

## Lecture 7 — Spectral methods

## 7.1 Linear algebra review

## 7.1.1 Eigenvalues and eigenvectors

**Definition 1.** A  $d \times d$  matrix  $\mathbf{M}$  has *eigenvalue*  $\lambda$  if there is a  $d$ -dimensional vector  $\mathbf{u} \neq \mathbf{0}$  for which  $\mathbf{M}\mathbf{u} = \lambda\mathbf{u}$ . This  $\mathbf{u}$  is the *eigenvector* corresponding to  $\lambda$ .

In other words, the linear transformation  $\mathbf{M}$  maps vector  $\mathbf{u}$  into the same direction. It is interesting that *any* linear transformation necessarily has directional fixed points of this kind. The following chain of implications helps in understanding this:

$$\begin{aligned} & \lambda \text{ is an eigenvalue of } \mathbf{M} \\ \Leftrightarrow & \text{ there exists } \mathbf{u} \neq \mathbf{0} \text{ with } \mathbf{M}\mathbf{u} = \lambda\mathbf{u} \\ \Leftrightarrow & \text{ there exists } \mathbf{u} \neq \mathbf{0} \text{ with } (\mathbf{M} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0} \\ \Leftrightarrow & (\mathbf{M} - \lambda\mathbf{I}) \text{ is singular (that is, not invertible)} \\ \Leftrightarrow & \det(\mathbf{M} - \lambda\mathbf{I}) = 0. \end{aligned}$$

Now,  $\det(\mathbf{M} - \lambda\mathbf{I})$  is a polynomial of degree  $d$  in  $\lambda$ . As such it has  $d$  roots (although some of them might be complex). This explains the existence of eigenvalues.

A case of great interest is when  $\mathbf{M}$  is real-valued and symmetric, because then the eigenvalues are real.

**Theorem 2.** Let  $\mathbf{M}$  be any real symmetric  $d \times d$  matrix. Then:

1.  $\mathbf{M}$  has  $d$  real eigenvalues  $\lambda_1, \dots, \lambda_d$  (not necessarily distinct).
2. There is a set of  $d$  corresponding eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_d$  that constitute an orthonormal basis of  $\mathbb{R}^d$ , that is,  $\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}$  for all  $i, j$ .

## 7.1.2 Spectral decomposition

The spectral decomposition recasts a matrix in terms of its eigenvalues and eigenvectors. This representation turns out to be enormously useful.

**Theorem 3.** Let  $\mathbf{M}$  be a real symmetric  $d \times d$  matrix with eigenvalues  $\lambda_1, \dots, \lambda_d$  and corresponding orthonormal eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_d$ . Then:

$$1. \mathbf{M} = \underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{\text{call this } \mathbf{Q}} \underbrace{\begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_d \end{pmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{pmatrix} \leftarrow & \mathbf{u}_1 & \rightarrow \\ \leftarrow & \mathbf{u}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{u}_d & \rightarrow \end{pmatrix}}_{\mathbf{Q}^T}.$$

$$2. \mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

*Proof.* A general proof strategy is to observe that  $\mathbf{M}$  represents a linear transformation  $\mathbf{x} \mapsto \mathbf{M}\mathbf{x}$  on  $\mathbb{R}^d$ , and as such, is completely determined by its behavior on *any* set of  $d$  linearly independent vectors. For instance,  $\{\mathbf{u}_1, \dots, \mathbf{u}_d\}$  are linearly independent, so any  $d \times d$  matrix  $\mathbf{N}$  that satisfies  $\mathbf{N}\mathbf{u}_i = \mathbf{M}\mathbf{u}_i$  (for all  $i$ ) is necessarily identical to  $\mathbf{M}$ .

Let's start by verifying (1). For practice, we'll do this two different ways.

*Method One:* For any  $i$ , we have

$$\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \mathbf{u}_i = \mathbf{Q}\mathbf{\Lambda}\mathbf{e}_i = \mathbf{Q}\lambda_i \mathbf{e}_i = \lambda_i \mathbf{Q}\mathbf{e}_i = \lambda_i \mathbf{u}_i = \mathbf{M}\mathbf{u}_i.$$

Thus  $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \mathbf{M}$ .

