6. Geometric applications

- localization from multiple camera views
- orthogonal Procrustes problem and polar decomposition
- fitting affine sets to points
- linear discriminant analysis
Introduction

applications in this lecture use matrix methods to solve problems in geometry

- $m \times n$ matrix is interpreted as collection of $m$ points in $\mathbb{R}^n$ or $n$ points in $\mathbb{R}^m$
- $m \times n$ matrices parametrize affine functions $f(x) = Ax + b$ from $\mathbb{R}^n$ to $\mathbb{R}^m$
- $m \times n$ matrices parametrize affine sets $\{x \mid Ax = b\}$ in $\mathbb{R}^n$
Multiple view geometry

- $n$ objects at positions $x_j \in \mathbb{R}^3$, $j = 1, \ldots, n$, are viewed by $l$ cameras
- $y_{ij} \in \mathbb{R}^2$ is the location of object $j$ in the image acquired by camera $i$
- each camera is modeled as an affine mapping:

$$y_{ij} = P_i x_j + q_i, \quad i = 1, \ldots, l, \quad j = 1, \ldots, n$$

define a $2l \times n$ matrix with the observations $y_{ij}$:

$$Y = \begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1n} \\
y_{21} & y_{22} & \cdots & y_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
y_{l1} & y_{l2} & \cdots & y_{ln}
\end{bmatrix} = \begin{bmatrix}
P_1 & q_1 \\
P_2 & q_2 \\
\vdots & \vdots \\
P_l & q_l
\end{bmatrix} \begin{bmatrix}
x_1 & x_2 & \cdots & x_n
\end{bmatrix}$$

- 2nd equality assumes noise-free observations and perfectly affine cameras
- the goal is to estimate the positions $x_j$ and the camera models $P_i, q_i$
Factorization algorithm

minimize Frobenius norm of error between model predictions and observations $Y$

$$\minimize \| PX + q1^T - Y \|^2_F$$

- $P$ is $2l \times 3$ matrix and $q$ is $2l$-vector with the $l$ camera models:

$$P = \begin{bmatrix} P_1 \\ \vdots \\ P_l \end{bmatrix}, \quad q = \begin{bmatrix} q_1 \\ \vdots \\ q_l \end{bmatrix}$$

- variables are the $3 \times n$ position matrix $X = [x_1 \cdots x_n]$ and camera models $P, q$

- variable $q$ can be eliminated: least squares estimate is $q = (1/n)(Y - PX)1$

- substituting expression for optimal $q$ gives

$$\minimize \| PX_c - Y_c \|^2_F$$
subject to $X_c 1 = 0$

here $Y_c = Y(I - (1/n)11^T)$ and the variable is $X_c = X(I - (1/n)11^T)$
Factorization algorithm

\[ \begin{align*}
\text{minimize} & \quad \|PX_c - Y_c\|_F^2 \\
\text{subject to} & \quad X_c1 = 0
\end{align*} \]

with variables \( P \) (a \( 2l \times 3 \) matrix) and \( X_c \) (a \( 3 \times 2n \) matrix)

- the solution follows from an SVD of \( Y_c \):

\[ Y_c = \sum_{i=1}^{\min\{2l,n\}} \sigma_i u_i v_i^T \]

- (assuming \( \text{rank}(Y_c) \geq 3 \)) truncate SVD after 3 terms and define:

\[ P = \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \sigma_3 u_3 \end{bmatrix}, \quad X_c = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T \]

- vectors \( v_1, v_2, v_3 \) are in the row space of \( Y_c \), hence orthogonal to \( 1 \), so \( X_c1 = 0 \)
- solution is not unique, since \( PX_c = (PT)(T^{-1}X_c) \) for any nonsingular \( T \)
- this ambiguity corresponds to the choice of coordinate system in \( \mathbb{R}^3 \)
References

  
  the original paper on the factorization method

  
  a more recent survey of the factorization method and extensions
Outline

• localization from multiple camera views

• **orthogonal Procrustes problem and polar decomposition**

• fitting affine sets to points

• linear discriminant analysis
Orthogonal Procrustes problem

given $m \times n$ matrices $A$, $B$, solve the optimization problem

$$\begin{align*}
\text{minimize} & \quad \|AX - B\|_F^2 \\
\text{subject to} & \quad X^T X = I
\end{align*}$$

