A short linear algebra review

MATH 5329
Department of mathematics, UTA

Optional reading: Ch2 of Rencher and Schaalje (2008). Linear Models in Statistics, 2nd Edition. Wiley

Vector norm

- Vector norm: a norm of a vector is a informal measure of the length of the vector.

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

- L1 norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- L2 norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$
- L-infinity norm: $\|x\|_\infty = \max_i |x_i|.$

Quadratic form

- A is a square matrix and y is a vector
  - quadratic form: $\mathbf{y}'\mathbf{A}\mathbf{y} = \sum_i a_{ii}y_i^2 + \sum_{i \neq j} a_{ij}y_iy_j$
  - bilinear form: $\mathbf{x}'\mathbf{A}\mathbf{y} = \sum_{ij} a_{ij}x_iy_j$
- Example

$$\begin{pmatrix} 2 & 3 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 4 & 5 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix} = 2\cdot 2^2 + 5\cdot 3^2 + 2\cdot 2\cdot 3 + 4\cdot 3\cdot 2$$

- The matrix of a quadratic form can be chosen to be symmetric.

$$y'Ay = (y'Ay)' = y'A'y = y'(\tfrac{1}{2}A + \tfrac{1}{2}A')y$$

Rank of matrix

- A set of vectors $a_1, a_2, \ldots, a_n$ is linearly dependent if scalars $c_1, c_2, \ldots, c_n$ (not all zero) can be found such that

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}.$$

- If no such coefficients exist, then the set of vectors is linearly independent.
- Columns of **A** are linearly independent if **A**c=0 implies c=0.
- rank(**A**) = # of linearly independent columns (rows) of **A**
- **A** is full-rank if and only if rank(**A**) = min(# columns, # rows)

Positive (semi)definite

- Positive definite matrix: the symmetric matrix $\mathbf{A}$ has the property $y'\mathbf{A}y > 0$ for any $y \neq 0$.

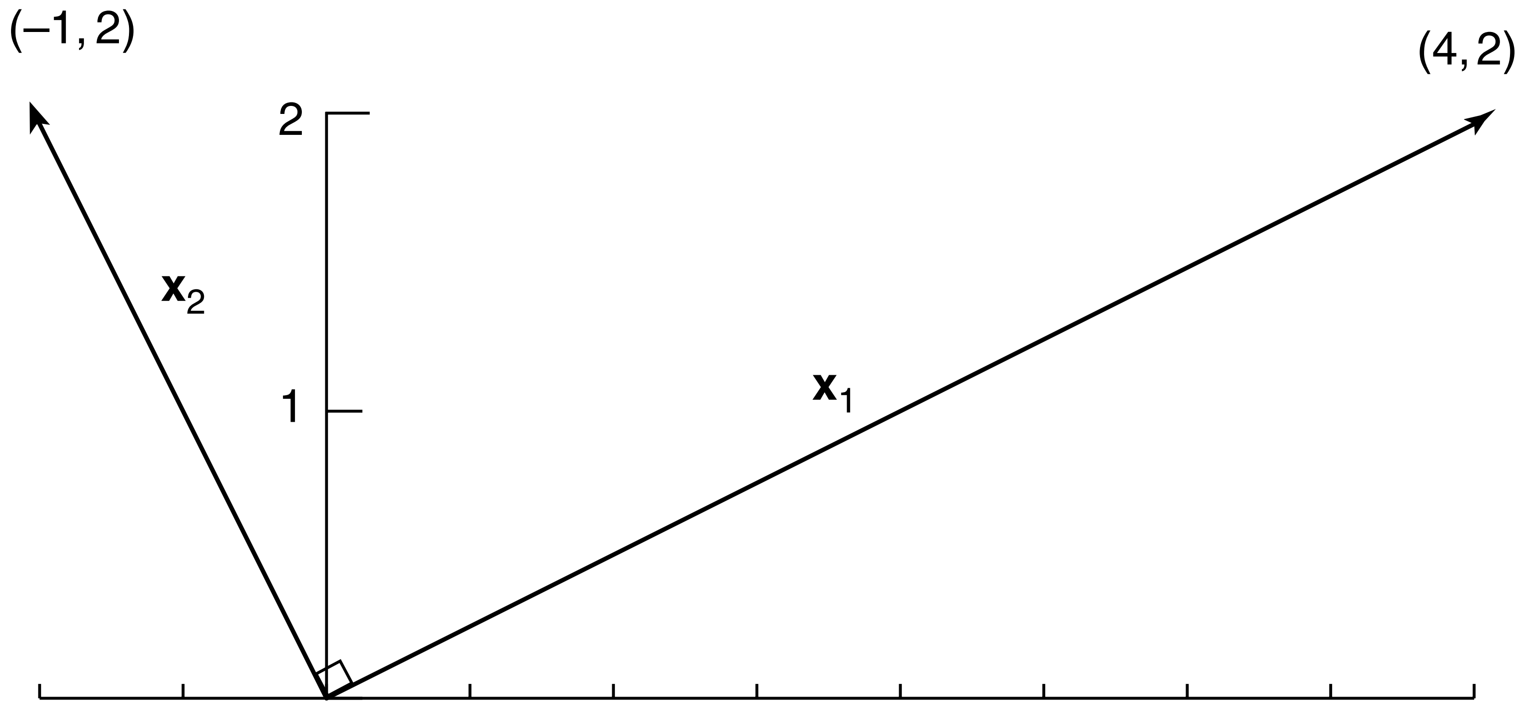
- Positive semidefinite matrix: the symmetric matrix $\mathbf{A}$ has the property $y'\mathbf{A}y \geq 0$ for any $y \neq 0$.