*Method Two:* Since the  $\mathbf{u}_i$  are orthonormal, we have  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ . Thus  $\mathbf{Q}$  is invertible, with  $\mathbf{Q}^{-1} = \mathbf{Q}^T$ ; whereupon  $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$ . For any  $i$ ,

$$\mathbf{Q}^T \mathbf{M}\mathbf{Q}\mathbf{e}_i = \mathbf{Q}^T \mathbf{M}\mathbf{u}_i = \mathbf{Q}^T \lambda_i \mathbf{u}_i = \lambda_i \mathbf{Q}^T \mathbf{u}_i = \lambda_i \mathbf{e}_i = \mathbf{\Lambda}\mathbf{e}_i.$$

Thus  $\mathbf{\Lambda} = \mathbf{Q}^T \mathbf{M}\mathbf{Q}$ , which implies  $\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ .

Now for (2). Again we use the same proof strategy. For any  $j$ ,

$$\left( \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T \right) \mathbf{u}_j = \lambda_j \mathbf{u}_j = \mathbf{M}\mathbf{u}_j.$$

Hence  $\mathbf{M} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ . □

### 7.1.3 Positive semidefinite matrices

We now introduce an important subclass of real symmetric matrices.

**Definition 4.** A real symmetric  $d \times d$  matrix  $\mathbf{M}$  is *positive semidefinite* (denoted  $\mathbf{M} \succcurlyeq 0$ ) if  $\mathbf{z}^T \mathbf{M}\mathbf{z} \geq 0$  for all  $\mathbf{z} \in \mathbb{R}^d$ . It is *positive definite* (denoted  $\mathbf{M} \succ 0$ ) if  $\mathbf{z}^T \mathbf{M}\mathbf{z} > 0$  for all nonzero  $\mathbf{z} \in \mathbb{R}^d$ .

**Example 5.** Consider any random vector  $X \in \mathbb{R}^d$ , and let  $\mu = \mathbb{E}X$  and  $\mathbf{S} = \mathbb{E}[(X - \mu)(X - \mu)^T]$  denote its mean and covariance, respectively. Then  $\mathbf{S} \succcurlyeq 0$  because for any  $\mathbf{z} \in \mathbb{R}^d$ ,

$$\mathbf{z}^T \mathbf{S}\mathbf{z} = \mathbf{z}^T \mathbb{E}[(X - \mu)(X - \mu)^T] \mathbf{z} = \mathbb{E}[(\mathbf{z}^T (X - \mu))((X - \mu)^T \mathbf{z})] = \mathbb{E}[(\mathbf{z} \cdot (X - \mu))^2] \geq 0.$$

Positive (semi)definiteness is easily characterized in terms of eigenvalues.

**Theorem 6.** Let  $\mathbf{M}$  be a real symmetric  $d \times d$  matrix. Then:

1.  $\mathbf{M}$  is positive semidefinite iff all its eigenvalues  $\lambda_i \geq 0$ .
2.  $\mathbf{M}$  is positive definite iff all its eigenvalues  $\lambda_i > 0$ .

*Proof.* Let's prove (1) (the second is similar). Let  $\lambda_1, \dots, \lambda_d$  be the eigenvalues of  $\mathbf{M}$ , with corresponding eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_d$ .

First, suppose  $\mathbf{M} \succcurlyeq 0$ . Then for all  $i$ ,  $\lambda_i = \mathbf{u}_i^T \mathbf{M}\mathbf{u}_i \geq 0$ .

Conversely, suppose that all the  $\lambda_i \geq 0$ . Then for any  $\mathbf{z} \in \mathbb{R}^d$ , we have

$$\mathbf{z}^T \mathbf{M}\mathbf{z} = \mathbf{z}^T \left( \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^T \right) \mathbf{z} = \sum_{i=1}^d \lambda_i (\mathbf{z} \cdot \mathbf{u}_i)^2 \geq 0.$$

□

### 7.1.4 The Rayleigh quotient

One of the reasons why eigenvalues are so useful is that they constitute the optimal solution of a very basic quadratic optimization problem.

**Theorem 7.** Let  $\mathbf{M}$  be a real symmetric  $d \times d$  matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ , and corresponding eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_d$ . Then:

$$\begin{aligned} \max_{\|\mathbf{z}\|=1} \mathbf{z}^T \mathbf{M} \mathbf{z} &= \max_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{M} \mathbf{z}}{\mathbf{z}^T \mathbf{z}} = \lambda_1 \\ \min_{\|\mathbf{z}\|=1} \mathbf{z}^T \mathbf{M} \mathbf{z} &= \min_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{M} \mathbf{z}}{\mathbf{z}^T \mathbf{z}} = \lambda_d \end{aligned}$$

and these are realized at  $\mathbf{z} = \mathbf{u}_1$  and  $\mathbf{z} = \mathbf{u}_d$ , respectively.

