7 Eigenvalue and Eigenvector decomposition of data matrix

- Recognition in general is to classify an observed data into one of the classes.
- Observed data often have multiple components and data of a same class are in general not exact same and have some variation. We thus represent the observed data by a random vector, a column vector $\mathbf{x}$.
- The probability distribution fully characterizes a random vector.
- In most cases, the probability distribution is unknown.
- Recognition via Machine Learning is to estimate the probability distribution directly or indirectly from the known training data/samples $\mathbf{x}_i$.
- A Gaussian PDF is uniquely specified by its mean and covariance matrix.

7 Eigenvalue and Eigenvector decomposition of data matrix

- Machine Learning is driven by training data, a collection of training samples
- A training sample (example) can be represented by a n-dimensional column vector $\mathbf{x}_i$.
- A training data set contains q n-dimensional training samples of c classes

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$$

- The number of training samples of class ω_j is denoted by q_j, j = 1, 2,… c.
- The covariance matrix of class ω_j , class-conditional covariance matrix is computed as

$$\boldsymbol{\Sigma}_j = \frac{1}{q_j}\sum_{X_i \in \omega_j} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T, \quad \text{where } \boldsymbol{\mu}_j = \frac{1}{q_j}\sum_{X_i \in \omega_j} \mathbf{x}_i$$

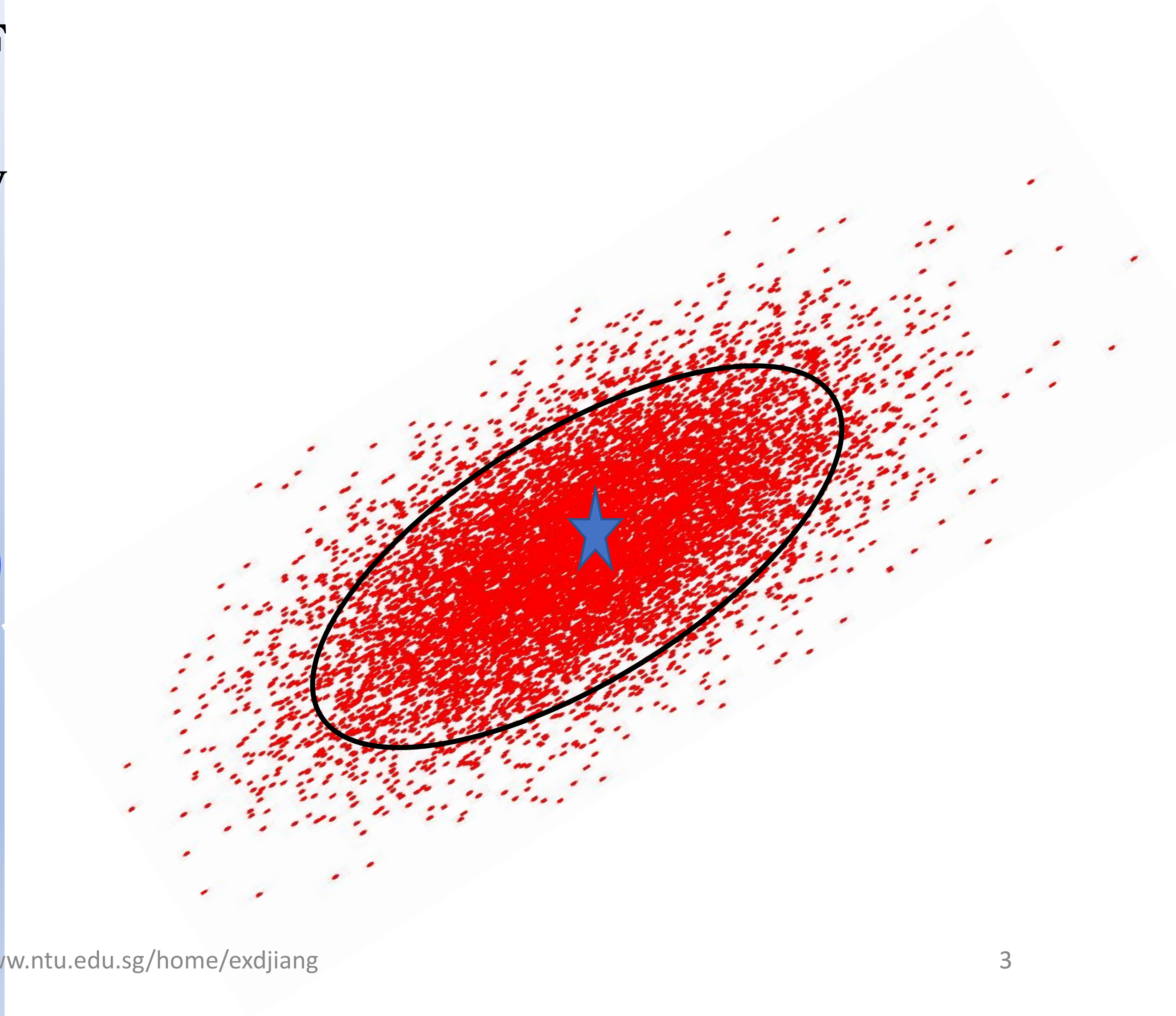
- The pooled covariance matrix can be defined as:

