3. Redução de Dimensionalidade

Uma das grandes questões inerentes à maioria dos problemas de reconhecimento de padrões está relacionada à dimensionalidade dos dados a serem analisados. Em geral, um padrão é representado por um número n de valores que pode ser caracterizado como um vetor num espaço n-dimensional. No reconhecimento de faces, em particular, este número n está associado à resolução da imagem a ser processada, por exemplo, uma imagem de 64x64 pontos (*pixels*) possui n igual a 4096. Desta forma, trabalhar com diversos exemplos de diferentes padrões requer um considerável armazenamento e processamento dos dados, podendo, com isto, inviabilizar qualquer implementação computacional. Assim sendo, faz-se necessário o estudo de técnicas de redução de dimensionalidade que permitam a transformação desses dados pertencentes a um espaço n-dimensional em um sub-espaço p-dimensional, onde p é muito menor que n ($p << n$). No entanto, toda redução de dimensionalidade implica numa perda de informação, e esta última pode vir a ser fundamental para a discriminação dos padrões. Por isto, o objetivo principal das técnicas de redução de dimensionalidade é preservar o máximo possível da informação relevante dos dados.

Duas técnicas de redução de dimensionalidade, a Análise de Componentes Principais e a Análise Multivariada de Discriminantes, são descritas a seguir.

3.1 Análise de Componentes Principais

3.1.1 Descrição

A técnica de Análise de Componentes Principais (ACP) ou *Principal Component Analysis* (PCA) foi descrita pela primeira vez por Karl Pearson, em 1901. Ele acreditava que a PCA poderia ser a solução para alguns problemas biométricos da época, embora não tivesse proposto nenhum método prático de cálculo que levasse em consideração mais de duas ou três variáveis. Este método foi descrito bem mais tarde, em 1933, por Hotteling, e a partir de então a técnica de análise de componentes principais ficou também conhecida como transformada de Hotteling ou *Hotteling´s Transform*. No entanto, os cálculos eram ainda desanimadores quando tratava-se de problemas que envolviam dezenas de variáveis. Foi nos tempos de hoje, com a disponibilidade e avanço dos computadores atuais, que esta técnica difundiu-se amplamente.

A análise de componentes principais tem por objetivo determinar uma transformação linear de um dado conjunto n de variáveis originais $X_1, X_2, ..., X_n$ em um novo conjunto p de variáveis $Y_1, Y_2, ..., Y_p$ de tal forma que estas últimas sejam descorrelacionadas. Além disso, essa transformação deve ser tal que as novas variáveis $Y_1, Y_2, ..., Y_p$ tenham a mesma variância total das variáveis originais e estejam ordenadas de tal forma que $\mathrm{Var}(Y_1) \geq \mathrm{Var}(Y_2) \geq ... \geq \mathrm{Var}(Y_p)$. Estas novas variáveis Y_i são chamadas de componentes principais.

Uma das principais aplicações da PCA é a redução de dimensionalidade através da eliminação das variáveis originais de menor variância. Embora a variabilidade total de um

sistema seja definida por n variáveis, geralmente muito desta variabilidade pode ser explicada por um número bem menor, p, de componentes principais. Se isto for verdade, ou seja, se as variáveis originais medirem explícita ou implicitamente informações redundantes, então a quantidade de dados relevantes contida nestas p componentes principais é equivalente àquela existente nas n variáveis originais. Assim, o conjunto de dados originais que consiste de k medidas de n variáveis é reduzido para um conjunto de k medidas de p componentes. Por isso, em muitas aplicações o PCA é utilizado como uma espécie de pré-processamento dos dados, servindo como entrada para outros modelos numéricos tais como regressão múltipla, redes neurais, etc. A vantagem, neste caso, está na redução do número de parâmetros do modelo imediatamente seguinte ao PCA.

3.1.2 Procedimento de Cálculo das Componentes Principais

Algebricamente, pode-se dizer que as componentes principais de um dado conjunto n de variáveis $X_1, X_2, ..., X_n$ nada mais são do que uma determinada transformação linear dessas mesmas variáveis X_i.

Seja um vetor aleatório $\mathbf{x}$ formado por um conjunto de variáveis aleatórias X_i de tal forma que $\mathbf{X}^T = [X_1, X_2, ..., X_n]$, onde $\mathbf{X} \in \Re^{\mathbf{n}}$. A matriz de dados de $\mathbf{X}$ é, portanto, composta por n colunas e k linhas onde k corresponde ao número de observações das variáveis aleatórias X_i. Assume-se que $\mathbf{X}$ população tenha matriz de covariância $\mathbf{\Sigma}$. Considera-se, então, a combinação linear abaixo:

$$\begin{array}{lll} Y_1 = & \boldsymbol{v}_1^T\mathbf{X} & = v_{11}X_1 + v_{21}X_2 + \ldots + v_{n1}X_n \\ Y_2 = & \boldsymbol{v}_2^T\mathbf{X} & = v_{12}X_1 + v_{22}X_2 + \ldots + v_{n2}X_n \\ \vdots & \vdots & \vdots \\ Y_p = & \boldsymbol{v}_p^T\mathbf{X} & = v_{1p}X_1 + v_{2p}X_2 + \ldots + v_{np}X_n \end{array}$$

