Chapter 2

Least squares problems

Least-squares and dimensionality reduction

Least-squares and dimensionality reduction

Given n data points in d dimensions:

$$X = \begin{bmatrix} - & \boldsymbol{x}_1^t & - \\ - & \boldsymbol{x}_2^t & - \\ - & \vdots & - \\ - & \boldsymbol{x}_n^t & - \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Want to reduce dimensionality from d to k. Choose k directions $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$, arrange them as columns in matrix W:

$$W = \begin{bmatrix} | & | & & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \ldots & \boldsymbol{w}_k \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{d \times k}$$

Project $\boldsymbol{x} \in \mathbb{R}^d$ down to $\boldsymbol{z} = W^t \boldsymbol{x} \in \mathbb{R}^k$. How to choose W?

Encoding–decoding model

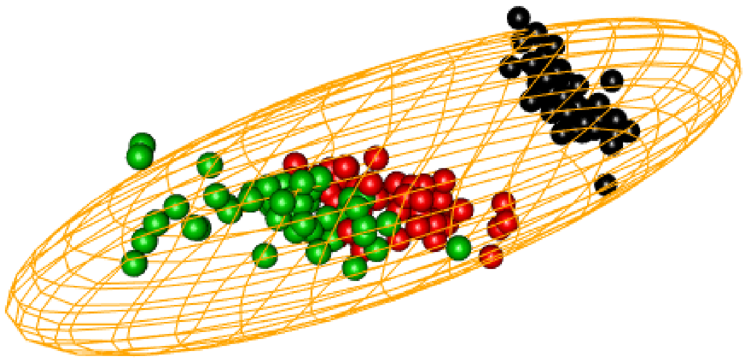
The projection matrix W serves two functions:

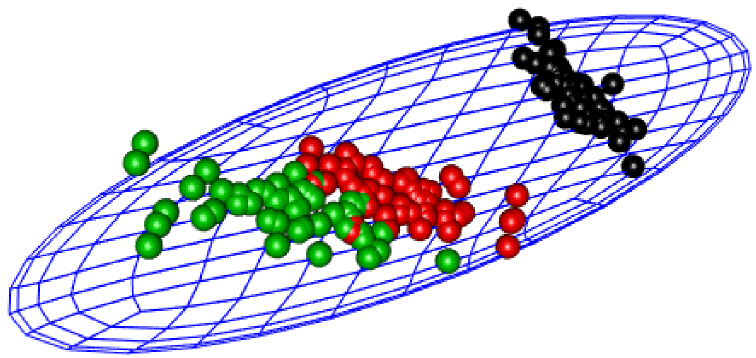
- **Encode:** $\boldsymbol{z} = W^t \boldsymbol{x}, \;\; \boldsymbol{z} \in \mathbb{R}^k, \;\; z_j = \boldsymbol{w}_j^t \boldsymbol{x}$.
  - The vectors $\boldsymbol{w}_j$ form a basis of the projected space.
  - We will require that this basis is orthonormal, i.e. $W^t W = I$.
- **Decode:** $\tilde{\boldsymbol{x}} = W \boldsymbol{z} = \sum_{j=1}^k z_j \boldsymbol{w}_j, \;\; \tilde{\boldsymbol{x}} \in \mathbb{R}^d$.
  - If $k = d$, the above orthonormality condition implies $W^t = W^{-1}$, and encoding can be undone without loss of information.
  - If $k < d$, the decoded $\tilde{\boldsymbol{x}}$ can only approximate $\boldsymbol{x}$ $\rightsquigarrow$ the reconstruction error will be nonzero.
- Note that we did not include an intercept term. Assumption: origin of coordinate system is in the sample mean, i.e. $\sum_i \boldsymbol{x}_i = 0$.

Principal Component Analysis (PCA)

We want the reconstruction error $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2$ to be small.