*Proof.* Denote the spectral decomposition by  $\mathbf{M} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$ . Then:

$$\begin{aligned} \max_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{M} \mathbf{z}}{\mathbf{z}^T \mathbf{z}} &= \max_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \mathbf{z}}{\mathbf{z}^T \mathbf{Q} \mathbf{Q}^T \mathbf{z}} \quad (\text{since } \mathbf{Q} \mathbf{Q}^T = \mathbf{I}) \\ &= \max_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T \mathbf{\Lambda} \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \quad (\text{writing } \mathbf{y} = \mathbf{Q}^T \mathbf{z}) \\ &= \max_{\mathbf{y} \neq \mathbf{0}} \frac{\lambda_1 y_1^2 + \dots + \lambda_d y_d^2}{y_1^2 + \dots + y_d^2} \leq \lambda_1, \end{aligned}$$

where equality is attained in the last step when  $\mathbf{y} = \mathbf{e}_1$ , that is,  $\mathbf{z} = \mathbf{Q} \mathbf{e}_1 = \mathbf{u}_1$ . The argument for the minimum is identical.  $\square$

**Example 8.** Suppose random vector  $X \in \mathbb{R}^d$  has mean  $\mu$  and covariance matrix  $\mathbf{M}$ . Then  $\mathbf{z}^T \mathbf{M} \mathbf{z}$  represents the variance of  $X$  in direction  $\mathbf{z}$ :

$$\text{var}(\mathbf{z}^T X) = \mathbb{E}[(\mathbf{z}^T (X - \mu))^2] = \mathbb{E}[\mathbf{z}^T (X - \mu)(X - \mu)^T \mathbf{z}] = \mathbf{z}^T \mathbf{M} \mathbf{z}.$$

Theorem 7 tells us that the direction of maximum variance is  $\mathbf{u}_1$ , and that of minimum variance is  $\mathbf{u}_d$ .

Continuing with this example, suppose that we are interested in the  $k$ -dimensional subspace (of  $\mathbb{R}^d$ ) that has the most variance. How can this be formalized?

To start with, we will think of a linear projection from  $\mathbb{R}^d$  to  $\mathbb{R}^k$  as a function  $\mathbf{x} \mapsto \mathbf{P}^T \mathbf{x}$ , where  $\mathbf{P}^T$  is a  $k \times d$  matrix with  $\mathbf{P}^T \mathbf{P} = \mathbf{I}_k$ . The last condition simply says that the rows of the projection matrix are orthonormal.

When a random vector  $X \in \mathbb{R}^d$  is subjected to such a projection, the resulting  $k$ -dimensional vector has covariance matrix

$$\text{cov}(\mathbf{P}^T X) = \mathbb{E}[\mathbf{P}^T (X - \mu)(X - \mu)^T \mathbf{P}] = \mathbf{P}^T \mathbf{M} \mathbf{P}.$$

Often we want to summarize the variance by just a single number rather than an entire matrix; in such cases, we typically use the *trace* of this matrix, and we write  $\text{var}(\mathbf{P}^T X) = \text{tr}(\mathbf{P}^T \mathbf{M} \mathbf{P})$ . This is also equal to  $\mathbb{E}\|\mathbf{P}^T X - \mathbf{P}^T \mu\|^2$ . With this terminology established, we can now determine the projection  $\mathbf{P}^T$  that maximizes this variance.

**Theorem 9.** Let  $\mathbf{M}$  be a real symmetric  $d \times d$  matrix as in Theorem 7. Pick any  $k \leq d$ .

$$\begin{aligned} \max_{\mathbf{P} \in \mathbb{R}^{d \times k}, \mathbf{P}^T \mathbf{P} = \mathbf{I}} \text{tr}(\mathbf{P}^T \mathbf{M} \mathbf{P}) &= \lambda_1 + \dots + \lambda_k \\ \min_{\mathbf{P} \in \mathbb{R}^{d \times k}, \mathbf{P}^T \mathbf{P} = \mathbf{I}} \text{tr}(\mathbf{P}^T \mathbf{M} \mathbf{P}) &= \lambda_{d-k+1} + \dots + \lambda_d. \end{aligned}$$

These are realized when the columns of  $\mathbf{P}$  span the  $k$ -dimensional subspace spanned by  $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$  and  $\{\mathbf{u}_{d-k+1}, \dots, \mathbf{u}_d\}$ , respectively.

