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 $\underline{x} \in \mathbb{R}^m$, compute $\underline{y} \in \mathbb{R}^\ell$, 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.

- The basic steps in the proof:
  - $J_3 = \text{trace}(S_w^{-1} S_m)$
  - $S_{yw} = A^T S_{xw} A, \quad S_{yb} = A^T S_{xb} A,$
  - $J_3(A) = \text{trace}\{(A^T S_{xw} A)^{-1} (A^T S_{xb} A)\}$
  - Compute A so that $J_3(A)$ is maximum.

- The solution:

  - Let B be the matrix that diagonalizes simultaneously matrices S_{yw}, S_{yb}, i.e:

    $$B^T S_{yw} B = I, \; B^T S_{yb} B = D$$

    where B, is a $\ell \times \ell$ matrix and D, a $\ell \times \ell$ diagonal matrix.

– Let $C = AB$ an $m \times \ell$ matrix. If A maximizes $J_3(A)$ then

$$\left(S_{xw}^{-1} S_{xb}\right) C = CD$$

The above is an eigenvalue-eigenvector problem. For an M-class problem, $S_{xw}^{-1} S_{xb}$ is of rank M-1.

- If $\ell = M$-1, choose C to consist of the M-1 eigenvectors, corresponding to the non-zero eigenvalues.

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

The above guarantees maximum J_3 value.
In this case: $J_{3,x} = J_{3,y}$.

- For a two-class problem, this results to the well known Fisher's linear discriminant

$$\underline{y} = \left(\underline{\mu}_1 - \underline{\mu}_2\right) S_{xw}^{-1} \underline{x}$$

For Gaussian classes, this is the optimal Bayesian classifier, with a difference of a threshold value .

- If $\ell < M\text{-}1$, choose the ℓ eigenvectors corresponding to the ℓ largest eigenvectors.
- In this case, $J_{3,y} < J_{3,x}$, that is there is loss of information.

– Geometric interpretation. The vector $\underline{y}$ is the projection of $\underline{x}$ onto the subspace spanned by the eigenvectors of $S_{xw}^{-1}S_{xb}$.

❖ Principal Components Analysis

(The Karhunen – Loève transform):

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

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

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:

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

– Define

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

– The Karhunen-Loève transform minimizes the square error:

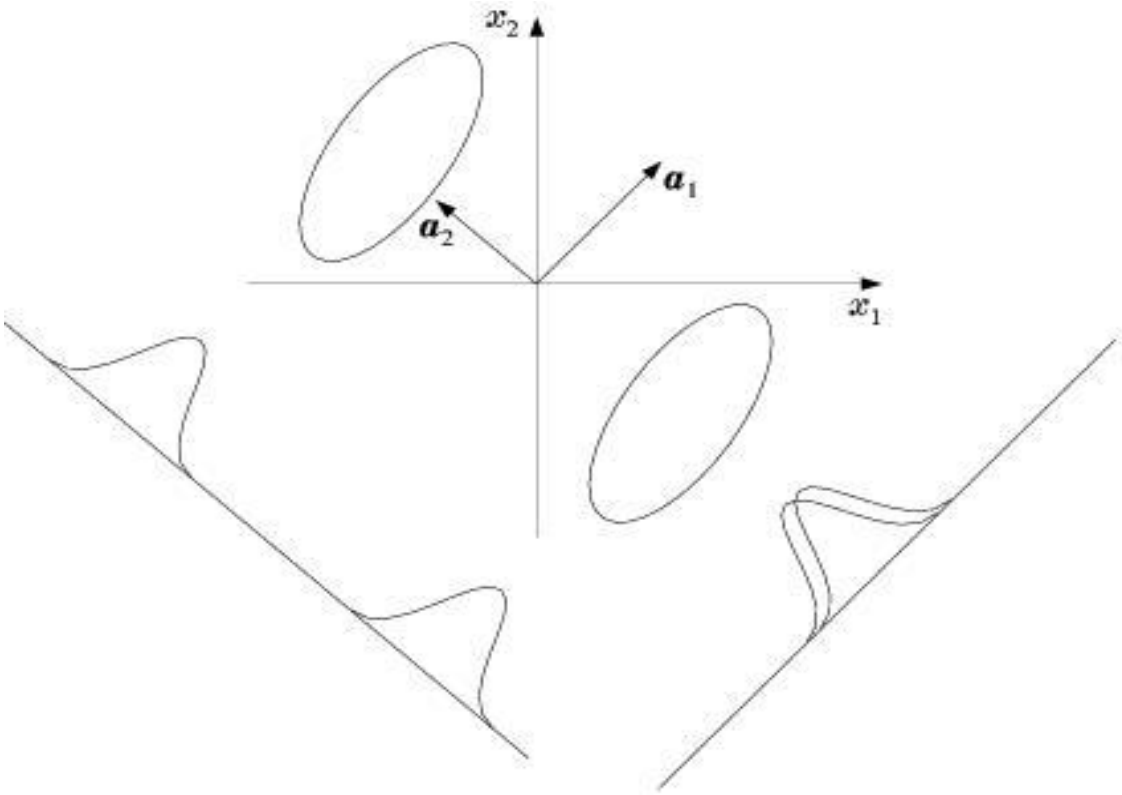
$$E\left[\left\|\underline{x} - \underline{\hat{x}}\right\|^2\right] = E\left[\left\|\sum_{i=\ell}^{m} y(i)\underline{a}_i\right\|^2\right]$$

