5. Applications to data fitting

- principal components
- canonical correlations
- dimension reduction
- rank-deficient least squares
- regularized least squares
- total least squares

Introduction

applications in this lecture use matrices to represent *data sets:*

- a set of examples (or samples, data points, observations, measurements)
- for each example, a list of attributes or features

an $m \times n$ *data matrix* A is used to represent the data

- rows are feature vectors for m examples
- columns correspond to n features
- rows are denoted by $a_1^T, \ldots, a_m^T$ with $a_i \in \mathbf{R}^n$
- in some applications, rows are interpreted as samples of a random n-vector

Outline

- **principal components**
- canonical correlations
- dimension reduction
- rank-deficient least squares
- regularized least squares
- total least squares

Principal components

recall the results from page 3.29

- we assume x is a random n-vector with mean μ and covariance matrix

$$C = \mathbf{E}((x-\mu)(x-\mu)^T)$$

  here we use notation C to avoid confusion with the matrix Σ in an SVD

- C is positive semidefinite with eigendecomposition

$$C = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

the *principal components* (p.c.'s) of x are the components of $y = Q^T x$:

$$y_1 = q_1^T x, \qquad y_2 = q_2^T x, \qquad \ldots, \qquad y_n = q_n^T x$$

coefficients of vector q_i are called the *loadings* for principal component y_i

Properties of principal components

the random vector y has mean $\bar{y} = Q^T \mu$ and covariance matrix Λ:

$$\begin{aligned}
\mathbf{E}((y-\bar{y})(y-\bar{y})^T) &= Q^T\,\mathbf{E}((x-\mu)(x-\mu)^T)Q \\
&= Q^T C Q \\
&= \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}
\end{aligned}$$

- principal components y_i are uncorrelated and have variances λ_i:

$$\mathbf{E}\left((y_i - \bar{y}_i)(y_j - \bar{y}_j)\right) = 0 \quad \text{if } i \neq j, \qquad \mathbf{E}(y_i - \bar{y}_i)^2 = \lambda_i$$

- principal components are ordered in order of decreasing variance

Example

multivariate normal (Gaussian) probability density function

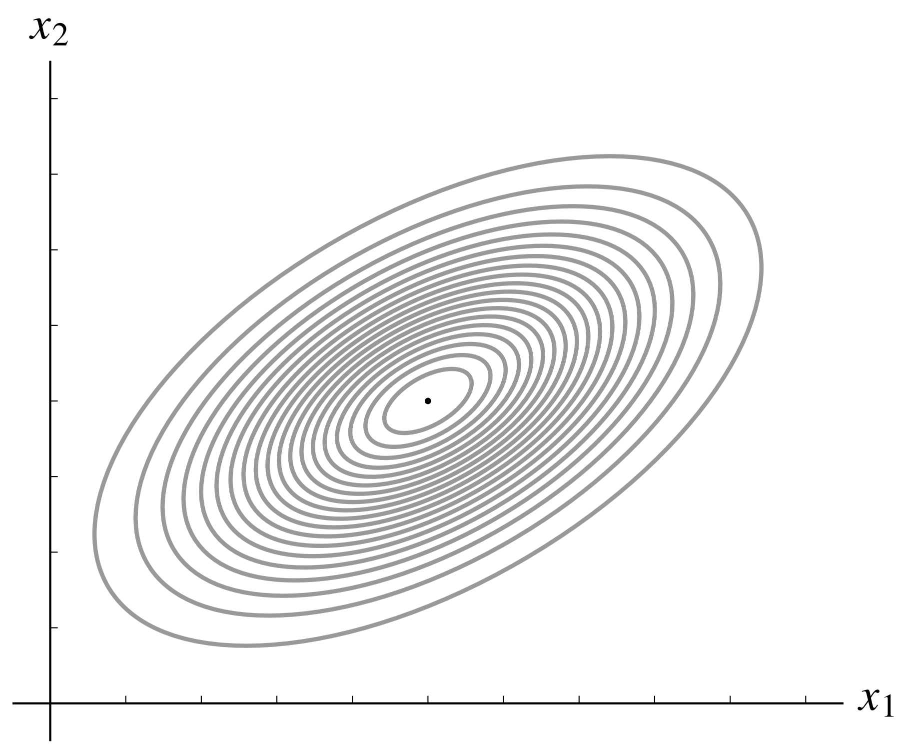
$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}$$

contour lines of density function for

$$C = \frac{1}{4}\begin{bmatrix} 7 & \sqrt{3} \\ \sqrt{3} & 5 \end{bmatrix}, \quad \mu = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

eigenvalues of Σ are $\lambda_1 = 2$, $\lambda_2 = 1$,

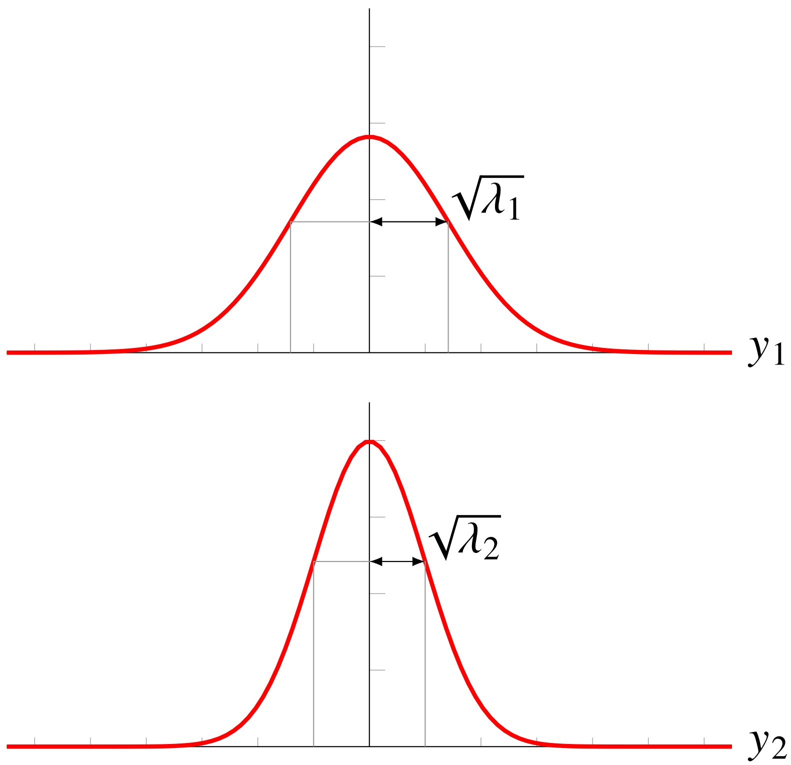
$$q_1 = \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \quad q_2 = \begin{bmatrix} 1/2 \\ -\sqrt{3}/2 \end{bmatrix}$$

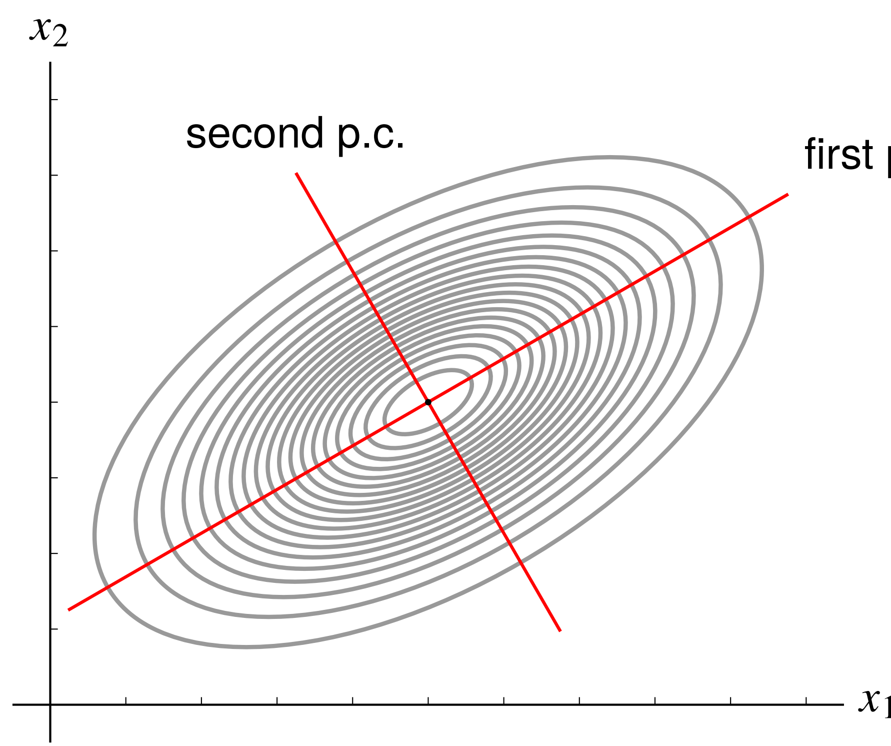


Multivariate normal distribution

the principal components $y_1 = q_1^T x, \ldots, y_n = q_n^T x$ have distribution

$$\tilde{p}(y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda_i}} \exp(-\frac{(y_i - \bar{y}_i)^2}{2\lambda_i})$$





