Machine Learning

Contents

10.1. Dimensionality Reduction

- High dimensional data - images, bag-of-words, protein-expressions etc.
- Data will typically lie close to a much lower dimensional space, provided underlying structure of data.

- What
  - Given d-dimensional data, reduce it to k-dimensional data, $k < d$, while preserving as much information as possible

- Why?
  - Time complexity: Less computation
  - Space complexity: Less parameters
  - The cost of observing the feature
  - Simpler models: robust on small datasets
  - More interpretable; simpler explanation
  - Data visualization if plotted in 2 or 3 dimensions

Feature Selection vs. Extraction

- Feature selection
  - Choose $k<d$ important features, ignoring the remaining $(d-k)$ features
  - Subset selection algorithms
- Feature extraction
  - Project the original d dimensions to new $k<d$ dimensions
  - Principal component analysis (PCA)..chap. 10
  - Linear discriminant analysis (LDA)..chap. 3
  - Factor analysis (FA)

Feature Selection

- 2^d possible subset selection from d features → cannot enumerate all of them
- Forward Search (set of features F)
  - Initial set of features ∅
  - Find the best new feature $j = \arg\min_i Err(F \cup x_i)$
  - Add x_j to F if $Err(F \cup x_j) < Err(F)$ else stop

- Backward Search
  - Start with all features
  - Remove one at each iteration, if possible

- Floating Search
  - Add some features and remove other features at each iteration

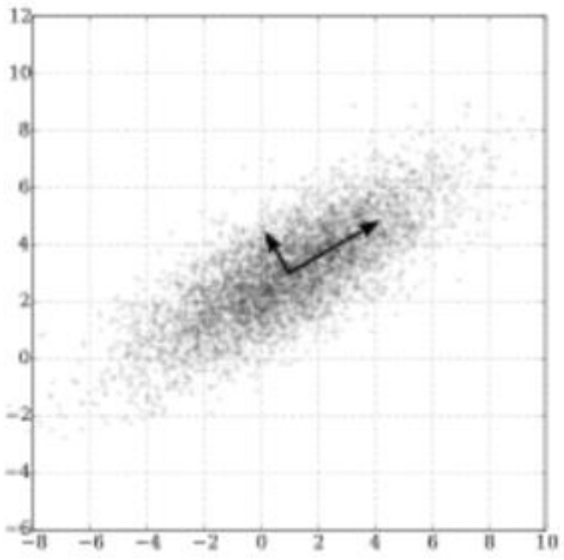
10.2. Principal Components Analysis (PCA)

- Idea:
  - Project *d*-dimensional data points onto lower dimensional space while preserving as much information as possible
  - In particular, choose projection that minimizes the squared error in reconstructing original data

- Given *d*-dimensional data **x**, learns the top *m*-dimensions where
  - the dimensions are orthogonal
  - the reconstructed data as a linear combination of the top *m*-dimensions minimizes reconstruction error (sum of squared errors)

Maximum co-variance and orthogonality

- Select a direction in *m*-dimensional space along which the variance in **x** is maximized.
- Find another direction along which variance is maximised, but restrict the search to all directions orthonormal to all previous selected directions.
- Repeat this until *m* vectors are selected.



PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.

10.3. Eigenstructure of Principal Component Analysis

zero - mean random vector :

$$\mathbf{x} = [x_1, x_2, \cdots, x_d]^T$$

projection unit vector :

$$\mathbf{w} = [w_1, w_2, \cdots, w_d]^T \text{ and } \quad \|\mathbf{w}\| = \left(\mathbf{w}^T \mathbf{w}\right)^{1/2} = 1$$

Projection :

$$z = \mathbf{w}^T \mathbf{x} \text{ and } E[z] = E[\mathbf{w}^T \mathbf{x}] = \mathbf{w}^T E[\mathbf{x}] = 0$$

Variance :

$$\sigma^2 = E[z^2] - E^2[z] = E[(\mathbf{w}^T \mathbf{x})(\mathbf{x}^T \mathbf{w})] = \mathbf{w}^T E[\mathbf{x}\mathbf{x}^T]\mathbf{w} = \mathbf{w}^T \mathbf{S}\mathbf{w}$$