Objective: minimize $\min_{W \in \mathbb{R}^{d \times k}:\, W^t W = I} \sum_{i=1}^{n} \|\boldsymbol{x}_i - W W^t \boldsymbol{x}_i\|^2$





Finding the principal components

Projection vectors are orthogonal $\rightsquigarrow$ can treat them separately:

$$\min_{\boldsymbol{w}:\,\|\boldsymbol{w}\|=1} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i\|^2$$

$$\begin{aligned}
\sum_i \|\boldsymbol{x}_i - \boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i\|^2 &= \sum_{i=1}^{n}[\boldsymbol{x}_i^t\boldsymbol{x}_i - 2\boldsymbol{x}_i^t\boldsymbol{w}\boldsymbol{w}^t\boldsymbol{x}_i + \boldsymbol{x}_i^t\boldsymbol{w}\underbrace{\boldsymbol{w}^t\boldsymbol{w}}_{=1}\boldsymbol{w}^t\boldsymbol{x}_i] \\
&= \sum_i [\boldsymbol{x}_i^t\boldsymbol{x}_i - \boldsymbol{x}_i^t\boldsymbol{w}\,\boldsymbol{w}^t\boldsymbol{x}_i] \\
&= \sum_i \boldsymbol{x}_i^t\boldsymbol{x}_i - \sum_i \boldsymbol{w}^t\boldsymbol{x}_i\,\boldsymbol{x}_i^t\boldsymbol{w} \\
&= \sum_i \boldsymbol{x}_i^t\boldsymbol{x}_i - \boldsymbol{w}^t\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^t\Big)\boldsymbol{w} \\
&= \underbrace{\sum_i \boldsymbol{x}_i^t\boldsymbol{x}_i}_{\text{const.}} - \boldsymbol{w}^t X^t X \boldsymbol{w}.
\end{aligned}$$

Finding the principal components

- Want to maximize $\boldsymbol{w}^t X^t X \boldsymbol{w}$ under the constraint $\|\boldsymbol{w}\| = 1$
- Can also maximize the ratio $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^t X^t X \boldsymbol{w}}{\boldsymbol{w}^t \boldsymbol{w}}$.
- Optimal projection $\boldsymbol{w}$ is the eigenvector of $X^t X$ with largest eigenvalue (compare handout on spectral matrix norm).
- We assumed $\sum_i \boldsymbol{x}_i = \mathbf{0}$, i.e. the columns of X sum to zero.
  $\rightsquigarrow$ compute SVD of "centered" matrix X_c
  $\rightsquigarrow$ column vectors in W are eigenvectors of $X_c^t X_c$
  $\rightsquigarrow$ they are the principal components.

Eigen-faces [Turk and Pentland, 1991]

- d = number of pixels
- Each $\boldsymbol{x}_i \in \mathbb{R}^d$ is a face image
- x_{ij} = intensity of the j-th pixel in image i

$$\boldsymbol{x}_i \approx WW^t\boldsymbol{x}_i = W\boldsymbol{z}_i$$

$$(X^t)_{d\times n} \approx W_{d\times k} \quad (Z^t)_{k\times n}$$

( $\cdots$ ) $\approx$ ()

$$\begin{bmatrix} | & & | \\ \boldsymbol{z}_1 & \ldots & \boldsymbol{z}_n \\ | & & | \end{bmatrix}$$

Conceptual: We can lean something about the structure of face images.

Computational: Can use $\boldsymbol{z}_i$ for efficient nearest-neighbor classification: Much faster when $k \ll d$.

Information retrieval: Latent Semantic Analysis [Deerwater, 1990]

- d = number of words in the vocabulary, say 10000.
- Each $\boldsymbol{x}_i \in \mathbb{R}^d$ is a vector of word counts
- x_{ij} = frequency of word j in document i

$$
\begin{array}{ccccc}
(X^t)_{d\times n} & \approx & W_{d\times k} & & (Z^t)_{k\times n} \\
\left[\begin{array}{cccc}
\text{stocks:} & 2 & \ldots\ldots & 0 \\
\text{chairman:} & 4 & \ldots\ldots & 1 \\
\text{the:} & 8 & \ldots\ldots & 7 \\
\ldots & \vdots & \ldots\ldots & \vdots \\
\text{wins:} & 0 & \ldots\ldots & 2 \\
\text{game:} & 1 & \ldots\ldots & 3
\end{array}\right]
& \approx &
\left[\begin{array}{ccc}
0.4 & \ldots & -0.001 \\
0.8 & \ldots & 0.03 \\
0.01 & \ldots & 0.04 \\
\vdots & \ldots & \vdots \\
0.002 & \ldots & 2.3 \\
0.003 & \ldots & 1.9
\end{array}\right]
& &
\left[\begin{array}{ccc}
| & & | \\
\boldsymbol{z}_1 & \ldots & \boldsymbol{z}_n \\
| & & |
\end{array}\right]
\end{array}
$$

How to measure similarity between two documents? Dot products $\boldsymbol{x}_i^t \boldsymbol{x}_j$

In such high-dimensional spaces most pairs of vectors are almost orthogonal $\rightsquigarrow$ scalar products tend to be "noisy".