First principal component

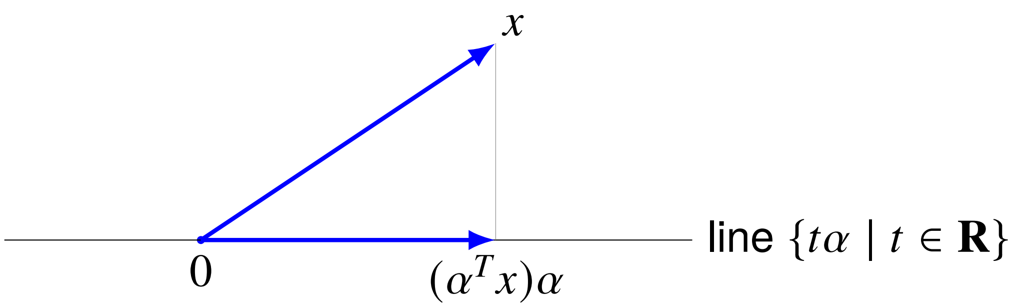
from page 3.24, the first eigenvector q_1 solves the optimization problem (in β)

$$\begin{array}{ll} \text{maximize} & \alpha^T C \alpha \\ \text{subject to} & \alpha^T \alpha = 1 \end{array} \qquad (1)$$

- cost function is the variance of scalar random variable $z = \alpha^T x$:

$$\mathbf{E}\, z = \alpha^T \mu, \qquad \mathbf{E}(z - \alpha^T \mu)^2 = \mathbf{E}(\alpha^T (x-\mu)(x-\mu)^T \alpha) = \alpha^T C \alpha$$

- with $\|\alpha\| = 1$, scalar z gives projection of x on the line in direction α



- in (1) we seek the direction α that maximizes the variance of z
- a solution of (1) is $\alpha = q_1$, the direction of the first principal component $y_1 = q_1^T x$

Second principal component

the second eigenvector q_2 solves the optimization problem

$$\begin{array}{ll} \text{maximize} & \alpha^T C\alpha \\ \text{subject to} & \alpha^T\alpha = 1 \\ & q_1^T\alpha = 0 \end{array} \qquad (2)$$

- cost function is again the variance of the scalar random variable $z = \alpha^T x$
- second constraint forces z to be uncorrelated with first principal component y_1:

$$\begin{aligned} \mathbf{E}((y_1 - q_1^T\mu)(z - \alpha^T\mu)) &= \mathbf{E}(q_1^T(x-\mu)(x-\mu)^T\alpha) \\ &= q_1^T C\alpha \\ &= \lambda_1 q_1^T\alpha \\ &= 0 \end{aligned}$$

- z gives projection of x on line in a direction orthogonal to direction q_1
- a solution of (2) is $\alpha = q_2$, direction of 2nd p.c. y_2 (see next page)

Second principal component

$$\begin{array}{ll} \text{maximize} & \alpha^T C \alpha \\ \text{subject to} & \alpha^T \alpha = 1 \\ & q_1^T \alpha = 0 \end{array}$$

- the 2nd constraint restricts α to the subspace orthogonal to $\text{span}\{q_1\}$
- the columns of $V = \begin{bmatrix} q_2 & \cdots & q_n \end{bmatrix}$ are an orthonormal basis for this subspace
- hence α must be of the form $\alpha = V\tilde{\alpha}$, with $\|\tilde{\alpha}\| = 1$, and problem is equivalent to

$$\begin{array}{ll} \text{maximize} & \tilde{\alpha}^T V^T C V \tilde{\alpha} \\ \text{subject to} & \tilde{\alpha}^T \tilde{\alpha} = 1 \end{array}$$

where

$$V^T C V = \begin{bmatrix} \lambda_2 & 0 & \cdots & 0 \\ 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

- $\tilde{\alpha} = (1, 0, \ldots, 0)$ is optimal, corresponding to $\alpha = q_2$ and $\alpha^T C \alpha = \lambda_2$

Plane defined by the first two principal components

- let $V = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}$ be an $n \times 2$ matrix with orthonormal columns α_1, α_2
- the projection of x on the plane spanned by α_1, α_2 is Vz where

$$z = V^T x$$

$z = (z_1, z_2)$ is a random vector with mean $(\bar{z}_1, \bar{z}_2) = (\alpha_1^T \mu, \alpha_2^T \mu)$ and covariance

$$\begin{bmatrix} \mathbf{E}(z_1 - \bar{z}_1)^2 & \mathbf{E}((z_1 - \bar{z}_1)(z_2 - \bar{z}_2)) \\ \mathbf{E}((z_1 - \bar{z}_1)(z_2 - \bar{z}_2)) & \mathbf{E}(z_2 - \bar{z}_2)^2 \end{bmatrix} = V^T C V$$

- from Courant–Fischer theorem (page 3.35) eigenvalues μ_1, μ_2 of $V^T C V$ satisfy

$$\mu_1 \leq \lambda_1, \qquad \mu_2 \leq \lambda_2 \qquad \text{with equality if } \alpha_1 = q_1, \alpha_2 = q_2$$

- plane of first two p.c. directions maximizes several useful quantities at once:

$$\lambda_{\max}(V^T C V), \quad \lambda_{\min}(V^T C V), \quad \mathrm{trace}(V^T C V), \quad \|V^T C V\|_F, \quad \det(V^T C V), \quad \ldots$$

Higher principal components

the interpretation of the second p.c. is easily extended to the other p.c.'s: consider

$$
\begin{array}{ll}
\text{maximize} & \alpha^T C \alpha \\
\text{subject to} & \alpha^T \alpha = 1 \\
 & q_1^T \alpha = \cdots = q_{k-1}^T \alpha
\end{array}
\qquad (3)
$$

- cost function is again the variance of the scalar random variable $z = \alpha^T x$
- second set of constraints forces z to be uncorrelated with $y_1, \ldots, y_{k-1}$
- a solution of (2) is $\alpha = q_k$, the direction of the kth principal component y_k
- Courant–Fischer theorem implies other optimality properties of first k p.c.'s

Sample principal components

if the covariance matrix is not known, we use the sample covariance matrix

$$\widehat{C} = \frac{1}{m} X_{\mathrm{c}}^T X_{\mathrm{c}} = \frac{1}{m} X^T (I - \frac{1}{m} \mathbf{1}\mathbf{1}^T) X$$

- X is $m \times n$ data matrix, containing m samples of the random n-vector x
- X_{c} is the centered data matrix

$$X_{\mathrm{c}} = (I - \frac{1}{m} \mathbf{1}\mathbf{1}^T) X = X - \mathbf{1}\hat{\mu}^T, \qquad \hat{\mu} = \frac{1}{m} X^T \mathbf{1}$$

- we distinguish *sample* (from $\widehat{C}$) and *population* (from C) principal components
- directions of sample principal components are the right singular vectors of X_{c}

