CS 8850: Advanced Machine Learning **Fall 2017**

Topic 1: Review of Linear Algebra

Instructor: Daniel L. Pimentel-Alarcón

1.1 Vector Spaces

In words, a *vector space* is a set of elements (usually called *vectors*) such that linear combinations of its elements are also in the set.

Example 1.1. The vector space that we will mostly use is $\mathbb{R}^{\mathrm{D}}$. Here are two vectors in $\mathbb{R}^3$:

$$\mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ \pi \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} e \\ 0 \\ 1 \end{bmatrix}.$$

We can see that for any real scalars a, b, the linear combination

$$\mathrm{a}\mathbf{x} + \mathrm{b}\mathbf{y} = \begin{bmatrix} \mathrm{a} + \mathrm{b}e \\ -2\mathrm{a} \\ \mathrm{a}\pi + \mathrm{b} \end{bmatrix}$$

is also an element of $\mathbb{R}^3$ (see Figure 1.1 to build some geometric intuition).

Why do I care about vector spaces?

If you are wondering this, you are asking yourself the right question! The reason is simple: in machine learning we have to deal with data, and it is useful to arrange it in vectors.

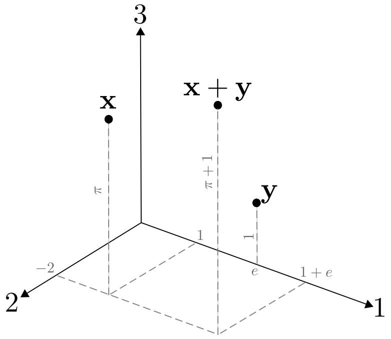
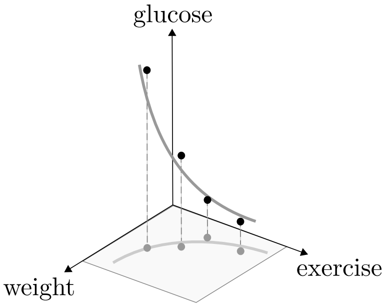


Figure 1.1: Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ from Example 1.1, and vector $\mathrm{a}\mathbf{x} + \mathrm{b}\mathbf{y} \in \mathbb{R}^3$, with $\mathrm{a} = \mathrm{b} = 1$.

Example 1.2 (Electronic health records)**.** Hospitals keep health records of their patients, containing information such as weight, amount of exercise they do, and glucose level. The information of the i^{th} patient can be arranged as a vector

$$\mathbf{x}_{\mathrm{i}} \;=\; \begin{bmatrix} \text{weight} \\ \text{exercise} \\ \text{glucose} \end{bmatrix} \in \mathbb{R}^3.$$

In this sort of problem we want to identify *causes* for diseases. This can be done by analyzing the patterns in the vectors of different patients. For example, if our data $\mathbf{x}_1, \ldots, \mathbf{x}_{\mathrm{N}}$ looks like:



then it is reasonable to conclude that overweight and lack of exercise are highly correlated with diabetes.

Of course, this is an oversimplified example. Not all correlations are as evident. Health records actually include much more comprehensive information, such as age, gender, ethnicity, cholesterol levels, etc. This would produce data vectors $\mathbf{x}_{\mathrm{i}}$ in higher dimensions:

$$\mathbf{x}_{\mathrm{i}} \;=\; \begin{bmatrix} \text{weight} \\ \text{exercise} \\ \text{glucoseage} \\ \text{gender} \\ \text{ethnicity} \\ \text{cholesterol} \\ \vdots \end{bmatrix} \in \mathbb{R}^{\mathrm{D}}.$$

Now you will have to use your imagination to decide how D-dimensional space looks like. In fact, it can be very challenging to visualize points in $\mathbb{R}^{\mathrm{D}}$ (with $\mathrm{D} > 3$, obviously). Luckily, using theory from vector spaces (and probability and statistics) we can find *lines*, *planes*, *curves*, etc. (similar to the gray curves depicted in the figure above, only in higher dimensions) that explain out data (just as the gray curves explain the correlations between weight, exercise and glucose).

