On Principal Component Analysis

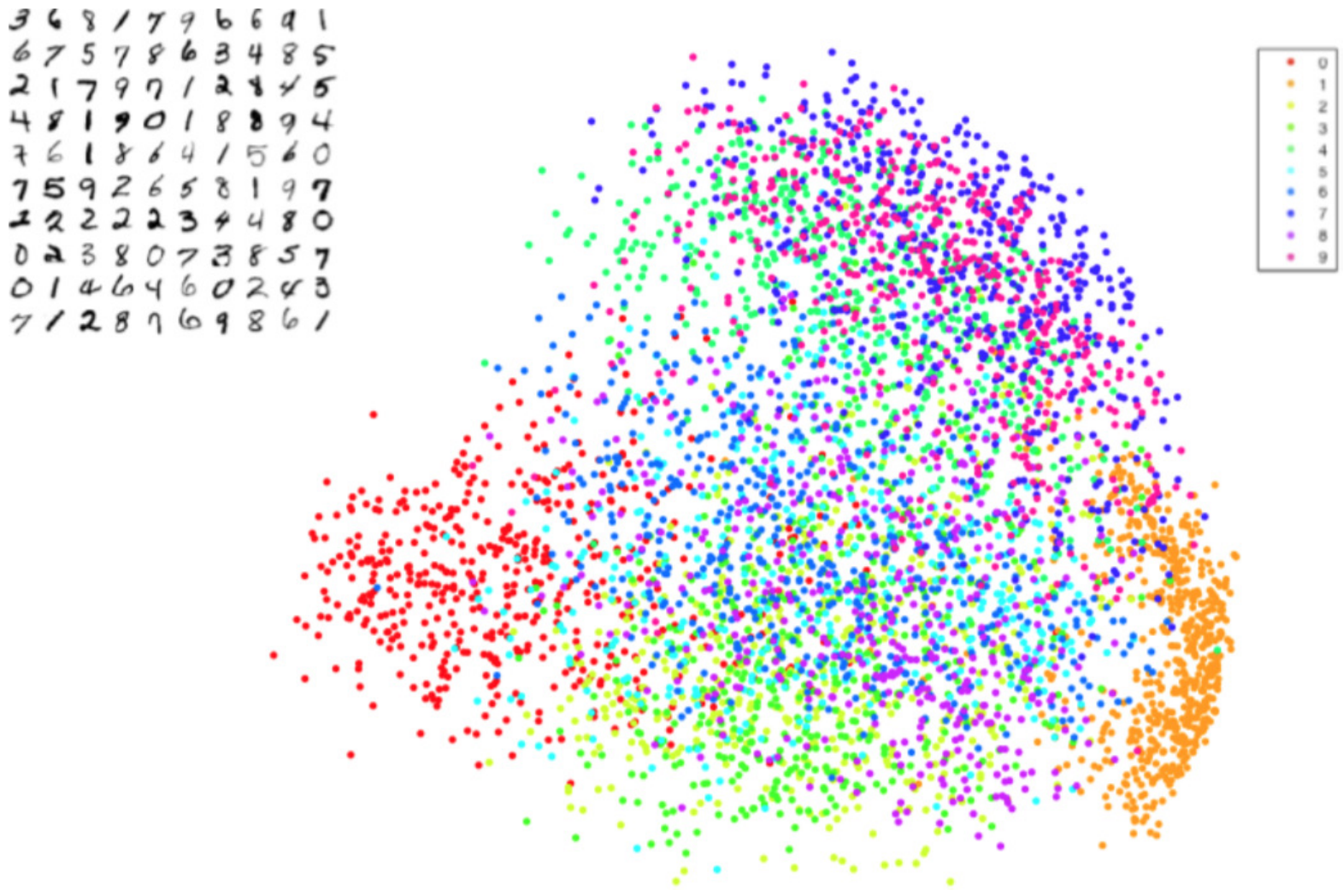
CS189/289A: Introduction to Machine Learning

Stella Yu

UC Berkeley

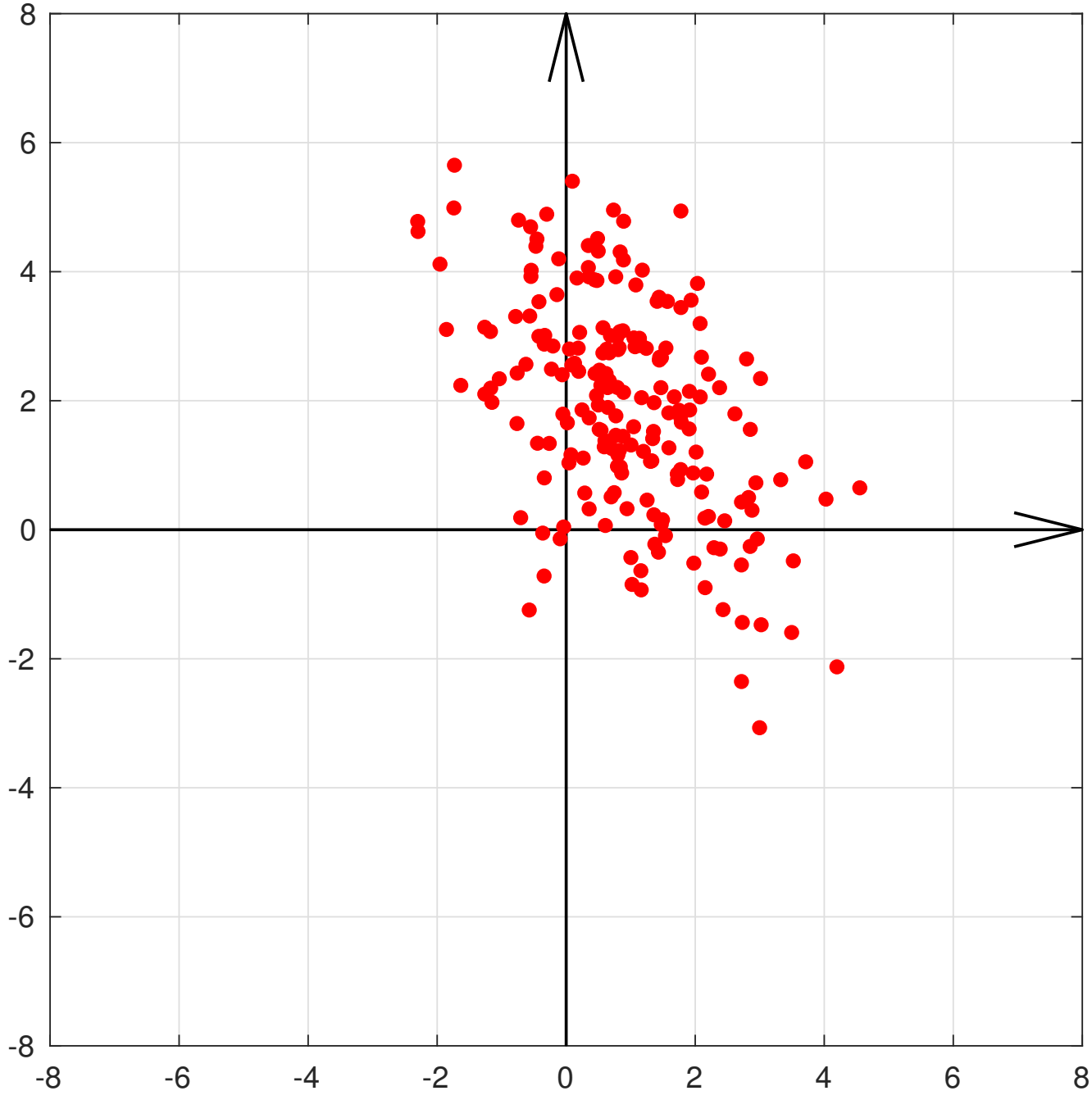
19 September 2017

Why Principal Component Analysis (PCA)?