Example

scatter plot shows $m = 500$ points from the normal distribution on page 5.5

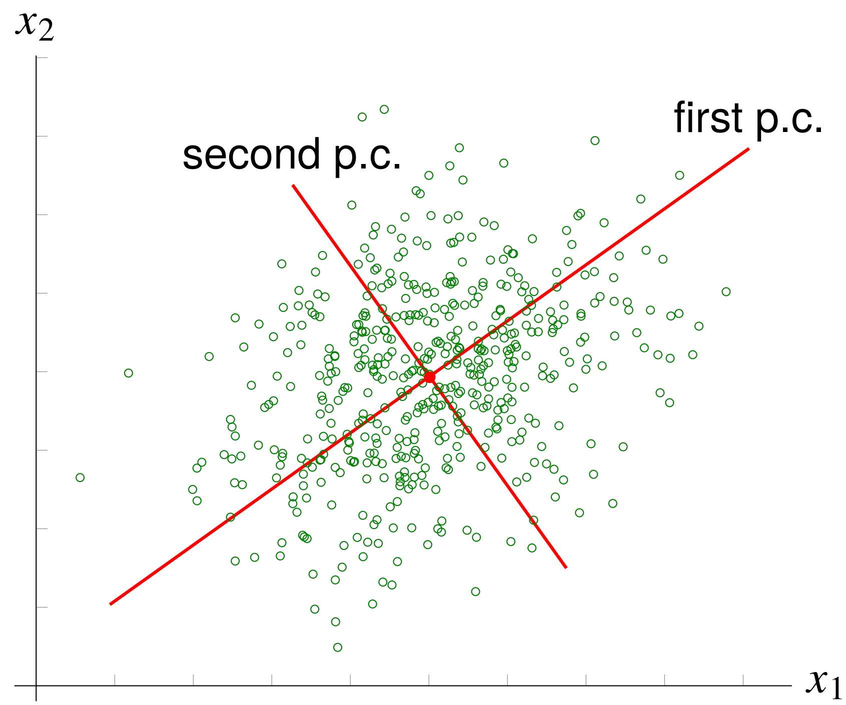
$$\mu = \begin{bmatrix} 5 \\ 4 \end{bmatrix}, \qquad C = \frac{1}{4}\begin{bmatrix} 7 & \sqrt{3} \\ \sqrt{3} & 5 \end{bmatrix}$$

sample estimate of mean is

$$\hat{\mu} = \frac{1}{m} X^T \mathbf{1} = \begin{bmatrix} 5.01 \\ 3.93 \end{bmatrix}$$

sample estimate of covariance is

$$\widehat{C} = \frac{1}{m} X_c^T X_c = \begin{bmatrix} 1.67 & 0.48 \\ 0.48 & 1.35 \end{bmatrix}$$



Outline

- principal components
- **canonical correlations**
- dimension reduction
- rank-deficient least squares
- regularized least squares
- total least squares

Correlation of two random variables

let w, z be two scalar random variables with means and covariance

$$\mathbf{E}\begin{bmatrix} w \\ z \end{bmatrix} = \begin{bmatrix} \bar{w} \\ \bar{z} \end{bmatrix}, \qquad \mathbf{E}\begin{bmatrix} w-\bar{w} \\ z-\bar{z} \end{bmatrix}\begin{bmatrix} w-\bar{w} \\ z-\bar{z} \end{bmatrix}^T = \begin{bmatrix} \sigma_w^2 & \sigma_{wz} \\ \sigma_{zw} & \sigma_z^2 \end{bmatrix}$$

recall the following definitions from lecture 2:

- σ_w^2, σ_z^2 are the variances of w, z
- σ_w, σ_z are the standard deviations of w, z
- $\sigma_{wz} = \sigma_{zw}$ is the covariance between w, z
- correlation between w, z is defined as

$$\rho_{wz} = \frac{\sigma_{wz}}{\sigma_w \sigma_z}$$

Exercise: show that $-1 \leq \rho_{wz} \leq 1$

Correlation of two vectors

in 133A we defined the correlation between non-constant m-vectors a, b as

$$\hat{\rho} = \frac{\tilde{b}^T \tilde{a}}{\|\tilde{a}\| \|\tilde{b}\|}$$

where $\tilde{a}$, $\tilde{b}$ are the *de-meaned* vectors

$$\tilde{a} = (I - \frac{1}{m}\mathbf{1}\mathbf{1}^T)a = a - \mathbf{avg}(a)\mathbf{1}, \qquad \tilde{b} = (I - \frac{1}{m}\mathbf{1}\mathbf{1}^T)b = b - \mathbf{avg}(b)\mathbf{1}$$

- $\hat{\rho}$ is the cosine of the angle between the de-meaned vectors $\tilde{a}, \tilde{b}$
- serves as an estimate of ρ_{wz} if a, b contain m samples of random scalars w, z

First canonical correlation

- assume $x \in \mathbf{R}^p$ and $y \in \mathbf{R}^q$ are random vectors with

$$\mathbf{E}\, x = \bar{x}, \qquad \mathbf{E}\, y = \bar{y}, \qquad \mathbf{E} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}^T = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$$

- consider two scalar linear combinations $w = \alpha^T x$ and $z = \beta^T y$:

$$\begin{bmatrix} \bar{w} \\ \bar{z} \end{bmatrix} = \begin{bmatrix} \alpha^T \bar{x} \\ \beta^T \bar{y} \end{bmatrix}, \qquad \begin{bmatrix} \sigma_w^2 & \sigma_{wz} \\ \sigma_{zw} & \sigma_z^2 \end{bmatrix} = \begin{bmatrix} \alpha^T C_{xx} \alpha & \alpha^T C_{xy} \beta \\ \beta^T C_{yx} \alpha & \beta^T C_{yy} \beta \end{bmatrix}$$

  (using the notation of p. 5.14)

- we are interested in determine a, b that maximize the correlation

$$\rho_{wz} = \frac{\sigma_{wz}}{\sigma_w \sigma_z} = \frac{\alpha^T C_{xy} \beta}{(\alpha^T C_{xx} \alpha)^{1/2} (\beta^T C_{yy} \beta)^{1/2}}$$

- maximum ρ_{wz} is called the first *canonical correlation*
- optimal $w = \alpha^T x$ and $z = \beta^T y$ are the 1st *canonical variates*

First canonical correlation via SVD

the directions α, β that maximize ρ_{wz} are solutions of the optimization problem

$$\begin{array}{ll} \text{maximize} & \alpha^T C_{xy} \beta \\ \text{subject to} & \alpha^T C_{xx} \alpha = \beta^T C_{yy} \beta = 1 \end{array}$$

- take Cholesky factorization of C_{xx}, C_{yy} (assuming they are positive definite)

$$C_{xx} = R_x^T R_x, \qquad C_{yy} = R_y^T R_y$$

- apply a change of variables $\tilde{\alpha} = R_x \alpha$ and $\tilde{\beta} = R_y \beta$:

$$\begin{array}{ll} \text{maximize} & \tilde{\alpha}^T R_x^{-T} C_{xy} R_y^{-1} \tilde{\beta} \\ \text{subject to} & \tilde{\alpha}^T \tilde{\alpha} = \tilde{\beta}^T \tilde{\beta} = 1 \end{array}$$

- from page 4.21, solution follows from SVD of $R_x^{-T} C_{xy} R_y^{-1}$:

$$\tilde{\alpha} = u_1, \qquad \tilde{\beta} = v_1, \qquad \alpha = R_x^{-1} u_1, \qquad \beta = R_y^{-1} v_1, \qquad \rho_{wz} = \sigma_1$$

u_1, v_1 are first left and right singular vectors, σ_1 is first singular value

Higher canonical correlations

- assume $p \geq q$ (where $x \in \mathbf{R}^p$ and $y \in \mathbf{R}^q$) and consider the reduced SVD

$$R_x^{-T} C_{xy} R_y^{-1} = U \Sigma V^T = \sum_{i=1}^{q} \sigma_i u_i v_i^T$$

- the *canonical correlations* between x and y are the singular values $\sigma_1, \ldots, \sigma_q$
- the kth *canonical variates* are the scalar variables w_k, z_k where

$$\begin{bmatrix} w_1 \\ \vdots \\ w_q \end{bmatrix} = \begin{bmatrix} u_1^T R_x^{-T} x \\ \vdots \\ u_q^T R_x^{-T} x \end{bmatrix}, \qquad \begin{bmatrix} z_1 \\ \vdots \\ z_q \end{bmatrix} = \begin{bmatrix} v_1^T R_y^{-T} y \\ \vdots \\ v_q^T R_y^{-T} y \end{bmatrix}$$

- interpretation: w_k, z_k are linear combinations $w = \alpha^T x$, $z = \beta^T y$ that solve

$$\begin{array}{ll} \text{maximize} & \rho_{wz} \\ \text{subject to} & w \text{ is uncorrelated with } w_1, \ldots, w_{k-1} \\ & z \text{ is uncorrelated with } z_1, \ldots, z_{k-1} \end{array}$$

Sample canonical correlations

if the covariance matrices are not known, we use the sample covariances

$$\begin{bmatrix} \widehat{C}_{xx} & \widehat{C}_{xy} \\ \widehat{C}_{yx} & \widehat{C}_{yy} \end{bmatrix} = \frac{1}{m} \begin{bmatrix} X_c^T X_c & X_c^T Y_c \\ Y_c^T X_c & Y_c^T Y_c \end{bmatrix}$$

- $X_c \in \mathbf{R}^{m \times p}$ and $Y_c \in \mathbf{R}^{m \times q}$ are centered data matrices for m samples of x, y
- first (sample) canonical variates $w = \alpha^T x$ and $z = \beta^T y$ maximize

$$\hat{\rho} = \frac{\alpha^T X_c^T Y_c \beta}{\|X_c \alpha\| \, \|Y_c \beta\|}$$

i.e., we find linear combinations of colums of X_c and Y_c with largest correlation

References

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning* (2013), §10.2.