If $k \ll d$, $\boldsymbol{z}_i^t \boldsymbol{z}_j$ is probably a better similarity measure than $\boldsymbol{x}_i^t \boldsymbol{x}_j$.

Appendix Chapters 1/2

The Gershgorin circle theorem

Gershgorin circle theorem

Every eigenvalue of $A_{n\times n}$ is in one or more of n circles in the complex plane. Each circle is centered at a diagonal entry a_{ii},
the radius is $r_i = \sum_{j\neq i} |a_{ij}| \rightsquigarrow$ "Gershgorin disk" $D(a_{ii}, r_i)$.

Proof: $A\boldsymbol{v} = \lambda \boldsymbol{v}$, assume i is the index for which $|v_i| \geq |v_j|, \ \forall j \neq i$

$$(A\boldsymbol{v})_i = \lambda v_i \quad \Leftrightarrow \quad \sum_j a_{ij} v_j = \lambda v_i$$

$$(\lambda - a_{ii}) v_i = \sum_{j\neq i} a_{ij} v_j$$

$$|\lambda - a_{ii}||v_i| = |\sum_{j\neq i} a_{ij} v_j|$$

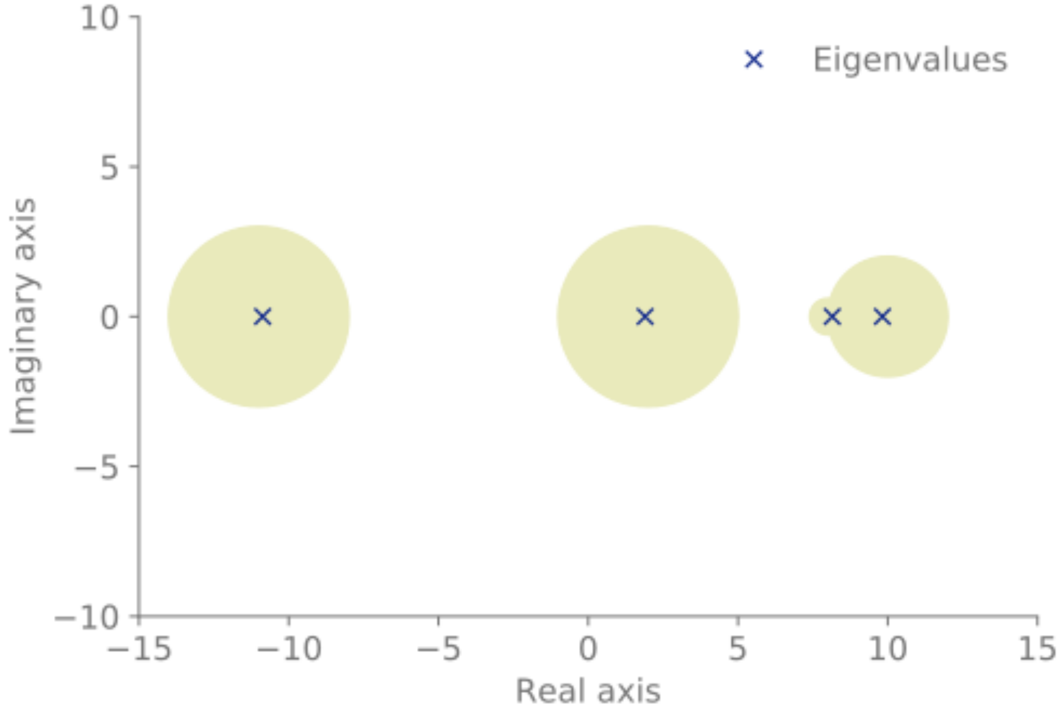
$\rightsquigarrow |\sum_{j\neq i} a_{ij} v_j| \leq \sum_{j\neq i} |a_{ij}||v_j| \leq \sum_{j\neq i} |a_{ij}||v_i| = r_i |v_i|$

$\rightsquigarrow |\lambda - a_{ii}||v_i| \leq r_i |v_i| \ \rightsquigarrow \ |\lambda - a_{ii}| \leq r_i$.