(1)

the variable is an $n \times n$ matrix $X$

- a matrix least squares problem with constraint that $X$ is orthogonal
- rows of $B$ are approximated by orthogonal linear function applied to rows of $A$

Solution: $X = UV^T$ with $U$, $V$ from an SVD of the $n \times n$ matrix $A^T B = U \Sigma V^T$
Solution of orthogonal Procrustes problem

- the problem is equivalent to maximizing \( \text{trace}(B^T AX) \) over orthogonal \( X \):

\[
\|AX - B\|_F^2 = \text{trace}((AX - B)(AX - B)^T) \\
= \text{trace}(AXX^TA^T) + \text{trace}(BB^T) - 2\text{trace}(AXB^T) \\
= \|A\|_F^2 + \|B\|_F^2 - 2\text{trace}(B^T AX)
\]

- compute \( n \times n \) SVD \( A^TB = U\Sigma V^T \) and make change of variables \( Y = U^T XV \):

\[
\text{maximize} \quad \text{trace}(\Sigma Y) = \sum_{i=1}^n \sigma_i Y_{ii} \\
\text{subject to} \quad Y^TY = I
\quad (2)
\]

- if \( Y \) is orthogonal, then \( |Y_{ii}| \leq 1 \) and \( \text{trace}(\Sigma Y) \leq \sum_{i=1}^n \sigma_i \):

\[
1 = (Y^TY)_{ii} = Y_{ii}^2 + \sum_{j\neq i} Y_{ji}^2 \geq Y_{ii}^2
\]

- hence \( Y = I \) is optimal for (2) and \( X = UYV^T = UV^T \) is optimal for (1)
Application

given two sets of points \( x_1, \ldots, x_m \) and \( y_1, \ldots, y_m \) in \( \mathbb{R}^n \), solve the problem

\[
\text{minimize} \quad \sum_{i=1}^{m} \|Qx_i + c - y_i\|^2
\]

subject to \( Q^T Q = I \)

- the variables are an \( n \times n \) matrix \( Q \) and \( n \)-vector \( c \)
- \( Q \) and \( c \) define a shape-preserving affine mapping \( f(x) = Qx + c \)
the problem is equivalent to an orthogonal Procrustes problem

- for given $Q$, optimal $c$ is
  \[
  c = \frac{1}{m} \sum_{i=1}^{m} (y_i - Qx_i)
  \]

- substitute expression for optimal $c$ in the cost function:
  \[
  \sum_{i=1}^{m} \|Qx_i + c - y_i\|^2 = \sum_{i=1}^{m} \|Q\tilde{x}_i - \tilde{y}_i\|^2 = \|Q\tilde{X} - \tilde{Y}\|^2_F
  \]

  where $\tilde{X} = [\tilde{x}_1 \cdots \tilde{x}_m]$, $\tilde{Y} = [\tilde{y}_1 \cdots \tilde{y}_m]$, and $\tilde{x}_i$, $\tilde{y}_i$ are the centered points
  \[
  \tilde{x}_i = x_i - \frac{1}{m} \sum_{j=1}^{m} x_j, \quad \tilde{y}_i = y_i - \frac{1}{m} \sum_{j=1}^{m} y_j,
  \]

- optimal $Q$ minimizes $\|Q\tilde{X} - \tilde{Y}\|^2_F = \|\tilde{X}^TQ^T - \tilde{Y}^T\|^2_F$ over orthogonal matrices

Geometric applications 6.10
Polar decomposition

every $m \times n$ matrix $A$ with $m \geq n$ can be factorized as

$$A = QH$$

- $Q$ is $m \times n$ with orthonormal columns ($Q^TQ = I$)
- $H$ is $n \times n$, symmetric, and positive semidefinite
- called polar decomposition (after the polar representation of complex numbers)