cujas variáveis Y_i possuem, respectivamente, variância e covariância dadas pelas fórmulas

$$\begin{array}{rlrl} \mathrm{Var}(Y_i) = & \boldsymbol{v}_i^T\boldsymbol{\Sigma}\boldsymbol{v}_i & i = & 1,2,\ldots,p \\ \mathrm{Cov}(Y_i, Y_j) = & \boldsymbol{v}_i^T\boldsymbol{\Sigma}\boldsymbol{v}_j & i,j = & 1,2,\ldots,p \end{array}$$

As componentes principais são as combinações lineares Y_i, descorrelacionadas, cujas variâncias são as maiores possíveis. Desta forma, a primeira componente principal Y_1 é dada pela combinação linear de máxima variância, ou seja, quando $\mathrm{var}(Y_1) = \max(\boldsymbol{v}_i^T\boldsymbol{\Sigma}\boldsymbol{v}_i)$. É importante ressaltar que o módulo do vetor de projeção $\boldsymbol{v}_i$ deve ser tal que não influencie o cálculo da respectiva variância. Por isso, $\|\boldsymbol{v}_i\| = \boldsymbol{v}_i^T\boldsymbol{v}_i = 1$. Define-se, portanto,

Primeira comp. principal $= Y_1 =$ combinação linear $\boldsymbol{v}_1^T\mathbf{X}$ que maximiza $\mathrm{Var}(\boldsymbol{v}_1^T\mathbf{X}) = \boldsymbol{v}_1^T\boldsymbol{\Sigma}\boldsymbol{v}_1$,

onde $\boldsymbol{v}_1^T\boldsymbol{v}_1 = 1$;

Segunda comp. principal $= Y_2 =$ combinação linear $\boldsymbol{v}_2^T\mathbf{X}$ que maximiza $\mathrm{Var}(\boldsymbol{v}_2^T\mathbf{X}) = \boldsymbol{v}_2^T\boldsymbol{\Sigma}\boldsymbol{v}_2$,

onde $\boldsymbol{v}_2^T\boldsymbol{v}_2 = 1$ e $\mathrm{Cov}(Y_1, Y_2) = 0$;

I-ésima comp. principal $= Y_i =$ combinação linear $\boldsymbol{v}_i^T\mathbf{X}$ que maximiza $\mathrm{Var}(\boldsymbol{v}_i^T\mathbf{X}) = \boldsymbol{v}_i^T\boldsymbol{\Sigma}\boldsymbol{v}_i$,

onde $\boldsymbol{v}_i^T\boldsymbol{v}_i = 1$ e $cov(Y_j, Y_i) = 0$ para todo $j < i$.

Então, admitindo que a matriz de covariância $\mathbf{\Sigma}$ de $\mathbf{X}$ tenha como autovalor-autovetor os respectivos pares $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, é um resultado provado [John92] que

$$\max_{v \perp e_1, e_2, \ldots, e_j} \boldsymbol{v}^T \mathbf{\Sigma} \boldsymbol{v} = \boldsymbol{\lambda}_{j+1} \quad \text{para } \boldsymbol{v} = \mathbf{e}_{j+1}, \qquad \text{onde } j = 1,2,\ldots, p-1$$

Portanto, a i-ésima componente principal é dada por

$$Y_i = \mathbf{e}_i^T \mathbf{X} = e_{1i} X_1 + e_{2i} X_2 + \cdots + e_{ni} X_n \qquad i = 1,2,\ldots, p$$

que implica em

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_i = \lambda_i \quad i = 1,2,\ldots, p \\ \mathrm{Cov}(Y_i, Y_j) &= \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_j = 0 \quad i \neq j \end{aligned}$$

Desta forma, as componentes principais Y_i são descorrelacionadas e possuem variâncias máximas iguais aos autovalores de $\mathbf{\Sigma}$. A soma desses autovalores é igual a variabilidade total das componentes principais e prova-se [John92] que esta variabilidade é também igual à variância total das variáveis originais X_i. Então,

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

e, consequentemente, a proporção da variância explicada pela j-ésima componente principal pode ser escrita como

$$\begin{pmatrix} \text{Variança explicada pela} \\ \text{j} - \text{ésima componente principal} \end{pmatrix} = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad j = 1,2,\ldots,p$$

Em muitas situações, principalmente quando as variáveis X_i possuem forte correlação entre si, quase 80% ou 90% da variância total dos dados originais pode ser explicada pelas duas ou três primeiras componentes principais, permitindo, assim, uma expressiva redução de dimensionalidade sem grandes perdas de informação relevante.

