Statistics for Data Science: Week 8

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Regression

In the beginning we distinguished between:

1. **Marginal Inference.** Here $(Y_1, ..., Y_n)^\top$ has i.i.d. entries each from the same distribution $F(y;\theta)$ with the same parameter θ.
    - In other words, all observations were obtained under identical experimental conditions, and thus depend in the same way on the same unknown θ.
2. **Regression.** Here $(Y_1, ..., Y_n)^\top$ has independent entries, each with distribution $F(y;\theta_i)$ of the same family but with different parameters.
    - Each observation was generated under slightly different experimental conditions. They depend in a similar way on different θ_i.
    - These θ_i correspond to different experimental conditions, say x_i.
    - Each x_i is called a covariate/feature, and is an input that the experimenter can vary. They are known. The index i reminds us that it corresponds to the ith observation Y_i.
    - Usually θ_i is postulated to have a special relationship to x_i, for example $\theta_i = \exp\{\alpha + \beta x_i\}$, for (α, β) uknown parameters.
    - The point here is to understand the effect of varying the covariate/feature on the distribution of the observable.

Statistical model for:

$$\boxed{Y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}}$$

Aim: understand the effect of x on the distribution of random variable Y

General formulation[1]:

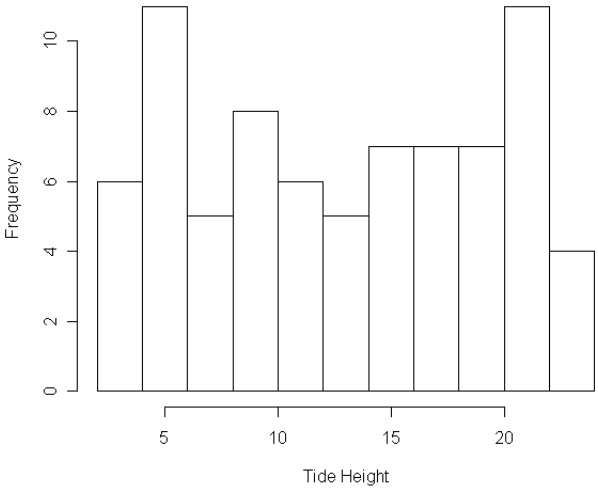
$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, \ldots, n.$$

Statistical Problem: Estimate (learn) $g(\cdot)$ from data $\{(x_i, Y_i)\}_{i=1}^n$. Use for:

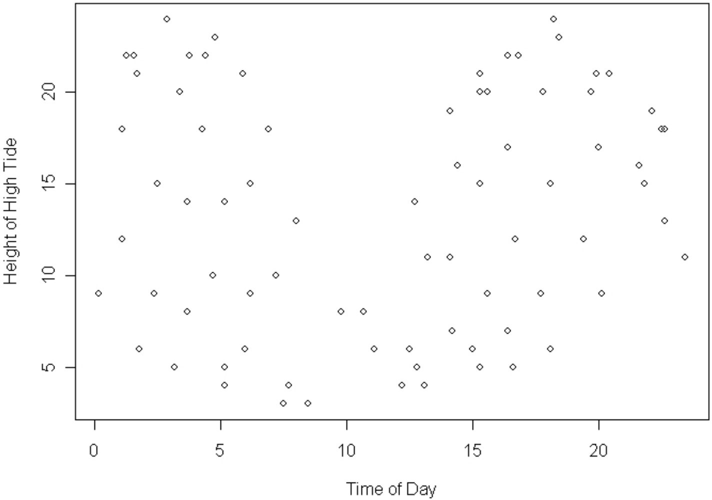
- Inference
- Prediction
- Data compression (parsimonious representations)

[1]Sometimes we write $Y_i | x_i \overset{\text{independent}}{\sim} \text{Distribution}\{g(x_i) = \theta_i\}$ to highlight that the distribution of Y depends on x, but without meaning that (X, Y) are jointly random; such an assumption is unnecessary (e.g., in a designed experiment we choose values for x).

Example: How to model the height of Honolulu tides throughout the day - Histogram



Example: Height of Honolulu tides as function of the time of day



A bewildering variety of models can be captured by the general specificaiton

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

x_i can be:

- continuous, discrete, categorical, vector ...
- arrive randomly, or be chosen by experimenter, or both
- however x arises, we treat it as constant in the analysis

Distribution can be:

- Gaussian, Laplace, Bernoulli, Poisson, gamma, general exponential family, ...