Correlation Matrix :

$$\mathbf{S} = E[\mathbf{x}\mathbf{x}^T]$$

The variance of the projection z is a function of the unit vector $\mathbf{w}$

$$\psi(\mathbf{w}) = \sigma^2 = \mathbf{w}^T \mathbf{S}\mathbf{w}$$

If $\mathbf{w}$ is a unit vector such that the variance probe $\psi(\mathbf{w})$ has an extremal value,
for any small perturbation $\delta\mathbf{w}$ of the unit vector $\mathbf{w}$

$$\psi(\mathbf{w} + \delta\mathbf{w}) = \psi(\mathbf{w})$$

$$\psi(\mathbf{w} + \delta\mathbf{w}) = (\mathbf{w} + \delta\mathbf{w})^T \mathbf{S}(\mathbf{w} + \delta\mathbf{w}) = \mathbf{w}^T \mathbf{S}\mathbf{w} + 2(\delta\mathbf{w})^T \mathbf{S}\mathbf{w} + (\delta\mathbf{w})^T \mathbf{S}\delta\mathbf{w}$$

Ignoring the second - order term $(\delta\mathbf{w})^T \mathbf{S}\delta\mathbf{w}$

$$\psi(\mathbf{w} + \delta\mathbf{w}) = \mathbf{w}^T \mathbf{S}\mathbf{w} + 2(\delta\mathbf{w})^T \mathbf{S}\mathbf{w} = \psi(\mathbf{w}) + 2(\delta\mathbf{w})^T \mathbf{S}\mathbf{w}$$

Hence

$$(\delta\mathbf{w})^T \mathbf{S}\mathbf{w} = 0$$

Restriction

$$\|\mathbf{w} + \delta\mathbf{w}\| = (\mathbf{w} + \delta\mathbf{w})^T (\mathbf{w} + \delta\mathbf{w}) = 1$$

$$(\delta\mathbf{w})^T \mathbf{w} = 0$$

$$(\delta\mathbf{w})^T \mathbf{S}\mathbf{w} - \lambda(\delta\mathbf{w})^T \mathbf{w} = 0 \quad \text{equivalently} \quad (\delta\mathbf{w})^T (\mathbf{S}\mathbf{w} - \lambda\mathbf{w}) = 0$$

Eigenvalue Problem $\quad \mathbf{S}\mathbf{w} = \lambda\mathbf{w}$

Eigenvalue Problem $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j, \quad j = 1,2,\ldots,d$

Let the corresponding eigenvalues be arranged in decreasing order

$$\lambda_1 > \lambda_2 > \lambda_3 \ldots\ldots > \lambda_d$$

Associated eigenvectors

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots\ldots \mathbf{w}_d]$$

Orthonormality

$$\mathbf{w}_i^T \mathbf{w}_j = 1 \,(\text{if } i{=}j) \text{ and } 0 \text{ otherwise} \qquad \mathbf{W}^T \mathbf{W} = \mathbf{I}$$

Basic data representation

$$z_j = \mathbf{w}_j^T \mathbf{x}, \qquad j = 1,2,\ldots,d : \text{principal component}$$

$$\mathbf{z} = [z_1, z_2, z_3, \ldots\ldots z_d]^T = \mathbf{W}^T \mathbf{x} \quad \text{and} \quad \mathbf{x} = \mathbf{W}\mathbf{z} = \sum_{j=1}^{d} z_j \mathbf{w}_j$$

Dimensionality Reduction

$$\mathbf{x}' = \sum_{j=1}^{k} z_j \mathbf{w}_j \qquad\qquad \mathbf{e} = \mathbf{x} - \mathbf{x}' = \sum_{j=k+1}^{d} z_j \mathbf{w}_j$$