– The error is:

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

It can be also shown that this is the minimum mean square error compared to **any** other representation of x by an ℓ-dimensional vector.

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



- Total variance: It is easily seen that

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

Thus Karhunen-Loève transform makes the total variance maximum.

- 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.

➢ 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_j, 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.

❖ Independent Component Analysis (ICA)

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 $\underline{y} \in \mathbb{R}^{\ell}$

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

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 $\underline{x}$:

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

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

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

unitary: $\hat{A}^* \hat{A} = \hat{A}\hat{A}^* = I$

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

$$\max_{\hat{A}\hat{A}^T = I} \Psi(\hat{A}) = \sum \kappa_4 \left(y(i) \right)^2$$

where, $\kappa_4 \left(\cdot \right)$ is the 4th order auto-cumulant.

Cummulants:

$$\kappa_1(y(i)) = E[y(i)] = 0$$

$$\kappa_2(y(i)y(j)) = E[y(i)y(j)]$$

$$\kappa_3(y(i)y(j)y(k)) = E[y(i)y(j)y(k)]$$

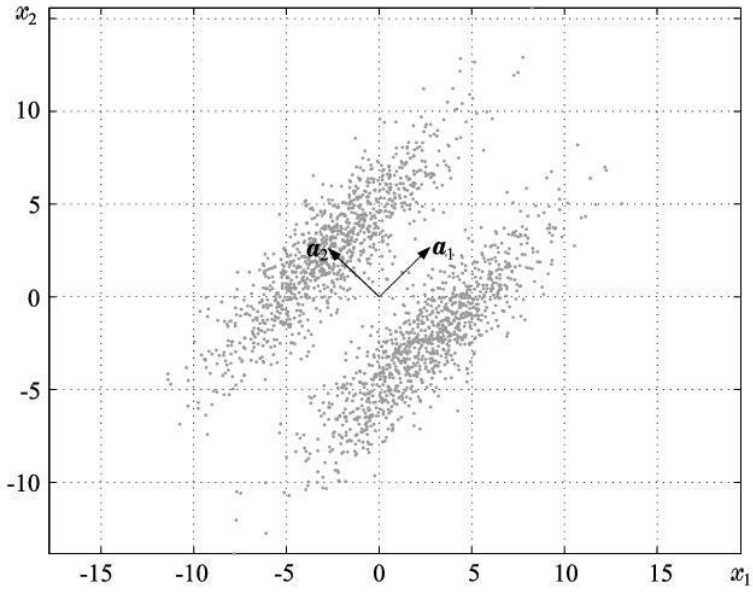
and the fourth-order cumulants are given by

$$\kappa_4(y(i)y(j)y(k)y(r)) = E[y(i)y(j)y(k)y(r)] - E[y(i)y(j)]E[y(k)y(r)]$$
$$- E[y(i)y(k)]E[y(j)y(r)]$$
$$- E[y(i)y(r)]E[y(j)y(k)]$$

- 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.

➢ Example:



The principal component is $\underline{\alpha}_1$, thus according to PCA one chooses as y the projection of $\underline{x}$ into $\underline{\alpha}_2$. 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 $\underline{\alpha}_2$, the statistics is bimodal. That is, no resemblance to Gaussian.