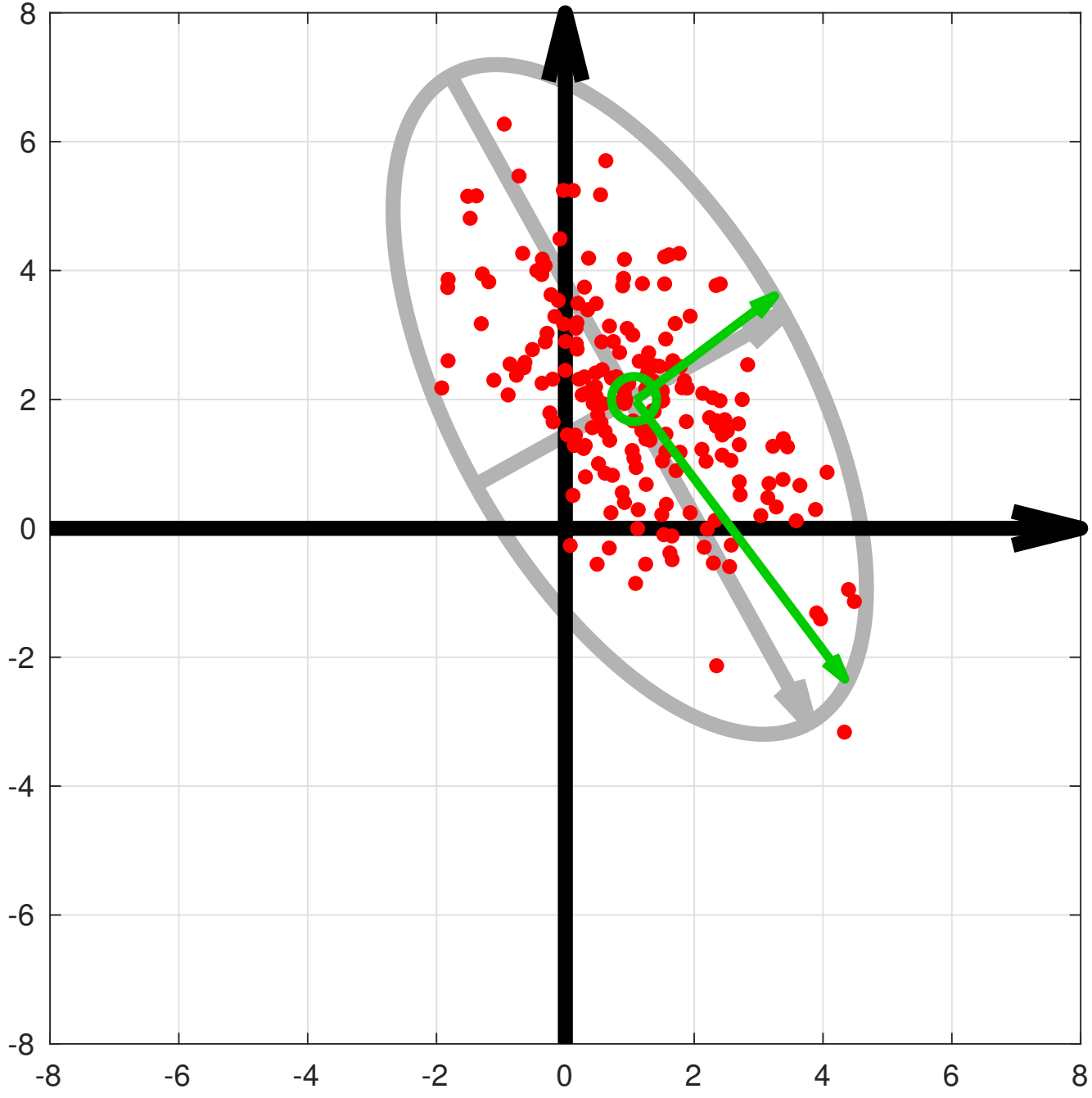


- Find a small basis for representing variations in complex data
- Reducing #dimensions makes computations cheaper
- Remove irrelevant dimensions to reduce overfitting in learning

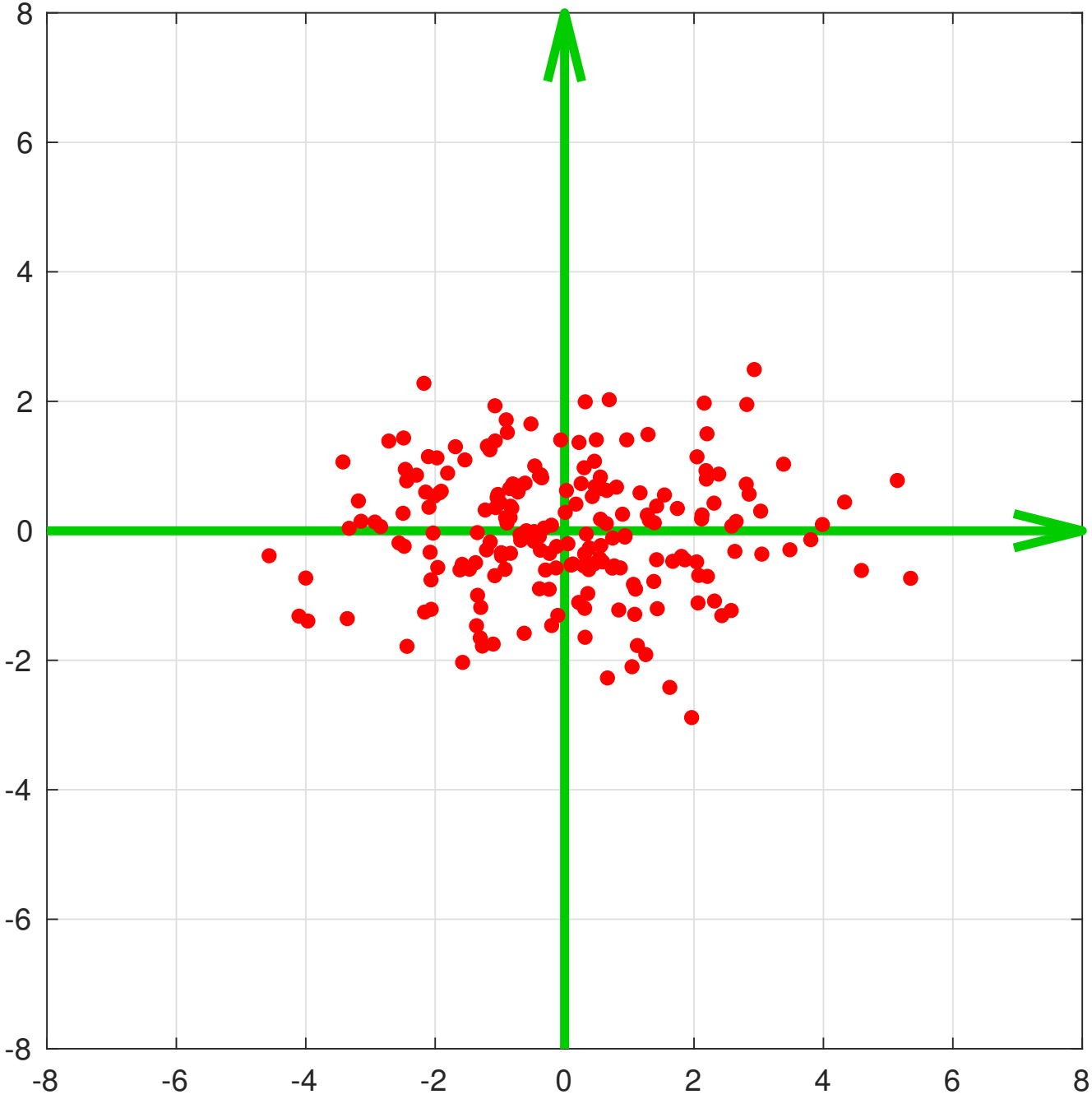
Point Set ~ Gaussian Probability Density



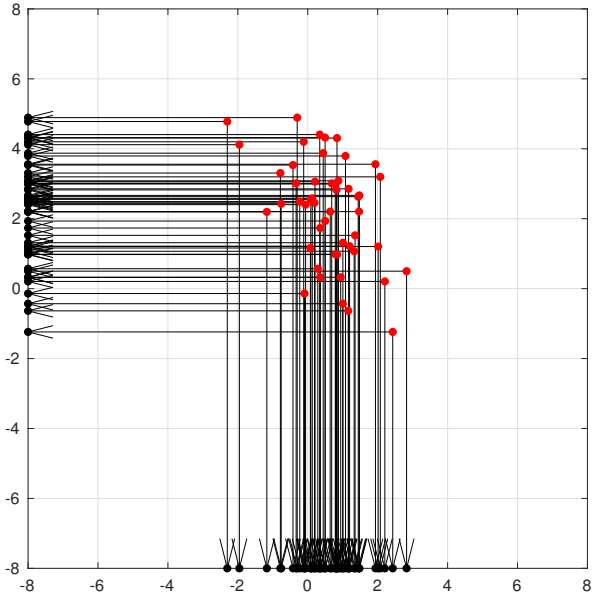
Principal Component Analysis (PCA)