$$\boldsymbol{\Sigma} = \frac{1}{c}\sum_{j=1}^{c} \boldsymbol{\Sigma}_j \quad \text{or better using} \quad \boldsymbol{\Sigma} = \sum_{j=1}^{c} \frac{q_j}{q} \boldsymbol{\Sigma}_j$$

7 Eigenvalue and Eigenvector decomposition of data matrix

- Data drawn from a Gaussian PDF are shown by red points. All points along the ellipse shaped by the covariance matrix has the same Mahalanobis distance to its center, the mean vector. It agrees to the contour of Gaussian PDF

$$d_{Maj} = (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)$$



How about the Euclidian distance?

$$d_{Euj} = ||\mathbf{x} - \boldsymbol{\mu}_j||^2 = (\mathbf{x} - \boldsymbol{\mu}_j)^T (\mathbf{x} - \boldsymbol{\mu}_j)$$

7 Eigenvalue and Eigenvector decomposition of data matrix

- The pooled covariance matrix $\mathbf{\Sigma}^{\mathrm{w}} = \sum_{j=1}^{c} \frac{q_j}{q} \mathbf{\Sigma}_j$ is also called within class scatter matrix

$$\mathbf{S}^{w} = \sum_{j=1}^{c} \frac{q_j}{q} \mathbf{\Sigma}_j$$

- The between class scatter matrix of c classes is defined as

$$\mathbf{S}^{b} = \sum_{j=1}^{c} \frac{q_j}{q} (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T, \text{ obviously: } \boldsymbol{\mu} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_i = \sum_{j=1}^{c} \frac{q_j}{q} \boldsymbol{\mu}_j$$

- It is the covariance matrix of class means.
- The covariance matrix over all training data $\mathbf{\Sigma}$ is also called total scatter matrix

$$\mathbf{S}^{t} = \frac{1}{q} \sum_{i=1}^{q} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{q} \sum_{i=1}^{q} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T = \frac{1}{q} \mathbf{X}\mathbf{X}^T$$

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}, \qquad \mathbf{X} = \begin{bmatrix} \tilde{\mathbf{x}}_1 & \tilde{\mathbf{x}}_2 & \dots & \tilde{\mathbf{x}}_q \end{bmatrix}$$

7 Eigenvalue and Eigenvector decomposition of data matrix

- It is easy to prove that the relationship of these matrices:

$$\mathbf{S}^t = \frac{1}{q}\sum_{i=1}^{q}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{q}\sum_{i=1}^{q}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T = \frac{1}{q}\mathbf{X}\mathbf{X}^T$$

$$\mathbf{S}^t = \mathbf{S}^w + \mathbf{S}^b = \frac{1}{q}\sum_{j=1}^{c} q_j \boldsymbol{\Sigma}_j + \mathbf{S}^b$$

- Obviously, the scatter matrix or covariance matrix is a *nxn* symmetry matrix.