- **Theorem 2.6d.** Let $\mathbf{B}$ be an $n \times p$ matrix.

  (i) If $\text{rank}(\mathbf{B}) = p$, then $\mathbf{B}'\mathbf{B}$ is positive definite.
  (ii) If $\text{rank}(\mathbf{B}) < p$, then $\mathbf{B}'\mathbf{B}$ is positive semidefinite.

Orthogonal vectors

- Two vectors a and b are said to be orthogonal if a dot product of a and b is 0.



- (-1,2)x(4,2)′ = -4 + 4 = 0.

- Orthogonal vectors are linearly independent.

Orthonormal vectors

- If a vector's L_2 norm (Euclidian norm) is 1, then we say the vector is normalized.

$$||b||_2 = \sqrt{b'b} = 1$$

- A set of vectors $c_1, c_2, \ldots, c_p$ that are normalized and mutually orthogonal is said to be an orthonormal set of vectors.

  - e.g. {(0 0 1)′, (0 1 0)′, (1 0 0)′}

Orthogonal matrix

- A square matrix **C** whose column vectors are orthonormal is said to be an orthogonal matrix.

- Row vectors in **C** are orthonormal too.

- An orthogonal matrix **C** is nonsingular and $\mathbf{C}^{-1}=\mathbf{C}'$.

- Multiplication of a vector by an orthogonal matrix has the effect of rotating axes. i.e., an orthogonal is a linear transformation that does not changes the length of vectors.

Eigenvector and eigenvalue

- For a square matrix A, a scalar λ is an eigenvalue of **A** and a nonzero vector x is an eigenvector of **A** if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ <=> $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$

- Eigenvalues can be found by solving $|\mathbf{A} - \lambda\mathbf{I}| = 0$ (characteristic equation).

  - **A**-λ**I** is singular.

- Eigenvectors are not unique. So, an eigenvector x is typically scaled to normalized.

Spectral decomposition (eigen-decomposition)

- **Theorem 2.12d.** If $\mathbf{A}$ is an $n \times n$ symmetric matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and normalized eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, then $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{CDC}' \iff \mathbf{D} = \mathbf{C}'\mathbf{AC} \tag{2.103}$$

$$= \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i', \tag{2.104}$$

where $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{C}$ is the orthogonal matrix $\mathbf{C} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. The result in either (2.103) or (2.104) is often called the *spectral decomposition* of $\mathbf{A}$.