- I.T. Jolliffe, *Principal Component Analysis* (2002).

Outline

- principal components
- canonical correlations
- **dimension reduction**
- rank-deficient least squares
- regularized least squares
- total least squares

Dimension reduction

low-rank approximation of data matrix can improve efficiency or performance

$$A \approx \tilde{A}Q^T \qquad \text{where } \tilde{A} \text{ is } m \times k \text{ and } Q \text{ is } n \times k$$

- we assume (without loss of generality) that Q has orthonormal columns
- columns of Q are a basis for a k-dimensional subspace in feature space $\mathbf{R}^n$
- $\tilde{A}$ is reduced data matrix; rows $\tilde{a}_i^T$ are reduced feature vectors:

$$a_i \approx Q\tilde{a}_i, \quad i = 1, \ldots, m$$

we discuss three choices for $\tilde{A}$ and Q

- truncated singular value decomposition
- truncated QR factorization
- k-means clustering

Truncated singular value decomposition

truncate SVD $A = U\Sigma V^T = \sum_i \sigma_i u_i v_i^T$ after k terms: $A \approx \tilde{A}Q^T$ with

$$\begin{aligned} \tilde{A} &= \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \cdots & \sigma_k u_k \end{bmatrix} \\ Q &= \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix} \end{aligned}$$

- $\tilde{A}Q^T$ is the best rank-k approximation of the data matrix A (see page 4.31)

$$\tilde{A}Q^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T \approx A$$

- rows $\tilde{a}_i^T$ of $\tilde{A}$ are (coordinates of) projections of the rows a_i^T on range of Q

$$\tilde{A} = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T \right) Q = AQ$$

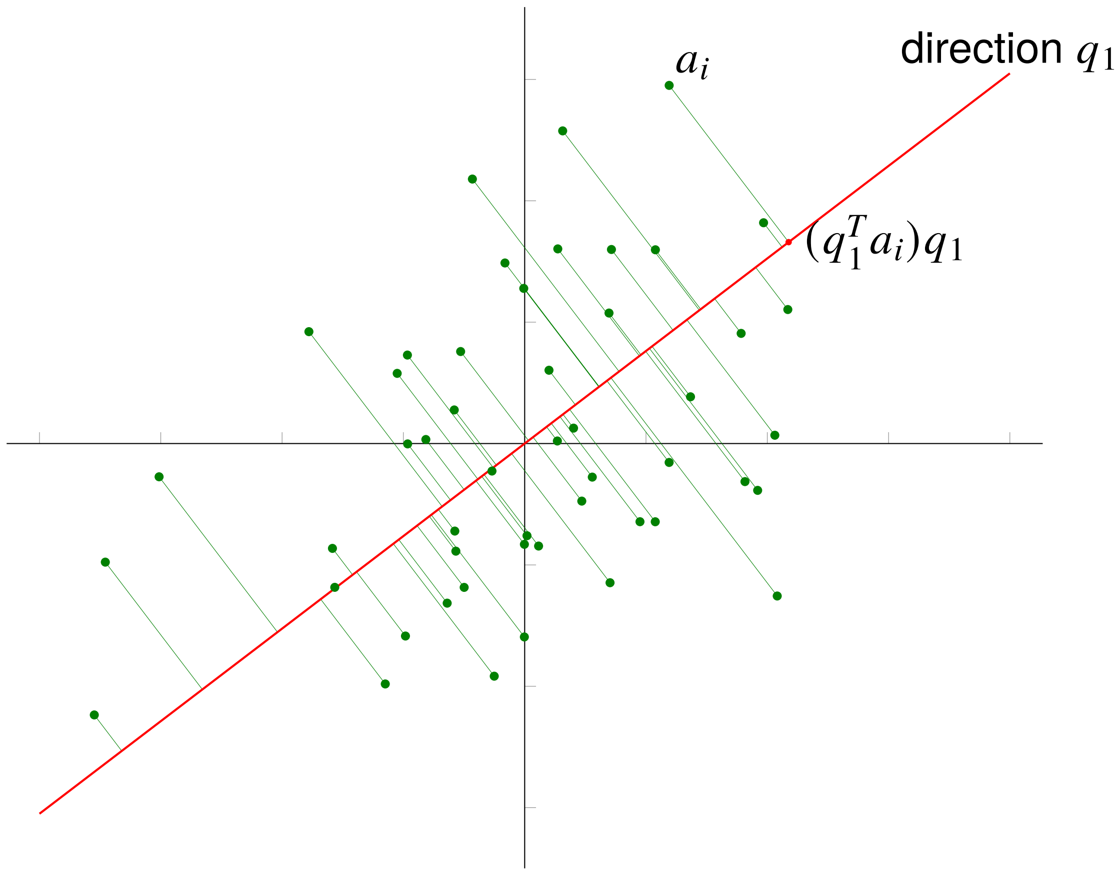
when A is centered ($\mathbf{1}^T A = 0$), columns in Q are the principal components

Interpretation

max–min properties of SVD give the columns of Q important optimality properties

First component: q_1 is the direction q that maximizes

$$\|Aq\|^2 = (q^T a_1)^2 + \cdots + (q^T a_m)^2$$

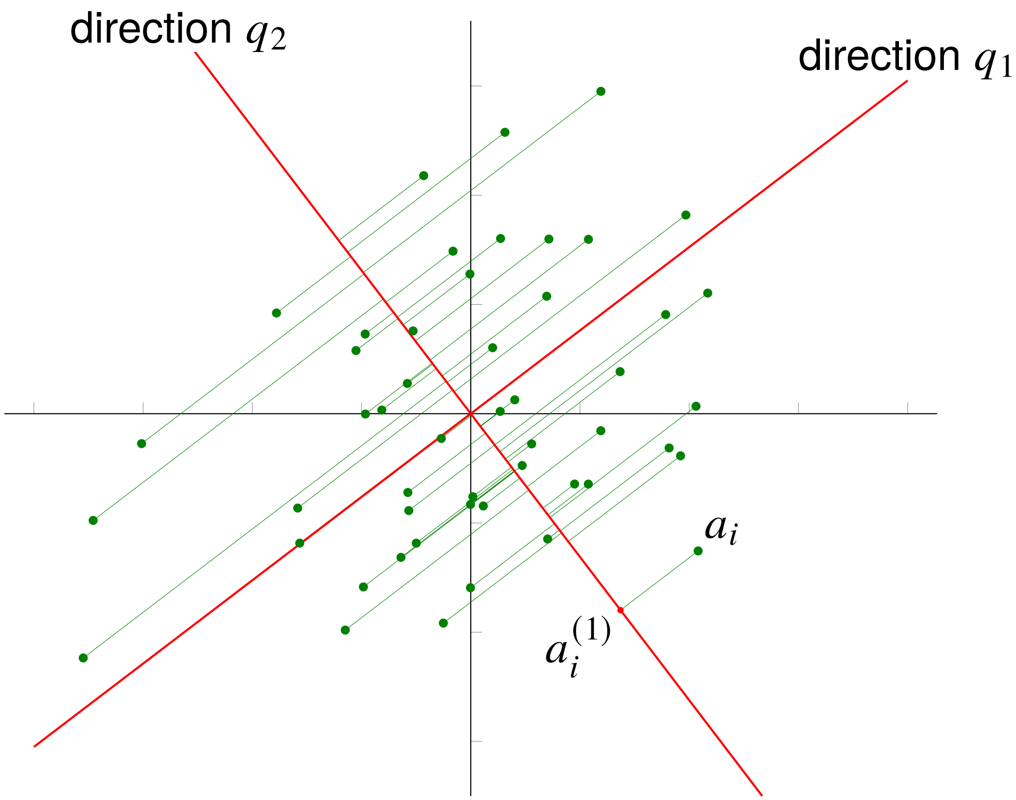


Interpretation

Second component: $q_2 = v_2$ is the first right singular vector of

$$A^{(1)} = A - \sigma_1 u_1 v_1^T = A(I - q_1 q_1^T)$$

- rows of $A^{(1)}$ are the rows of A projected on the orthogonal complement of q_1
- q_2 is the direction q that maximizes $\|A^{(1)}q\|^2$



