Ch6: Optimal Feature Generation

- In general, feature generation is a problem-dependent task. However, there are a few general directions common in a number of applications. We focus on three such alternatives.

  - Optimized features based on Scatter matrices (Fisher's linear discrimination).
    - The goal: Given an original set of m measurements $\underline{x} \in \Re^m$, compute $\underline{y} \in \Re^\ell$, by the linear transformation

$$\underline{y} = A^T \underline{x}$$

so that the J_3 scattering matrix criterion involving S_w, S_b is maximized. A^T is an $\ell \times m$ matrix.

❖ Principal Components Analysis (PR-ch3-part3 pp19)

(The Karhunen – Loève transform):

- The goal: Given an original set of m measurements $\underline{x} \in \Re^m$ compute $\underline{y} \in \Re^\ell$

$$\underline{y} = A^T \underline{x}$$

for an orthogonal A, so that the elements of $\underline{y}$ are optimally mutually uncorrelated.

That is

$$E\left[y(i)y(j)\right] = 0, i \neq j$$

- Sketch of the proof:

$$R_y = E\left[\underline{y}\,\underline{y}^T\right] = E\left[A^T \underline{x}\,\underline{x}^T A\right] = A^T R_x A$$

- If A is chosen so that its columns $\underline{a}_i$ are the orthogonal eigenvectors of R_x, then

$$R_y = A^T R_x A = \Lambda$$

where Λ is diagonal with elements the respective eigenvalues λ_i.

- Observe that this is a sufficient condition but not necessary. It **imposes** a specific orthogonal structure on A.

- Properties of the solution
  - Mean Square Error approximation.
  - Due to the orthogonality of A:

$$\underline{x} = \sum_{i=0}^{N-1} y(i)\underline{a}_i, \quad y(i) = \underline{a}_i^T \underline{x}$$

– Define a new vector in the m-dimensional subspace

$$\underline{\hat{x}} = \sum_{i=0}^{m-1} y(i)\underline{a}_i$$

– The Karhunen – Loève transform minimizes the square error:

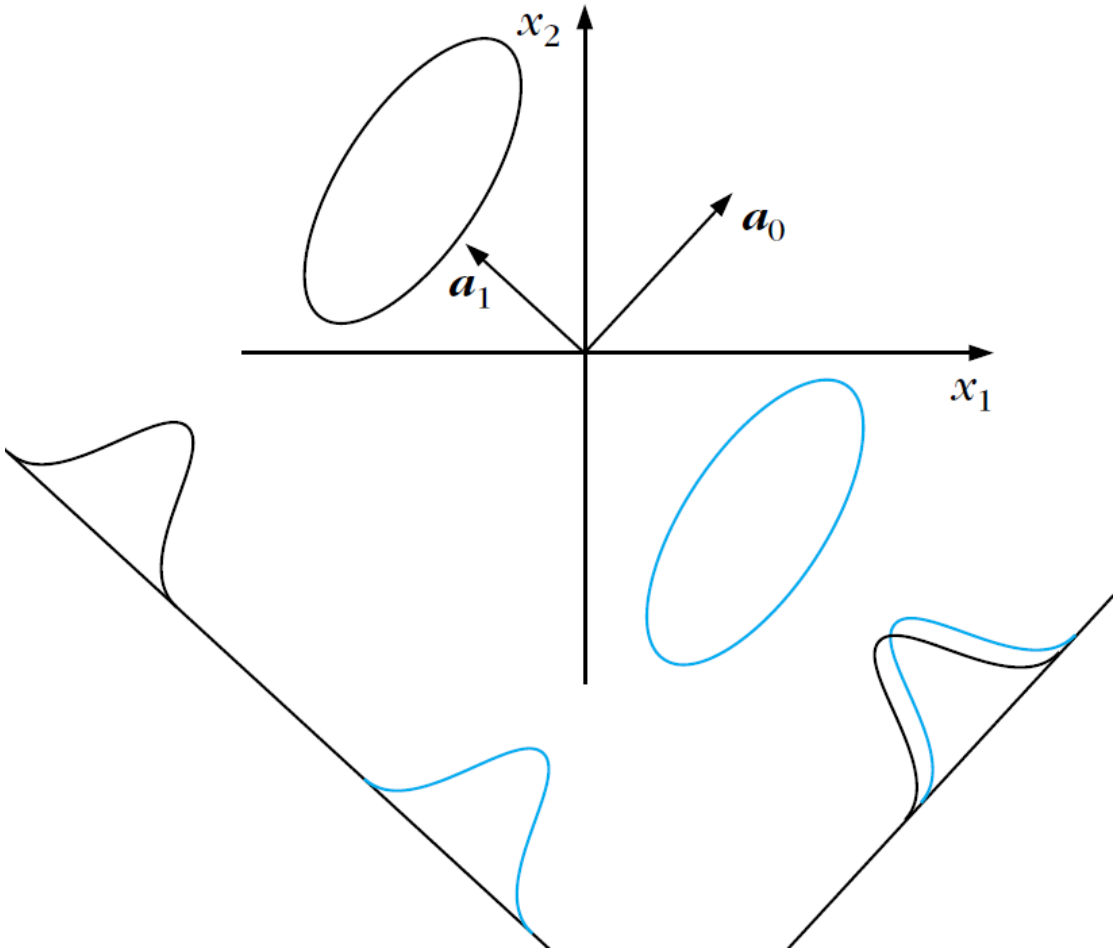
$$E\left[\left\|\underline{x} - \underline{\hat{x}}\right\|^2\right] = E\left[\left\|\sum_{i=m}^{N-1} y(i)\underline{a}_i\right\|^2\right]$$