$$\mathbf{S}^t = (\mathbf{S}^t)^T$$

7 Eigenvalue and Eigenvector decomposition of data matrix

- Eigenvalue and eigenvector of a nxn matrix Σ is defined mathematically by:

$$\Sigma\phi_i = \lambda_i \phi_i, \qquad i = 1,\ 2,\ \ldots,\ n$$

- If Σ is a symmetry matrix, eigenvectors corresponding to the distinct eigenvalues λ_1, λ_2, …, λ_n are orthogonal. Take the unit length of ϕ_1, ϕ_2, …, ϕ_n

$$\phi_i^T \phi_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- Prove it ! (orthogonal+unit length is called orthonormal)

7 Eigenvalue and Eigenvector decomposition of data matrix

We have $\boldsymbol{\Sigma}\phi_k = \lambda_k \phi_k$ and $\boldsymbol{\Sigma}\phi_j = \lambda_j \phi_j, \quad k \neq j$

$\therefore \quad \phi_j^T \left(\boldsymbol{\Sigma}\phi_k\right) = \phi_j^T \left(\lambda_k \phi_k\right)$ and $\left(\boldsymbol{\Sigma}\phi_j\right)^T \phi_k = \left(\lambda_j \phi_j\right)^T \phi_k$

$\therefore \quad \phi_j^T \boldsymbol{\Sigma}\phi_k = \lambda_k \phi_j^T \phi_k$ and $\phi_j \boldsymbol{\Sigma}^T \phi_k = \lambda_j \phi_j^T \phi_k$

$\because \boldsymbol{\Sigma}^T = \boldsymbol{\Sigma},$ we have $\lambda_k \phi_j^T \phi_k = \lambda_j \phi_j^T \phi_k$

$\therefore (\lambda_k - \lambda_j)\phi_j^T \phi_k = 0$

$\because (\lambda_k - \lambda_j) \neq 0$

$\therefore \phi_j^T \phi_k = 0$

7 Eigenvalue and Eigenvector decomposition of data matrix

- Let $\mathbf{\Phi}$ be the orthonormal matrix formed by the eigenvectors

$$\mathbf{\Phi} = [\phi_1 \;\; \phi_2 \;\; \ldots \;\; \phi_n]$$

$$\text{Obviously: } \mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{I} \qquad \therefore \mathbf{\Phi}^{-1} = \mathbf{\Phi}^T, \qquad \therefore \mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{I}$$

Let Λ be a diagonal matrix:

$$\mathbf{\Lambda} = diag(\lambda_1 \;\; \lambda_2 \;\; \ldots \;\; \lambda_n) = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

From $$\mathbf{\Sigma}\phi_i = \lambda_i\phi_i, \qquad i = 1,\; 2,\; \ldots,\; n$$

- We have $$\mathbf{\Sigma\Phi} = \mathbf{\Phi\Lambda} \qquad \therefore \mathbf{\Sigma} = \mathbf{\Phi\Lambda\Phi}^T, \text{ or } \mathbf{\Lambda} = \mathbf{\Phi}^T\mathbf{\Sigma\Phi}$$