참조: 차원축소 오차

식 (10.3.25)를 유도해보자. (10.3.24)를 대입하면

$$E[\boldsymbol{e}^T\boldsymbol{e}] = E[\sum_{j=k+1}^{d} z_j \boldsymbol{w}_j^T \boldsymbol{w}_j z_j] = \sum_{j=k+1}^{d} E[z_j^2] = \sum_{j=k+1}^{d} \boldsymbol{w}_j^T E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w}_j \qquad (10.3.26)$$

$$= \sum_{j=k+1}^{d} \boldsymbol{w}_j^T \boldsymbol{S}\boldsymbol{w}_j = \sum_{j=k+1}^{d} \boldsymbol{w}_j^T \lambda_j \boldsymbol{w}_j = \sum_{j=k+1}^{d} \lambda_j$$

를 얻게 된다.

예제 10.3-1

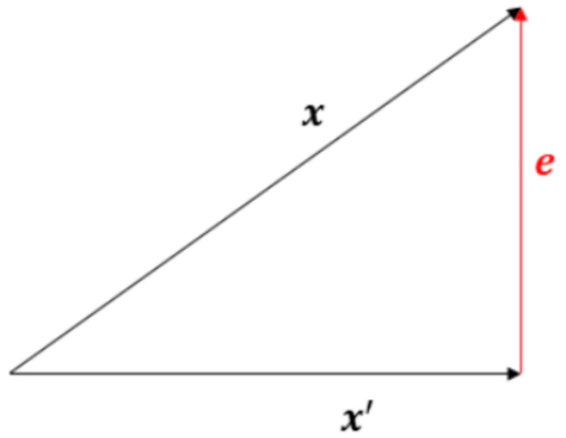
PCA에 의해 차원 축소 후 복원된 벡터 $\boldsymbol{x}'$과 복원 오차 $\boldsymbol{e}$가 각각 식 (10.3.23)과 (10.3.24)로 주어졌다. 두 벡터 $\boldsymbol{x}'$과 $\boldsymbol{e}$가 직교함을 보이고, 이들 사이의 관계를 $\boldsymbol{x}$와 함께 그림으로 나타내어라.

풀이

먼저, 두 벡터 $\boldsymbol{x}'$과 $\boldsymbol{e}$가 직교함을 보이겠다. 이를 위하여 두 벡터의 내적을 계산하면

$$\boldsymbol{e}^T\boldsymbol{x}' = \left(\sum_{j=k+1}^{d} z_j \boldsymbol{w}_j^T\right)\left(\sum_{i=1}^{k} z_i \boldsymbol{w}_i\right) = \sum_{j=k+1}^{d}\sum_{i=1}^{k} z_j \boldsymbol{w}_j^T \boldsymbol{w}_i z_i = 0$$