Applied to A^t: λ_i must also lie within circles corresponding to the columns of A.

Example

$$A = \begin{bmatrix} 10 & -1 & 0 & 1 \\ 0.2 & 8 & 0.2 & 0.2 \\ 1 & 1 & 2 & 1 \\ -1 & -1 & -1 & -11 \end{bmatrix}$$



For every row, a_{ii} is the center for the disc with radius $\sum_{j \neq i} |a_{ij}| = r_i$.
Discs: $D(10, 2)$, $D(8, 0.6)$, $D(2, 3)$, $D(-11, 3)$.
Can improve the accuracy of last two discs by applying the formula to the columns: $D(2, 1.2)$ and $D(-11, 2.2)$. True eigenvalues are $9.8218, 8.1478, 1.8995, -10.86$.

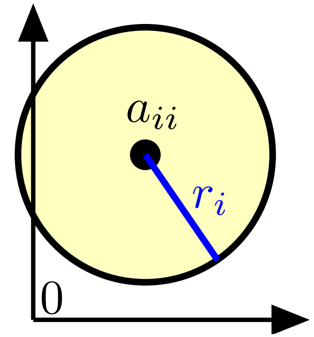
Note that A^t is diagonal dominant: $|a_{ii}| > \sum_{j \neq i} |a_{ji}| \rightsquigarrow$ most of the matrix is in the diagonal $\rightsquigarrow$ explains why the eigenvalues are so close to the centers.

Gershgorin circle theorem and diagonal dominance

A diagonal dominant matrix (i.e. $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$) is **non-singular**.

$\lambda \in \mathbb{C}$ is in at least one of the Gershgorin discs $D(a_{ii}, r_i)$ in the complex plane, but none of these discs contains 0:

$|a_{ii}| - r_i = |a_{ii}| - \sum_{j \neq i} |a_{ij}| > 0$, so each disc center a_{ii} is further away from 0 than the disc radius, and the point $\lambda = 0$ can't belong to any circle.



A **symmetric** diagonal dominant matrix that has **positive** values on its diagonal (i.e. $a_{ii} > \sum_{j \neq i} |a_{ij}|$) is **positive definite.**

Eigenvalues of symmetric matrices are real.
$\lambda \in \mathbb{R}$ is in at least one of the intervals $[a_{ii} - r_i, a_{ii} + r_i]$, but all intervals contain only positive numbers: $a_{ii} - r_i = a_{ii} - \sum_{j \neq i} |a_{ij}| > 0$.

Consequences: Jacobi iterations

- Assume that all diagonal entries of A are nonzero.
- Write $A = D + L + U$

  where $D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$ and $L+U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}$
- So $A\boldsymbol{x} = \boldsymbol{b} \quad \rightsquigarrow \quad (L + D + U)\boldsymbol{x} = \boldsymbol{b}$.
- Define $J = D^{-1}(L + U)$ as the **iteration matrix.**
- The solution is then obtained iteratively via

$$\boldsymbol{x}_{(i+1)} = -J\boldsymbol{x}_{(i)} + D^{-1}\boldsymbol{b}.$$

- Error $\boldsymbol{\epsilon}_{(i+1)} = -J\boldsymbol{\epsilon}_{(i)} = \cdots = (-1)^{i+1}J^{i+1}\boldsymbol{\epsilon}_{(0)}$.
- Arrange eigenvalues of J in diagonal matrix Λ.

Consequences: Jacobi iterations

If all the eigenvalues of J have magnitude < 1,
then $\Lambda^n \to 0$ and consequently $J^n \to 0 \rightsquigarrow$ convergence.

A diagonally dominant $\rightsquigarrow$ Jacobi method converges.

Assume rows of A are rescaled such that diagonal entries are all 1.
If $A = L + I + U$ is diagonal dominant, i.e. $1 \geq$ row sums of abs$(L + U)$,
then $L \pm \lambda I + U$ is also diagonally dominant if $|\lambda| \geq 1$,
because $|\lambda| \geq 1 \geq$ row sums of abs$(L + U)$.

Let λ be an eigenvalue of J.

$$\Rightarrow \quad det(J - \lambda I) = det(L + U - \lambda I) = 0.$$

But if $|\lambda| \geq 1$, then $L + U - \lambda I$ is diagonal dominant as well, so it is non-singular and $det = 0$ is not possible. So $|\lambda| < 1$.