- Matrix power using spectral decomposition
  - $\mathbf{A}^2 = \mathbf{CDC}'\mathbf{CDC}' = \mathbf{CD}^2\mathbf{C}'$
  - $\mathbf{A}^k = \mathbf{CDC}' = \mathbf{CD}^k\mathbf{C}'$

❖ **Theorem 2.12e.** If $\mathbf{A}$ is any $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then

(i)
$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i. \tag{2.107}$$

(ii)
$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i. \tag{2.108}$$

□

❖ We can prove these using the spectral decomposition.

Theorem 2.12f. Let $\mathbf{A}$ be $n \times n$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.

(i) If $\mathbf{A}$ is positive definite, then $\lambda_i > 0$ for $i = 1, 2, \ldots, n$.
(ii) If $\mathbf{A}$ is positive semidefinite, then $\lambda_i \geq 0$ for $i = 1, 2, \ldots, n$. The number of eigenvalues λ_i for which $\lambda_i > 0$ is the rank of $\mathbf{A}$.

❖ sketch of pf) $\mathbf{x}_i$ is an eigenvector corresponding to λ_i. Hence, $\mathbf{x}_i'\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i'\mathbf{x}_i$ >(or ≥) 0 where $\mathbf{x}_i'\mathbf{x}_i \geq 0$. Thus, λ_i >(or ≥) 0. The proof for 2nd part of (ii) is lengthy.

Square root of matrix

- For a symmetric matrix **A** is positive (semi)definite,

$$A^{1/2} = CD^{1/2}C'$$

$$D^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})$$

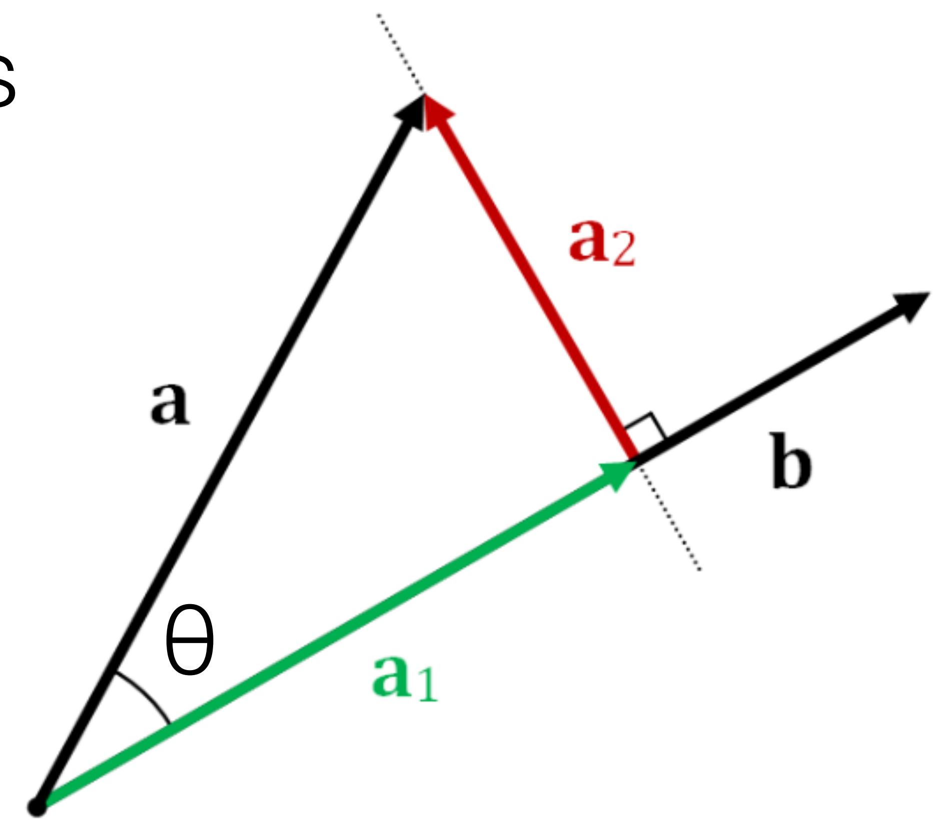
- Idempotent: a square matrix **A** is said to be idempotent if $\mathbf{A}^2 = \mathbf{A}$.