Uma outra consideração algébrica importante quanto ao cálculo das componentes principais se refere à normalização das variáveis originais X_i. Comumente, obtem-se as componentes principais a partir da matriz de correlação $\mathbf{R}$ das variáveis $X_1, X_2, \ldots, X_n$ e não da respectiva matriz de covariância $\mathbf{\Sigma}$. Isto é feito através da transformação das variáveis originais X_i em variáveis Z_i com média igual a 0 e variância igual a 1. As fórmulas abaixo explicitam este cálculo:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}, Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}, \cdots, Z_p = \frac{(X_n - \mu_n)}{\sqrt{\sigma_{nn}}}$$

onde μ_i e σ_{ii} são, respectivamente, média e variância da variável X_i. Portanto, dado $\mathbf{Z}^T = [Z_1, Z_2, \ldots, Z_n]$, tem-se $Cov(\mathbf{Z}) = \mathbf{R}$. As componentes principais são obtidas a partir desta nova matriz de dados – a matriz de correlação $\mathbf{R}$ de X_i. Esta normalização se faz

necessária quando as variáveis originais X_i representam valores em escalas bem distintas. Por exemplo, supondo $n = 2$, se a variável X_1 medir valores na escala de 10^3 ou 10^4 e a variável X_2 representar valores entre 0 e 1, então a variância total será fortemente influenciada pela variável de maior peso, ou seja, pela variável X_1. Isto implicará numa análise de componentes principais incorreta.

Pode-se resumir, então, o procedimento de cálculo das componentes principais na seguinte ordem de passos:

1. Se as variáveis originais $X_1, X_2, ..., X_n$ representarem valores em escalas bem distintas, então essas variáveis devem ser normalizadas de tal forma que tenham média 0 e variância 1.
2. Calcula-se a matriz de covariância das variáveis. Se a condição do passo 1 for satisfeita, então esta matriz de covariância é, na verdade, a matriz de correlação das variáveis originais.
3. Calcula-se os pares de autovalor-autovetor da matriz do passo 2. Estes pares devem ser ordenados em ordem decrescente de autovalores.
4. Eliminam-se as componentes principais cujos autovalores são desprezíveis com relação a variância total dos dados originais.

3.1.3 Interpretação Geométrica das Componentes Principais

A análise de componentes principais é uma transformação linear algébrica de variáveis. Sabe-se que toda transformação linear eqüivale, em termos geométricos, a uma rotação nos eixos coordenados definidos pelas variáveis originais do sistema. Para a análise de componentes principais (PCA), isto se dá de tal forma que esses eixos, rotacionados e

definidos pelas componentes principais, representam as direções de máxima variabilidade dos dados originais fazendo com que os valores iniciais projetados nestes novos eixos sejam descorrelacionadas entre si.

Para facilitar a visualização das operações implícitas no PCA, convém considerar o caso gaussiano. Supõe-se que os dados $\mathbf{x_1, x_2, ..., x_k}$ representem $\mathbf{k}$ observações independentes de uma população $\mathbf{X}$ n-dimensional cuja distribuição é normal com média $\boldsymbol{\mu}$ e covariância $\boldsymbol{\Sigma}$, ou seja, $\mathbf{X} \sim \mathrm{N_n}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Logo, as amostras de $\mathbf{X}$ representadas nas p respectivas componentes principais possuem média 0 e matriz de covariância $\boldsymbol{\Lambda}$, isto é, $\mathbf{N_p(0, \Lambda)}$. Demonstra-se [John92] que a matriz $\boldsymbol{\Lambda}$ é uma matriz diagonal composta pelos elementos $\lambda_1, \lambda_2, ..., \lambda_p$ e os pares $(\lambda_i, \mathbf{e}_i)$ são descritos pelos autovalores-autovetores de $\boldsymbol{\Sigma}$. Pode-se aproximar $\boldsymbol{\mu}$ por $\bar{\mathbf{x}}$ e $\boldsymbol{\Sigma}$ por $\mathbf{S}$, onde $\bar{\mathbf{x}}$ é a vetor-média das k observações de $\mathbf{X}$ e $\mathbf{S}$ a correspondente matriz de covariância. Se a matriz $\mathbf{S}$ for definida positiva, ou melhor, se $\mathbf{S}$ for inversível e possuir todos os autovalores maiores que zero, então os valores de $\mathbf{x}$ n-dimensionais que satisfazem

$$(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = \mathrm{c}^2 \qquad (3.1.3.1)$$

são uma estimativa do contorno de densidade da população $\mathbf{X}$. Esta aproximação é comumente desenhada em um gráfico do tipo *scatterplot*, permitindo a visualização da distribuição normal gerada pelos dados amostrais. Um exemplo no final desta seção irá ilustrar este tipo de gráfico.

Há situações em que esta suposição de que os dados amostrais possuem distribuição normal nem sempre é verdadeira e, consequentemente, a estimativa apresentada no gráfico

scatterplot pode diferir de um formato elipsóide. Mesmo assim, é possível calcular os autovalores da matriz de covariância amostral $\mathbf{S}$ e obter suas correspondentes componentes principais. Geometricamente, esses dados são apresentados num gráfico como k pontos em um espaço n-dimensional. Esses pontos podem ser representados em um novo eixo de coordenadas, que coincide com aqueles definidos pela equação 3.1.3.1.