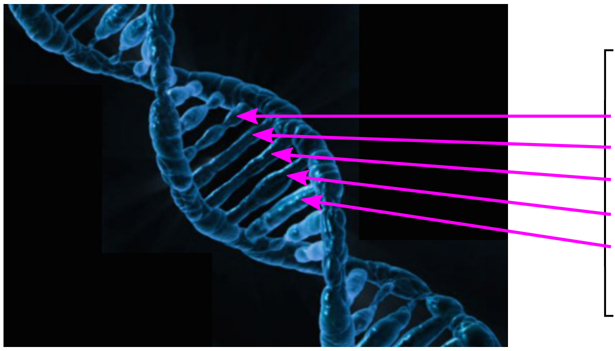
Example 1.3 (Recommender systems)**.** Similarly, Amazon, Netflix, Pandora, Spotify, Pinterest, Yelp, Apple, etc., keep information of their users, such as age, gender, income level, and very importantly, ratings of their products. The information of the i^{th} user can be arranged as a vector

$$\mathbf{x}_\mathrm{i} \;=\; \begin{bmatrix} \text{age} \\ \text{gender} \\ \text{income} \\ \text{rating of item 1} \\ \text{rating of item 2} \\ \vdots \\ \text{rating of item } \mathrm{D}-3 \end{bmatrix} \in \mathbb{R}^{\mathrm{D}}.$$

In this sort of problem we want to analyze these data vectors to predict which users will like which items, in order to make good recommendations. If Amazon recommends you an item you will like, you are more likely to buy it. You can see why all these companies have a great interest in this problem, and they are paying *a lot* of money to people who work on this.

This can be done by finding structures (e.g., *lines* or *curves*) in high-dimensions that explain the data. As in Example 1.2, where we discovered that weight and exercise are good predictors for diabetes, here we want to discover which variables (e.g., gender, age, income, etc.) can predict which items (e.g., movies, shoes, songs, etc.) you would like.

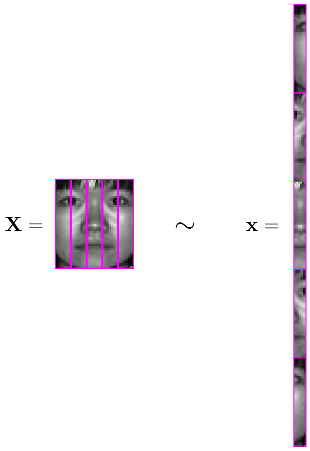
Example 1.4 (Genomics)**.** The genome of each individual can be stored as a vector containing its corresponding sequence of nucleotides, e.g., Adenine, Thymine, Guanine, Cytosine, Thymine, . . .



$$\begin{bmatrix} \vdots \\ A \\ T \\ G \\ C \\ T \\ \vdots \end{bmatrix} \sim \mathbf{x}_\mathrm{i} \;=\; \begin{bmatrix} \vdots \\ 1 \\ 4 \\ 3 \\ 2 \\ 4 \\ \vdots \end{bmatrix} \in \mathbb{R}^{\mathrm{D}}$$

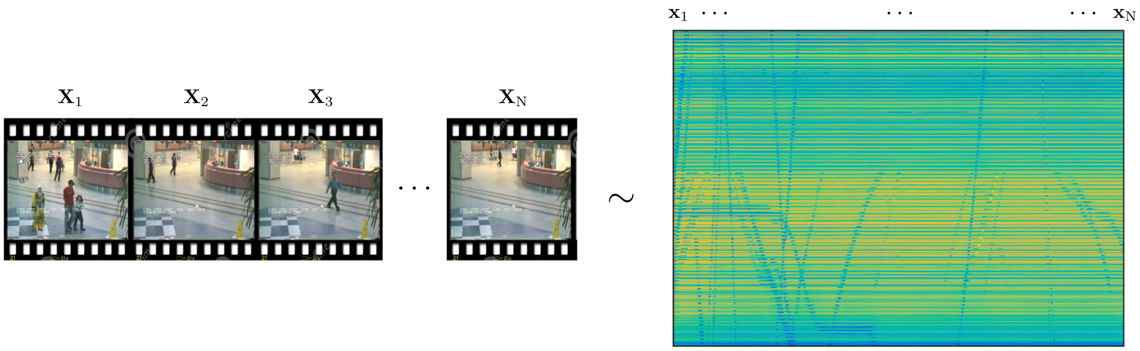
In this sort of problem we want to analyze these data vectors to determine which genes are correlated to which diseases (or features, like height or weight).