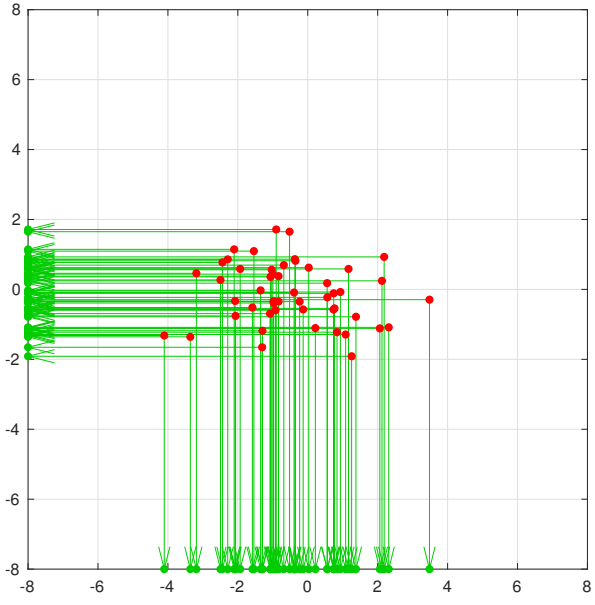


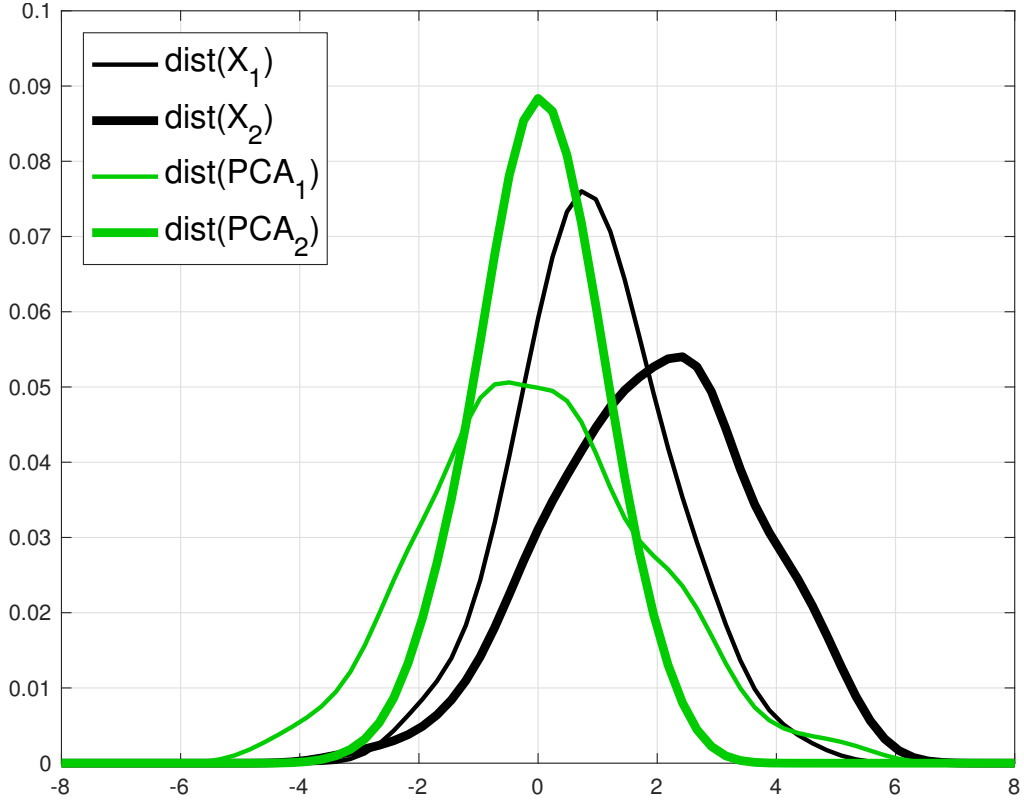
PCA Projection: Major/ Minor Axes of Variation



PCA: Max Variance of Projected Data in A Subspace







PCA: Centered Data

- Consider row data matrix $X_{n\times d}$, n points, d dimensions.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{1}$$

- Assume zero-mean:

$$\text{mean}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = 0 \tag{2}$$

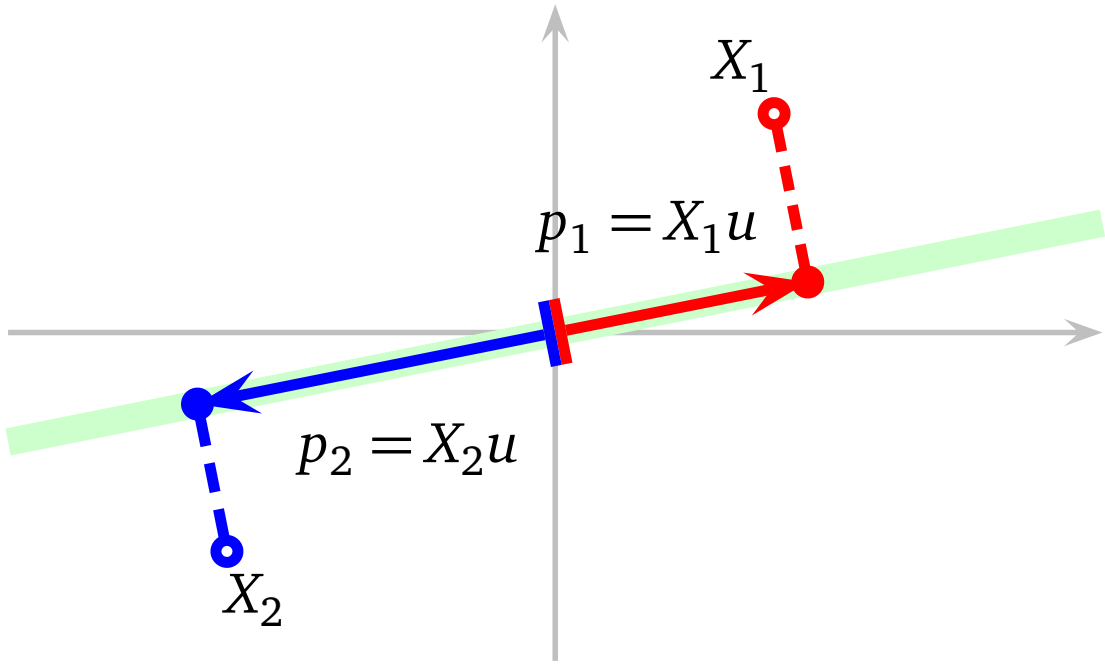
Otherwise, subtract the mean of the entire point set:

$$X \Leftarrow X - \text{mean}(X) \tag{3}$$

$$\text{mean}(X) = 0 \tag{4}$$

PCA: Maximum Variance of Projected Data

- Projection of X along direction u, where $\|u\| = 1$, is $P = Xu$.