Function $g(\cdot)$ can be:

- $g(x) = \beta_0 + \beta_1 x$, $g(x) = \sum_{k=-K}^{K} \beta_k e^{-ikx}$, cubic spline, neural net...

Table: A coarse classification of regression models we will consider

Distribution / Function g	$g(\boldsymbol{x}_i^\top) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	g nonparametric
Gaussian	Linear Regression	Smoothing
Exponential Family	GLM	GAM

GLM: Generalized Linear Model and GAM: Generalized Additive Model

We start with a very standard model: Linear Regression with $Y|x$ being Gaussian.

- $Y, x \in \mathbb{R}$, $g(x) = \beta_0 + \beta_1 x$

$$Y \mid x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$
$$\Updownarrow$$
$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

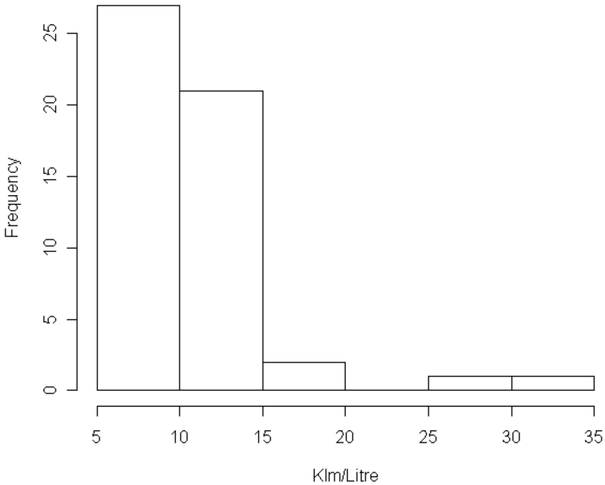
The second version is useful for mathematical work, but is puzzling statistically, since we don't observe ϵ.

- Also, covariate could be vector ($Y, \beta_0 \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^p$):

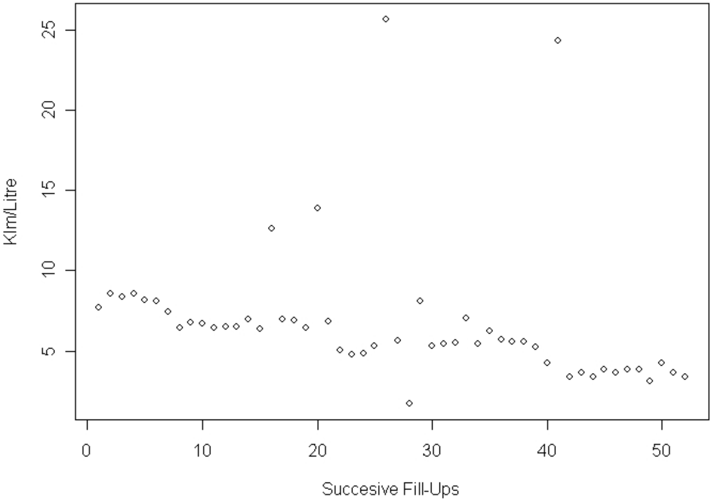
$$Y \mid x \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$$
$$\Updownarrow$$
$$Y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Example: How to model my van's consumption of gas

Example: Histogram of consumption of gas (km/L)



Example: Gas consumption as function of successive fill-ups



Start from Gaussian linear regression then gradually generalise . . .

Obviously: important features of Gaussian linear model are

- Gaussian distribution
- Linearity

These two combine well and give geometric insights to solve the estimation problem. Thus we need to revise some probabilistic linear algebra. . .

- Subpsaces and projection matrices
- Multivariate Gaussian Distribution
- Optimal dimension reduction
- Random quadratic forms

Linear Algebra Intermezzo

Linear Subspaces, Orthogonal Projections, Gaussian Vectors

If $\boldsymbol{Q}$ is an $n \times p$ real matrix, we define the column space (or *range*) of $\boldsymbol{Q}$ to be the set spanned by its columns:

$$\mathcal{M}(\boldsymbol{Q}) = \{\boldsymbol{y} \in \mathbb{R}^n : \ \exists \boldsymbol{\beta} \in \mathbb{R}^p, \ \boldsymbol{y} = \boldsymbol{Q}\boldsymbol{\beta}\}.$$

- Recall that $\mathcal{M}(\boldsymbol{Q})$ is a subspace of $\mathbb{R}^n$.
- The columns of $\boldsymbol{Q}$ provide a coordinate system for the subspace $\mathcal{M}(\boldsymbol{Q})$
- If $\boldsymbol{Q}$ is of full column rank (p), then the coordinates $\boldsymbol{\beta}$ corresponding to a $\boldsymbol{y} \in \mathcal{M}(\boldsymbol{Q})$ are unique.
- Allows interpretation of system of linear equations

$$\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{y}.$$

[existence of solution $\leftrightarrow$ is $\boldsymbol{y}$ an element of $\mathcal{M}(\boldsymbol{Q})$?]
[uniqueness of solution $\leftrightarrow$ is there a unique coordinate vector $\boldsymbol{\beta}$?]