Example 1.5 (Image processing)**.** A $\mathrm{m} \times \mathrm{n}$ grayscale image can be stored in a data matrix $\mathbf{X} \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ whose $(\mathrm{i},\mathrm{j})^{\mathrm{th}}$ entry contains the gray intensity of pixel (i,j). Furthermore, $\mathbf{X}$ can be *vectorized*, i.e., we can stack its columns to form a vector $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$, with $\mathrm{D} = \mathrm{mn}$.



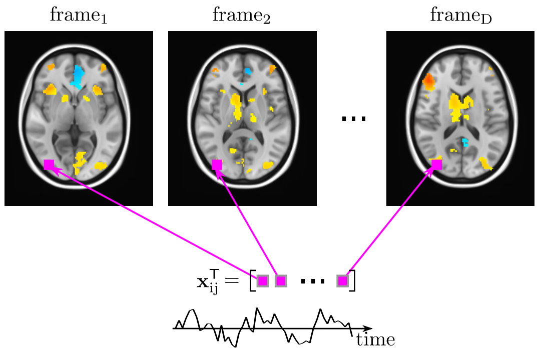
We want to analyze these vectors to interpret the image. For example, identify the objects that appear in the image, classifying faces, etc.

Example 1.6 (Computer vision)**.** The images $\mathbf{X}_1, \ldots, \mathbf{X}_\mathrm{N} \in \mathbb{R}^{\mathrm{m} \times \mathrm{n}}$ that form a video can be vectorized to obtain vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{N} \in \mathbb{R}^\mathrm{D}$.



Similar to image processing, we want to analyze these vectors to interpret the video. For example, be able to distinguish background from foreground, track objects, etc. This has applications in surveillance, defense, robotics, etc.

Example 1.7 (Neural activity)**.** *Functional magnetic resonance imaging* (fMRI) generates a series of MRI images over time. Because oxygenated and deoxygenated hemoglobin have slightly different magnetic characteristics, variations in the MRI intensity indicate areas of the brain with increased blood flow and hence neural activity. The central task in fMRI is to reliably detect neural activity at different spatial locations (pixels) in the brain. The measurements over time at the $(\mathrm{i}, \mathrm{j})^{\mathrm{th}}$ pixel can be stored in a data vector $\mathbf{x}_{\mathrm{ij}} \in \mathbb{R}^\mathrm{D}$.

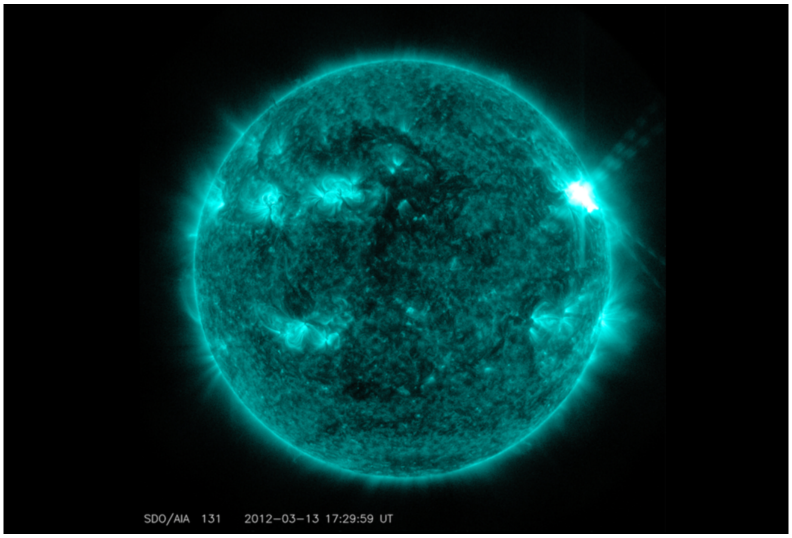


The idea is to analyze these vectors to determine the active pixels.

Example 1.8 (Sun flares)**.** The Sun, like all active stars, is constantly producing huge electromagnetic *flares*. Every now and then, these flares hit the Earth. Last time this happened was in 1859, and all that happened was that you could see the northern lights all the way down to Mexico — not a bad secondary effect! However, back in 1859 we didn't have a massive power grid, satellites, wireless communications, GPS, airplanes, space stations, etc. If a flare hit the Earth now, all these systems would be crippled, and repairing them could take *years* and would cost *trillions* of dollars to the U.S. alone! To make things worse, it turns out that these flares are not rare at all! It is estimated that the chance that a flare hits the earth in the next decade is about 12%.