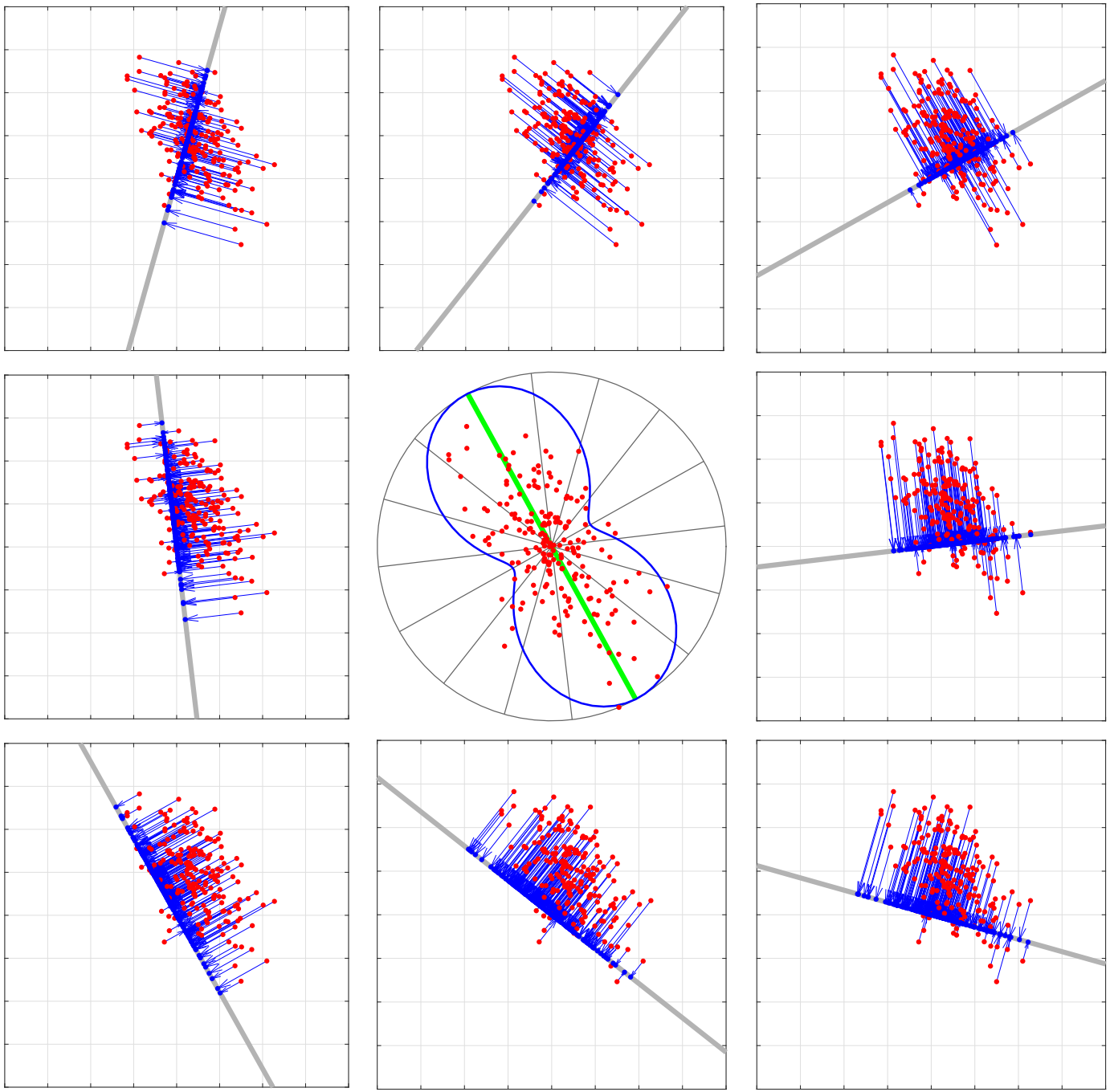


- The principal component is the direction that captures most of the variance in the data.

$$E[P] = 0, \quad V[P] \propto \sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} (X_i u)^2 = u'(X'X)u \tag{5}$$

$$\max_{\|u\|=1} \varepsilon_{PCA\_Var}(u) = \max_{\|u\|=1} u'(X'X)u = \max_{u} \frac{u'(X'X)u}{u'u} \tag{6}$$

What Does the PCA Criterion $\frac{u'(X'X)u}{u'u}$ Look Like?



PCA: Maximum Variance in A Subspace

- The first principal component

$$u_{(1)} = \arg\max_{\|u\|=1} u'(X'X)u = \arg\max_{u} \frac{u'(X'X)u}{u'u}. \tag{7}$$

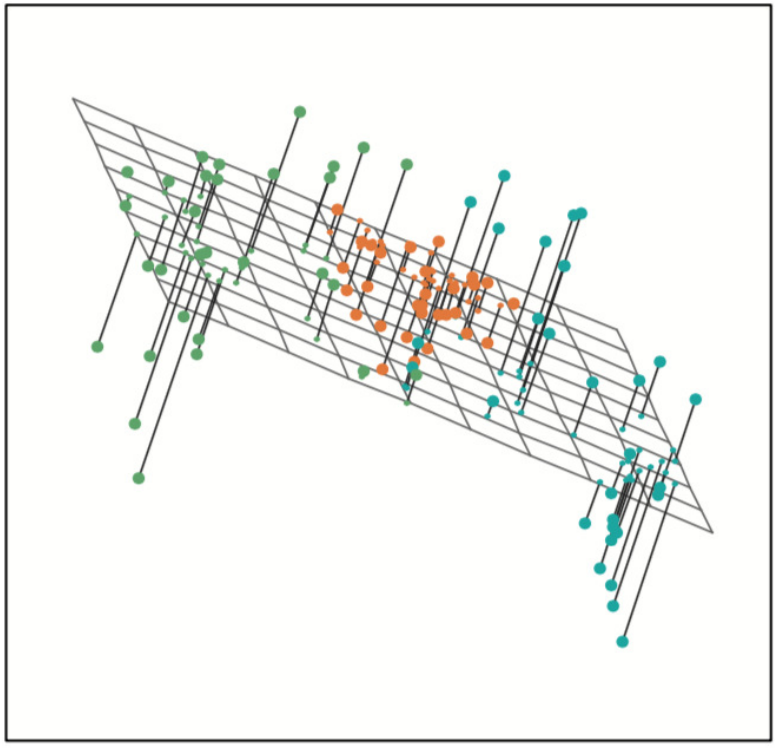
- The residue after subtracting the first $k-1$ projections is:

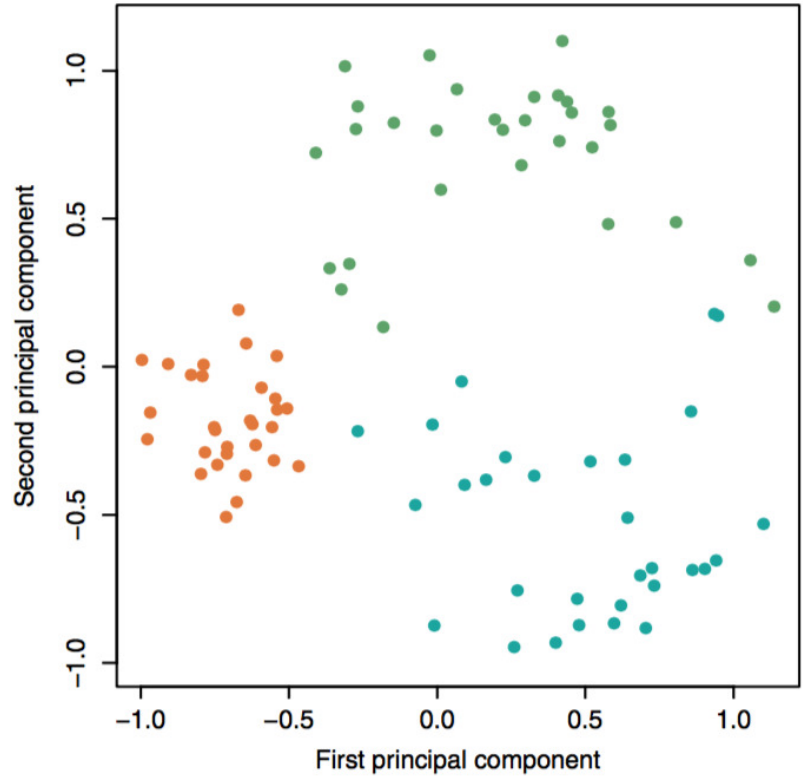
$$\hat{X} = X - \sum_{j=1}^{k-1} X u_{(j)} \cdot u'_{(j)}. \tag{8}$$

- The k-th principal component is the direction that captures most of the variance in the residue:

$$u_{(k)} = \arg \max_{\substack{u'u_{(j)}=0 \\ j=1,\ldots,k-1}} \frac{u'(\hat{X}'\hat{X})u}{u'u} = \arg \max_{\substack{u'u_{(j)}=0 \\ j=1,\ldots,k-1}} \frac{u'(X'X)u}{u'u} \tag{9}$$