A equação 3.1.3.1 descreve um hiper-elipsóide centralizado em $\bar{\mathbf{x}}$ cujos eixos são dados pelos autovetores de $\mathbf{S}^{-1}$. Como $\mathbf{S}$ é uma matriz positiva definida e simétrica então os autovetores de $\mathbf{S}^{-1}$ são iguais aos autovetores de $\mathbf{S}$. Os tamanhos dos eixos do hiper-elipsóide são proporcionais a $\sqrt{\lambda_i}$, onde $i = 1,2,...,p$ e $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ são os autovalores de $\mathbf{S}$. Como o módulo do vetor $\mathbf{e_i}$ é igual a 1, então o módulo da i-ésima componente principal, dado pela fórmula $|\mathbf{y_i}| = |\mathbf{e_i^T}(\mathbf{x} - \bar{\mathbf{x}})|$, representa o tamanho da projeção do vetor $(\mathbf{x} - \bar{\mathbf{x}})$ sobre o vetor unitário $\mathbf{e_i}$. Assim, pode-se dizer que as componentes principais $\mathbf{y_i} = \mathbf{e_i^T}(\mathbf{x} - \bar{\mathbf{x}})$ têm a mesma direção dos eixos definidos pelo hiper-elipsóide de contorno de $\mathbf{X}$ e que os módulos dessas componentes representam os tamanhos dessas projeções nas direções dos eixos $\mathbf{e_i}$. Consequentemente, as componentes principais podem ser entendidas como o resultado de duas operações geométricas básicas:

1. Uma translação na origem do sistema de eixos coordenados para o ponto definido pelo vetor-média $\bar{\mathbf{x}}$ e
2. Uma rotação nos eixos coordenados originais de tal forma que os mesmos se alinhem segundo as direções de máxima variância dos dados.

A interpretação geométrica das componentes principais calculadas a partir das amostras de $\mathbf{X}$ está exemplificada na figura 3.1, para $n = 2$.

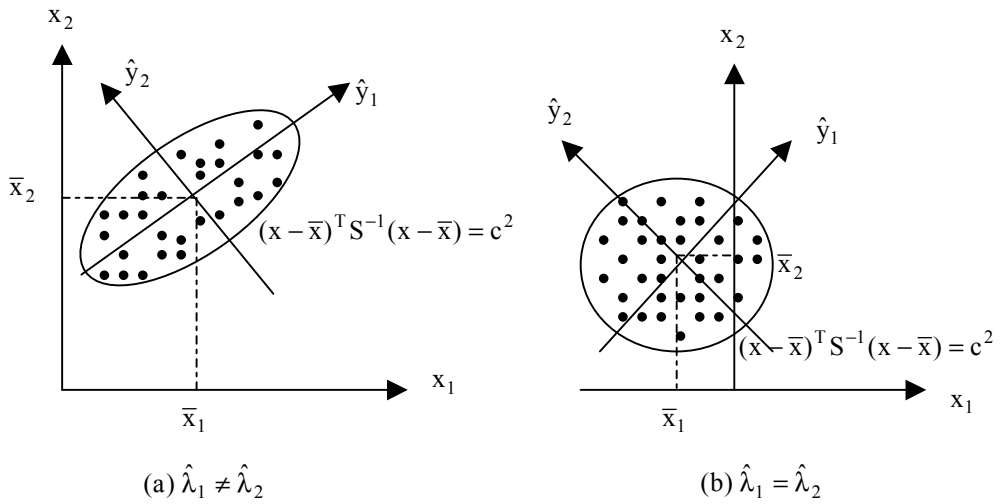

Figura 3.1 – Gráficos tipo *scatterplot* para interpretação geométrica das componentes principais

A figura 3.1(a) mostra uma elipse centrada em $\bar{\mathbf{x}}$ com autovalores $\lambda_1 \neq \lambda_2$. As componentes principais $\hat{\mathbf{y}}_1$ e $\hat{\mathbf{y}}_2$ são bem definidas e sobrepõem os eixos da elipse, apontando para as direções de máxima variância dos dados amostrados. A figura 3.1(b) apresenta uma outra elipse, centrada em $\bar{\mathbf{x}}$ e com autovalores $\lambda_1 = \lambda_2$. Neste caso, pode-se verificar que os eixos da elipse não são determinados de forma única, podendo apontar para quaisquer duas direções perpendiculares, inclusive aquelas definidas pelos eixos originais. De maneira similar, as componentes principais também podem apontar para quaisquer duas direções perpendiculares, incluindo também as direções definidas pelos eixos $\mathbf{x_1}$ e $\mathbf{x_2}$. Portanto, quando o contorno definido pela equação 3-1 é aproximadamente circular ou, equivalentemente, quando os autovalores de $\mathbf{S}$ têm valores bem próximos, a variância dos

dados é praticamente homogênea em todas as direções. Nesta circunstância, não é possível representar os dados originais em espaços de dimensão menor.