Interpretation

Component i

$q_i = v_i$ is the first singular vector of

$$A^{(i-1)} = A - \sum_{j=1}^{i-1} \sigma_j u_j v_j^T = A(I - q_1 q_1^T - \cdots - q_{i-1} q_{i-1}^T)$$

- rows of $A^{(i-1)}$ are the rows of A projected on $\mathrm{span}\{q_1, \ldots, q_{i-1}\}^\perp$
- q_i is the direction q that maximizes

$$\|A^{(i-1)} q\|^2 = \left(q^T a_1^{(i-1)}\right)^2 + \left(q^T a_2^{(i-1)}\right)^2 + \cdots + \left(q^T a_m^{(i-1)}\right)^2$$

Truncated QR factorization

truncate the pivoted QR factorization of A^T after k steps

- partial QR factorization after k steps (see page 1.26)

$$PA = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T + \begin{bmatrix} 0 \\ B^T \end{bmatrix}, \qquad B^T Q = 0$$

  P a permutation, R_1 is $k \times k$ and upper triangular, Q has orthonormal columns

- to define a rank-k reduced data matrix we drop B and use the first term

$$PA \approx \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T$$

this does not have the optimality properties of the SVD but is cheaper to compute

Reduced data matrix

$$PA = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \approx \begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix} Q^T$$

- $A_1 = R_1^T Q^T$: a subset of k examples from the original data matrix A
- the k-dimensional reduced feature subspace is

$$\mathrm{range}(Q) = \mathrm{range}(QR_1) = \mathrm{range}(A_1^T)$$

  reduced subspace is spanned by the feature vectors in A_1

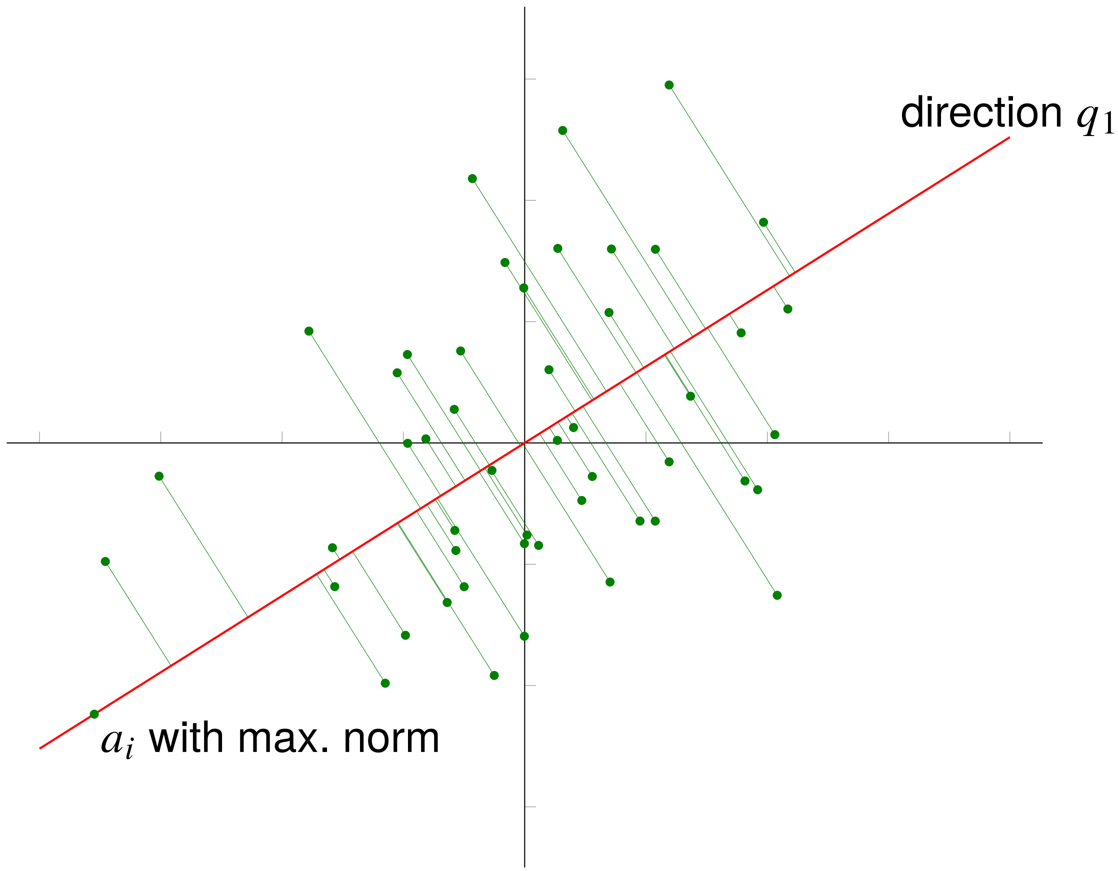
- the rows of $R_2^T Q^T$ are the rows of A_2 projected on $\mathrm{range}(Q)$:

$$A_2 Q Q^T = (R_2^T Q^T + B^T) Q Q^T = R_2^T Q^T$$

Interpretation

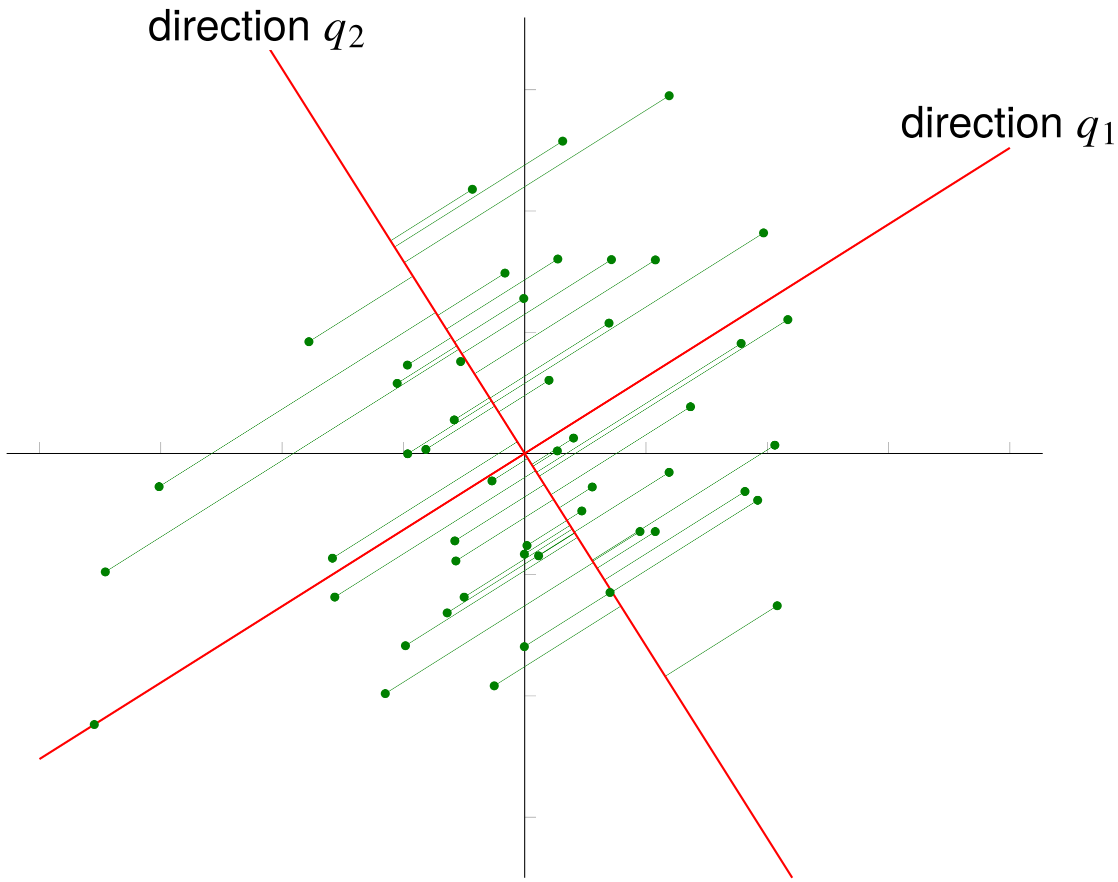
we use the pivoting rule of page 1.26

First component: q_1 is direction of largest row in A



Interpretation

Second component: q_2 is direction of largest row of $A^{(1)} = A(I - q_1 q_1^T)$