*Proof.* We will prove the result for the maximum; the other case is symmetric. Let  $\mathbf{p}_1, \dots, \mathbf{p}_k$  denote the columns of  $\mathbf{P}$ . Then

$$\text{tr}(\mathbf{P}^T \mathbf{M} \mathbf{P}) = \sum_{i=1}^k \mathbf{p}_i^T \mathbf{M} \mathbf{p}_i = \sum_{i=1}^k \mathbf{p}_i^T \left( \sum_{j=1}^d \lambda_j \mathbf{u}_j \mathbf{u}_j^T \right) \mathbf{p}_i = \sum_{j=1}^d \lambda_j \sum_{i=1}^k (\mathbf{p}_i \cdot \mathbf{u}_j)^2.$$

We will show that this quantity is at most  $\lambda_1 + \dots + \lambda_k$ . To this end, let  $z_j$  denote  $\sum_{i=1}^k (\mathbf{p}_i \cdot \mathbf{u}_j)^2$ ; clearly it is nonnegative. We will show that  $\sum_j z_j = k$  and that each  $z_j \leq 1$ ; the desired bound is then immediate.

First,

$$\sum_{j=1}^d z_j = \sum_{i=1}^k \sum_{j=1}^d (\mathbf{p}_i \cdot \mathbf{u}_j)^2 = \sum_{i=1}^k \sum_{j=1}^d \mathbf{p}_i^T \mathbf{u}_j \mathbf{u}_j^T \mathbf{p}_i = \sum_{i=1}^k \mathbf{p}_i^T \mathbf{Q} \mathbf{Q}^T \mathbf{p}_i = \sum_{i=1}^k \|\mathbf{p}_i\|^2 = k.$$

To upper-bound an individual  $z_j$ , start by extending the  $k$  orthonormal vectors  $\mathbf{p}_1, \dots, \mathbf{p}_k$  to a full orthonormal basis  $\mathbf{p}_1, \dots, \mathbf{p}_d$  of  $\mathbb{R}^d$ . Then

$$z_j = \sum_{i=1}^k (\mathbf{p}_i \cdot \mathbf{u}_j)^2 \leq \sum_{i=1}^d (\mathbf{p}_i \cdot \mathbf{u}_j)^2 = \sum_{i=1}^d \mathbf{u}_j^T \mathbf{p}_i \mathbf{p}_i^T \mathbf{u}_j = \|\mathbf{u}_j\|^2 = 1.$$

It then follows that

$$\text{tr}(\mathbf{P}^T \mathbf{M} \mathbf{P}) = \sum_{j=1}^d \lambda_j z_j \leq \lambda_1 + \dots + \lambda_k,$$

and equality holds when  $\mathbf{p}_1, \dots, \mathbf{p}_k$  span the same space as  $\mathbf{u}_1, \dots, \mathbf{u}_k$ .  $\square$

## 7.2 Principal component analysis

Let  $X \in \mathbb{R}^d$  be a random vector. We wish to find the single direction that captures as much as possible of the variance of  $X$ . Formally: we want  $\mathbf{p} \in \mathbb{R}^d$  (the direction) such that  $\|\mathbf{p}\| = 1$ , so as to maximize  $\text{var}(\mathbf{p}^T X)$ .

**Theorem 10.** *The solution to this optimization problem is to make  $\mathbf{p}$  the principal eigenvector of  $\text{cov}(X)$ .*

*Proof.* Denote  $\mu = \mathbb{E}X$  and  $\mathbf{S} = \text{cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^T]$ . For any  $\mathbf{p} \in \mathbb{R}^d$ , the projection  $\mathbf{p}^T X$  has mean  $\mathbb{E}[\mathbf{p}^T X] = \mathbf{p}^T \mu$  and variance

$$\text{var}(\mathbf{p}^T X) = \mathbb{E}[(\mathbf{p}^T X - \mathbf{p}^T \mu)^2] = \mathbb{E}[\mathbf{p}^T (X - \mu)(X - \mu)^T \mathbf{p}] = \mathbf{p}^T \mathbf{S} \mathbf{p}.$$

By Theorem 7, this is maximized (over all unit-length  $\mathbf{p}$ ) when  $\mathbf{p}$  is the principal eigenvector of  $\mathbf{S}$ .  $\square$

Likewise, the  $k$ -dimensional subspace that captures as much as possible of the variance of  $X$  is simply the subspace spanned by the top  $k$  eigenvectors of  $\text{cov}(X)$ ; call these  $\mathbf{u}_1, \dots, \mathbf{u}_k$ .