Conclui-se, assim, que se os últimos (ordem decrescente) autovalores λ_i forem suficientemente pequenos a ponto de as variações nas direções dos respectivos autovetores $\mathbf{e_i}$ puderem ser consideradas desprezíveis, então as correspondentes componentes principais $\mathbf{y_i}$ podem ser ignoradas e os dados originais podem ser adequadamente representados em um sub-espaço de dimensão reduzida formado pelas primeiras componentes principais.

3.2 Análise Multivariada de Discriminantes

3.2.1 Descrição

A Análise Multivariada de Discriminantes (AMD) ou *Multivariate Discriminant Analysis* (MDA) é uma técnica linear de análise multivariada de dados. Foi introduzida em 1936, por R. A. Fisher, e, por isso, o MDA é frequentemente referenciado na literatura de estatística multivariada como Método Linear de Discriminantes de Fisher.

A idéia central da técnica de Análise Multivariada de Discriminantes consiste em substituir um conjunto de n variáveis que expressam as características de objetos pertencentes a g diferentes grupos, por m combinações lineares dessas mesmas variáveis originais, onde $m < n$, de tal forma que os g grupos existentes fiquem tão diferenciados ou *separados* quanto possível. Em termos algébricos, pode-se dizer que o MDA tem por objetivo encontrar uma nova base que define um sub-espaço m-dimensional no qual a

variabilidade inter-grupos e intra-grupos dos objetos sejam, respectivamente, máxima e mínima. A formulação matemática dessa análise é detalhada a seguir.

3.2.2 Procedimento de Cálculo das Componentes Discriminantes

Supõe-se que as variáveis originais $X_1, X_2, ..., X_n$ tenham distribuição multivariada, não necessariamente normal, e que seus g grupos tenham matrizes de covariância iguais e inversíveis, ou seja, $\mathbf{\Sigma_1 = \Sigma_2 = ... = \Sigma_g = \Sigma}$. Define-se π_i e $\boldsymbol{\mu}_i$ como sendo, respectivamente, a população e o vetor média de cada grupo, onde $i = 1,2,...g$ grupos. Seja $\overline{\boldsymbol{\mu}}$ o vetor média de todas as populações e $\mathbf{B}$ a soma dos produtos inter-grupos, ou seja,

$$\mathbf{B} = \sum_{i=1}^{g} (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^T, \qquad \text{onde} \quad \overline{\boldsymbol{\mu}} = \frac{1}{g} \sum_{i=1}^{g} \boldsymbol{\mu}_i .$$

Considera-se, então, a combinação linear $\mathbf{Y} = \boldsymbol{v}^T \mathbf{X}$, cujo valor esperado é dado por

$$E(\mathbf{Y}) = E(\boldsymbol{v}^T \mathbf{X}) = \boldsymbol{v}^T E(\mathbf{X} \mid \pi_i) = \boldsymbol{v}^T \boldsymbol{\mu}_i$$

para a população π_i, e cuja variância é

$$\mathrm{Var}(\mathbf{Y}) = \sigma_Y^2 = \boldsymbol{v}^T \mathrm{Cov}(\mathbf{X})\boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v}$$

para todas as populações. Consequentemente, o valor esperado $\mu_{iY} = \boldsymbol{v}^T \boldsymbol{\mu}_i$ é potencialmente diferente para cada um dos g grupos. Define-se então a média global $\overline{\mu}_Y$ como

$$\overline{\mu}_Y = \frac{1}{g} \sum_{i=1}^{g} \mu_{iY} = \frac{1}{g} \sum_{i=1}^{g} \boldsymbol{v}^T \boldsymbol{\mu}_i = \boldsymbol{v}^T (\frac{1}{g} \sum_{i=1}^{g} \boldsymbol{\mu}_i) = \boldsymbol{v}^T \overline{\boldsymbol{\mu}}$$

e a razão entre a soma do quadrado das distâncias das médias dos grupos à média global e a variância de $\mathbf{Y}$ como

$$\frac{\sum_{i=1}^{g} (\mu_{iY} - \overline{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g} (\boldsymbol{v}^T \boldsymbol{\mu}_i - \boldsymbol{v}^T \overline{\boldsymbol{\mu}})^2}{\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v}} = \frac{\boldsymbol{v}^T \left(\sum_{i=1}^{g} (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^T \right) \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v}} = \frac{\boldsymbol{v}^T \mathbf{B} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v}}$$

Esta razão mede, por conseguinte, a variabilidade entre os grupos de $\mathbf{Y}$ com relação a variabilidade intra-grupos, pois, conforme hipótese inicial, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. Pode-se, assim, encontrar $\boldsymbol{v}$ de tal forma que esta razão seja maximizada.