PCA: Maximum Variance in A Subspace





- The first k principal components capture most of the variance in a k-dimensional subspace $U = [u_{(1)}, \ldots, u_{(k)}]$:

$$U = \arg\max_{U'U=I} \sum_{j=1}^{k} u'_{(j)}(X'X)u_{(j)} = \arg\max_{U'U=I} \operatorname{tr}(U'(X'X)U) \quad (10)$$

Rayleigh Quotient

- PCA maximizes the function $\frac{u'(X'X)u}{u'u}$, a Rayleigh quotient.
- Rayleigh quotient of $d \times d$ real symmetric matrix M:

$$R(u;M) = \frac{u'Mu}{u'u} = (u'Mu)_{\text{where}\|u\|=1} \tag{11}$$

- Consider M's eigen-decomposition:

$$\text{eigenvalue:} \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \tag{12}$$

$$\text{eigenvector:} \qquad MV_j = \lambda_j V_j, \quad j = 1, 2, \ldots, d \tag{13}$$

$$\text{eigenspace:} \qquad V = [V_1, V_2, \ldots, V_d], \quad V'V = I \tag{14}$$

$$\text{diagonization:} \qquad V'MV = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \tag{15}$$

Rayleigh Quotient Optimization: Single

- Rayleigh quotient of M represented in M's eigenspace:

$$u = \sum_{j=1}^{d} \alpha_j V_j = V\alpha \tag{16}$$

$$u'u = 1 \Rightarrow \alpha'\alpha = 1 \Rightarrow \sum_{j=1}^{n} \alpha_j^2 = 1 \tag{17}$$

$$R(u;M) = u'Mu \tag{18}$$

$$= \alpha'V'MV\alpha \tag{19}$$

$$= \alpha'\Lambda\alpha = \sum_{j=1}^{d} \lambda_j \alpha_j^2 \tag{20}$$

- Any weighted average is bounded by the extrema:

$$\lambda_1 = R(V_1;M) \geq R(u;M) \geq R(V_d;M) = \lambda_d \tag{21}$$

- Projection along the largest eigenvector of M maximizes the Rayleigh quotient of M.

Rayleigh Quotient Optimization: Multiple

- Bounds of k Rayleigh quotients over a subspace:

$$\begin{aligned} \lambda_1 + \lambda_2 + \ldots + \lambda_k =& R([V_1, V_2, \ldots, V_k]; M) && (22) \\ \geq& R(U; M) \geq \\ R([V_{d-k+1}, \ldots, V_{d-1}, V_d]; M) =& \lambda_{d-k+1} + \ldots + \lambda_{d-1} + \lambda_d && (23) \end{aligned}$$

- Projection to the subspace spanned by the first k leading eigenvectors maximizes the variance of projected data.
- Rayleigh quotient of a matrix pair (M, D):

$$\arg\max_u \varepsilon(u) = \frac{u'Mu}{u'Du} = \text{eig}(M, D) \quad (24)$$

$$Mu = \lambda Du \quad (25)$$

$$\text{special case: } Mu = \lambda u \text{ when } D = I \quad (26)$$

Projecton to the largest generalized eigenvectors of (M, D) maximizes the Rayleigh quotient of (M, D).

PCA: Eigenvectors of Covariance and Data Matrices

- PCA solution according to the Rayleigh quotient optimization:

$$\max_u \varepsilon_{PCA\_Var}(u) = \frac{u'X'Xu}{u'u} \quad \Rightarrow \quad (X'X)u = \lambda u \tag{27}$$

- Data covariance matrix $C_{d\times d}$: correlation between features

$$C = \frac{1}{n}X'X \propto X'X \tag{28}$$

- $X'X$, XX', and X all have the same (left or right) eigenvectors:

$$X = U_{n\times n}S_{n\times d}V'_{d\times d}, \quad UU' = I, V'V = I, S = \text{Diag} \tag{29}$$

$$C = X'X = VSU'USV' = VS^2V' \tag{30}$$

$$XX' = USV'VSU' = US^2U' \quad \rightarrow \quad V = X'US^{-1} \tag{31}$$

- Algorithm: If $d \ll n$, compute eigs(C), otherwise compute svds(X), or compute eigs(XX'), then derive V.

PCA for Minimum Reconstruction Error, TLS vs. LS

- PCA: minimizing reconstruction error = maximizing variance:

$$\arg\min_u \varepsilon_{PCA\_Err}(u) = \|X - Xuu'\|^2 \tag{32}$$

$$= \arg\min_u \text{constant} - \varepsilon_{PCA\_Var}(u) = \arg\max_u \varepsilon_{PCA\_Var}(u) \tag{33}$$

- In terms of predicting y from X, both PCA and LS regression minimize the sum of the squares of the projection distances. The distance is measured vertically for LS regression, and orthogonal to the projection hyperplane for PCA.

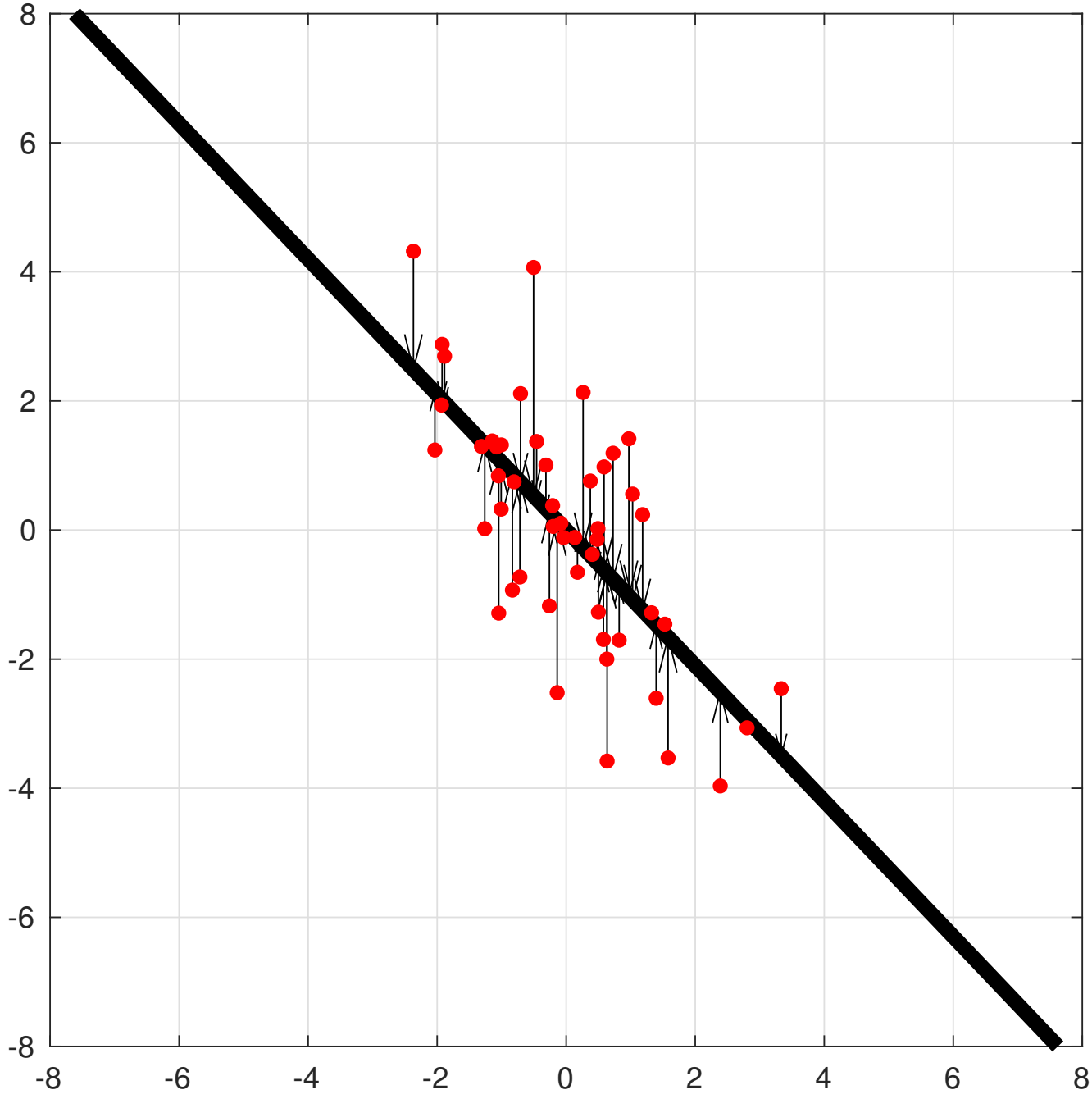
$$\min \varepsilon_{LS}(u) = \|y - Xu\|^2 = \left\| [y, X] \begin{bmatrix} 1 \\ u \end{bmatrix} \right\|^2 \tag{34}$$