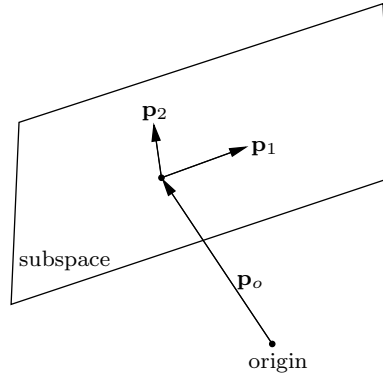
Projection onto these eigenvectors is called *principal component analysis* (PCA). It can be used to reduce the dimension of the data from  $d$  to  $k$ . Here are the steps:

- Compute the mean  $\mu$  and covariance matrix  $\mathbf{S}$  of the data  $X$ .
- Compute the top  $k$  eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_k$  of  $\mathbf{S}$ .
- Project  $X \mapsto \mathbf{P}^T X$ , where  $\mathbf{P}^T$  is the  $k \times d$  matrix whose rows are  $\mathbf{u}_1, \dots, \mathbf{u}_k$ .

### 7.2.1 The best approximating affine subspace

We've seen one optimality property of PCA. Here's another: it is the  $k$ -dimensional affine subspace that best approximates  $X$ , in the sense that the expected squared distance from  $X$  to the subspace is minimized.

Let's formalize the problem. A  $k$ -dimensional affine subspace is given by a displacement  $\mathbf{p}_o \in \mathbb{R}^d$  and a set of (orthonormal) basis vectors  $\mathbf{p}_1, \dots, \mathbf{p}_k \in \mathbb{R}^d$ . The subspace itself is then  $\{\mathbf{p}_o + \alpha_1 \mathbf{p}_1 + \dots + \alpha_k \mathbf{p}_k : \alpha_i \in \mathbb{R}\}$ .



The projection of  $X \in \mathbb{R}^d$  onto this subspace is  $\mathbf{P}^T X + \mathbf{p}_o$ , where  $\mathbf{P}^T$  is the  $k \times d$  matrix whose rows are  $\mathbf{p}_1, \dots, \mathbf{p}_k$ . Thus, the expected squared distance from  $X$  to this subspace is  $\mathbb{E}\|X - (\mathbf{P}^T X + \mathbf{p}_o)\|^2$ . We wish to find the subspace for which this is minimized.

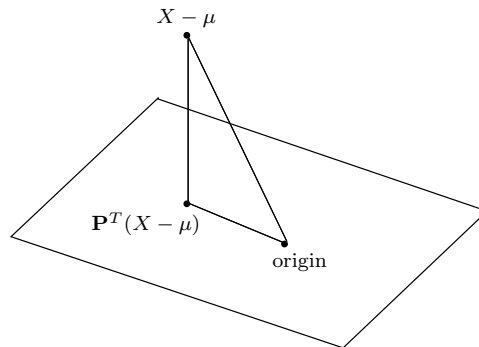
**Theorem 11.** *Let  $\mu$  and  $\mathbf{S}$  denote the mean and covariance of  $X$ , respectively. The solution of this optimization problem is to choose  $\mathbf{p}_1, \dots, \mathbf{p}_k$  to be the top  $k$  eigenvectors of  $\mathbf{S}$  and to set  $\mathbf{p}_o = (\mathbf{I} - \mathbf{P}^T)\mu$ .*

*Proof.* Fix any matrix  $\mathbf{P}$ ; the choice of  $\mathbf{p}_o$  that minimizes  $\mathbb{E}\|X - (\mathbf{P}^T X + \mathbf{p}_o)\|^2$  is (by calculus)  $\mathbf{p}_o = \mathbb{E}[X - \mathbf{P}^T X] = (\mathbf{I} - \mathbf{P}^T)\mu$ .

Now let's optimize  $\mathbf{P}$ . Our cost function is

$$\mathbb{E}\|X - (\mathbf{P}^T X + \mathbf{p}_o)\|^2 = \mathbb{E}\|(\mathbf{I} - \mathbf{P}^T)(X - \mu)\|^2 = \mathbb{E}\|X - \mu\|^2 - \mathbb{E}\|\mathbf{P}^T(X - \mu)\|^2,$$

where the second step is simply an invocation of the Pythagorean theorem.