Of course, we cannot stop these flares any more than we can stop an earthquake. If it hits us, it hits us. However, like with an earthquake, we can act ahead. If we know that one flare is coming, we can turn everything off, let it pass, and then turn everything back on, like nothing happened. Hence the NASA and other institutions are investing a great deal of time, effort and money to develop techniques that enable us to *predict* that a flare is coming.

So essentially, we want to device a sort of flares *radar* or *detector*. This radar would receive, for example, an image $\mathbf{X}$ of the sun (or equivalently, a vectorized image $\mathbf{x} \in \mathbb{R}^D$), and would have to decide whether a flare is coming or not.



These are only a few examples that I hope help convince you that vector spaces are the backbone of machine learning. Studying vector spaces will allow us to use the powerful machinery of vector spaces that has been developed over centuries (e.g., principal component analysis) in order to tackle the modern problems in machine learning.

Formal definition

Definition 1.1 (Vector space)**.** A set $\mathcal{X}$ is a *vector space* if it satisfies the following *additive* properties:

($a1$) **Closure:** For every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\mathbf{x} + \mathbf{y} \in \mathcal{X}$.

($a2$) **Commutative law:** For every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.

($a3$) **Associative law:** For every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$, $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.

($a4$) **Additive identity:** There exists an element in $\mathcal{X}$, denoted by $\mathbf{0}$, such that for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{x} + \mathbf{0} = \mathbf{x}$.

($a5$) **Additive inverse:** For every $\mathbf{x} \in \mathcal{X}$, there exists a unique element in $\mathcal{X}$, denoted by $-\mathbf{x}$, such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.

and the following *multiplicative* properties:

($m1$) **Closure:** For each $\mathrm{a} \in \mathbb{R}$ and $\mathbf{x} \in \mathcal{X}$, $\mathrm{a}\mathbf{x} \in \mathcal{X}$.

($m2$) **Associative law:** For every $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ and any $\mathbf{x} \in \mathcal{X}$, $\mathrm{a}(\mathrm{b}\mathbf{x}) = (\mathrm{ab})\mathbf{x}$.

($m3$) **First distributive law:** For any $\mathrm{a} \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\mathrm{a}(\mathbf{x} + \mathbf{y}) = \mathrm{a}\mathbf{x} + \mathrm{a}\mathbf{y}$.

($m4$) **Second distributive law:** For any $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ and any $\mathbf{x} \in \mathcal{X}$, $(\mathrm{a} + \mathrm{b})\mathbf{x} = \mathrm{a}\mathbf{x} + \mathrm{b}\mathbf{x}$.

($m5$) **Multiplicative identity:** For every $\mathbf{x} \in \mathcal{X}$, $1\mathbf{x} = \mathbf{x}$.

Example 1.9. $\mathbb{R}^{\mathrm{D}}$ is a vector space.

How would we show that $\mathbb{R}^{\mathrm{D}}$ is a vector space? We would have to show that $\mathbb{R}^{\mathrm{D}}$ satisfies all the properties of a vector space. For example, let us show property ($a1$).

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathrm{D}}$. Since $\mathbf{x} + \mathbf{y} \in \mathbb{R}^{\mathrm{D}}$, $\mathbb{R}^{\mathrm{D}}$ satisfies property ($a1$), as desired. □

In the proof we are taking an arbitrary $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^{\mathrm{D}}$, and we show that $\mathbf{x} + \mathbf{y}$ is also in $\mathbb{R}^{\mathrm{D}}$. For example, with the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ in Example 1.1, we can see that

$$\mathbf{x} + \mathbf{y} \;=\; \begin{bmatrix} 1 + e \\ -2 \\ \pi + 1 \end{bmatrix} \in \mathbb{R}^3.$$

Definition 1.2 (Linear combination, coefficients)**.** A vector $\mathbf{y}$ is a *linear combination* of $\{\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{R}\}$ if it can be written as

$$\mathbf{y} \;=\; \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_\mathrm{r} \mathbf{x}_\mathrm{r} \tag{1.1}$$

for some $\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{R} \in \mathbb{R}$. The scalars $\{\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{R}\}$ are called the *coefficients* of $\mathbf{y}$ with respect to (w.r.t.) $\{\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{R}\}$.