- Idempotent matrix is a projection matrix (not necessarily orthogonal projection).

(orthogonal) Projection

- Projection is special linear transformation which is extremely useful in linear models

- Projection of a onto b
  - $a_1 = b(b'b)^{-1}b'a$, $a_2 = a - a_1$
  - a_2 and a_1 are orthogonal.
  - $\mathbf{P_b} = b(b'b)^{-1}b'$ is a projection matrix. (projection onto a subspace spanned by b). Then, $\mathbf{P_b}a_1 = a_1$, $\mathbf{P_b}a_2 = 0$.



(orthogonal) Projection matrix

- Projection matrix $\mathbf{P}$ is symmetric and idempotent.
    - $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P} = \mathbf{P}'$.
- A projection matrix $\mathbf{P}$ that projects a vector onto subspace V is denoted by $\mathbf{P}_V$.
- Orthogonal complement of V is denoted by $V^\perp$.
    - All vectors in $V^\perp$ are orthogonal to all vectors in V.
    - If V is in $\mathbf{R}^n$, $V+V^\perp=\mathbf{R}^n$, where "+" denotes the disjoint union.

Subspace and orthogonal complement

- Example

    - $\mathrm{V} = \{(x_1, x_2, 0)' \text{ for any } x_1, x_2 \in \mathrm{R}\}$.

      Then, $\mathrm{V}^{\perp} = \{(0,0,x_3)' \text{ for any } x_3 \in \mathrm{R}\}$.

      Also, $\mathrm{V} + \mathrm{V}^{\perp} = \mathrm{R}^3$.

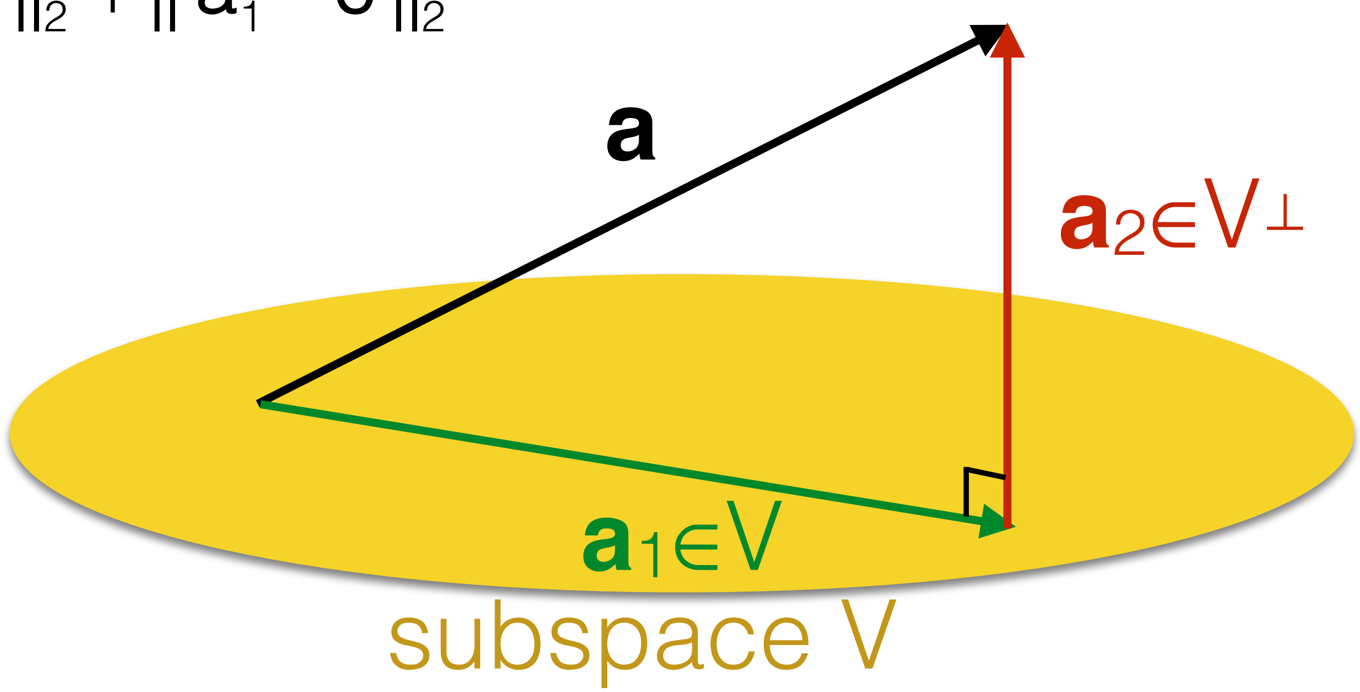
- Proposition 1. For a full rank matrix $\mathbf{X} \in \mathbf{R}^{n \mathrm{x} p}$ $(p \leq n)$, $\mathbf{P}_V = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projection matrix, and V is the column space of $\mathbf{X}$.