Component i: q_i is direction of largest row of

$$A^{(i-1)} = A(I - q_1 q_1^T) \cdots (I - q_{i-1} q_{i-1})^T$$

k-means clustering

run k-means on the rows of A to cluster them in k groups with representatives

$$b_1, \quad b_2, \quad \ldots, \quad b_k \in \mathbf{R}^n$$

- this can be interpreted as a rank-k approximation of A:

$$A \approx CB^T, \qquad C_{ij} = \begin{cases} 1 & \text{row } i \text{ of } A \text{ is assigned to group } j \\ 0 & \text{otherwise} \end{cases}$$

  in other words, in CB^T each row a_i^T is replaced by its group representative

- QR factorization $B = QR$ gives an orthonormal basis for $\mathrm{range}(B)$
- $\tilde{A} = CR^T$ is a possible choice of reduced data matrix
- alternatively, to improve approximation one computes $\tilde{A}$ by minimizing

$$\|A - \tilde{A}Q^T\|_F^2$$

Example: document analysis

a collection of documents is represented by a *term–document matrix* D

- each row corresponds to a word in a dictionary
- each column corresponds to a document

entries give frequencies of word in documents, usually weighted, for example, as

$$D_{ij} = f_{ij} \log(m/m_i)$$

- f_{ij} is frequency of term i in document j
- m is number of documents
- m_i is number of documents that contain term i

for consistency with the earlier notation, we define

$$A = D^T$$

A is $m \times n$ (number of documents $\times$ number of words)

Comparing documents and queries

Comparing documents: as measure of document similarity, we can use

$$\frac{a_i^T a_j}{\|a_i\| \|a_j\|}$$

- a_i^T and a_j^T are the rows of $A = D^T$ corresponding to documents i and j
- this is called the *cosine similarity:* cosine of the angle beween a_i and a_j

Query matching: find the most relevant documents based on keywords in a query

- we treat the query as a simple document, represented by an n-vector x:

$$x_j = 1 \quad \text{if term } j \text{ appears in the query,} \qquad x_j = 0 \quad \text{otherwise}$$

- we rank documents according to their cosine similiarity with x:

$$\frac{a_i^T x}{\|a_i\| \|x\|}, \quad j = 1, \ldots, m$$

Dimension reduction

it is common to make a low-rank approximation of the term–document matrix

$$D^T = A \approx \tilde{A} Q^T$$

- if the truncated SVD is used, this is called *latent semantic indexing* (LSI)
- LSI is early technique for search engines (anno 1990)
- cosine similarity of query vector x with ith row $Q\tilde{a}_i$ of reduced data matrix is

$$\frac{\tilde{a}_i^T Q^T x}{\|Q\tilde{a}_i\| \|x\|} = \frac{\tilde{a}_i^T Q^T x}{\|\tilde{a}_i\| \|x\|}$$

- an alternative is to compute the angle between $\tilde{a}_i$ and $Q^T x$:

$$\frac{\tilde{a}_i^T Q^T x}{\|\tilde{a}_i\| \|Q^T x\|}$$

References

- Lars Eldén, *Matrix Methods in Data Mining and Pattern Recognition* (2007), chapter 11.

  describes the document analysis application, including Latent Semantic Indexing and k-means clustering

- Michael W. Berry, Zlatko Drmač, Elizabeth R. Jessup, *Matrices, Vector Spaces, and Information Retrieval*, SIAM Review (1999).

  also discusses the QR factorization method

Outline

- principal components
- canonical correlations
- dimension reduction
- **rank-deficient least squares**
- regularized least squares
- total least squares

Minimum-norm least squares solution

least squares problem with $m \times n$ matrix A and $\mathrm{rank}(A) = r$ (possibly $r < n$)

$$\text{minimize} \quad \|Ax - b\|^2$$

- on page 1.42 we showed that the minimum-norm least squares solution is

$$\hat{x} = A^{\dagger} b$$

- other (not minimum-norm) LS solutions are $\hat{x} + v$ for nonzero $v \in \mathrm{null}(A)$

if A has rank r and SVD $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, the formulas for $A^{\dagger}$ and $\hat{x}$ are

$$A^{\dagger} = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T, \qquad \hat{x} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

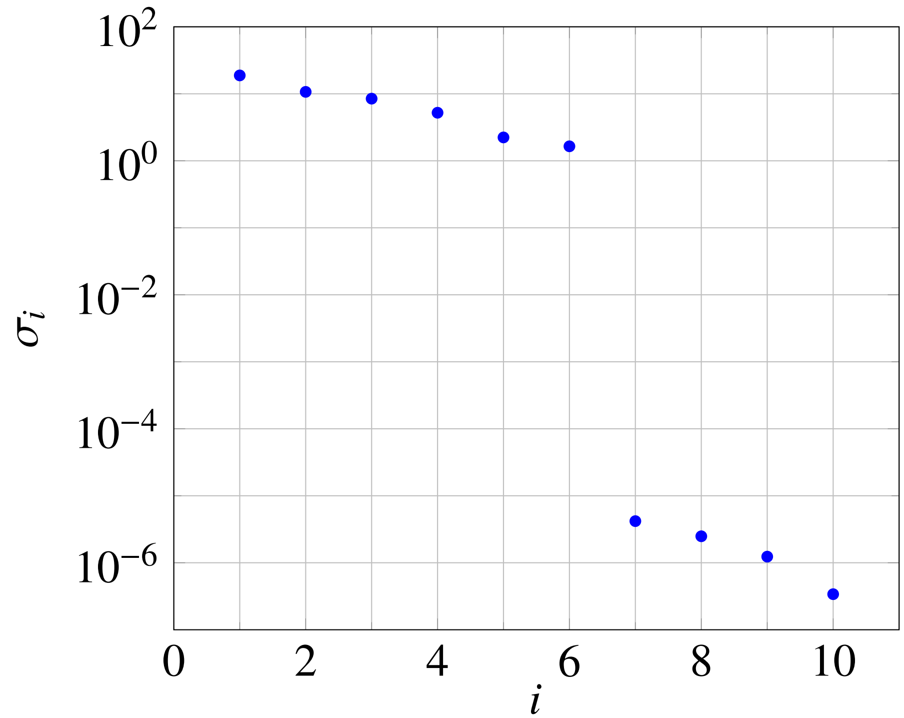
(see page 4.14 for expresson of the pseudo-inverse)

Estimating rank

a perturbation of a rank-deficient matrix will make all singular values nonzero

Example (10×10 matrix)

singular values suggest matrix is a perturbation of a matrix with rank 6



- the *numerical rank* is the number of singular values above a certain threshold
- good value of threshold is application-dependent
- truncating after numerical rank $\tilde{r}$ removes influence of small singular values

$$\hat{x} = \sum_{i=1}^{\tilde{r}} \frac{u_i^T b}{\sigma_i} v_i$$

Outline

- principal components
- canonical correlations
- low-rank matrix representations
- rank-deficient least squares
- **regularized least squares**
- total least squares

Tikhonov regularization

least squares problem with quadratic regularization

$$\text{minimize} \quad \|Ax - b\|^2 + \lambda\|x\|^2$$

- known as *Tikhonov regularization* or *ridge regression*
- weight λ controls trade-off between two objectives $\|Ax - b\|^2$ and $\|x\|^2$
- regularization term can help avoid over-fitting
- equivalent to standard least squares problem with a stacked matrix:

$$\text{minimize} \quad \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2$$

- for positive λ, the regularized problem always has a unique solution

$$\hat{x}_\lambda = (A^T A + \lambda I)^{-1} A^T b$$

Exercise

regularized least squares problem with a column of ones in the coefficient matrix:

$$\text{minimize} \quad \left\| \begin{bmatrix} \mathbf{1} & A \end{bmatrix} \begin{bmatrix} v \\ x \end{bmatrix} - b \right\|^2 + \lambda \|x\|^2$$