Therefore, we need to maximize  $\mathbb{E}\|\mathbf{P}^T(X - \mu)\|^2 = \text{var}(\mathbf{P}^T X)$ , and we've already seen how to do this in Theorem 10 and the ensuing discussion.  $\square$

### 7.2.2 The projection that best preserves interpoint distances

Suppose we want to find the  $k$ -dimensional projection that minimizes the expected distortion in interpoint distances. More precisely, we want to find the  $k \times d$  projection matrix  $\mathbf{P}^T$  (with  $\mathbf{P}^T \mathbf{P} = \mathbf{I}_k$ ) such that, for i.i.d. random vectors  $X$  and  $Y$ , expected squared distortion  $\mathbb{E}[\|X - Y\|^2 - \|\mathbf{P}^T X - \mathbf{P}^T Y\|^2]$  is minimized (of course, the term in brackets is always positive).

**Theorem 12.** *The solution is to make the rows of  $\mathbf{P}^T$  the top  $k$  eigenvectors of  $\text{cov}(X)$ .*

*Proof.* This time we want to maximize

$$\mathbb{E}\|\mathbf{P}^T X - \mathbf{P}^T Y\|^2 = 2\mathbb{E}\|\mathbf{P}^T X - \mathbf{P}^T \mu\|^2 = 2\text{var}(\mathbf{P}^T X),$$

and once again we're back to our original problem.  $\square$

This is emphatically not the same as finding the *linear transformation* (that is, not necessarily a projection)  $\mathbf{P}^T$  for which  $\mathbb{E}[\|X - Y\|^2 - \|\mathbf{P}^T X - \mathbf{P}^T Y\|^2]$  is minimized. The random projection method that we saw earlier falls in this latter camp, because it consists of a projection *followed by a scaling by  $\sqrt{d/k}$* .

### 7.2.3 A prelude to $k$ -means clustering

Suppose that for random vector  $X \in \mathbb{R}^d$ , the optimal  $k$ -means centers are  $\mu_1^*, \dots, \mu_k^*$ , with cost

$$\text{OPT} = \mathbb{E}\|X - (\text{nearest } \mu_i^*)\|^2.$$

If instead, we project  $X$  into the  $k$ -dimensional PCA subspace, and find the best  $k$  centers  $\mu_1, \dots, \mu_k$  *in that subspace*, how bad can these centers be?

**Theorem 13.**  $\text{cost}(\mu_1, \dots, \mu_k) \leq 2 \cdot \text{OPT}$ .

*Proof.* Without loss of generality  $\mathbb{E}X = \mathbf{0}$  and the PCA mapping is  $X \mapsto \mathbf{P}^T X$ . Since  $\mu_1, \dots, \mu_k$  are the best centers for  $\mathbf{P}^T X$ , it follows that

$$\mathbb{E}\|\mathbf{P}^T X - (\text{nearest } \mu_i)\|^2 \leq \mathbb{E}\|\mathbf{P}^T(X - (\text{nearest } \mu_i^*))\|^2 \leq \mathbb{E}\|X - (\text{nearest } \mu_i^*)\|^2 = \text{OPT}.$$

Let  $X \mapsto \mathbf{A}^T X$  denote projection onto the subspace spanned by  $\mu_1^*, \dots, \mu_k^*$ . From our earlier result on approximating affine subspaces, we know that  $\mathbb{E}\|X - \mathbf{P}^T X\|^2 \leq \mathbb{E}\|X - \mathbf{A}^T X\|^2$ . Thus

$$\begin{aligned} \text{cost}(\mu_1, \dots, \mu_k) &= \mathbb{E}\|X - (\text{nearest } \mu_i)\|^2 \\ &= \mathbb{E}\|\mathbf{P}^T X - (\text{nearest } \mu_i)\|^2 + \mathbb{E}\|X - \mathbf{P}^T X\|^2 \quad (\text{Pythagorean theorem}) \\ &\leq \text{OPT} + \mathbb{E}\|X - \mathbf{A}^T X\|^2 \\ &\leq \text{OPT} + \mathbb{E}\|X - (\text{nearest } \mu_i^*)\|^2 = 2 \cdot \text{OPT}. \end{aligned}$$