Example 1.10. If $\mathcal{X} = \mathbb{R}^\mathrm{D}$, we can write (1.1) in matrix form as $\mathbf{y} = \mathbf{X}\mathbf{c}$, where $\mathbf{X} = [\mathbf{x}_1 \;\cdots\; \mathbf{x}_\mathrm{R}] \in \mathbb{R}^{\mathrm{D} \times \mathrm{R}}$ and $\mathbf{c} = [\mathrm{c}_1 \;\cdots\; \mathrm{c}_\mathrm{R}]^\mathsf{T} \in \mathbb{R}^\mathrm{R}$.

Definition 1.3 (Linear independence)**.** A set of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{R}\}$ is *linearly independent* if

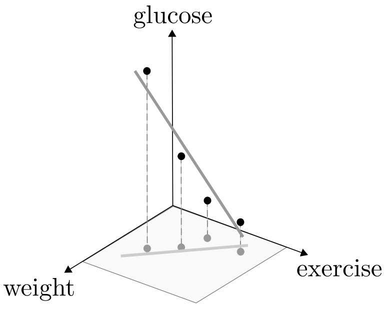
$$\sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_\mathrm{r} \mathbf{x}_\mathrm{r} \;=\; \mathbf{0}$$

implies $\mathrm{c}_\mathrm{r} = 0$ for every $\mathrm{r} = 1, \dots, \mathrm{R}$. Otherwise we say it is *linearly dependent.*

1.2 Subspaces

Subspaces are essentially high-dimensional lines. A 1-dimensional subspace is a line, a 2-dimensional subspaces is a plane, and so on. Subspaces are useful because data often lies near subspaces.

Example 1.11. The health records data in Example 1.2 lies near a 1-dimensional subspace (line):



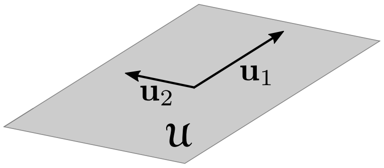


Figure 1.2: Subspace $\mathcal{U}$ (plane) spanned by two vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$.

In higher dimensions subspaces may be harder to visualize, so you will have to use imagination to decide how a higher-dimensional subspace looks. Luckily, we have a precise and formal mathematical way to define them:

Definition 1.4 (Subspace)**.** A subset $\mathcal{U} \subseteq \mathcal{X}$ is a *subspace* if for every $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ and every $\mathbf{u}, \mathbf{v} \in \mathcal{U}$, $\mathrm{a}\mathbf{u} + \mathrm{b}\mathbf{v} \in \mathcal{U}$.

Definition 1.5 (Span)**.** $\mathrm{span}[\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}]$ is the set of all linear combinations of $\{\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}\}$. More formally,

$$\mathrm{span}[\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}] \ := \ \left\{ \mathbf{x} \in \mathcal{X} \ : \ \mathbf{x} = \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_\mathrm{r} \mathbf{u}_\mathrm{r} \ \text{ for some } \mathrm{c}_1, \dots, \mathrm{c}_\mathrm{R} \in \mathbb{R} \right\}.$$

Example 1.12. Let $\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R} \in \mathbb{R}^\mathrm{D}$. Then $\mathrm{span}[\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}]$ is a subspace.

Definition 1.6 (Basis)**.** A set of linearly independent vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_\mathrm{R}\}$ is a *basis* of a subspace $\mathcal{U}$ if each $\mathbf{v} \in \mathcal{U}$ can be written as

$$\mathbf{v} \ = \ \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{c}_\mathrm{r} \mathbf{u}_\mathrm{r}$$

for a *unique* set of coefficients $\{\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{R}\}$.

1.3 Inner Products

To analyze vectors we often need to study the relationship between two or more vectors. One useful way to measure these relationships is through their angle. Luckily, even when vectors in high dimensions can be hard to visualize, inner products allow us to study their relationships. In words, the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ is a proxy for the angle between $\mathbf{x}$ and $\mathbf{y}$ vectors (see equation (1.2) below). More formally:

Definition 1.7 (Inner product)**.** A mapping $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an *inner product in* $\mathcal{X}$ if it satisfies:

(i) For every $\mathbf{x} \in \mathcal{X}$, $0 \leq \langle \mathbf{x}, \mathbf{x} \rangle < \infty$, with $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

(ii) For every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.

(iii) For every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$, and every $\mathrm{a}, \mathrm{b} \in \mathbb{R}$, $\langle \mathrm{a}\mathbf{x} + \mathrm{b}\mathbf{y}, \mathbf{z} \rangle = \mathrm{a}\langle \mathbf{x}, \mathbf{z} \rangle + \mathrm{b}\langle \mathbf{y}, \mathbf{z} \rangle$.