    - pf) $\mathbf{P}_V^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_V$, and $\mathbf{P}_V' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_V$. We skip the proof of 2nd part.

Geometric interpretation

- Proposition 2. $\mathbf{a}_1 = \mathbf{P}_V\mathbf{a}$ is the closest vector in V to $\mathbf{a}$. In other words, $\min_{\mathbf{c}} \|\mathbf{a}-\mathbf{c}\|_2 = \|\mathbf{a}_2\|_2$ for all $\mathbf{c}$ in V.

  - $$\begin{aligned} & \| a - c \|_2^2 \\ = & \| a - a_1 + a_1 - c \|_2^2 \\ = & \| a - a_1 \|_2^2 + \| a_1 - c \|_2^2 + 2(a - a_1)'(a_1 - c) \\ = & \| a - a_1 \|_2^2 + \| a_1 - c \|_2^2 \end{aligned}$$



Relation to linear regression

- Let $y \in \mathbf{R}^n$ be a vector of outputs, $\mathbf{X} \in \mathbf{R}^{n \times p}$ be a matrix of inputs, and $\beta \in \mathbf{R}^p$ be a vector of regression coefficients.
- Model: $y = \mathbf{X}\beta + \varepsilon$. Typically, there is no β that satisfy this equation. So, we want to approximate β so as to $\|y - \mathbf{X}\beta\|_2$ is minimized.
- $\mathbf{X}\beta$ is in the column space (V) of $\mathbf{X}$. So, $\mathbf{X}\beta$ that minimizes $\|y - \mathbf{X}\beta\|_2$ is $\mathbf{X}\beta = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \in V$.
- $y - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \in V^{\perp}$ is residual.

Covariance matrix

- $x = (x_1, x_2, \ldots, x_p)'$ is a p-dimensional random vector.
- Let $\sigma_{ij} = \mathrm{cov}(x_i, x_j)$. The covariance matrix of x is

$$\Sigma = \mathrm{cov}(\mathbf{x}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{21} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \cdots & \cdots & \sigma_{pp} \end{pmatrix}$$

- For $a \in \mathbf{R}^p$, $a'x = (a_1x_1 + \ldots + a_px_p)'$ is an affine transformation.
- $\mathrm{var}(a'x) = a'\mathbf{\Sigma}a \geq 0$. i.e., Covariance matrix is symmetric positive semidefinite.

- Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ be eigenvalues of $\mathbf{\Sigma}$, and $c_1, c_2, \ldots, c_p$ be their associated (normalized) eigenvectors.
  - $\lambda_1 = \max a'\mathbf{\Sigma}a$ subject to $||a||=1$.
  - c_1 is a direction that maximizes the variance of the affine transformation: $\mathrm{var}(c_1'x) = c_1'\mathbf{\Sigma}c_1 = \lambda_1 c_1'c_1 = \lambda_1$
  - $\lambda_2 = \max a'\mathbf{\Sigma}a$ subject to $||a||=1$ and $a \perp c_1$.
    - c_2 is a direction orthogonal to c_1, and $\mathrm{var}(c_2'x)$ is the 2nd largest.
  - In general, c_i is a direction orthogonal to $(c_1, \ldots, c_{i-1})$ while $\mathrm{var}(c_i'x)$ is the i-th largest.

Singular value decomposition (SVD)

- Let $\mathbf{X} \in \mathbf{R}^{n\times p}$ $(n \geq p)$ be a matrix. Then, there exists an orthogonal matrix $\mathbf{V} \in \mathbf{R}^{p\times p}$, a matrix $\mathbf{U} \in \mathbf{R}^{n\times p}$ whose columns are orthonormal, and a diagonal matrix $D = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbf{R}^{p\times p}$ such that $\mathbf{X} = \mathbf{UDV}'$ (If $\mathbf{X}$ is symmetric ($\in \mathbf{R}^{p\times p}$), $\mathbf{U} = \mathbf{V}$).

  - $\mathbf{XX}' = \mathbf{UDV}'\mathbf{VDU} = \mathbf{UD}^2\mathbf{U}'$ and $\mathbf{X}'\mathbf{X} = \mathbf{VDU}'\mathbf{UDV}' = \mathbf{VD}^2\mathbf{V}'$.

- $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$ are called singular values, and $\sigma_i^2 = \lambda_i$ where λ_i is a eigenvalue of $\mathbf{X}'\mathbf{X}$.

Low rank approximation

- Recall **X** is n by p matrix. By SVD,

$$X = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i v_i{}' \text{ where } \mathbf{u}_i \text{ and } v_i \text{ are i}th \text{ column of } \mathbf{U} \text{ and V, respectively.}$$

- Reconstruct **X** using first r singular values:

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i{}' \ (r << p) \Rightarrow A = \underset{rank(A)=r}{\operatorname{argmin}} \| X - A \|_F^2$$

- This is the best rank r approximation of matrix **X**, and can be used to compress **X**.