$\square$

## 7.3 Singular value decomposition

For any real symmetric  $d \times d$  matrix  $\mathbf{M}$ , we can find its eigenvalues  $\lambda_1 \geq \dots \geq \lambda_d$  and corresponding orthonormal eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_d$ , and write

$$\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

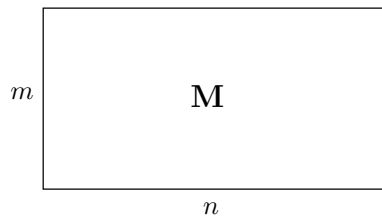
The best rank- $k$  approximation to  $\mathbf{M}$  is

$$\mathbf{M}_k = \sum_{i=1}^k \lambda_i \mathbf{u}_i \mathbf{u}_i^T,$$

in the sense that this minimizes  $\|\mathbf{M} - \mathbf{M}_k\|_F^2$  over all rank- $k$  matrices. (Here  $\|\cdot\|_F$  denotes Frobenius norm; it is the same as  $L_2$  norm if you imagine the matrix rearranged into a very long vector.)

In many applications,  $\mathbf{M}_k$  is an adequate approximation of  $\mathbf{M}$  even for fairly small values of  $k$ . And it is conveniently compact, of size  $O(kd)$ .

But what if we are dealing with a matrix  $\mathbf{M}$  that is not square; say it is  $m \times n$  with  $m \leq n$ :



To find a compact approximation in such cases, we look at  $\mathbf{M}^T \mathbf{M}$  or  $\mathbf{M} \mathbf{M}^T$ , which are square. Eigendecompositions of these matrices lead to a good representation of  $\mathbf{M}$ .

### 7.3.1 The relationship between $\mathbf{M} \mathbf{M}^T$ and $\mathbf{M}^T \mathbf{M}$

**Lemma 14.**  $\mathbf{M}^T \mathbf{M}$  and  $\mathbf{M} \mathbf{M}^T$  are symmetric positive semidefinite matrices.

*Proof.* We'll do  $\mathbf{M}^T \mathbf{M}$ ; the other is similar. First off, it is symmetric:

$$(\mathbf{M}^T \mathbf{M})_{ij} = \sum_k (\mathbf{M}^T)_{ik} \mathbf{M}_{kj} = \sum_k \mathbf{M}_{ki} \mathbf{M}_{kj} = \sum_k (\mathbf{M}^T)_{jk} \mathbf{M}_{ki} = (\mathbf{M}^T \mathbf{M})_{ji}.$$

Next,  $\mathbf{M}^T \mathbf{M} \succcurlyeq 0$  since for any  $\mathbf{z} \in \mathbb{R}^n$ , we have  $\mathbf{z}^T \mathbf{M}^T \mathbf{M} \mathbf{z} = \|\mathbf{M} \mathbf{z}\|^2 \geq 0$ .  $\square$

Which one should we use,  $\mathbf{M}^T \mathbf{M}$  or  $\mathbf{M} \mathbf{M}^T$ ? Well, they are of different sizes,  $n \times n$  and  $m \times m$  respectively.



Ideally, we'd prefer to deal with the smaller of two,  $\mathbf{M} \mathbf{M}^T$ , especially since eigenvalue computations are expensive. Fortunately, it turns out the two matrices have the same (non-zero) eigenvalues!

**Lemma 15.** If  $\lambda$  is an eigenvalue of  $\mathbf{M}^T \mathbf{M}$  with eigenvector  $\mathbf{u}$ , then

- **either:** (i)  $\lambda$  is an eigenvalue of  $\mathbf{M} \mathbf{M}^T$  with eigenvector  $\mathbf{M} \mathbf{u}$ ,
- **or** (ii)  $\lambda = 0$  and  $\mathbf{M} \mathbf{u} = \mathbf{0}$ .

*Proof.* Say  $\lambda \neq 0$ ; we'll prove that condition (i) holds. First of all,  $\mathbf{M}^T \mathbf{M} \mathbf{u} = \lambda \mathbf{u} \neq \mathbf{0}$ , so certainly  $\mathbf{M} \mathbf{u} \neq \mathbf{0}$ . It is an eigenvector of  $\mathbf{M} \mathbf{M}^T$  with eigenvalue  $\lambda$ , since

$$\mathbf{M} \mathbf{M}^T (\mathbf{M} \mathbf{u}) = \mathbf{M} (\mathbf{M}^T \mathbf{M} \mathbf{u}) = \mathbf{M} (\lambda \mathbf{u}) = \lambda (\mathbf{M} \mathbf{u}).$$