$$\min \varepsilon_{PCA\_Err}(u) = \|[y, X] - [y, X]uu'\|^2 \tag{35}$$

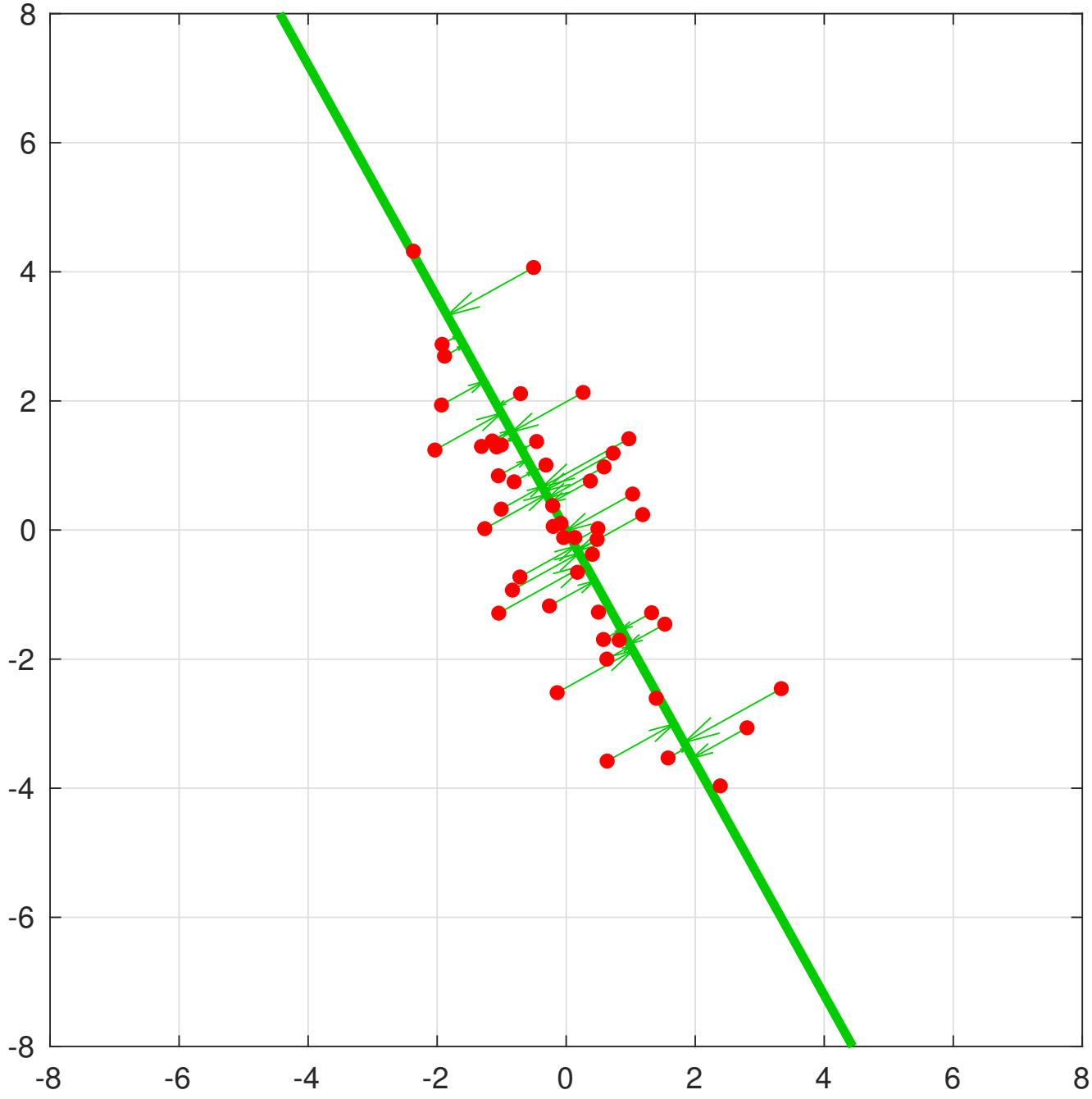
$$= \|[y, X](I - uu')\|^2, \quad u'u = 1 \tag{36}$$

- TLS = PCA, although TLS seeks the normal direction of a line, whereas PCA seeks the direction of the line.