– The error is:

$$E\left[\left\|\underline{x} - \underline{\hat{x}}\right\|^2\right] = \sum_{i=m}^{N-1} \lambda_i$$

It can be also shown that this is the minimum mean square error compared to **any** other representation of $\underline{x}$ by an m-dimensional vector.

– In other words, $\hat{\underline{x}}$ is the projection of $\underline{x}$ into the subspace spanned by the principal m eigenvectors. However, for Pattern Recognition this is not always the best solution.



The KL transform is not always best for pattern recognition. In this example, projection on the eigenvector with the larger eigenvalue makes the two classes coincide. On the other hand, projection on the other eigenvector keeps the classes separated.

- Total variance: It is easily seen that

$$\sigma^2_{y(i)} = E\left[y^2(i)\right] = \lambda_i$$

That is, the eigenvalues of the input correlation matrix are equal to the variances of the transformed features.

- Thus Karhunen – Loève transform makes the total variance maximum.

Entropy

- The entropy of a process is defined as

$$H_y = -E\left[\ln P_{\underline{y}}(\underline{y})\right]$$

and it is a measure of the randomness of the process.

- Assuming $\underline{y}$ to be a zero mean multivariate Gaussian, then the K-L transform maximizes the entropy:

$H_y = -E\left[\ln P_{\underline{y}}(\underline{y})\right]$ of the resulting $\underline{y}$ process.

Note:

For a zero mean Gaussian multivariable m-dimensional process, the entropy becomes

$$H_y = \frac{1}{2}E[\boldsymbol{y}^T R_y^{-1}\boldsymbol{y}] + \frac{1}{2}\ln|R_y| + \frac{m}{2}\ln(2\pi)$$

$$E[\boldsymbol{y}^T R_y^{-1}\boldsymbol{y}] = E[\text{trace}\{\boldsymbol{y}^T R_y^{-1}\boldsymbol{y}\}] = E[\text{trace}\{R_y^{-1}\boldsymbol{y}\boldsymbol{y}^T\}] = \text{trace}(I) = m$$

$$\ln|R_y| = \ln(\lambda_0\lambda_1\ldots\lambda_{m-1})$$

In words, selection of the m features that correspond to the m largest eigenvalues maximizes the entropy of the process. This is expected because variance and randomness are directly related.

- Subspace Classification. Following the idea of projecting in a subspace, the subspace classification classifies an unknown $\underline{X}$ to the class whose subspace is closer to $\underline{X}$.

  The following steps are in order:

  - For each class, estimate the autocorrelation matrix R_i, and compute the m largest eigenvalues. Form A_i, by using respective eigenvectors as columns.

  - Classify $\underline{X}$ to the class ω_i, for which the norm of the subspace projection is maximum

    $$\left\| A_i^T \underline{x} \right\| > \left\| A_j^T \underline{x} \right\| \quad \forall i \neq j$$

    According to Pythagoras theorem, this corresponds to the subspace to which $\underline{x}$ is closer.

Example 6.2

The correlation matrix of a vector $\boldsymbol{x}$ is given by $\boldsymbol{R}_x$, Compute the KL transform of the input vector.

$$R_x = \begin{bmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{bmatrix}$$

The eigenvalues of R_x are $\lambda_0 = \lambda_1 = 0.4$, $\lambda_2 = 0.1$. Since the matrix R_x is symmetric, we can always construct orthonormal eigenvectors. For this case we have

$$\boldsymbol{a}_0 = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{a}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{a}_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

The KL transform is then given by

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \end{bmatrix} = \begin{bmatrix} 2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{3} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix}$$

where $y(0)$, $y(1)$ correspond to the two largest eigenvalues.

Example 6.3

Figure 6.2 shows 100 points in the two-dimensional space. The points spread around the $x_2 = x_1$ line, and they have been generated by the model $x_2 = x_1 + \epsilon$, where ϵ is a noise source following the uniform distribution in $[-0.5, 0.5]$.