Proof: from (reduced) SVD $A = U\Sigma V^T$

- $U$ is $m \times n$ with orthonormal columns, $\Sigma$ is $n \times n$, $V$ is $n \times n$ and orthogonal
- write SVD in the form of the polar decomposition:

$$A = U\Sigma V^T = (UV^T)(V\Sigma V^T) = QH$$

where $Q = UV^T$ and $H = V\Sigma V^T$

- $Q$ has orthonormal columns because $Q^TQ = VU^TUV^T = VV^T = I$
- $H$ is symmetric, positive semidefinite, with eigenvalues $\sigma_1, \ldots, \sigma_n$
Orthogonal Procrustes problem

\[
\begin{align*}
\text{minimize} & \quad \|AX - B\|_F^2 \\
\text{subject to} & \quad X^TX = I
\end{align*}
\]

- \(A, B\) are matrices of the same dimensions
- \(X\) is square and constrained to be orthogonal
- from page 6.7, solution \(X\) is the Q-factor in polar decomposition \(A^TB = QH\)

Nearest matrix with orthonormal columns

\[
\begin{align*}
\text{minimize} & \quad \|X - B\|_F^2 \\
\text{subject to} & \quad X^TX = I
\end{align*}
\]

- \(B\) is an \(m \times n\) matrix with \(m \geq n\)
- \(X\) is \(m \times n\) and constrained to have orthonormal columns
- optimal \(X\) is Q-factor in polar decomposition of \(B\) (proof on next page)
Proof

- the problem is equivalent to maximizing \( \text{trace}(B^TX) \) subject to \( X^TX = I \):

\[
\|X - B\|_F^2 = \text{trace}(X^TX) + \text{trace}(B^TB) - 2 \text{trace}(B^TX)
\]
\[
= n + \|B\|_F^2 - 2 \text{trace}(B^TX)
\]

- consider full and reduced SVDs of \( B \)

\[
B = U \Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^T = U_1 \Sigma_1 V^T
\]

(where \( U \) is \( m \times m \) and \( U_1 \) is \( m \times n \))

- make change of variables \( Y = U^TXV \), where \( Y \) is \( m \times n \):

maximize \( \text{trace}(\Sigma^TY) = \sum_{i=1}^{n} \sigma_i Y_{ii} \)
subject to \( Y^TY = I \)

- optimal \( Y \) and \( X \) are

\[
Y = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad X = UYV^T = U_1 V^T
\]
Exercise

suppose $A, B$ are $m \times n$ matrices that satisfy

$$AA^T = BB^T$$

we show that $B = AX$ for some orthogonal matrix $X$

• show that $A$ and $B$ have SVDs of the form

$$A = U\Sigma V_1^T, \quad B = U\Sigma V_2^T$$

(these are full SVDs, i.e., with $U, V_1, V_2$ square and orthogonal)

• show that $A^T B$ has a polar decomposition

$$A^T B = QH \quad \text{where} \quad Q = V_1 V_2^T \quad \text{and} \quad H = V_2 \Sigma^T \Sigma V_2^T$$

• show that $B = AX$ for $X = Q$
Outline

- localization from multiple camera views
- orthogonal Procrustes problem and polar decomposition
- fitting affine sets to points
- linear discriminant analysis
Affine set

a subset $S$ of $\mathbb{R}^n$ is affine if

$$\alpha x + \beta y \in S$$

for all vectors $x, y \in S$ and all scalars $\alpha, \beta$ with $\alpha + \beta = 1$

- affine combinations of elements of $S$ are in $S$
- if $x \neq y$ are two points in $S$, then the entire line through $x, y$ is in $S$