7 Eigenvalue and Eigenvector decomposition of data matrix

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

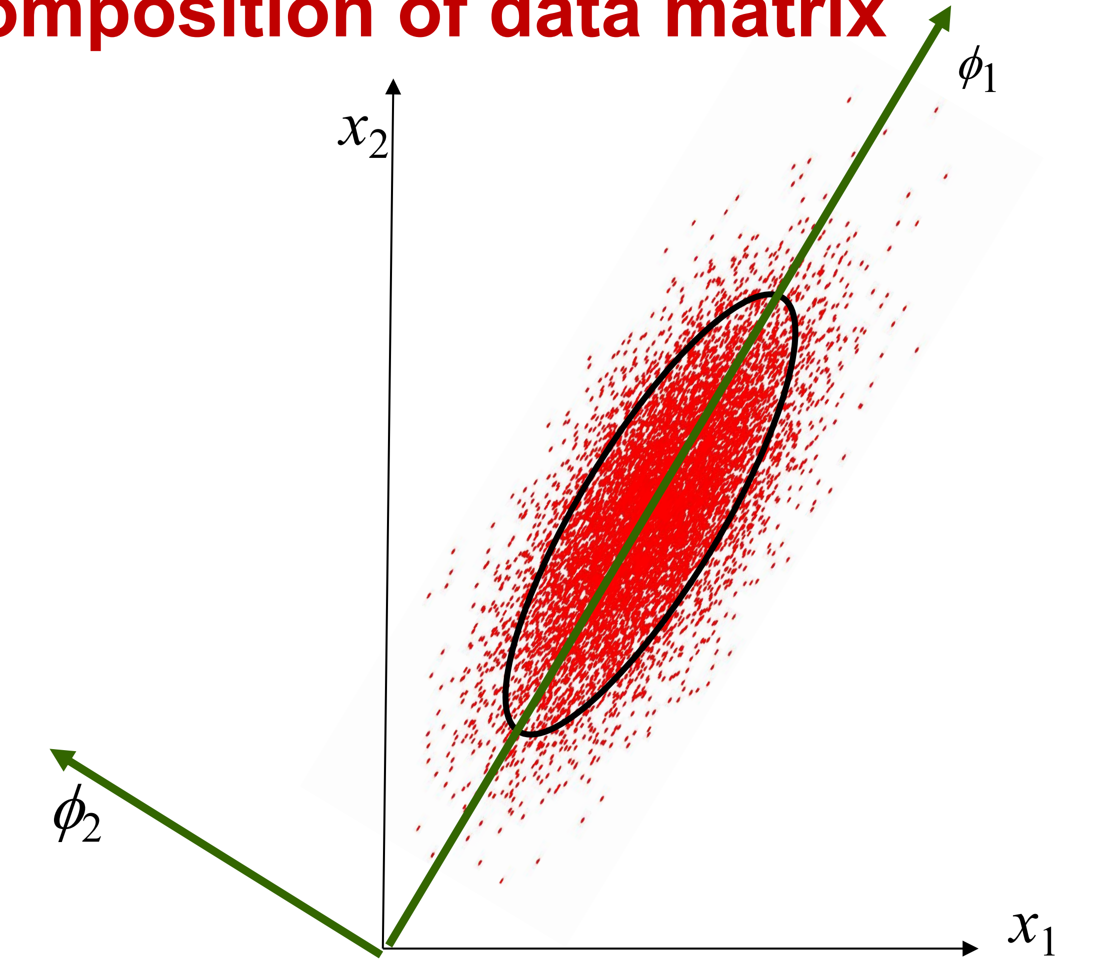
$$= [\phi_1 \quad \phi_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} [\phi_1 \quad \phi_2]^T$$

$$= \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \boldsymbol{\Phi}^T \boldsymbol{\Sigma}_x \boldsymbol{\Phi}$$

$$= \boldsymbol{\Phi}^T \mathbf{X}\mathbf{X}^T \boldsymbol{\Phi} = \boldsymbol{\Phi}^T \mathbf{X} \left(\boldsymbol{\Phi}^T \mathbf{X} \right)^T = \mathbf{Y}\mathbf{Y}^T = \boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}$$

where: $\mathbf{Y} = \boldsymbol{\Phi}^T \mathbf{X} = \mathbf{W}^T \mathbf{X}$



Eigenvalue/Eigenvector decomposition for classification

- We have learned that if data of all classes obey Gaussian distribution, the Bayes optimal decision becomes to evaluate:

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

  The first part is Mahalanobis distance between X and μ_i, where b_i is a threshold for user to control the error rate of a class at a price of other classes.
- Problem is that human knowledge cannot provide the class mean and covariance matrix of the data population, which can only be estimated or learned by machine from the available training samples.
- These estimates may be different from the ground truth. To achieve robust accurate classification we need to study how these estimation errors impact to the classification accuracy.

Eigenvalue/Eigenvector decomposition for classification

- However, the inverse of the *n* by *n* covariance matrix is a great burden for us to clearly see the problems and find solutions.

- It is very difficult to study the problems of the covariance matrix directly as it carries two different kinds of information by n^2 estimates: data variations and correlations, which could have millions of parameters estimated by a limited number of training samples.

- Eigen-decomposition provides an effective tool to simplify the problem. It provides us a powerful tools to study, analyze and find problems of the high dimensional matrix.

- As the covariance matrix is symmetric, its eigenvectors provide an orthogonal basis for *n*-space.