Example 1.13. Let $\mathbf{x}, \mathbf{y} \in \mathcal{X} = \mathbb{R}^{\mathrm{D}}$. Then $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^{\mathsf{T}}\mathbf{y} \ = \sum_{\mathrm{d}=1}^{\mathrm{D}} \mathrm{x}_{\mathrm{d}}\mathrm{y}_{\mathrm{d}}$ defines an inner product.

Definition 1.8 (Orthogonal)**.** A collection of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_{\mathrm{R}}\}$ is *orthogonal* if $\langle \mathbf{x}_r, \mathbf{x}_k \rangle = 0$ for every $\mathrm{r} \neq \mathrm{k}$.

1.4 Norms

The norm $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is essentially its size. Norms are also useful because they allow us to measure distance between vectors (through their difference).

Example 1.14. Consider the following images:

and vectorize them as in Example 1.5 to produce vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$. We want to do face clustering, i.e., we want to know which images correspond to the same person. If $\|\mathbf{x} - \mathbf{y}\|$ is small (i.e., $\mathbf{x}$ is similar to $\mathbf{y}$), it is reasonable to conclude that the first two images correspond to the same person. If $\|\mathbf{x} - \mathbf{z}\|$ is large (i.e., $\mathbf{x}$ is very different from $\mathbf{z}$), it is reasonable to conclude that the first and second images corresponds to different persons.

Definition 1.9 (Norm)**.** A mapping $\| \cdot \| : \mathcal{X} \to \mathbb{R}$ is a *norm in* $\mathcal{X}$ if it satisfies:

(i) For every $\mathbf{x} \in \mathcal{X}$, $0 \leq \|\mathbf{x}\| < \infty$, with $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

(ii) For every $\mathbf{x} \in \mathcal{X}$ and every $\mathrm{a} \in \mathbb{R}$, $\|\mathrm{a}\mathbf{x}\| = |\mathrm{a}|\|\mathbf{x}\|$.

(iii) For every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. This is known as the **triangle inequality**.

These properties of norms (in particular property (iii)) allow you to draw intuitive conclusions. For instance, in Example 1.14, knowing that $\|\mathbf{x}-\mathbf{y}\|$ is small and that $\|\mathbf{x}-\mathbf{z}\|$ is large allows us to conclude that $\|\mathbf{y}-\mathbf{z}\|$ is also large. Intuitively, this allows us to conclude that if $\mathbf{x}, \mathbf{y}$ correspond to the same person, and $\mathbf{x}$ and $\mathbf{z}$ corresponds to different persons, then $\mathbf{y}$ and $\mathbf{z}$ also correspond to different persons. In other words, nothing weird will happen.

Also, using inner products and norms we can compute angle θ between $\mathbf{x}$ and $\mathbf{y}$ as:

$$\cos\theta \;=\; \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{1.2}$$

Example 1.15. Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^{\mathrm{R}}$. Then $\|\mathbf{x}\| := \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$ is a norm.

Example 1.16. A norm satisfies the inequality $|\|\mathbf{x}\| - \|\mathbf{y}\|| \;\leq\; \|\mathbf{x}-\mathbf{y}\|$.

If $\|\mathbf{x}\| = 1$, we say $\mathbf{x}$ is a *unit* vector, or that it is *normalized.* Similarly, a collection of normalized, orthogonal vectors is called *orthonormal.* As you can see, there is a tight relation between inner products and norms. The following is one of the most important and useful inequalities that describe this relationship.

Proposition 1.1 (Cauchy-Schwartz inequality)**.** For every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \;\leq\; \|\mathbf{x}\|\|\mathbf{y}\|.$$

Furthermore, if $\mathbf{y} \neq \mathbf{0}$, then equality holds if and only if $\mathbf{x} = \mathrm{a}\mathbf{y}$ for some $\mathrm{a} \in \mathbb{R}$.

1.5 Projections

In words, the projection $\hat{\mathbf{x}}$ of a vector $\mathbf{x}$ onto a subspace $\mathcal{U}$ is the vector in $\mathcal{U}$ that is closest to $\mathbf{x}$. More formally,

Definition 1.10 (Projection)**.** The *projection* of $\mathbf{x} \in \mathcal{X}$ onto subspace $\mathcal{U}$ is the vector $\hat{\mathbf{x}} \in \mathcal{U}$ satisfying

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \;\leq\; \|\mathbf{x} - \mathbf{u}\| \qquad \text{for every } \mathbf{u} \in \mathcal{U}.$$

Notice that if $\mathbf{x} \in \mathcal{U}$, then $\hat{\mathbf{x}} = \mathbf{x}$. The following proposition tells us exactly how to compute projections.