Examples

- a subspace $\mathcal{V}$ is an affine set: if $x, y \in \mathcal{V}$ then $\alpha x + \beta y \in \mathcal{V}$ for all $\alpha, \beta$
- subspace plus vector: $\{x + a \mid x \in \mathcal{V}\}$ where $\mathcal{V}$ is a subspace and $a$ a vector
- solution set of linear equation $\{x \mid Ax = b\}$
- the empty set is affine (but not a subspace)
Parallel subspace

suppose $S$ is a nonempty affine set, $x_0$ is a point in $S$, and define

$$\mathcal{V} = \{x - x_0 \mid x \in S\}$$

- $\mathcal{V}$ is a subspace: if $x \in \mathcal{V}$, $y \in \mathcal{V}$, then $x + x_0 \in S$, $y + x_0 \in S$, and

$$\alpha x + \beta y + x_0 = \alpha(x + x_0) + \beta(y + x_0) + (1 - \alpha - \beta)x_0 \in S \quad \text{for all } \alpha, \beta$$

(right-hand side is affine combination of 3 points $x + x_0$, $y + x_0$, and $x_0$ in $S$)

- $\mathcal{V}$ does not depend on the choice of $x_0 \in S$: if $x + x_0 \in S$ and $y_0 \in S$, then

$$x + y_0 = (x + x_0) - x_0 + y_0 \in S$$

(right-hand side is affine combination of 3 points $x + x_0$, $x_0$, $y_0$ in $S$)

- the dimension of $S$ is defined as the dimension of the parallel subspace $\mathcal{V}$
Range representation

every nonempty affine set \( S \subseteq \mathbb{R}^m \) can be represented as

\[
S = \{ Ax + b \mid x \in \mathbb{R}^n \}
\]

- \( b \) is any vector in \( S \)
- \( A \) is any matrix with range equal to the parallel subspace: \( S = \text{range}(A) + b \)
- \( \text{dim}(S) = \text{rank}(A) \)
Nullspace representation

every affine set $S \subseteq \mathbb{R}^n$ (including the empty set) can be represented as

$$S = \{ x \in \mathbb{R}^n \mid Ax = b \}$$

for a nonempty affine set $S$:

- $b = Ax_0$ where $x_0$ is any vector in $S$
- $A$ is any matrix with nullspace equal to the parallel subspace: $S = \text{null}(A) + x_0$
- $\dim(S) = \text{rank}(A) - n$

the empty set is the solution set of an inconsistent equation (e.g., $A = 0, b \neq 0$)
Distance to affine set

suppose $S$ is the affine set $S = \{y \mid Ay = b\}$

**Projection:** projection of $x$ on $S$ is the solution $y$ of the “least-distance” problem

\[
\begin{align*}
\text{minimize} & \quad \|y - x\| \\
\text{subject to} & \quad Ay = b
\end{align*}
\]

- if $A$ has linearly independent rows, $y = x + A^\dagger(b - Ax)$
- if $A$ has orthonormal rows, $y = x + A^T(b - Ax)$

**Distance:** we denote the distance of $x$ to $S$ by $d(x, S)$

- if $A$ has linearly independent rows, $d(x, S) = \|A^\dagger(Ax - b)\|$
- if $A$ has orthonormal rows, $d(x, S) = \|Ax - b\|$
Least squares fit of affine set to points

fit an affine set $S$ of specified dimension $k$ to $N$ points $x_1, \ldots, x_N$ in $\mathbb{R}^n$:

$$\text{minimize } \sum_{i=1}^{N} d(x_i, S)^2$$

Example: $k = 1$, $N = 50$, $n = 2$
Least squares fit of affine set to points

use nullspace representation \( S = \{ x \mid Ax = b \} \), where \( A \) has orthonormal rows:

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{N} \|Ax_i - b\|^2 \\
\text{subject to} & \quad AA^T = I
\end{align*}
\]

the variables are the \( m \times n \) matrix \( A \) and \( m \)-vector \( b \), where \( m = n - k \)

**Algorithm** (assuming \( m \leq n \leq N \)):

- compute center \( \bar{x} = (1/N)(x_1 + \cdots + x_N) \)
- rows of optimal \( A \) are the last \( m \) left singular vectors of matrix of centered points \( X = [x_1 - \bar{x} \quad x_2 - \bar{x} \quad \cdots \quad x_N - \bar{x}] \)
- optimal \( b \) is \( b = A\bar{x} \)

we derive this solution on the next page
Least squares fit of affine set to points