Next, suppose  $\lambda = 0$ ; we'll establish condition (ii). Notice that

$$\|\mathbf{M} \mathbf{u}\|^2 = \mathbf{u}^T \mathbf{M}^T \mathbf{M} \mathbf{u} = \mathbf{u}^T (\mathbf{M}^T \mathbf{M} \mathbf{u}) = \mathbf{u}^T (\lambda \mathbf{u}) = 0.$$

Thus it must be the case that  $\mathbf{M} \mathbf{u} = \mathbf{0}$ . □

### 7.3.2 A spectral decomposition for rectangular matrices

Let's summarize the consequences of Lemma 15. We have two square matrices, a large one ( $\mathbf{M}^T \mathbf{M}$ ) of size  $n \times n$  and a smaller one ( $\mathbf{M} \mathbf{M}^T$ ) of size  $m \times m$ . Let the eigenvalues of the large matrix be  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ , with corresponding orthonormal eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_n$ . From the lemma, we know that at most  $m$  of the eigenvalues are nonzero.

The smaller matrix  $\mathbf{M} \mathbf{M}^T$  has eigenvalues  $\lambda_1, \dots, \lambda_m$ , and corresponding orthonormal eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_m$ . The lemma suggests that  $\mathbf{v}_i = \mathbf{M} \mathbf{u}_i$ ; this is certainly a valid set of eigenvectors, but they are not necessarily normalized to unit length. So instead we set

$$\mathbf{v}_i = \frac{\mathbf{M} \mathbf{u}_i}{\|\mathbf{M} \mathbf{u}_i\|} = \frac{\mathbf{M} \mathbf{u}_i}{\sqrt{\mathbf{u}_i^T \mathbf{M}^T \mathbf{M} \mathbf{u}_i}} = \frac{\mathbf{M} \mathbf{u}_i}{\sqrt{\lambda_i}}.$$

This finally gives us the *singular value decomposition*, a spectral decomposition for general matrices.

**Theorem 16.** Let  $\mathbf{M}$  be a rectangular  $m \times n$  matrix with  $m \leq n$ . Define  $\lambda_i, \mathbf{u}_i, \mathbf{v}_i$  as above. Then

$$\mathbf{M} = \underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_m \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{\mathbf{Q}_1, \text{ size } m \times m} \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & 0 & \left| & 0 \\ & \sqrt{\lambda_2} & & & \\ & & \ddots & & \\ 0 & & & \sqrt{\lambda_m} & \end{pmatrix}}_{\mathbf{\Sigma}, \text{ size } m \times n} \underbrace{\begin{pmatrix} \leftarrow & \mathbf{u}_1 & \rightarrow \\ \leftarrow & \mathbf{u}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{u}_n & \rightarrow \end{pmatrix}}_{\mathbf{Q}_2^T, \text{ size } n \times n}.$$

*Proof.* We will check that  $\mathbf{\Sigma} = \mathbf{Q}_1^T \mathbf{M} \mathbf{Q}_2$ . By our proof strategy from Theorem 3, it is enough to verify that both sides have the same effect on  $\mathbf{e}_i$  for all  $1 \leq i \leq n$ . For any such  $i$ ,

$$\mathbf{Q}_1^T \mathbf{M} \mathbf{Q}_2 \mathbf{e}_i = \mathbf{Q}_1^T \mathbf{M} \mathbf{u}_i = \begin{cases} \mathbf{Q}_1^T \sqrt{\lambda_i} \mathbf{v}_i & \text{if } i \leq m \\ \mathbf{0} & \text{if } i > m \end{cases} = \begin{cases} \sqrt{\lambda_i} \mathbf{e}_i & \text{if } i \leq m \\ \mathbf{0} & \text{if } i > m \end{cases} = \mathbf{\Sigma} \mathbf{e}_i.$$

□

The alternative form of the singular value decomposition is

$$\mathbf{M} = \sum_{i=1}^m \sqrt{\lambda_i} \mathbf{v}_i \mathbf{u}_i^T,$$

which immediately yields a rank- $k$  approximation

$$\mathbf{M}_k = \sum_{i=1}^k \sqrt{\lambda_i} \mathbf{v}_i \mathbf{u}_i^T.$$

As in the square case,  $\mathbf{M}_k$  is the rank- $k$  matrix that minimizes  $\|\mathbf{M} - \mathbf{M}_k\|_F^2$ .