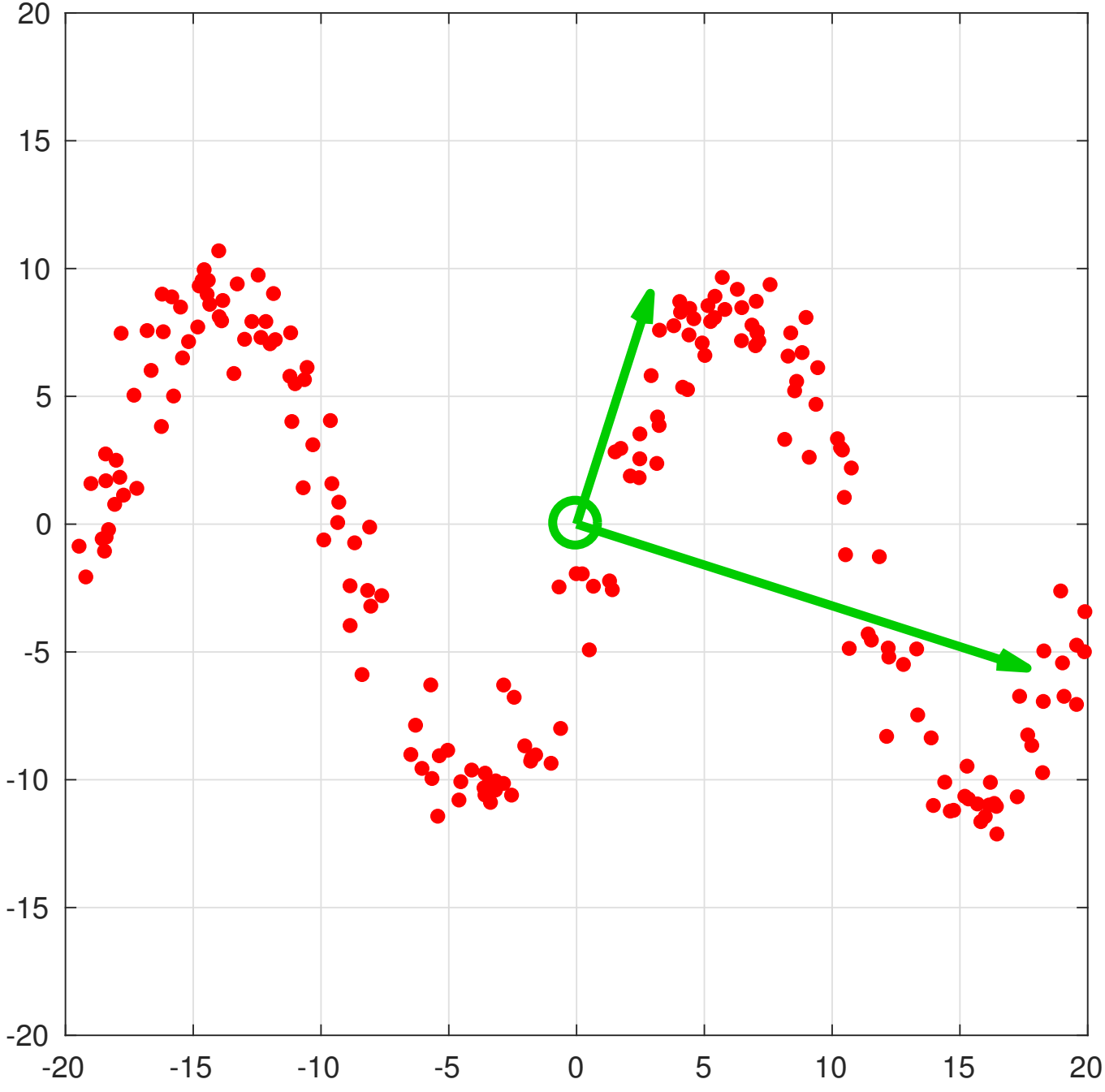
Linear Regression: LS Result



Linear Regression: PCA / TLS Result



PCA on Non-Gaussian Data

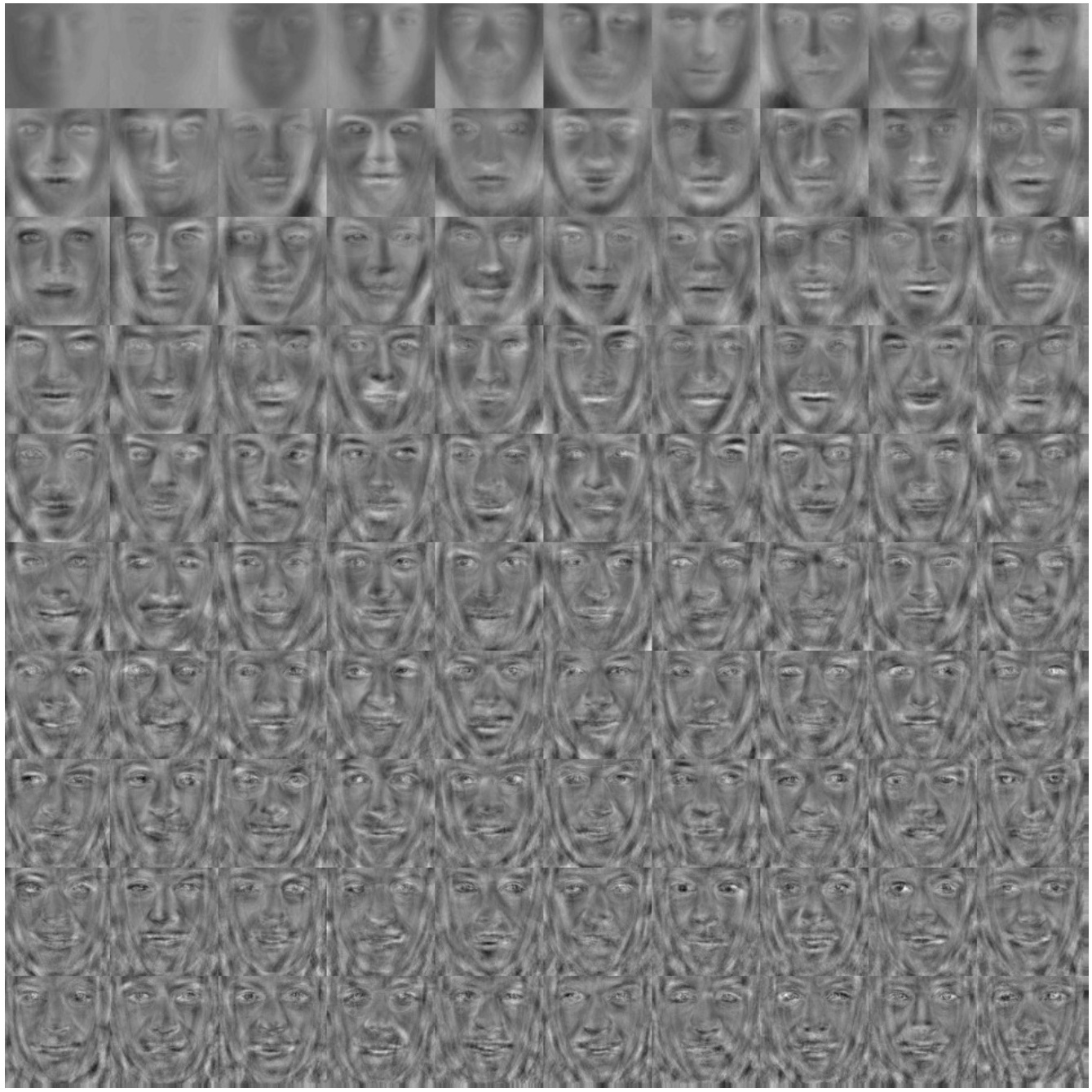


Face Dataset

Mean Image: Mean Value Per Pixel

Std Image: Standard Deviation Per Pixel

Eigen-Faces



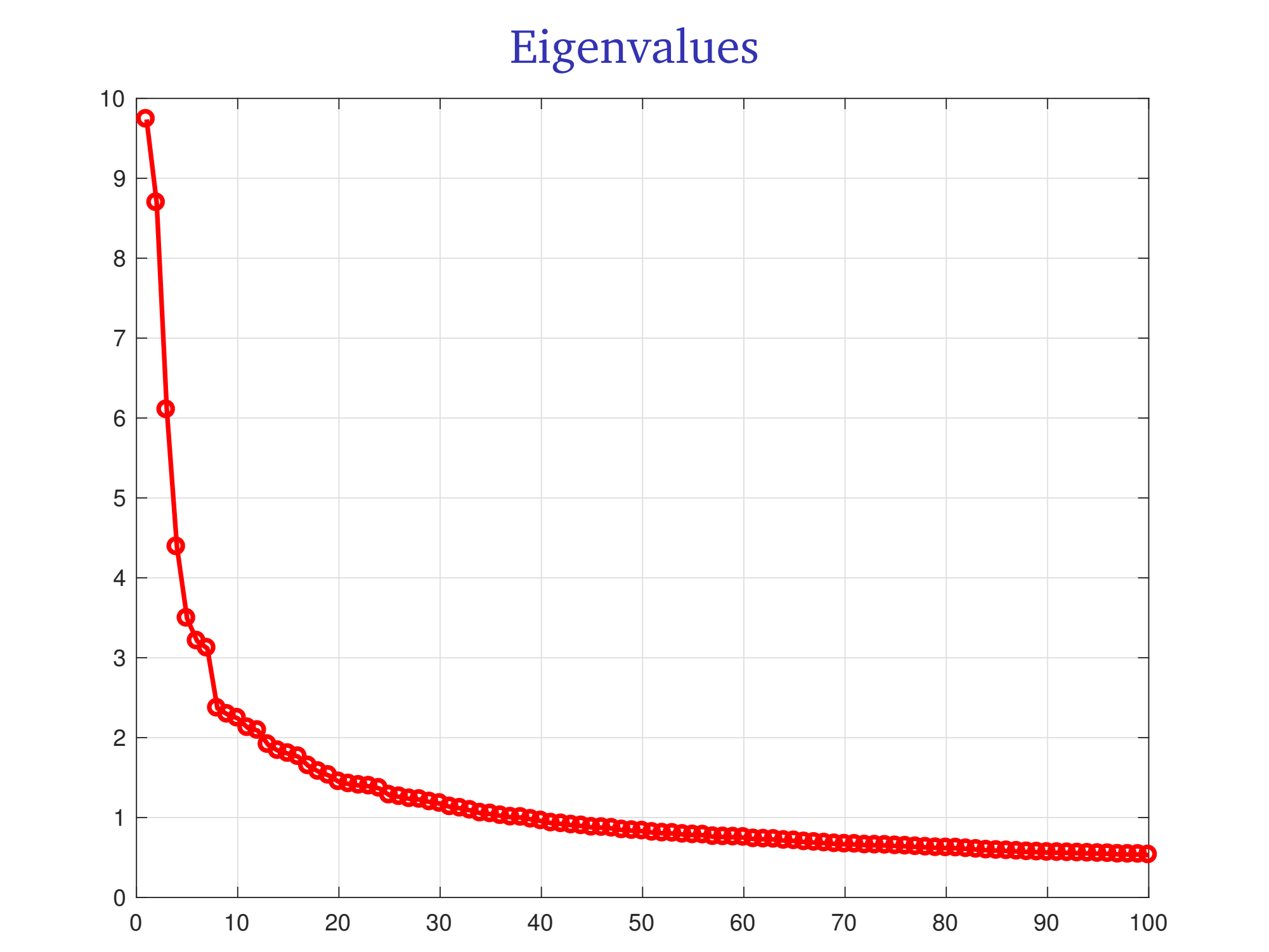
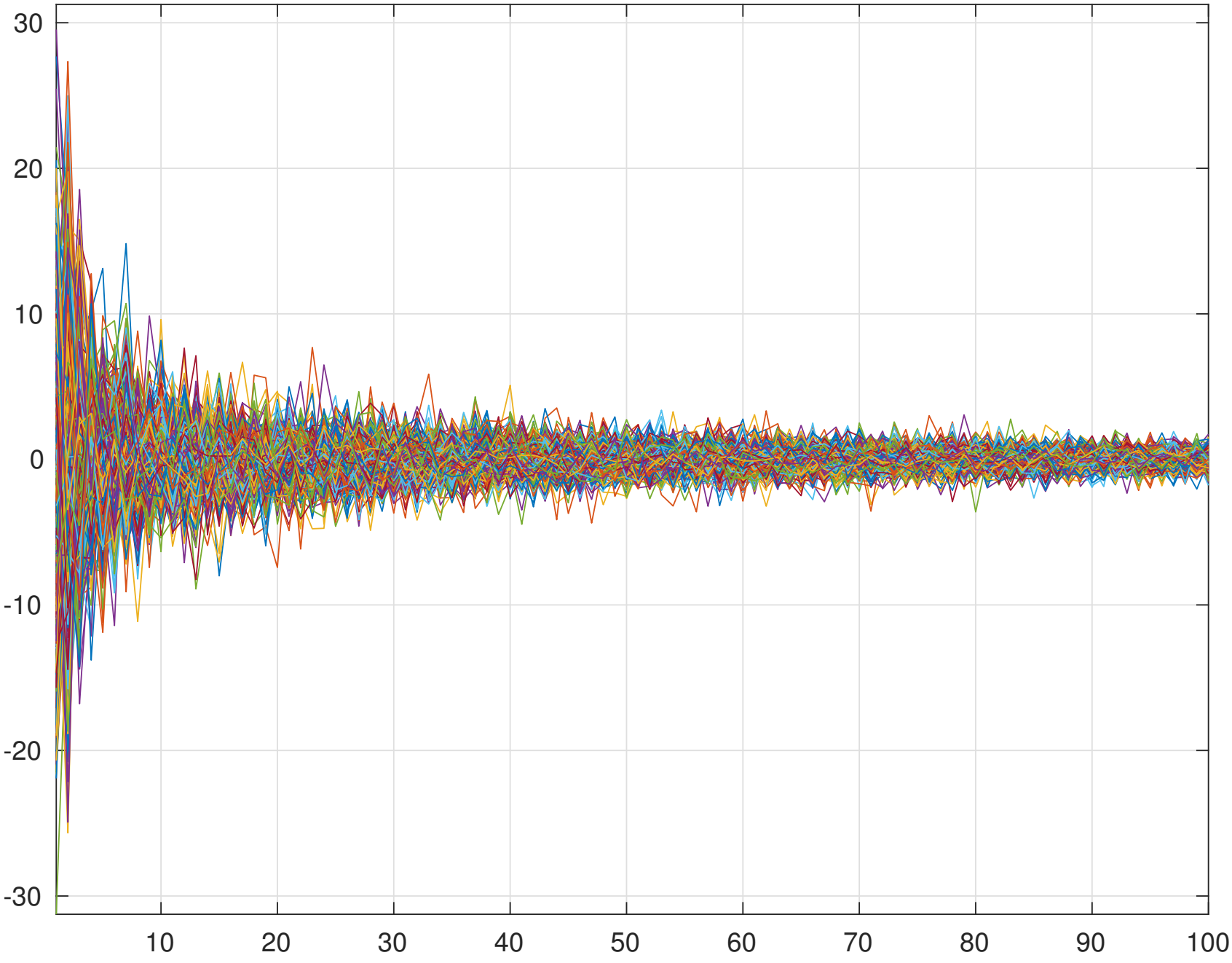

Eigenvalues
10
9
8
7
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
80
90
100

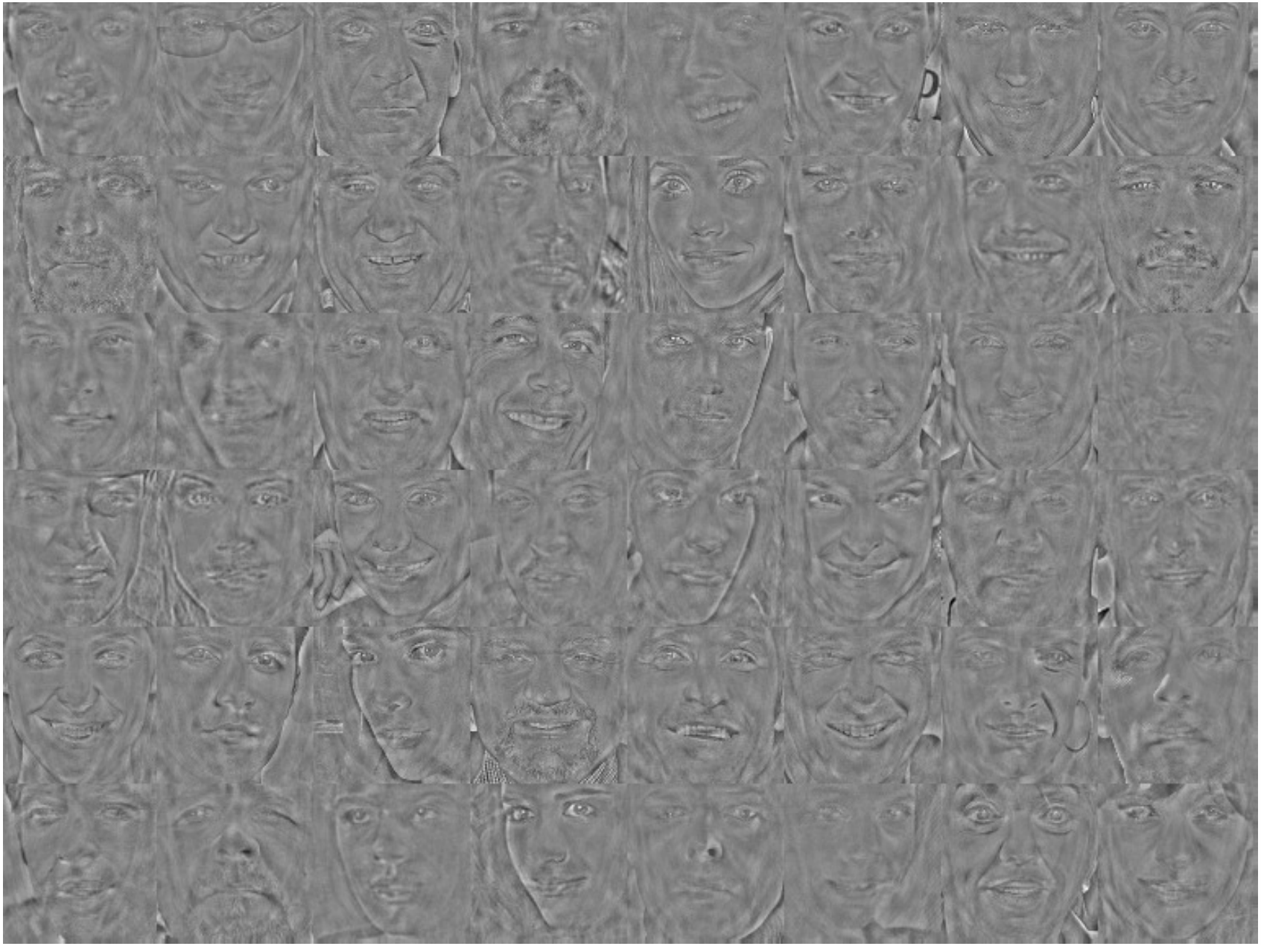
Image PCA Coefficients



Original Images: 96×96

PCA Reconstruction Results

PCA Reconstruction Error



PCA Reconstruction of A Single Image

PCA Running Reconstruction Example

Summary

- PCA reveals the internal structure of the data in a way that best explains the variance in the data.
- PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, after mean centering.
- The first few principal components provide a lower-dimensional picture of high-dimensional data points, viewed from a most informative viewpoint.
- PCA is related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.