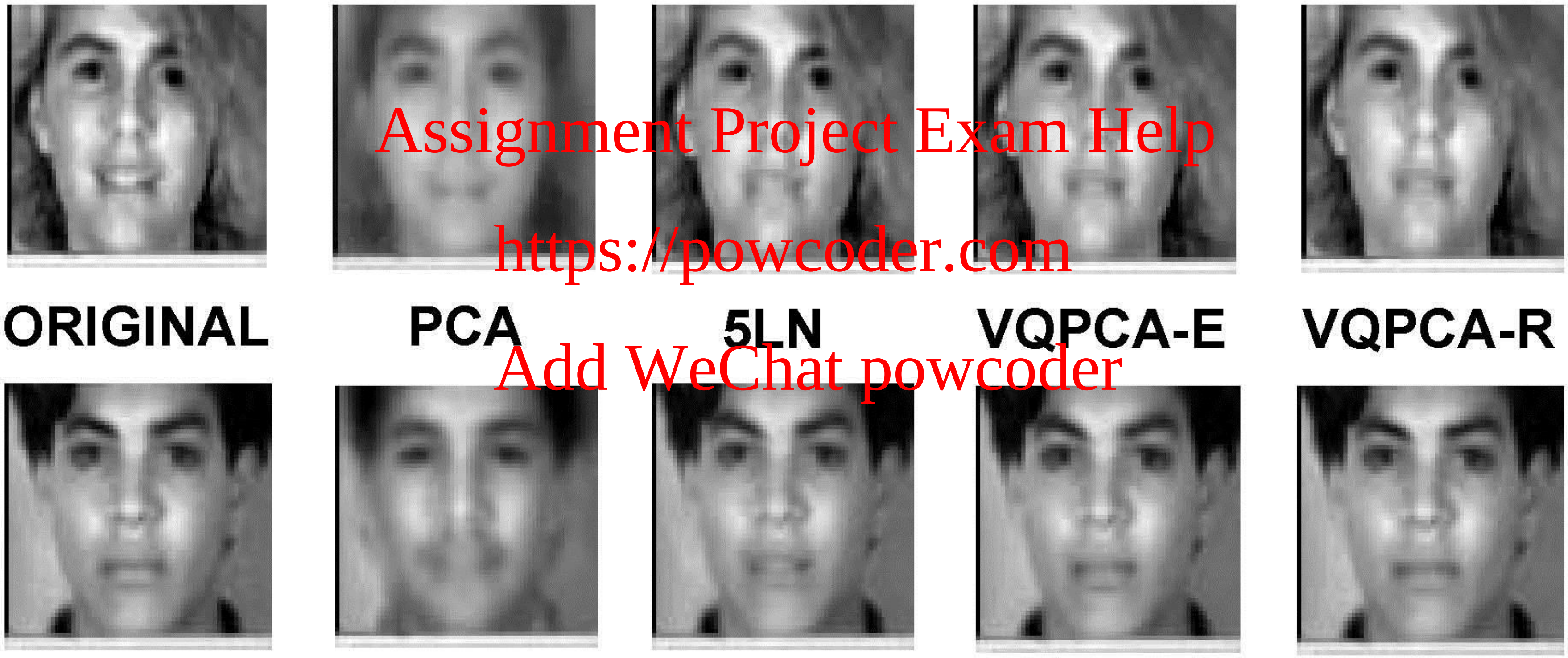
Nonlinear PCA Curved Manifold Model
(Deep learning --- cf 1995)



Approximating the Manifold By Local Tangent Spaces



PCA & Nonlinear Dimension Reduction



1789

GEORGETOWN UNIVERSITY