- for given \( A \), the optimal \( b \) is the average \( (1/N)A(x_1 + \cdots + x_N) = A\bar{x} \)
- eliminating \( b \) reduces the problem to an optimization over \( m \times n \) variable \( A \)

\[
\begin{align*}
\text{minimize} & \quad \|AX\|_F^2 \\
\text{subject to} & \quad AA^T = I
\end{align*}
\]

- denote singular values and left singular vectors of \( n \times N \) matrix \( X \) by

\[
\sigma_1 \geq \cdots \geq \sigma_n, \quad u_1, \ldots, u_n
\]

- from page 4.28, singular values \( \tau_1 \geq \cdots \geq \tau_m \) of the \( m \times N \) matrix \( AX \) satisfy

\[
\tau_1 \geq \sigma_{n-m+1}, \quad \tau_2 \geq \sigma_{n-m+2}, \quad \ldots, \quad \tau_{m-1} \geq \sigma_{n-1}, \quad \tau_m \geq \sigma_n
\]

all inequalities are equalities if \( A = \left[u_{n-m+1} \cdots u_n\right]^T \)

- this choice of \( A \) also minimizes

\[
\|AX\|_F^2 = \tau_1^2 + \cdots + \tau_m^2
\]

Geometric applications 6.22
**k**-means clustering with affine sets

Partition \( N \) points \( x_1, \ldots, x_N \) in \( k \) classes

- in the \( k \)-means algorithm, clusters are represented by representative vectors \( s_j \)
- the \( k \)-means algorithm is a heuristic for minimizing the clustering objective

\[
J_{\text{clust}} = \frac{1}{N} \sum_{i=1}^{N} \|x_i - s_{j_i}\|^2 \quad (j_i \text{ is the index of the cluster that point } i \text{ is assigned to})
\]

by alternating minimization over assignment and over representatives

As an extension, we can use affine sets as representatives

Geometric applications 6.23
**$k$-means clustering with affine sets**

- represent the $k$ clusters by affine sets $S_1, \ldots, S_k$ of specified dimension

- use the $k$-means alternating minimization heuristic to minimize the objective

\[
J_{\text{clust}} = \frac{1}{N} \sum_{i=1}^{N} d(x_i, S_{j_i})^2 \quad (j_i \text{ is the index of the cluster that point } i \text{ is assigned to})
\]

- to update partition we assign each point $x_i$ to nearest representative

- to update each group representative $S_j$ we fit affine set to points in group $j$

- standard $k$-means is a special case with affine sets of dimension zero
Example: iteration 1

we start with a random initial assignment

fit representatives to groups

update assignment

Geometric applications
Example: iteration 2

fit representatives to groups

update assignment
Example: iteration 3

fit representatives to groups
update assignment
Example: iteration 8

fit representatives to groups

update assignment

Geometric applications 6.28
Example: iteration 9

fit representatives to groups

update assignment
Example: iteration 10

fit representatives to groups

update assignment
Outline

- localization from multiple camera views
- orthogonal Procrustes problem and polar decomposition
- fitting affine sets to points
- linear discriminant analysis
Motivation

principal components are not necessarily good features for classification

- the two sets of points (large dots) are linearly separable
- their projections on the 1st principal component direction (small circles) are not
Classification problem

we are given a training set with examples of \( K \) classes

- \( C_k \): set of examples for class \( k \)
- \( N_k \): number of examples for class \( k \)
- \( C \): set of all training examples \( C = C_1 \cup \cdots \cup C_K \)
- \( N \): total number of training examples \( N = N_1 + \cdots + N_K \)

- \( \bar{x}_k \) denotes the mean for class \( k \), \( \bar{x} \) denotes the mean for the entire set:

\[
\bar{x}_k = \frac{1}{N_k} \sum_{x \in C_k} x, \quad \bar{x} = \frac{1}{N} \sum_{x \in C} x = \frac{1}{N} (N_1 \bar{x}_1 + \cdots + N_K \bar{x}_K)
\]