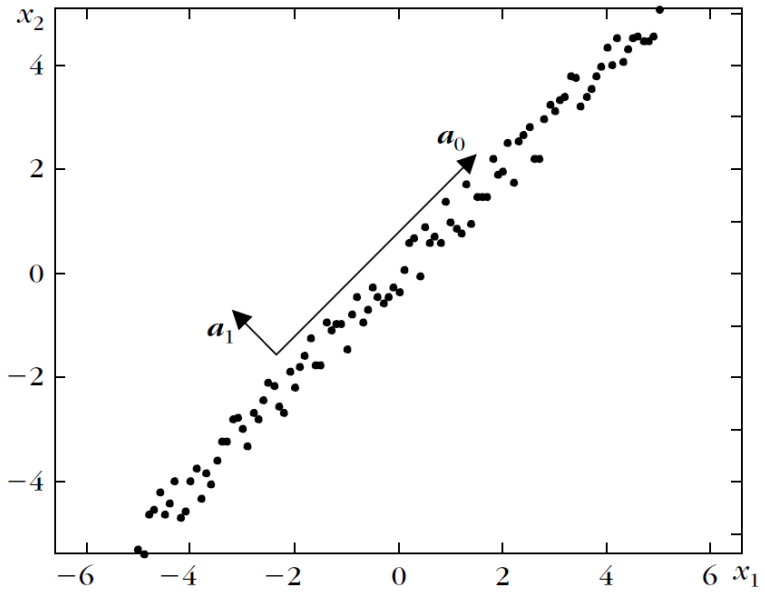
We first compute the covariance matrix and perform an eigendecomposition. The resulting eigenvectors are

$$\boldsymbol{a}_0 = [0.7045, 0.7097]^T, \quad \boldsymbol{a}_1 = [-0.7097, 0.7045]^T$$

corresponding to the eigenvalues

$$\lambda_0 = 17.26, \quad \lambda_1 = 0.04$$

respectively. Observe that $\lambda_0 >> \lambda_1$. Figure 6.2 shows the two eigenvectors. $\boldsymbol{a}_0$, which correspond to the largest eigenvalue, points in the direction where data show maximum variability. Projecting along this direction retains most of the variance. Moreover, according to PCA, the dimensionality of the set is approximately one, due to the large gap between λ_0 and λ_1, which is the correct answer. Also, note, that $\boldsymbol{a}_0$, is (approximately) parallel to the line $x_2 = x_1$.



❖ Independent Component Analysis (ICA) PR-Ch3-part4

In contrast to PCA, where the goal was to produce uncorrelated features, the goal in ICA is to produce statistically independent features. This is a much stronger requirement, involving higher to second order statistics. In this way, one may overcome the problems of PCA, as exposed before.

- The goal: Given $\underline{x}$, compute $\underline{y} \in \Re^{\ell}$

$$\underline{y} = W\,\underline{x}$$

so that the components of $\underline{y}$ are statistically independent. In order the problem to have a solution, the following assumptions must be valid:

- Assume that $\underline{x}$ is indeed generated by a linear combination of independent components

$$\underline{x} = \Phi\,\underline{y}$$

- Φ is known as the mixing matrix and W as the demixing matrix.
- Φ must be invertible or of full column rank.
- Identifiability condition: All independent components, $y(i)$, must be non-Gaussian. Thus, in contrast to PCA that can always be performed, ICA is meaningful for non-Gaussian variables.
- Under the above assumptions, $y(i)$'s can be uniquely estimated, within a scalar factor.

- Common's method: Given $\underline{x}$, and under the previously stated assumptions, the following steps are adopted:
  - Step 1: Perform PCA on $\underline{x}$:

$$\underline{\hat{y}} = A^T \underline{x}$$

  - Step 2: Compute a unitary matrix, $\hat{A}$, so that the fourth order cross-cummulants of the transform vector

$$\underline{y} = \hat{A}^T \hat{y}$$

are zero. This is equivalent to searching for an $\hat{A}$ that makes the squares of the auto-cummulants maximum,

$$\max_{\hat{A}\hat{A}^T = I} \Psi(\hat{A}) = \sum_{i=0}^{N-1} k_4\left(y(i)\right)^2$$

where, $k_4(\cdot)$ is the 4th order auto-cumulant.