Two further important subspaces associated with a real $n \times p$ matrix $\boldsymbol{Q}$:

- the null space (or *kernel*), $\ker(\boldsymbol{Q})$, of $\boldsymbol{Q}$ is the subspace defined as

$$\ker(\boldsymbol{Q}) = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{Q}\boldsymbol{x} = 0\};$$

- the orthogonal complement of $\mathcal{M}(\boldsymbol{Q})$, $\mathcal{M}^\perp(\boldsymbol{Q})$, is the subspace defined as

$$\begin{aligned} \mathcal{M}^\perp(\boldsymbol{Q}) &= \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y}^\top \boldsymbol{Q}\boldsymbol{x} = 0, \ \forall \boldsymbol{x} \in \mathbb{R}^p\} \\ &= \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y}^\top \boldsymbol{v} = 0, \ \forall \boldsymbol{v} \in \mathcal{M}(\boldsymbol{Q})\}. \end{aligned}$$

The orthogonal complement may be defined for arbitrary subspaces by using the second equality.

Theorem (Spectral Theorem)

A $p \times p$ matrix $\boldsymbol{Q}$ is symmetric if and only if there exists a $p \times p$ orthogonal matrix[a] *$\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that*

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top.$$

In particular:

1. *the columns of $\boldsymbol{U} = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_p)$ are eigenvectors of $\boldsymbol{Q}$, i.e. there exist λ_j such that*
$$\boldsymbol{Q}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \qquad j = 1, \ldots, p;$$
2. *the entries of $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ are the corresponding eigenvalues of $\boldsymbol{Q}$, which are real; and*
3. *the rank of $\boldsymbol{Q}$ is the number of non-zero eigenvalues.*

[a] *A matrix is orthogonal if $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = I_p$*

Note: if the eigenvalues are distinct, the eigenvectors are unique (up to changes in signs).

Theorem (Singular Value Decomposition)

Any $n \times p$ real matrix can be factorised as

$$\underset{n\times p}{\boldsymbol{Q}} = \underset{n\times n}{\boldsymbol{U}} \underset{n\times p}{\Sigma} \underset{p\times p}{\boldsymbol{V}}^\top,$$

where $\boldsymbol{U}$ and $\boldsymbol{V}^\top$ are orthogonal with columns called **left singular vectors** *and* **right singular vectors***, respectively, and Σ is diagonal with real entries called* **singular values***.*

1. The left singular vectors are eigenvectors of $\boldsymbol{Q}\boldsymbol{Q}^\top$.[2]
2. The right singular vectors are eigenvectors of $\boldsymbol{Q}^\top\boldsymbol{Q}$.
3. The squares of the singular values are eigenvalues of both $\boldsymbol{Q}\boldsymbol{Q}^\top$ and $\boldsymbol{Q}^\top\boldsymbol{Q}$.
4. The left singular vectors corresponding to non-zero singular values form an orthonormal basis for $\mathcal{M}(\boldsymbol{Q})$.
5. The left singular vectors corresponding to zero singular values form an orthonormal basis for $\mathcal{M}^\perp(\boldsymbol{Q})$.

[2] *hint: compute $\boldsymbol{Q}\boldsymbol{Q}^\top U_i = \lambda_i^2 U_i$ for all $i \leq p$. And similarly with $\boldsymbol{Q}\boldsymbol{Q}^\top V_i = \lambda_i^2 V_i$*

A matrix $\boldsymbol{Q}$ is called idempotent if $\boldsymbol{Q}^2 = \boldsymbol{Q}$.

An orthogonal projection (henceforth projection) onto a subspace $\mathcal{V}$ is a symmetric idempotent matrix $\boldsymbol{H}$ such that $\mathcal{M}(\boldsymbol{H}) = \mathcal{V}$, i.e. the column space is generated by the subspace $\mathcal{V}$.