Habitualmente, $\boldsymbol{\Sigma}$ e $\boldsymbol{\mu}_i$ são valores não disponíveis. Faz-se, portanto, uso de um conjunto de treinamento composto de observações com classificações previamente conhecidas. Supõe-se que este conjunto de treinamento tenha um número k_i de amostras pertencentes à população ou ao grupo π_i, onde $i = 1,2,...g$ grupos. Denota-se $\mathbf{X}^{\mathbf{i}}$ como o conjunto de dados de dimensão n x k_i da população π_i, onde a respectiva j-ésima coluna é definida como $\mathbf{x}_{ij}$. Calcula-se o vetor média de cada grupo $\overline{\mathbf{x}}_i$ como

$$\overline{\mathbf{x}}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbf{x}_{ij}$$

e a correspondente matriz de covariância $\mathbf{S_i}$ de dimensão n x n como

$$\mathbf{S_i} = \frac{1}{(k_i - 1)} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T .$$

Estima-se, o vetor média global $\overline{\mathbf{x}}$

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{g} k_i \overline{\mathbf{x}}_i}{\sum_{i=1}^{g} k_i} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{k_i} \mathbf{x}_{ij}}{\sum_{i=1}^{g} k_i}$$

que tem dimensão n x 1 e corresponde à média de todas as observações do conjunto de treinamento. Pode-se calcular, também, a estimativa da matriz inter-grupos $\mathbf{B}$

$$\hat{\mathbf{B}} = \sum_{i=1}^{g} (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T \qquad (3.2.2.1)$$

e a estimativa da matriz de covariância $\mathbf{\Sigma}$ baseada na matriz intra-grupos $\mathbf{W}$, dada por

$$\mathbf{W} = \sum_{i=1}^{g} (k_i - 1)\mathbf{S_i} = \sum_{i=1}^{g} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T . \qquad (3.2.2.2)$$

Consequentemente, pode-se dizer que a matriz de covariância média $\mathbf{S}_{\mathbf{pooled}}$ para as g populações, calculada como

$$\mathbf{S}_{\mathbf{pooled}} = \frac{\mathbf{W}}{(k_1 + k_2 + \ldots + k_g - g)},$$

é uma estimativa da matriz de covariância $\boldsymbol{\Sigma}$.

É importante verificar que como $\mathbf{W}$ é a matriz $\mathbf{S}_{\mathbf{pooled}}$ multiplicada por uma constante, então o mesmo vetor $\hat{\boldsymbol{v}}$ que maximiza a razão $\hat{\boldsymbol{v}}^T\hat{\mathbf{B}}\hat{\boldsymbol{v}}/\hat{\boldsymbol{v}}^T\mathbf{S}_{\mathbf{pooled}}\hat{\boldsymbol{v}}$ também maximiza $\hat{\boldsymbol{v}}^T\hat{\mathbf{B}}\hat{\boldsymbol{v}}/\hat{\boldsymbol{v}}^T\mathbf{W}\hat{\boldsymbol{v}}$. Assim, supondo m autovalores não nulos $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m > 0$ de $\mathbf{W}^{-1}\hat{\mathbf{B}}$, onde $m \leq \min(g-1, n)$, e respectivos autovetores $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_m$ normalizados de tal forma que $\hat{\mathbf{e}}^T\mathbf{S}_{\mathbf{pooled}}\hat{\mathbf{e}} = \mathbf{1}$, prova-se [John92] que o vetor de coeficientes $\hat{\boldsymbol{v}}$ que maximiza a razão

$$\frac{\hat{\boldsymbol{v}}^T\hat{\mathbf{B}}\hat{\boldsymbol{v}}}{\hat{\boldsymbol{v}}^T\mathbf{W}\hat{\boldsymbol{v}}} = \frac{\hat{\boldsymbol{v}}^T\left(\sum_{i=1}^{g}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T\right)\hat{\boldsymbol{v}}}{\hat{\boldsymbol{v}}^T\left[\sum_{i=1}^{g}\sum_{j=1}^{k_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T\right]\hat{\boldsymbol{v}}}$$

é dado por $\hat{\boldsymbol{v}}_1 = \hat{\mathbf{e}}_1$. A combinação linear $\hat{\boldsymbol{v}}_1^T\mathbf{X}$ é chamada estimativa do primeiro discriminante. A escolha de $\hat{\boldsymbol{v}}_2 = \hat{\mathbf{e}}_2$ produz a estimativa do segundo discriminante, dada por $\hat{\boldsymbol{v}}_2^T\mathbf{X}$. Por conseguinte, de forma consecutiva, $\hat{\boldsymbol{v}}_s^T\mathbf{X} = \hat{\mathbf{e}}_s^T\mathbf{X}$ é a estimativa do s-ésimo discriminante, para $s \leq m$.

Assim sendo, observa-se que o problema de encontrar a base para o sub-espaço m-dimensional que produz a máxima separação entre os grupos a partir da matriz de dados formada pelas n variáveis originais $X_1, X_2, ..., X_n$ e as k amostras dos g grupos resume-se ao cálculo dos autovetores da matriz produto dada por $\mathbf{W}^{-1}\mathbf{B}$.