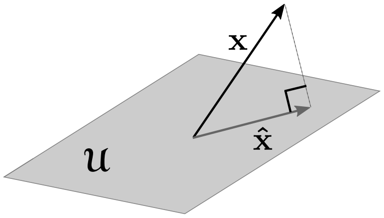


Figure 1.3: Projection $\hat{\mathbf{x}}$ of vector $\mathbf{x}$ onto subspace $\mathcal{U}$.

Proposition 1.2. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{R}\}$ be an orthonormal basis of $\mathcal{U}$. The projection of $\mathbf{x} \in \mathcal{X}$ onto $\mathcal{U}$ is given by

$$\hat{\mathbf{x}} \;=\; \sum_{\mathrm{r}=1}^{\mathrm{R}} \langle \mathbf{x}, \mathbf{u}_\mathrm{r} \rangle \mathbf{u}_\mathrm{r}.$$

In other words, the coefficient of $\mathbf{x}$ w.r.t. $\mathbf{u}_\mathrm{r}$ is given by $\langle \mathbf{x}, \mathbf{u}_\mathrm{r} \rangle$.

Furthermore, the following proposition tells us that we can compute projections very efficiently: just using a simple matrix multiplication! This makes projections very attractive in practice. For example, as we saw before, data often lies near subspaces. We can measure how close using the norm of the *residual* $\mathbf{x} - \hat{\mathbf{x}}$.

Proposition 1.3 (Projector operator)**.** Let $\mathbf{U} \in \mathbb{R}^{\mathrm{D} \times \mathrm{R}}$ be a basis of $\mathcal{U}$. The *projection operator* $\mathbf{P}_\mathbf{U} : \mathcal{X} \to \mathcal{U}$ that maps any vector $\mathbf{x} \in \mathcal{X}$ to its projection $\hat{\mathbf{x}} \in \mathcal{U}$ is given by:

$$\mathbf{P}_\mathbf{U} \;=\; \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}.$$

Notice that if $\mathbf{U}$ is orthonormal, then $\mathbf{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^\mathsf{T}$.

Proof. Since $\hat{\mathbf{x}} \in \mathcal{U}$, that means we can write $\hat{\mathbf{x}}$ as $\mathbf{U}\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^\mathrm{R}$. We thus want to find the $\mathbf{c}$ that minimizes:

$$\|\mathbf{x} - \mathbf{U}\mathbf{c}\|_2^2 \;=\; (\mathbf{x} - \mathbf{U}\mathbf{c})^\mathsf{T}(\mathbf{x} - \mathbf{U}\mathbf{c}) \;=\; \mathbf{x}^\mathsf{T} - 2\mathbf{c}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{x} + \mathbf{c}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{c}.$$

Since this is convex in $\mathbf{c}$, we can use elemental optimization to find the desired minimizer, i.e., we will take derivative w.r.t. $\mathbf{c}$, set to zero and solve for $\mathbf{c}$. To learn more about how to take derivatives w.r.t. vectors and matrices see *Old and new matrix algebra useful for statistics* by Thomas P. Minka. The derivative w.r.t. $\mathbf{c}$ is given by:

$$-2\mathbf{U}^\mathsf{T}\mathbf{x} + 2\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{c}.$$

Setting to zero and solving for $\mathbf{c}$ we obtain:

$$\hat{\mathbf{c}} \;:=\; \operatorname*{arg\,min}_{\mathbf{c} \in \mathbb{R}^\mathrm{R}} \; \|\mathbf{x} - \mathbf{U}\mathbf{c}\|_2^2 \;=\; (\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}\mathbf{x},$$

where we know $\mathbf{U}^\mathsf{T}\mathbf{U}$ is invertible because $\mathbf{U}$ is a basis by assumption, so its columns are linearly independent. It follows that

$$\hat{\mathbf{x}} \;=\; \mathbf{U}\hat{\mathbf{c}} \;=\; \underbrace{\mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}}_{\mathbf{P_U}}\mathbf{x},$$

as claimed. If $\mathbf{U}$ is orthonormal, then $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$, and hence $\mathbf{P_U}$ simplifies to $\mathbf{U}\mathbf{U}^\mathsf{T}$. Notice that $\hat{\mathbf{c}}$ are the coefficients of $\hat{\mathbf{x}}$ w.r.t. the basis $\mathbf{U}$. □

1.6 Gram-Schmidt Orthogonalization

Orthonormal bases have very nice and useful properties. For example, in Proposition 1.3, if the basis $\mathbf{U}$ is orthonormal, then the projection operator is simplified into $\mathbf{U}\mathbf{U}^\mathsf{T}$, which requires much less computations than $\mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$. The following procedure tells us exactly how to transform an arbitrary basis into an orthonormal basis.