Proposition

The only possible eigenvalues of a projection matrix are 0 and 1.

Proposition

Let $\mathcal{V}$ be a subspace and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then $\boldsymbol{I} - \boldsymbol{H}$ is the projection matrix onto $\mathcal{V}^{\perp}$.

Proof (∗).

We first prove that $\boldsymbol{I} - \boldsymbol{H}$ is a projection matrix (idempotent and symmetric).
$(\boldsymbol{I} - \boldsymbol{H})^{\top} = \boldsymbol{I} - \boldsymbol{H}^{\top} = \boldsymbol{I} - \boldsymbol{H}$ since $\boldsymbol{H}$ is symmetric and,
$(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I}^2 - 2\boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - \boldsymbol{H}$.

It remains to identify the column space of $\boldsymbol{I} - \boldsymbol{H}$. Let $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$ be the spectral decomposition of $\boldsymbol{H}$.

Then $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{U}\boldsymbol{U}^{\top} - \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top} = \boldsymbol{U}(\boldsymbol{I} - \boldsymbol{\Lambda})\boldsymbol{U}^{\top}$.

Hence the column space of $\boldsymbol{I} - \boldsymbol{H}$ is spanned by the eigenvectors of $\boldsymbol{H}$ corresponding to zero eigenvalues of $\boldsymbol{H}$, which coincides with $\mathcal{M}^{\perp}(\boldsymbol{H}) = \mathcal{V}^{\perp}$. □

Proposition

Let $\mathcal{V}$ be a subspace and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}$ for all $\boldsymbol{y} \in \mathcal{V}$.

Proposition

If $\boldsymbol{P}$ and $\boldsymbol{Q}$ are projection matrices onto a subspace $\mathcal{V}$, then $\boldsymbol{P} = \boldsymbol{Q}$.

Proposition

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are linearly independent[a] and are such that span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \mathcal{V}$, then the projection onto $\mathcal{V}$ can be represented as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

where $\boldsymbol{X}$ is a matrix with columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$.

[a] $\sum_{i \leq p} a_i \boldsymbol{x}_i = 0$ iff. $a_i = 0$, for all $i \leq p$

Proposition

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then

$$\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{v}\|, \qquad \forall \boldsymbol{v} \in \mathcal{V}.$$

Proof $(*)$.

Let $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $\boldsymbol{H}$, $\boldsymbol{U} = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_n)$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Letting $p = \dim(\mathcal{V})$, then

by assumption of $\boldsymbol{H}$

1. $\lambda_1 = \cdots = \lambda_p = 1$ and $\lambda_{p+1} = \cdots = \lambda_n = 0$, (*by definition of a projection matrix s19*)
2. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is an orthonormal basis of $\mathbb{R}^n$,
3. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ is an an orthonormal basis of $\mathcal{V}$.

Let's us it in the following computations

$$
\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{x}\|^2 &= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - (\boldsymbol{H}\boldsymbol{x})^\top \boldsymbol{u}_i)^2 \qquad \text{[orthonormal basis]} \\
&= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{u}_i)^2 \qquad [H \text{ is symmetric}] \\
&= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - \lambda_i \boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad [u\text{'s are eigenvectors of } H] \\
&= 0 + \sum_{i=p+1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad \text{[eigenvalues 0 or 1]} \\
&\leq \sum_{i=1}^{p} (\boldsymbol{x}^\top \boldsymbol{u}_i - \boldsymbol{v}^\top \boldsymbol{u}_i)^2 + \sum_{i=p+1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad \forall \boldsymbol{v} \in \mathcal{V} \\
&= \|\boldsymbol{x} - \boldsymbol{v}\|^2.
\end{aligned}
$$

□

Proposition

Let $\mathcal{V}_1 \subseteq \mathcal{V} \subseteq \mathbb{R}^n$ be two nested linear subspaces. If $\boldsymbol{H}_1$ is the projection onto $\mathcal{V}_1$ and $\boldsymbol{H}$ is the projection onto $\mathcal{V}$, then

$$\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1 = \boldsymbol{H}_1\boldsymbol{H}.$$

Proof (∗).