- data matrix includes a constant feature $\mathbf{1}$ (parameter v is the offset or intercept)
- associated variable v is excluded from regularization term

show that the problem is equivalent to

$$\text{minimize} \quad \|A_{\text{c}}x - b\|^2 + \lambda \|x\|^2$$

where A_{c} is the centered data matrix

$$A_{\text{c}} = (I - \frac{1}{m}\mathbf{1}\mathbf{1}^T)A = A - \frac{1}{m}\mathbf{1}(\mathbf{1}^T A)$$

Regularization path

suppose A has full SVD

$$A = U\Sigma V^T = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T$$

substituting the SVD in the formula for $\hat{x}_\lambda$ shows the effect of λ:

$$\begin{aligned} \hat{x}_\lambda = (A^TA + \lambda I)^{-1}A^Tb &= (V\Sigma^T\Sigma V^T + \lambda I)^{-1}V\Sigma^T U^T b \\ &= V(\Sigma^T\Sigma + \lambda I)^{-1}V^T V\Sigma^T U^T b \\ &= V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T U^T b \\ &= \sum_{i=1}^{\min\{m,n\}} \frac{\sigma_i(u_i^T b)}{\sigma_i^2 + \lambda} v_i \end{aligned}$$

this expression is valid for any matrix shape and rank

Interpretation

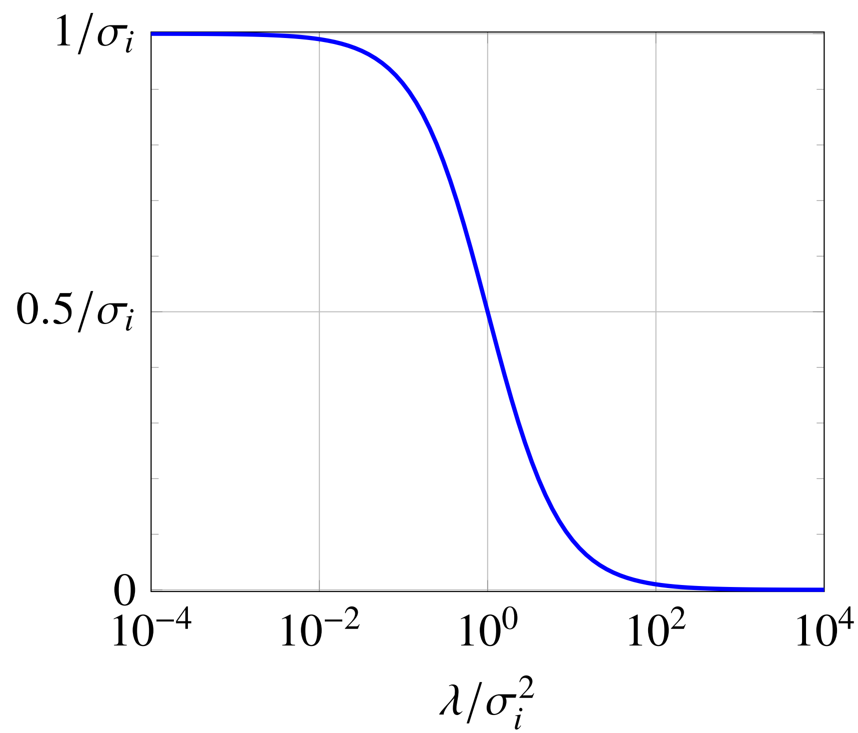
$$\hat{x}_\lambda = \sum_{i=1}^{\min\{m,n\}} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i (u_i^T b)$$

- positive λ reduces (shrinks) all terms in the sum
- terms for small σ_i are suppressed more
- all terms with $\sigma_i = 0$ are removed

plot shows the weight function

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} = \frac{1/\sigma_i}{1 + \lambda/\sigma_i^2}$$

versus λ, for a term with $\sigma_i > 0$



Truncated SVD as regularization

- suppose we determine numerical rank of A by comparing σ_i with threshold τ
- truncating SVD of A gives approximation $\tilde{A} = \sum_{\sigma_i > \tau} \sigma_i u_i v_i^T$
- minimum-norm least squares solution for truncated matrix is (page 5.36)

$$\hat{x}_{\text{trunc}} = \sum_{\sigma_i > \tau} \frac{1}{\sigma_i} v_i (u_i^T b)$$

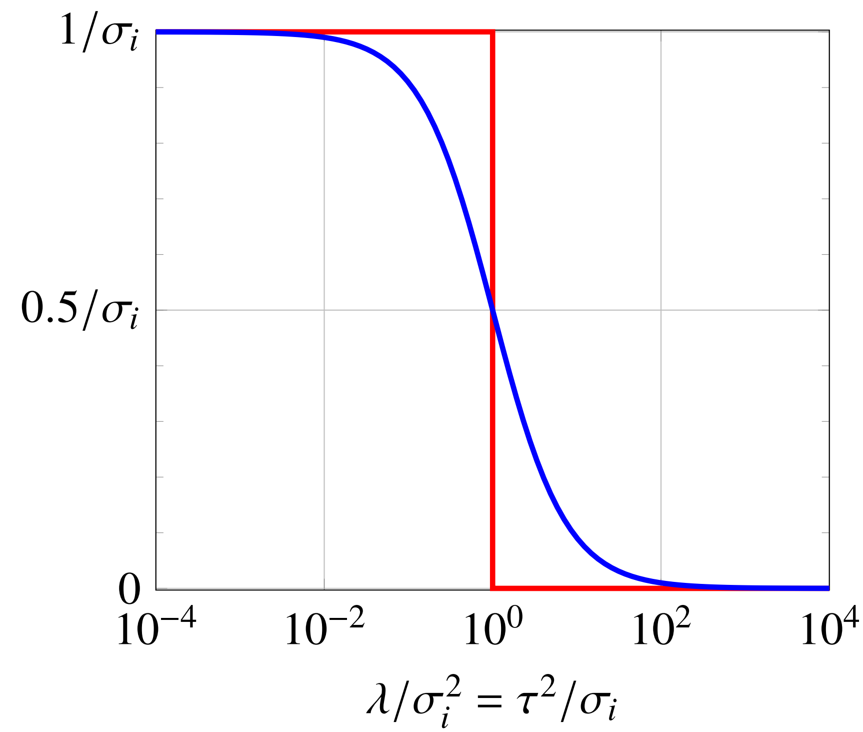
plot shows two weight functions

- Tikhonov regularization:

$$\frac{1/\sigma_i}{1 + \lambda/\sigma_i^2}$$

- truncated SVD solution with $\tau = \sqrt{\lambda}$:

$$\begin{cases} 1/\sigma_i & \sigma_i > \sqrt{\lambda} \\ 0 & \sigma_i \leq \sqrt{\lambda} \end{cases}$$



Limit for $\lambda = 0$

Regularized least squares solution

$$\hat{x}_\lambda = \sum_{i=1}^{\min\{m,n\}} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i (u_i^T b) = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i (u_i^T b)$$

- the limit for $\lambda \to 0$ is

$$\lim_{\lambda \to 0} \hat{x}_\lambda = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i (u_i^T b)$$

- this is the minimum-norm solution of the unregularized problem (page 5.35)

Pseudo-inverse: this gives a new interpretation of the pseudo-inverse

$$\begin{aligned} A^\dagger = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T &= \lim_{\lambda \to 0} \sum_{i=1}^{\min\{m,n\}} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i u_i^T \\ &= \lim_{\lambda \to 0} (A^T A + \lambda I)^{-1} A^T \end{aligned}$$