- \( S_k \) is the covariance matrix for class \( k \):

\[
S_k = \frac{1}{N_k} \sum_{x \in C_k} (x - \bar{x}_k)(x - \bar{x}_k)^T = \frac{1}{N_k} \sum_{x \in C_k} xx^T - \bar{x}_k\bar{x}_k^T
\]

- \( S \) is the covariance matrix for the entire set:

\[
S = \frac{1}{N} \sum_{x \in C} (x - \bar{x})(x - \bar{x})^T = \frac{1}{N} \sum_{x \in C} xx^T - \bar{x}\bar{x}^T
\]
Principal components

the principal component directions are the eigenvectors of the covariance matrix

\[ S = \sum_{i=1}^{n} \lambda_i v_i v_i^T \]

- principal component directions can be defined recursively: \( v_k \) solves

\[
\begin{align*}
\text{maximize} & \quad x^T S x \\
\text{subject to} & \quad ||x|| = 1 \quad v_i^T x = 0 \quad \text{for } i = 1, \ldots, k - 1
\end{align*}
\]

- max–min characterization: the matrix of first \( k \) eigenvectors \([v_1 \cdots v_k]\) solves

\[
\begin{align*}
\text{maximize} & \quad \lambda_{\text{min}}(X^T S X) \\
\text{subject to} & \quad X^T X = I_k
\end{align*}
\]

PCA does not distinguish between variance within and between classes
Within-class and between-class covariance

the covariance of the entire set can be written as a sum of two terms

\[ S = S_w + S_b \]

**Within-class covariance**

\[
S_w = \frac{1}{N} \sum_{k=1}^{K} \frac{N_k}{N} S_k = \frac{1}{N} \left( \sum_{x \in C} xx^T - \sum_{k=1}^{K} N_k \bar{x}_k \bar{x}_k^T \right)
\]

- \( S_w \) is the weighted average of the class covariance matrices \( S_k \)
- describes the variability of points within the same class

**Between-class covariance**

\[
S_b = \frac{1}{N} \sum_{k=1}^{K} N_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T = \frac{1}{N} \sum_{k=1}^{K} N_k \bar{x}_k \bar{x}_k^T - \bar{x} \bar{x}^T
\]

- \( S_b \) is the covariance matrix of the class means (weighted by class size)
- describes the variability between classes

Geometric applications
Linear discriminant analysis (LDA)

- good directions for classification make $v^T S_b v$ large while keeping $v^T S_w v$ small
- instead of maximizing $(v^T S_v)/(v^T v)$ as in PCA, it is better to maximize

$$\frac{v^T S_b v}{v^T S_w v}$$

LDA directions: a sequence of vectors $v_1, v_2, \ldots$

- first direction $v_1$ maximizes $(x^T S_b x)/(x^T S_w x)$ or, equivalently, solves

$$\begin{align*}
\text{maximize} & \quad x^T S_b x \\
\text{subject to} & \quad x^T S_w x = 1
\end{align*}$$

- other directions are defined recursively: $v_k$ is the solution $x$ of

$$\begin{align*}
\text{maximize} & \quad x^T S_b x \\
\text{subject to} & \quad x^T S_w x = 1 \\
& \quad v_i^T S_w x = 0 \quad \text{for } i = 1, \ldots, k - 1
\end{align*}$$
Computation via eigendecomposition

the $k$th LDA direction $v_k$ is the solution $x$ of

$$\begin{align*}
\text{maximize} & \quad x^T S_b x \\
\text{subject to} & \quad x^T S_w x = 1 \\
& \quad v_i^T S_w x = 0 \quad \text{for } i = 1, \ldots, k - 1
\end{align*}$$

we assume $S_w$ has full rank (is positive definite)

- compute Cholesky factorization $S_w = R^T R$
- make a change of variables $y = Rx$:

$$\begin{align*}
\text{maximize} & \quad y^T (R^{-T} S_b R^{-1}) y \\
\text{subject to} & \quad y^T y = 1 \\
& \quad v_i^T R^T y = 0 \quad \text{for } i = 1, \ldots, k - 1
\end{align*}$$

the vectors $w_k = R v_k$ are the eigenvectors of $R^{-T} S_b R^{-1}$
Generalized eigenvectors

Suppose $A$ and $B$ are symmetric, and $B$ is positive definite.