First we show that $\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1$, and then that $\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}_1$. For all $\boldsymbol{y} \in \mathbb{R}^n$ we have $\boldsymbol{H}_1\boldsymbol{y} \in \mathcal{V}_1$. But then $\boldsymbol{H}_1\boldsymbol{y} \in \mathcal{V}$, since $\mathcal{V}_1 \subseteq \mathcal{V}$.
Therefore $\boldsymbol{H}\boldsymbol{H}_1\boldsymbol{y} = \boldsymbol{H}_1\boldsymbol{y}$. We have shown that $(\boldsymbol{H}\boldsymbol{H}_1 - \boldsymbol{H}_1)\boldsymbol{y} = 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$, so that $\boldsymbol{H}\boldsymbol{H}_1 - \boldsymbol{H}_1 = 0$, as its kernel is all $\mathbb{R}^n$. Hence $\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1$.

To prove that $\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}_1$, note that symmetry of projection matrices and the first part of the proof give

$$\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}_1^\top\boldsymbol{H}^\top = (\boldsymbol{H}\boldsymbol{H}_1)^\top = (\boldsymbol{H}_1)^\top = \boldsymbol{H}_1 = \boldsymbol{H}\boldsymbol{H}_1.$$

□

Definition (Quadratic Form Definition)

A $p \times p$ real symmetric matrix $\mathbf{\Omega}$ is called **non-negative definite** (written $\mathbf{\Omega} \succeq 0$) if and only if $\boldsymbol{x}^\top \mathbf{\Omega} \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^p$. If $\boldsymbol{x}^\top \mathbf{\Omega} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^p \setminus \{0\}$, then we call $\mathbf{\Omega}$ **positive definite** (written $\mathbf{\Omega} \succ 0$).

Definition (Spectral Definition)

A $p \times p$ real symmetric matrix $\mathbf{\Omega}$ is called **non-negative definite** (written $\mathbf{\Omega} \succeq 0$) if and only the eigenvalues of $\mathbf{\Omega}$ are non-negative. If the eigenvalues of $\mathbf{\Omega}$ are strictly positive, then $\mathbf{\Omega}$ is called **positive definite** (written $\mathbf{\Omega} \succ 0$).

Lemma (Little exercise)

The two definitions are equivalent.

Proposition (Non-Negative and Covariance Matrices)

Let $\mathbf{\Omega}$ *be a real symmetric matrix.*

Then $\mathbf{\Omega}$ *is non-negative definite iff.* $\mathbf{\Omega}$ *is the covariance matrix of some random vector* $\boldsymbol{Y}$.

We want to find the subspace that explains the most a random vector $\boldsymbol{Y}$ in $\mathbb{R}^d$ with covariance matrix $\boldsymbol{\Omega}$.

- *Step $j = 1$:* Find direction $\boldsymbol{v}_1 \in \mathbb{S}^{d-1}$ such that the projection of $\boldsymbol{Y}$ onto $\boldsymbol{v}_1$ has maximal variance.
- *Steps $j = 2, 3, \ldots, d$:* Find direction $\boldsymbol{v}_j \perp \{\boldsymbol{v}_1, ..., \boldsymbol{v}_{j-1}\}$ such that projection of $\boldsymbol{Y}$ onto $\boldsymbol{v}_j$ has maximal variance.

- First, by Proposition s26, $\boldsymbol{\Omega}$ is symmetric, non-negative definite of size $d \times d$.

- *Step $j = 1$:* Maximise $\mathrm{var}(\boldsymbol{v}_1^\top \boldsymbol{Y}) = \boldsymbol{v}_1^\top \boldsymbol{\Omega} v_1$ over $\|\boldsymbol{v}_1\| = 1$

$$\boldsymbol{v}_1^\top \boldsymbol{\Omega} \boldsymbol{v}_1 = \boldsymbol{v}_1^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{v}_1 = \|\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top \boldsymbol{v}_1\|^2 = \sum_{i=1}^{d} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{v}_1)^2 \qquad \text{[change of basis]}$$

  Now $\sum_{i=1}^{d} (\boldsymbol{u}_i^\top \boldsymbol{v}_1)^2 = \|\boldsymbol{v}_1\|^2 = 1$ so we have a convex combination of $\{\lambda_j\}_{j=1}^{d}$,

$$\sum_{i=1}^{d} p_i \lambda_i, \qquad \sum_i p_i = 1, \quad p_i \geq 0, \quad i = 1, \dots, d.$$

  But $\lambda_1 \geq \lambda_i \geq 0$ so clearly this sum is maximised when $p_1 = 1$ and $p_j = 0$ $\forall j \neq 1$, i.e. $\boldsymbol{v}_1 = \pm \boldsymbol{u}_1$.

- *Steps $j = 2, 3, \dots, d$:* Iteratively, $\boldsymbol{v}_j = \pm \boldsymbol{u}_j$, i.e. principal components are eigenvectors of $\boldsymbol{\Omega}$.