Este processo de cálculo é similar ao empregado na Análise de Componentes Principais (PCA) discutida na seção anterior, diferindo apenas no critério de determinação da nova base. No caso do MDA, a matriz em questão reflete a separação máxima entre os grupos e não a variância máxima da totalidade da população, como acontece no PCA.

3.2.3 Interpretação Geométrica das Componentes Discriminantes

A análise multivariada de discriminantes (MDA) é, como a análise de componentes principais (PCA), uma transformação linear algébrica de variáveis. Conforme visto na seção 3.1.3 que descreve a interpretação geométrica das componentes principais, uma transformação linear eqüivale a uma rotação nos eixos coordenados definidos pelas variáveis originais do sistema. Para a MDA, no entanto, esta transformação acontece de um modo diferente. Os eixos originais são rotacionados de tal forma que os mesmos representem as direções de máxima separação entre os grupos e não mais as maiores diferenças entre todas as amostras de todos os grupos, como acontece na PCA.

Um exemplo de separação de grupos para um caso bi-dimensional pode ser visto na figura 3.2.

Neste exemplo, tem-se como componentes principais das amostras os eixos $\mathbf{Y_1}$ e $\mathbf{Y_2}$ e como componente discriminante o eixo $\mathbf{Z_1}$. Verifica-se que a projeção dos pontos sobre a componente principal $\mathbf{Y_1}$ é incapaz de separar os dois grupos existentes. Entretanto, uma

projeção desses pontos sobre a componente discriminante $\mathbf{Z}_1$ faz com que os mesmos tenham as suas classes claramente diferenciadas. Em termos de projeção linear, os eixos definidos pela análise multivariada de discriminantes representam a solução ótima de separação entre os grupos definidos pelo conjunto de treinamento considerado.

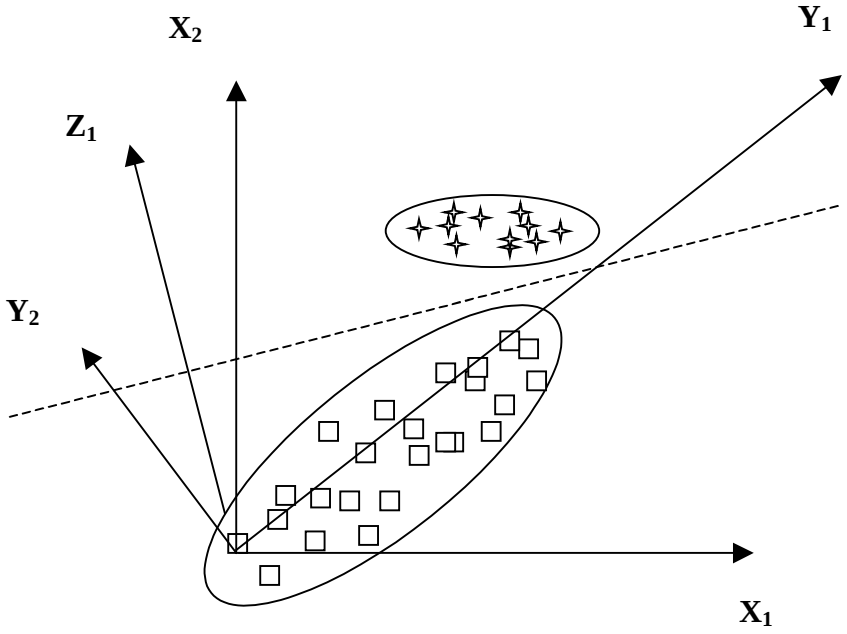

Figura 3.2 – Comparação geométrica entre as componentes discriminantes e as componentes principais

Alguns aspectos importantes, particulares à analise multivariada de discriminantes, devem ser destacados. Como a matriz de determinação da nova base formada pelos eixos discriminantes, a matriz $\mathbf{W}^{-1}\mathbf{B}$, é uma matriz não simétrica seus respectivos autovetores são descorrelacionados mas não necessariamente ortogonais. Além disso, sabe-se que o número máximo de autovetores ou eixos discriminantes é definido pelo número de linhas (ou colunas) linearmente independentes ou posto da matriz $\mathbf{W}^{-1}\mathbf{B}$ em questão. Como, na prática, o número de grupos g é menor que o número de variáveis originais n, então o número máximo de eixos é, em geral, menor ou igual ao número de grupos do problema menos um. Diz-se

menor pois em determinadas situações nem todos os $g-1$ eixos possuem autovalores maiores que zero, ou seja, apresentam variabilidade estatística com relação aos grupos existentes, permitindo uma maior redução de dimensionalidade.

Por fim, pode-se dizer que na análise multivariada de discriminantes também expressam-se os autovalores λ_i de $\mathbf{W}^{-1}\mathbf{B}$ como porcentagens da variância total explicada pelos autovetores correspondentes, semelhantemente ao que se tem na análise de componentes principais. Pode-se, assim, calcular a variânça explicada pelos primeiros eixos discriminantes. Consequentemente, se essas variâncias forem bastante significativas com relação à variância total dos dados, pode-se desprezar os demais autovetores, reduzindo ainda mais a dimensionalidade dos dados originais.