Eigenvalue/Eigenvector decomposition for classification

Recall that from the basic definition of eigen-decomposition:

$$\mathbf{\Sigma\Phi} = \mathbf{\Phi\Lambda} \qquad \mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{I} = \mathbf{\Phi\Phi}^T \qquad \mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$$

$$\mathbf{\Phi}^T\mathbf{\Sigma\Phi} = \mathbf{\Lambda}, \ \text{or} \ \mathbf{\Sigma} = \mathbf{\Phi\Lambda\Phi}^T$$

- We have

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{\Phi}_i\mathbf{\Lambda}_i\mathbf{\Phi}_i^T)^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{\Phi}_i^T)^{-1}\mathbf{\Lambda}_i^{-1}\mathbf{\Phi}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_{ii}) + b_i$$

$$= -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Phi}_i\mathbf{\Lambda}_i^{-1}\mathbf{\Phi}_i^T(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\tfrac{1}{2}[\mathbf{\Phi}_i^T(\mathbf{x}-\boldsymbol{\mu}_i)]^T\mathbf{\Lambda}_i^{-1}[\mathbf{\Phi}_i^T(\mathbf{x}-\boldsymbol{\mu}_i)] + b_i$$

For symbolic simplicity, the index i is omitted where k is necessary.

$$= -\tfrac{1}{2}\sum_{k=1}^{n}\frac{(z_k-\bar{z}_k)^2}{\lambda_k} + b_i$$

$$z_k = \phi_k^T\mathbf{x}, \quad \bar{z}_k = \phi_k^T\boldsymbol{\mu}_i,$$

$$\mathbf{\Phi}_i = [\phi_1, \phi_1, \ldots, \phi_n]$$

Eigenvalue/Eigenvector decomposition for classification

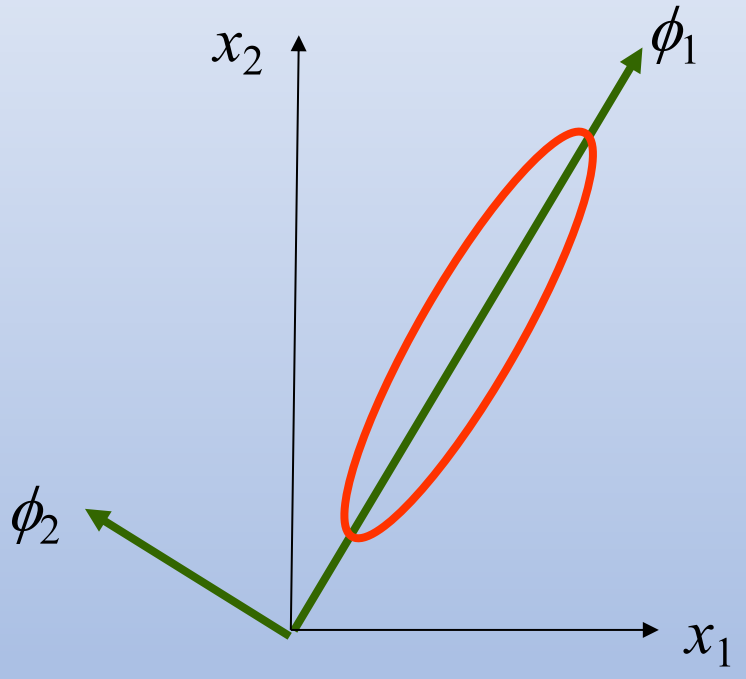
- In the eigen-space spanned by ϕ_k, we can handle the complex inverse of the n by n covariance matrix in a convenient scalar form!

$$\mathbf{z} = \{z_k\}_1^n = \mathbf{\Phi}_i^T \mathbf{x} = \{\phi_k\}_1^{n\ T} \mathbf{x}$$

$$\bar{z}_k = \phi_k^T \boldsymbol{\mu}_i$$

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\tfrac{1}{2}\sum_{k=1}^{n}\frac{(z_k-\bar{z}_k)^2}{\lambda_k} + b_i$$



- Problem is now evident if some eigenvalues are unreliable, especially if some eigenvalues are small and approach to zero.