- Step 3: $W = \left(A\hat{A}\right)^T$

➢ A hierarchy of components: which ℓ to use? In PCA one chooses the principal ones. In ICA one can choose the ones with the least resemblance to the Gaussian pdf.

characteristic fun. of $p(\boldsymbol{x})$: $\Phi(\Omega) = \int_{-\infty}^{+\infty} p(x)\exp(j\Omega x)\,dx \equiv E[\exp(j\Omega x)]$

the moment generating function:

If jΩ is changed into s $\Phi(s) = \int_{-\infty}^{+\infty} p(x)\exp(sx)\,dx \equiv E[\exp(sx)]$

the 2nd characteristic function of $\boldsymbol{x}$: $\Psi(\Omega) = \ln\Phi(\Omega)$

$\frac{d^n\Phi(s)}{ds^n} \equiv \Phi^{(n)}(s) = E[x^n\exp(sx)]$ ⇒ the nth-order moment of x: $\Phi^{(n)}(0) = E[x^n] \equiv m_n$

theTaylor expansion of the second generating function results in:

$\Psi(s) = \sum_{n=1}^{+\infty} \frac{\kappa_n}{n!}s^n$ where $\kappa_n \equiv \frac{d^n\Psi(0)}{ds^n}$

are known as the cumulants of the random variable x.

For a zero mean random variable

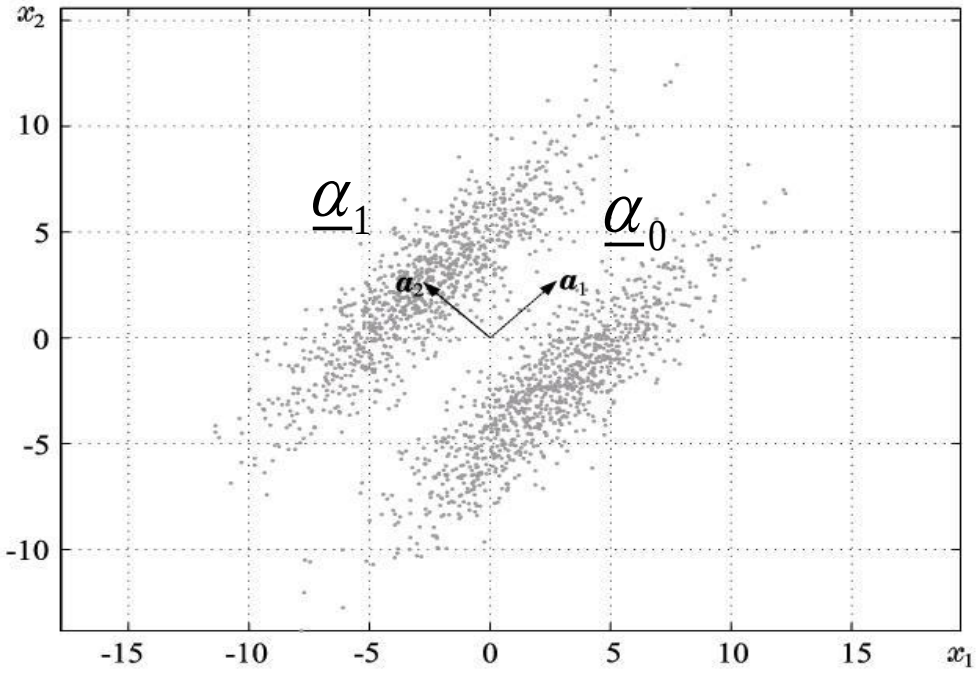
$$\kappa_1(x) = E[x] = 0$$

$$\kappa_2(x) = E[x^2] = \sigma^2$$

$$\kappa_3(x) = E[x^3]$$

$$\kappa_4(x) = E[x^4] - 3\sigma^4$$

(kurtosis)



➢ Example:

$$\boldsymbol{\mu}_1 = [-2.6042, \quad 2.5]^T$$

$$\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$$

$$\Sigma = \begin{bmatrix} 10.5246 & 9.6313 \\ 9.6313 & 11.3203 \end{bmatrix} \Rightarrow W = \begin{bmatrix} -0.7088 & 0.7054 \\ 0.7054 & 0.7088 \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_0^T \end{bmatrix}$$

The principal component is $\underline{\alpha}_0$, thus according to PCA one chooses as y the projection of $\underline{x}$ into $\underline{\alpha}_0$. According to ICA, one chooses as y the projection on $\underline{\alpha}_1$. This is the least Gaussian. Indeed: $K_4(y_1) =$ -1.7, $K_4(y_2) =$ 0.1

Observe that across $\underline{\alpha}_1$, the statistics is bimodal. That is, no resemblance to Gaussian.

Other Feature Generation Methods

- The Singular Value Decomposition (SVD)
- The Discrete Fourier Transform (DFT)
- The Discrete Cosine And Sine Transforms
- The Hadamard Transform
- The Haar Transform
- Discrete Time Wavelet Transform (DTWT)
- The Multiresolution Interpretation
- Wavelet Packets
- Regional Features
- Features For Shape And Size Characterization
- Typical Features For Speech And Audio Classification
- …