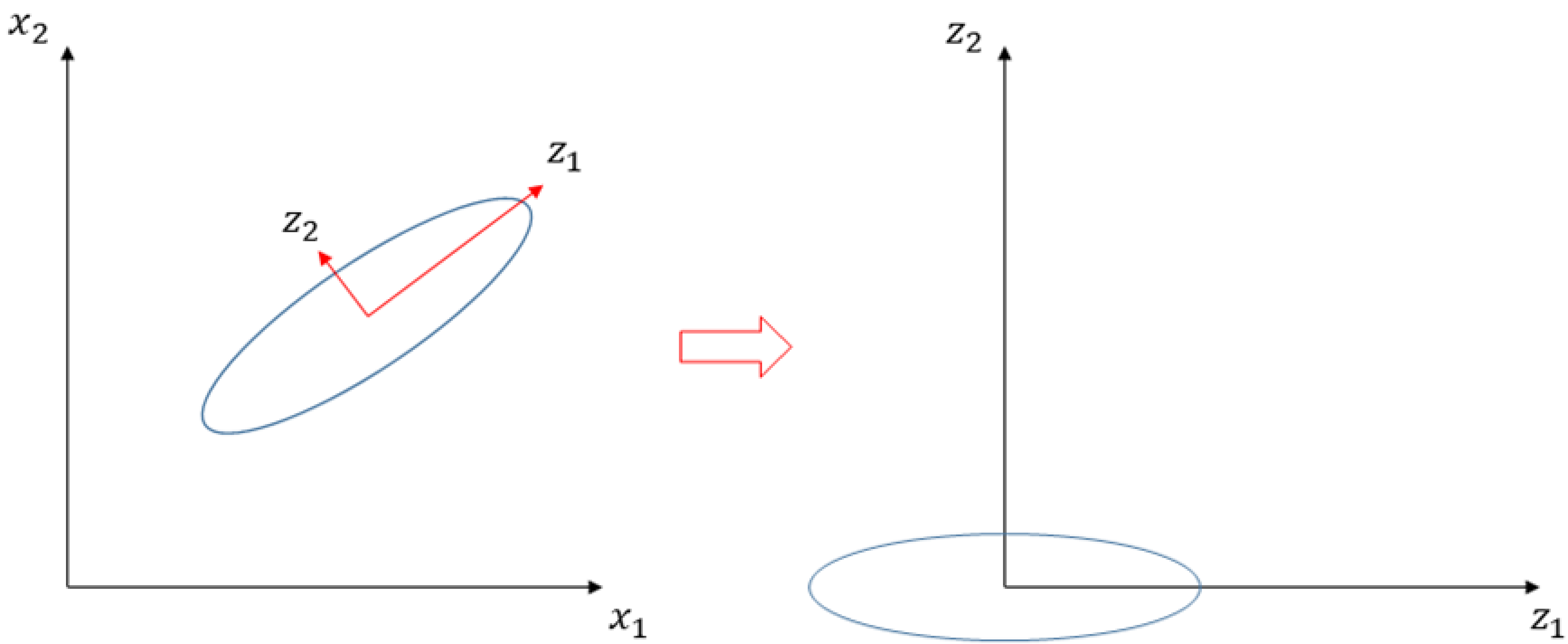
이 된다. 그 이유는 $\boldsymbol{w}_i^T\boldsymbol{w}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$와 같이 정규직교하기 때문이다. $\boldsymbol{x}$, $\boldsymbol{x}'$, $\boldsymbol{e}$의 관계를 그림으로 그리면 아래와 같다.



10.4. Understanding PCA

Define $\mathbf{z} = W^T(\mathbf{x} - \mathbf{m})$

- W: Columns as eigenvectors of S (estimator to Σ)
- **m**: Sample mean
- Centers the data at the origin and rotates the axis



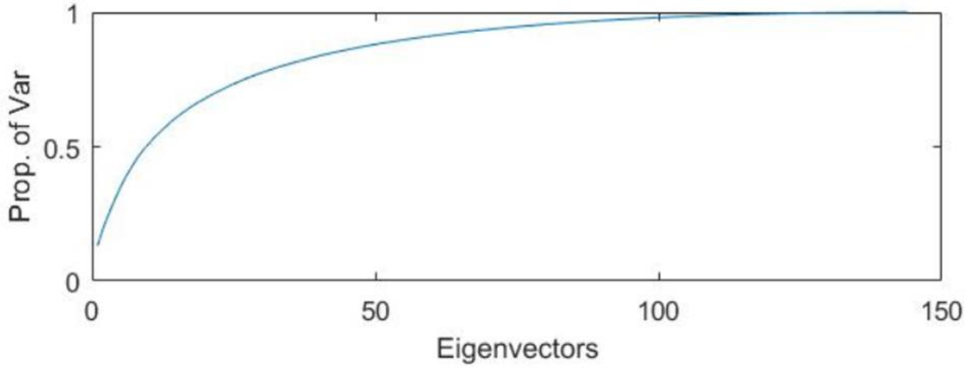
The Size of Reduced Dimension

- Larger eigenvalues contribute more to variance of $\boldsymbol{z}$

① POV(Proportion of Variance)