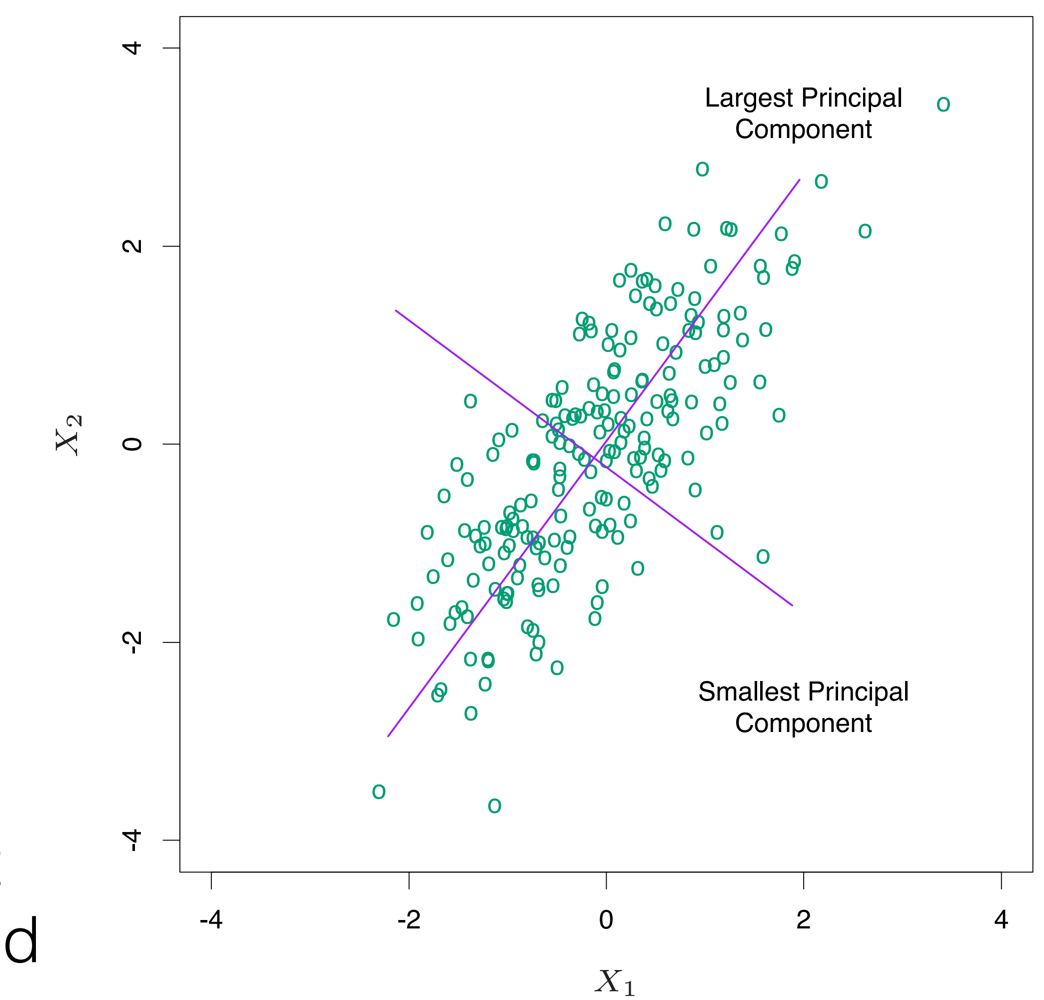
SVD on feature matrix **X**

- SVD on $\mathbf{X}=\mathbf{UDV}'$ leads to principal component analysis (PCA).

- Assuming a feature matrix **X** is centered, $\mathbf{X}'\mathbf{X}$ / (n-1) is a sample covariance matrix.

- Singular values are squared root of eigenvalues of the sample covariance matrix. i.e., v_1 is a direction with the largest variance of an affine transformation.

- v_1 is first column of **V** (principal component direction), u_1 is first column of **U** (normalized principal component), and $z_1=\mathbf{X}v_1$ is first principal component.



Remarks on PCA

- Principal component ($z_i = \mathbf{X}v_i$) is the linear combination of the variables in **X**, and the 1st principal component has largest variance.
- The second principal component has 2nd largest variance, and it is uncorrelated with the 1st, and so on… -> principal components are uncorrelated.
- The principal components (or SVD) provides the best rank r approximation of a matrix.
- $$A = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i' \approx \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i' \ (r << p).$$

SVD for image processing

- Suppose **A** is an image matrix. Each entry of **A** represents pixels in the image.

- Based on the SVD on **A**, one can select first few terms of the basis representation to compress the image without losing the quality much.

Orignal

first 5 terms

first 10 terms

first 50 terms

Take home messages

- The PCA extracts the important information from the data and to express this information as a set of new uncorrelated (orthogonal) variables.

- The PCA gives the best low rank approximation of a matrix. i.e., it compresses the data size keeping only important information.

Vector and matrix calculus

- Let u=f(**x**):**R**^{p}->**R**. We define the partial derivative as

$$\frac{\partial u}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ \vdots \\ \frac{\partial u}{\partial x_p} \end{pmatrix}.$$

- **Theorem 2.14a.** Let $u = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$, where $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$ is a vector of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}. \tag{2.112}$$

- **Theorem 2.14b.** Let $u = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric matrix of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}. \tag{2.113}$$