3.3 O Paradoxo da Redução de Dimensionalidade

Conforme dito no início do capítulo e exemplificado nas seções anteriores, toda técnica de redução de dimensionalidade procura extrair o máximo de informação relevante dos dados originais, realçando suas principais características discriminatórias.

Sabe-se, também, que é inerente a todo processo de redução de dimensionalidade a perda de informação. Desta forma, parece natural pensar que quanto maior o número de características utilizadas pelos modelos de classificação, melhor a discriminação dos respectivos grupos. Este raciocínio, no entanto, nem sempre se confirma na prática. Em muitas situações, constata-se que a partir de um determinado ponto a adição de novas características, ou seja, um aumento de dimensionalidade, diminui a separação entre os grupos. Este efeito é comumente chamado na estatística de análise multivariada de *"the curse of dimensionality"*, denominado aqui como o "paradoxo da redução de dimensionalidade".

Considere-se, como exemplo, a seguinte técnica (não recomendada na prática) para modelar um mapeamento não-linear de um conjunto de n variáveis de entrada X_i, onde $i = 1,2,...,n$, em uma variável de saída Y tomando-se como base um determinado conjunto de treinamento. Divide-se a faixa de valores de cada variável de entrada em um número k de intervalos de tal forma que o valor de saída dessa variável pode ser especificado, aproximadamente, pelo intervalo em que a mesma se encontra. Em termos geométricos, este procedimento pode ser entendido como a divisão do espaço original de variáveis em diversas caixas ou células, como mostra a figura 3.3 abaixo para o caso bidimensional.

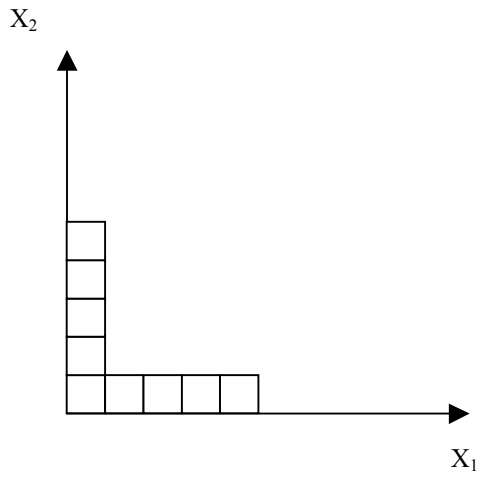

Figura 3.3 – Mapeamento de um espaço bidimensional X_1, X_2 em uma variável de saída Y

Cada amostra de treinamento corresponde a um ponto em uma dessas células, possuindo um respectivo valor para a variável de saída Y. Pode-se determinar o valor de saída Y para um novo ponto neste espaço calculando-se a média dos valores de Y referentes a todos os exemplos de treinamento que pertencem à mesma célula em que recai este ponto. Como a faixa de valores de cada variável possui k intervalos, tem-se ao todo k^n células. Este número cresce exponencialmente com a dimensionalidade n do espaço. Verifica-se, portanto,

o seguinte problema. Como cada célula deve conter ao menos um ponto, isto implica em que a quantidade de exemplos de treinamento necessária para modelar o mapeamento em questão deve também crescer exponencialmente. Como comumente se dispõe de um número limitado de dados, um aumento de precisão, ou melhor, um aumento de dimensionalidade, pode fazer com que os dados fiquem relativamente escassos, produzindo uma fraca representação do mapeamento.

Este exemplo mostra que um aumento de dimensionalidade não necessariamente melhora a representação dos dados. Observa-se que em problemas cuja quantidade de informação é limitada, ao se reduzir o número de variáveis de entrada melhora-se a representação dos dados, mesmo com perda de informação. O que ocorre nestas situações é que as variáveis originais do modelo possuem alguma correlação entre si e, assim, não descrevem todo o espaço original. Na realidade, os pontos definidos por essas variáveis possuem uma dimensionalidade intrínseca, concentrando-se apenas em uma região específica do espaço, ou melhor, em um sub-espaço de dimensão menor. Assim, esta aparente perda de informação está relacionada ou à redundância de informação caracterizada pelas variáveis originais ou à limitação de representação imposta pela quantidade de dados disponível, prejudicando uma melhor discriminação dos dados .

Evidencia-se, portanto, a importância do uso de técnicas de redução de dimensionalidade, como as descritas anteriormente neste capítulo, em sistemas de mapeamento de variáveis. Estas técnicas devem ser empregadas antes dos modelos de classificação, como uma espécie de pré-processamento dos dados originais. Este pré-processamento além de diminuir consideravelmente os parâmetros a serem otimizados, possibilita uma melhor representação dos dados originais, aumentando a performance dos modelos de classificação a serem utilizados posteriormente. Alguns modelos de classificação

usualmente empregados em sistemas de reconhecimento de padrões serão descritos no próximo capítulo.