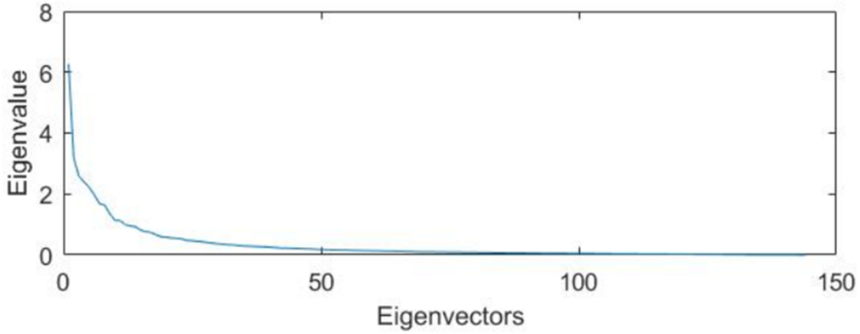
- Rule of thumb : POV>0.9

$$POV = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_k + \dots + \lambda_d}$$

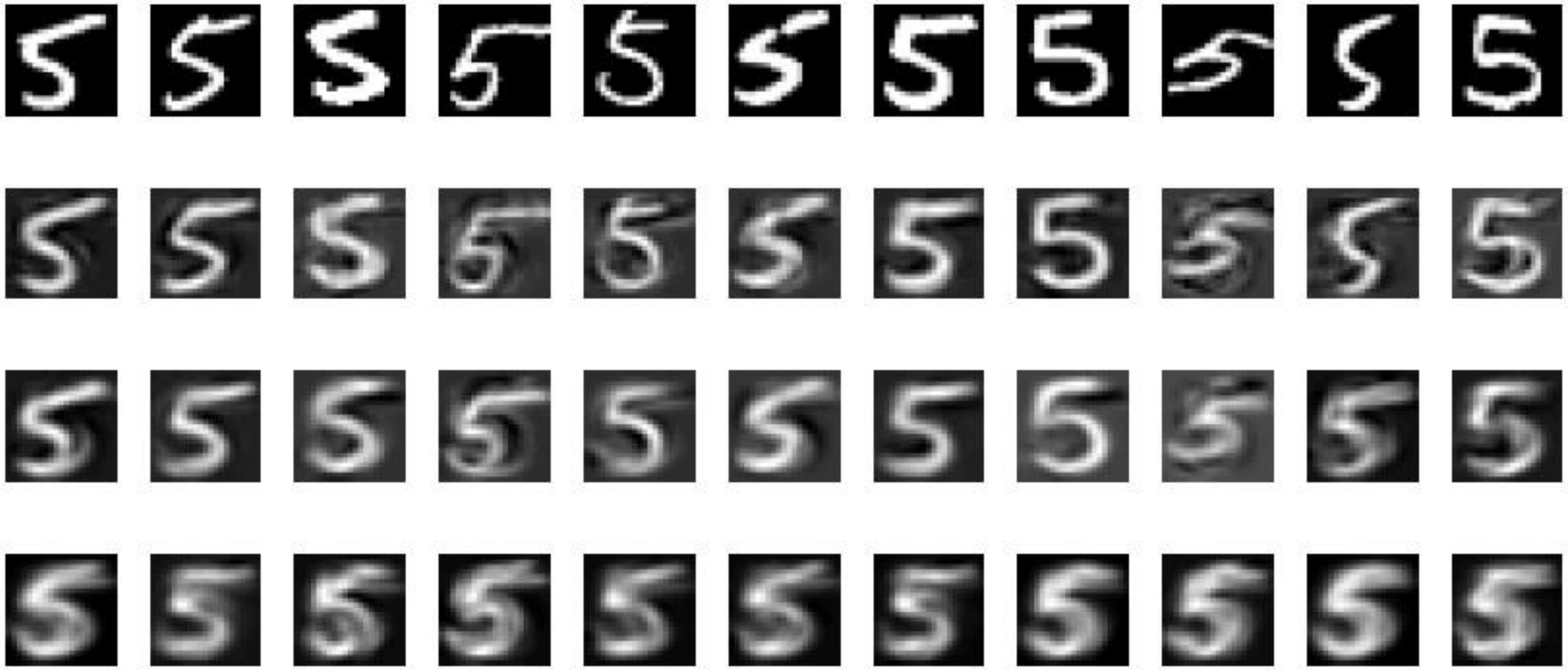


② Stop at elbow of λ_k

③ Keep the eigenvectors whose eigenvalues are larger than average of eigenvalue



Reducing the dimension of digits



Top row : a selection of the digit 5 taken from the database of 892 examples. Plotted beneath each digit is the reconstruction using 100, 30 and 5 eigenvectors (from top to bottom). Note how the reconstructions for fewer eigenvectors express less variability from each other, and resemble more a mean 5 digit.