Example

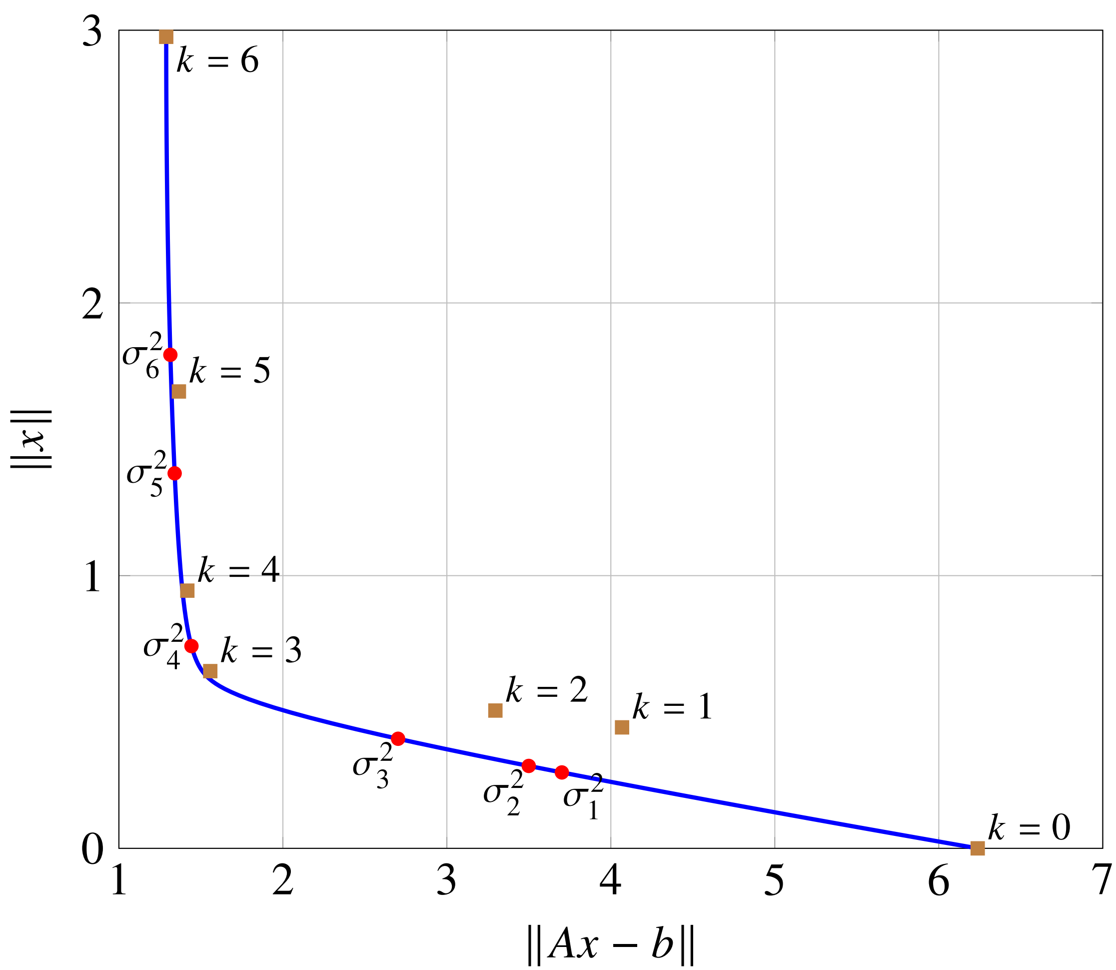
10×6 matrix with singular values

$$\sigma_1 = 10.66, \quad \sigma_2 = 9.86, \quad \sigma_3 = 7.11, \quad \sigma_4 = 0.94, \quad \sigma_5 = 0.27, \quad \sigma_6 = 0.18$$

solid line is trade-off curve

●: solution $\hat{x}_\lambda$ with $\lambda = \sigma_i^2$

■: truncate SVD after k terms



Outline

- principal components
- canonical correlations
- low-rank matrix representations
- rank-deficient least squares
- regularized least squares
- **total least squares**

Total least squares

Least squares problem

$$\text{minimize} \quad \|Ax - b\|^2$$

- can be written as constrained least squares problem with variables x and e

$$\begin{array}{ll} \text{minimize} & \|e\|^2 \\ \text{subject to} & Ax = b + e \end{array}$$

- e is the smallest adjustment to b that makes the equation $Ax = b + e$ solvable

Total least squares (TLS) problem

$$\begin{array}{ll} \text{minimize} & \|E\|_F^2 + \|e\|^2 \\ \text{subject to} & (A + E)x = b + e \end{array}$$

- variables are n-vector x, m-vector e, and $m \times n$ matrix E (where A is $m \times n$)
- E and e are the smallest adjustments to A, b that make the equation solvable
- eliminating e gives a nonlinear LS problem: minimize $\|E\|_F^2 + \|(A + E)x - b\|^2$

TLS solution via singular value decomposition

$$\begin{array}{ll} \text{minimize} & \|E\|_F^2 + \|e\|^2 \\ \text{subject to} & (A+E)x = b + e \end{array}$$

we assume $m \geq n+1$ and $\sigma_{\min}(A) > \sigma_{\min}(C) > 0$ where $C = \begin{bmatrix} A & -b \end{bmatrix}$

- compute an SVD of the $m \times (n+1)$ matrix C:

$$C = \begin{bmatrix} A & -b \end{bmatrix} = \sum_{i=1}^{n+1} \sigma_i u_i v_i^T$$

- partition the right singular vector v_{n+1} of C as

$$v_{n+1} = \begin{bmatrix} w \\ z \end{bmatrix} \qquad \text{with } w \in \mathbf{R}^n \text{ and } z \in \mathbf{R}$$

- the solution of the TLS problem is

$$E = -\sigma_{n+1} u_{n+1} w^T, \qquad e = \sigma_{n+1} u_{n+1} z, \qquad x = w/z$$

Proof:

$$\begin{array}{ll} \text{minimize} & \|E\|_F^2 + \|e\|^2 \\ \text{subject to} & \begin{bmatrix} A+E & -(b+e) \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = 0 \end{array}$$

- the matrix of rank n closest to C and its difference with C are

$$\begin{bmatrix} A+E & -(b+e) \end{bmatrix} = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \qquad \begin{bmatrix} E & -e \end{bmatrix} = -\sigma_{n+1} u_{n+1} v_{n+1}^T$$

- $v_{n+1} = (w, z)$ spans the nullspace of this matrix
- if $z \neq 0$ we can normalize v_{n+1} to get a solution $x = w/z$ that satisfies

$$\begin{bmatrix} A+E & -(b+e) \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = 0$$

- the assumption $\sigma_{\min}(A) > \sigma_{\min}(C)$ implies that z is nonzero: $z = 0$ contradicts

$$\sigma_{\min}(A) = \min_{\|y\|=1} \|Ay\| > \sigma_{\min}(C) = \|Aw - bz\|$$

Extension

$$\begin{array}{ll} \text{minimize} & \|E\|_F^2 + \|e\|^2 \\ \text{subject to} & A_1x_1 + (A_2 + E)x_2 = b + e \end{array} \qquad (4)$$

- variables are E, e, x_1, x_2
- we make the smallest adjustment to A_2 and b that makes the equation solvable
- no adjustment is made to A_1
- eliminating e gives a nonlinear least squares problem in E, x_1, x_2:

$$\text{minimize} \quad \|E\|_F^2 + \|A_1x_1 + (A_2 + E)x_2 - b\|^2$$

- we will assume that A_1 has linearly independent columns

Solution

- assume A_1 has QR factorization $A_1 = Q_1 R$ and $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ is orthogonal
- multiply the constraint in (4) on the left with Q^T:

$$R x_1 + (Q_1^T A_2 + E_1) x_2 = Q_1^T b + e_1, \qquad (Q_2^T A_2 + E_2) x_2 = Q_2^T b + e_2 \qquad (5)$$

  where $E_1 = Q_1^T E, \;\; E_2 = Q_2^T E, \;\; e_1 = Q_1^T e, \;\; e_2 = Q_2^T e$
- cost function in (4) is

$$\|E\|_F^2 + \|e\|^2 = \|E_1\|_F^2 + \|E_2\|_F^2 + \|e_1\|^2 + \|e_2\|^2$$

- first equation in (5) is solvable for any E_1, e_1, so $E_1 = 0$, $e_1 = 0$ are optimal
- for the 2nd equation we solve the TLS problem in E_2, e_2, x_2:

$$\begin{array}{ll} \text{minimize} & \|E_2\|_F^2 + \|e_2\|^2 \\ \text{subject to} & (Q_2^T A_2 + E_2) x_2 = Q_2^T b + e_2 \end{array}$$

- after computing x_2, we find x_1 by solving $R x_1 = Q_1^T b - Q_1^T A_2 x_2$

Example: orthogonal distance regression

fit an affine function $f(t) = x_1 + x_2 t$ to m points (a_i, b_i)

$$\begin{array}{ll} \text{minimize} & \|\delta a\|^2 + \|\delta b\|^2 \\ \text{subject to} & x_1 \mathbf{1} + x_2(a + \delta a) = b + \delta b \end{array}$$

- the variables are m-vectors δa, δb and scalars x_1, x_2
- we fit the line by minimizing the sum of squared distances to the line