- Nonzero $x$ is a generalized eigenvector of $A$, $B$, with generalized eigenvalue $\lambda$, if
  \[ Ax = \lambda Bx \]

- Via the Cholesky factorization $B = R^T R$ this can be written as
  \[ (R^{-T} A R^{-1})(Rx) = \lambda (Rx) \]

- Generalized eigenvalues of $A$, $B$ are eigenvalues of $R^{-T} A R^{-1}$
- $x$ is a generalized eigenvector if and only if $Rx$ is eigenvector of $R^{-T} A R^{-1}$

LDA directions are generalized eigenvectors of $S_b$, $S_w$
Number of LDA directions

the between-class covariance matrix has rank at most \( K - 1 \)

\[
S_b = \frac{1}{N} \sum_{k=1}^{K} N_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T = \frac{1}{N} YY^T
\]

where \( Y \) is the \( n \times K \) matrix

\[
Y = \begin{bmatrix}
\sqrt{N_1} (\bar{x}_1 - \bar{x})^T \\
\vdots \\
\sqrt{N_K} (\bar{x}_K - \bar{x})^T
\end{bmatrix}
\]

the rank of \( Y \) is at most \( K - 1 \) because the rows of \( Y \) are linearly dependent:

\[
Y^T \begin{bmatrix}
\sqrt{N_1} \\
\vdots \\
\sqrt{N_K}
\end{bmatrix} = N_1\bar{x}_1 + N_2\bar{x}_2 + \cdots + N_K\bar{x}_K - (N_1 + \cdots + N_K)\bar{x} = 0
\]

• therefore \( R^{-T}S_bR^{-1} \) has at most \( K - 1 \) nonzero eigenvalues
• there are at most \( K - 1 \) LDA directions (other directions are in \( \text{null}(S_b) \))

Geometric applications
LDA for Boolean classification ($K = 2$)

in the Boolean case, $\bar{x} = (N_1\bar{x}_1 + N_2\bar{x}_2)/N$ and

$$S_b = \frac{N_1}{N} (\bar{x}_1 - \bar{x}) (\bar{x}_1 - \bar{x})^T + \frac{N_2}{N} (\bar{x}_2 - \bar{x}) (\bar{x}_2 - \bar{x})^T$$

$$= \frac{2N_1N_2}{N^2} (\bar{x}_1 - \bar{x}_2) (\bar{x}_1 - \bar{x}_2)^T$$

• the LDA direction $v$ is defined as the solution $x$ of

$$\text{maximize} \quad x^T S_b x$$

$$\text{subject to} \quad x^T S_w x = 1$$

• via the change of variable $y = Rx$, where $S_w = R^T R$, we find the solution

$$y = \frac{R^{-T}(\bar{x}_1 - \bar{x}_2)}{\|R^{-T}(\bar{x}_1 - \bar{x}_2)\|}, \quad v = R^{-1} y = \frac{S_w^{-1}(\bar{x}_1 - \bar{x}_2)}{((\bar{x}_1 - \bar{x}_2)^T S_w^{-1}(\bar{x}_1 - \bar{x}_2))^{1/2}}$$

the LDA direction is the direction of $S_w^{-1}(\bar{x}_1 - \bar{x}_2)$
the example of page 6.31

projections on LDA direction (small circles) are separable
Fisher’s Iris flower data set

- 50 examples of each of the three classes, 4 features
- first LDA direction separates the classes better than first PCA direction
- second LDA direction does not add much information
- eigenvalues of $R^{-T}S_bR^{-1}$ are (32.19, 0.29, 0, 0) (see page 6.36)
Reference


  discusses PCA and LDA for face recognition