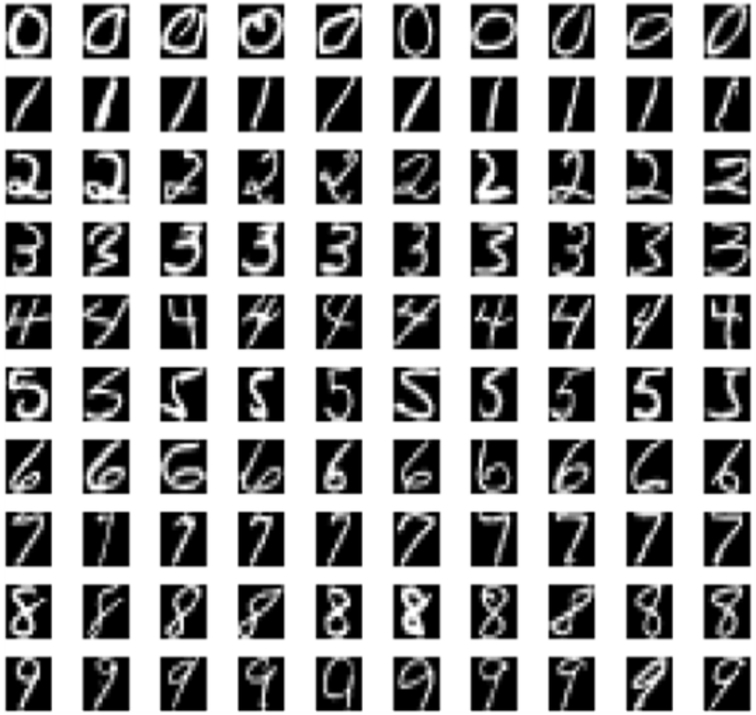


그림 10.4. 숫자 영상 예시

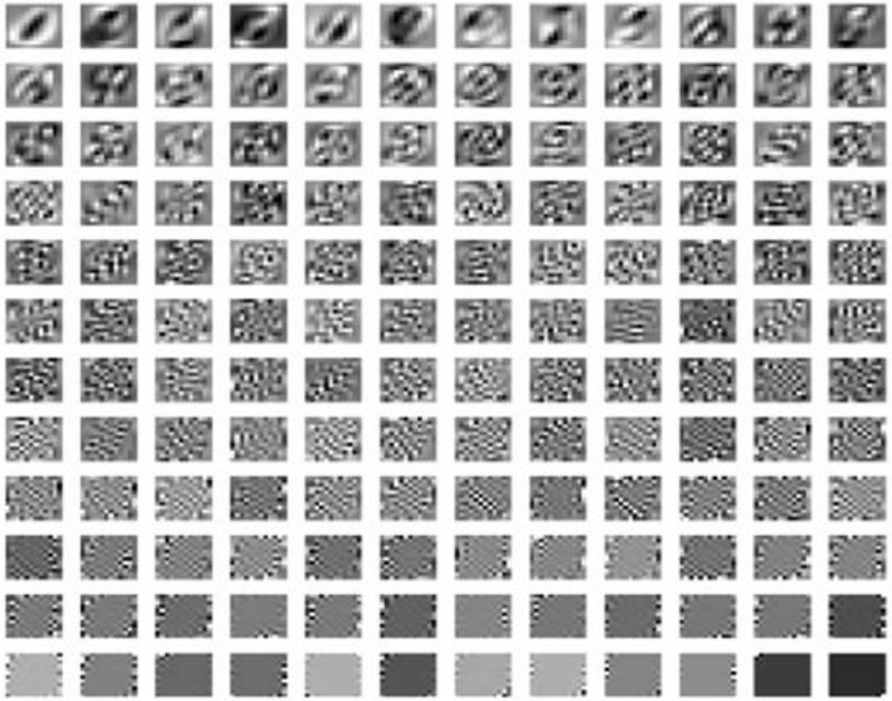


그림 10.5. 숫자영상에서 얻어진 고유벡터


그림 10.6. 주성분 분석에 의한 영상복원 예시

Reducing the dimension of faces

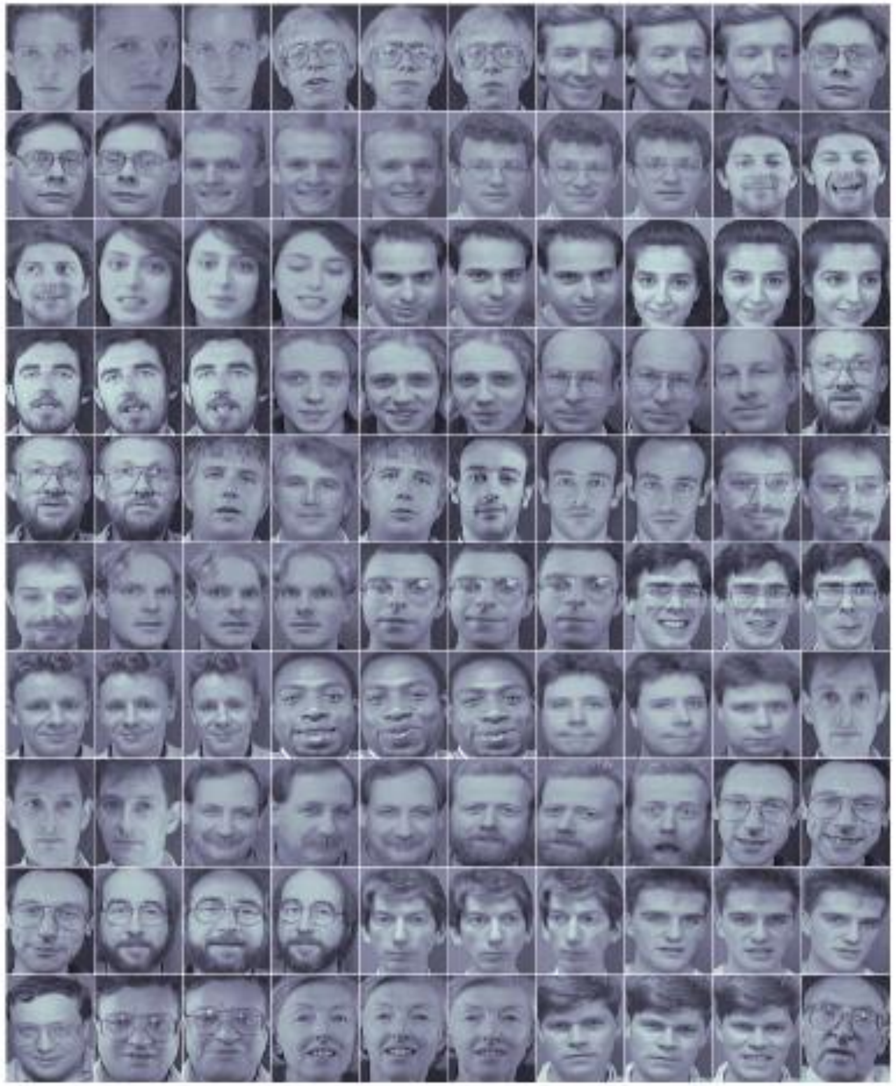
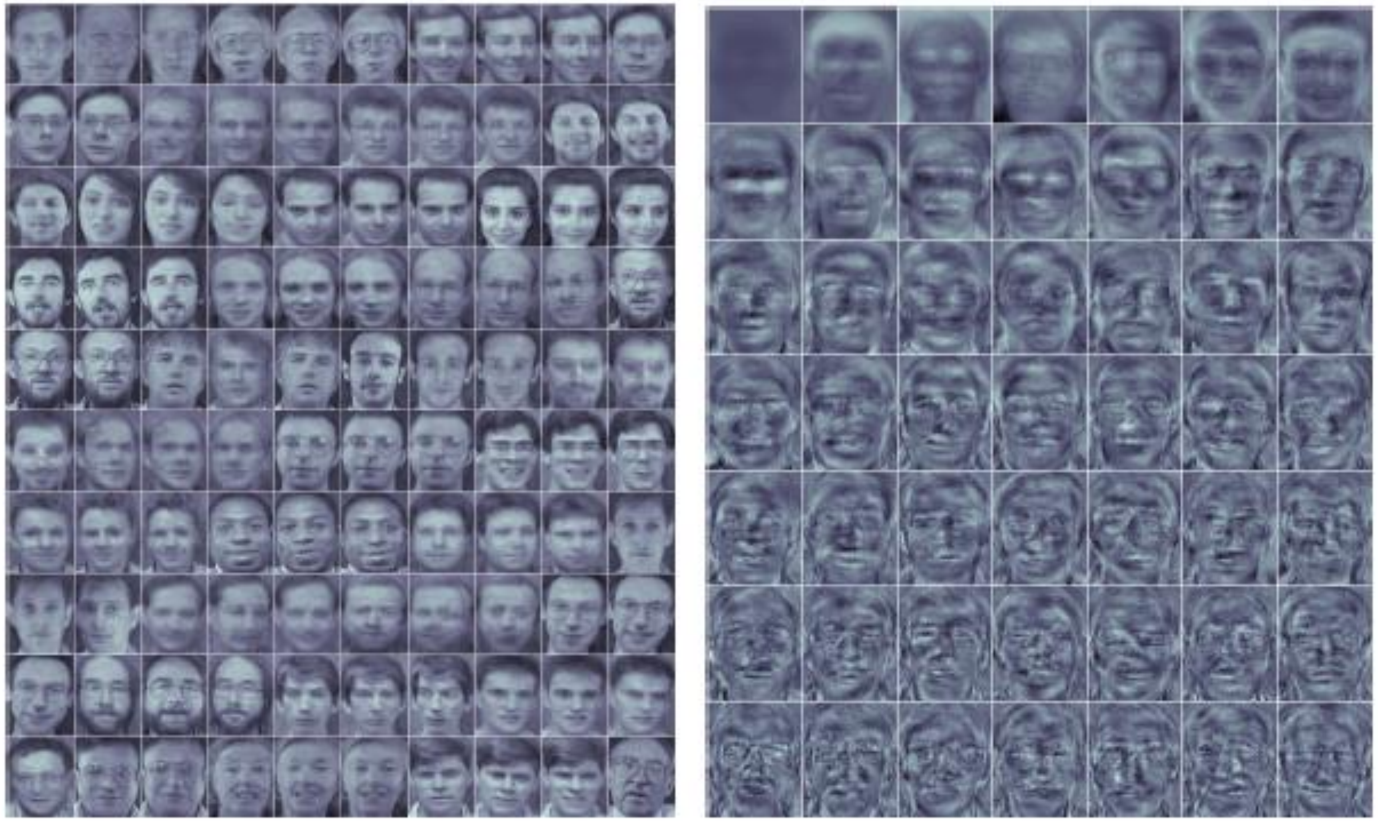


Figure : 100 of the 120 training images (40 people, with 3 images of each person). Each image consists of $92 \times 112 = 10304$ non-negative greyscale pixels. The data is scaled so that, represented as an image, the components of each image sum to 1. The average value of each pixel across all images is 9.70×10^{-5}. This is a subset of the 400 images in the full Olivetti Research Face Database

Reducing the dimension of faces



(a) (b)

Figure : **(a)**: SVD reconstruction of the images using a combination of the 49 eigen-images. **(b)**: The eigen-images are found using SVD of the above data and taking the 49 eigenvectors with largest eigenvalue. The images corresponding to the largest eigenvalues are contained in the first row, and the next 7 in the row below, *etc.*