Proposition 1.4 (Gram-Schmidt procedure)**.** Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{R}\}$ be a basis of $\mathcal{U}$. Let

$$\begin{aligned} \mathbf{v}'_\mathrm{r} \;&=\; \begin{cases} \mathbf{u}_1 & \mathrm{r} = 1, \\ \mathbf{u}_\mathrm{r} - \sum_{\mathrm{k}=1}^{\mathrm{r}-1} \langle \mathbf{u}_\mathrm{r}, \mathbf{v}_\mathrm{k} \rangle \mathbf{v}_\mathrm{k} & \mathrm{r} = 2, \ldots, \mathrm{R}, \end{cases} \\ \mathbf{v}_\mathrm{r} \;&=\; \mathbf{v}'_\mathrm{r} / \|\mathbf{v}'_\mathrm{r}\|. \end{aligned}$$

Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{R}\}$ are an orthonormal basis of $\mathcal{U}$.

Proof. We know from Proposition 1.2 that $\sum_{\mathrm{k}=1}^{\mathrm{r}-1} \langle \mathbf{u}_\mathrm{r}, \mathbf{v}_\mathrm{k} \rangle \mathbf{v}_\mathrm{k}$ is the projection of $\mathbf{u}_\mathrm{r}$ onto $\mathrm{span}[\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}]$. This implies $\mathbf{v}'_\mathrm{r}$ is the orthogonal residual of $\mathbf{u}_\mathrm{r}$ onto $\mathrm{span}[\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}]$, and hence it is orthogonal to $\{\mathbf{v}_1, \ldots, \mathbf{v}_{\mathrm{r}-1}\}$, as desired. $\mathbf{v}_\mathrm{r}$ is simply the normalized version of $\mathbf{v}'_\mathrm{r}$. □

1.7 Singular Value Decomposition (SVD)

We often store a collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{N} \in \mathbb{R}^\mathrm{D}$ as columns in an $\mathrm{D} \times \mathrm{N}$ matrix $\mathbf{X}$. These vectors may be skewed towards certain directions, often known as *principal components*. The singular value decomposition produces a basis of $\mathbb{R}^\mathrm{D}$ containing such directions.

Proposition 1.5 (Singular value decomposition)**.** Let $\mathbf{X} \in \mathbb{R}^{\mathrm{D}\times\mathrm{N}}$. Then there exist matrices with orthonormal columns $\mathbf{U} \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$ and $\mathbf{V} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ and a matrix $\mathbf{\Sigma} \in \mathbb{R}^{\mathrm{D}\times\mathrm{N}}$ of the form:

$$\mathbf{\Sigma} = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_\mathrm{D} & \end{array}\right] \qquad \text{or} \qquad \mathbf{\Sigma} = \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_\mathrm{N} \\ \hline & \mathbf{0} & \end{array}\right],$$

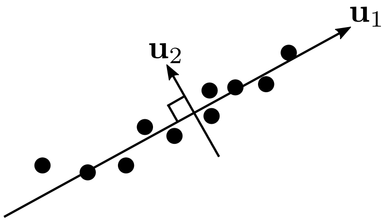


Figure 1.4: Principal components, $\mathbf{u}_1$ and $\mathbf{u}_2$ of a dataset.

depending on whether $\mathrm{D} \leq \mathrm{N}$ or $\mathrm{D} > \mathrm{N}$, such that

$$\mathbf{X} \;=\; \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}.$$

The first R columns in $\mathbf{U}$, often called the left-singular vectors, span the R-dimensional subspace that captures most of the variance in $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{N}\}$. σ_r captures the variance in the r^{th} direction, and the i^{th} column of $\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}$ gives the coefficients of $\mathbf{x}_\mathrm{i}$ w.r.t. the basis in $\mathbf{U}$.