Theorem (Optimal (Linear) Dimension Reduction Theorem)

Let $\boldsymbol{Y}$ be a mean-zero random variable in $\mathbb{R}^d$ with $d \times d$ covariance $\boldsymbol{\Omega}$. Let $\boldsymbol{H}$ be the projection matrix onto the span of the first k eigenvectors of $\boldsymbol{\Omega}$. Then

$$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}\|^2 \leq \mathbb{E}\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{Y}\|^2$$

for any $d \times d$ projection matrix $\boldsymbol{Q}$ or rank at most k.

Intuitively: if you want to approximate a mean-zero random variable taking values $\mathbb{R}^d$ by a random variable that ranges over a subspace of dimension at most $k \leq d$, the optimal choice is the projection of the random variable onto the space spanned by its first k principal components (eigenvectors of the covariance). "Optimal" is with respect to the mean squared error.

For the proof, use lemma below (follows immediately from spectral decomposition)

Lemma

$\boldsymbol{Q}$ is a rank k projection matrix iff. there exist orthonormal vectors $\{\boldsymbol{v}_j\}_{j=1}^k$ such that $\boldsymbol{Q} = \sum_{j=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$.

Proof of Optimal Linear Dimension Reduction ($*$).

Write $\boldsymbol{Q} = \sum_{j=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^\top$ for some orthonormal $\{\boldsymbol{v}_j\}_{j=1}^{k}$. Then

$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{QY}\|^2 =$

$$
\begin{aligned}
&= \mathbb{E}\left[\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{Q})^\top (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{Y}\right] = \mathbb{E}\left[\mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{YY}^\top (\boldsymbol{I} - \boldsymbol{Q})^\top\}\right] \\
&= \mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\mathbb{E}\left[\boldsymbol{YY}^\top\right](\boldsymbol{I} - \boldsymbol{Q})^\top\} = \mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{Q})^\top (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{\Omega}\} \\
&= \mathrm{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{\Omega}\} = \mathrm{tr}\{\boldsymbol{\Omega}\} - \mathrm{tr}\{\boldsymbol{Q\Omega}\} = \sum_{i=1}^{d} \lambda_i - \mathrm{tr}\left\{\sum_{j=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{\Omega}\right\} \\
&= \sum_{i=1}^{d} \lambda_i - \sum_{j=1}^{k} \mathrm{tr}\left\{\boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{\Omega}\right\} = \sum_{i=1}^{d} \lambda_i - \sum_{j=1}^{k} \boldsymbol{v}_i \boldsymbol{\Omega} \boldsymbol{v}_i^\top \\
&= \sum_{i=1}^{d} \lambda_i - \sum_{j=1}^{k} \mathrm{var}[\boldsymbol{v}_i^\top \boldsymbol{Y}]
\end{aligned}
$$

If we can minimise this expression over all $\{\boldsymbol{v}_j\}_{j=1}^{k}$ with $\boldsymbol{v}_i^\top \boldsymbol{v}_j = \mathbf{1}\{i = j\}$, then we're done. By PCA, this is done by choosing the top k eigenvectors of $\boldsymbol{\Omega}$. □

Recall that for any matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, we have $tr(\boldsymbol{ABC}) = tr(\boldsymbol{BCA}) = tr(\boldsymbol{CAB})$ under conditions (cf A3W8).

Corollary (Deterministic Version)

Let $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\} \subset \mathbb{R}^d$ be such that $\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_p = 0$, and let $\boldsymbol{X}$ be the matrix with columns $\{\boldsymbol{x}_i\}_{i=1}^p$. The best approximating k-hyperplane to the points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\}$ is given by the span of the first k eigenvectors of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$, i.e. if $\boldsymbol{H}$ is the projection onto this span, it holds that

$$\sum_{i=1}^{p} \|\boldsymbol{x}_i - \boldsymbol{H}\boldsymbol{x}_i\|^2 \leq \sum_{i=1}^{p} \|\boldsymbol{x}_i - \boldsymbol{Q}\boldsymbol{x}_i\|^2$$

for any $d \times d$ projection operator $\boldsymbol{Q}$ or rank at most k.

Proof.

Define the discrete random vector $\boldsymbol{Y}$ by $\mathbb{P}[\boldsymbol{Y} = \boldsymbol{x}_i] = 1/p$, and use optimal linear dimension reduction as stated earlier. □