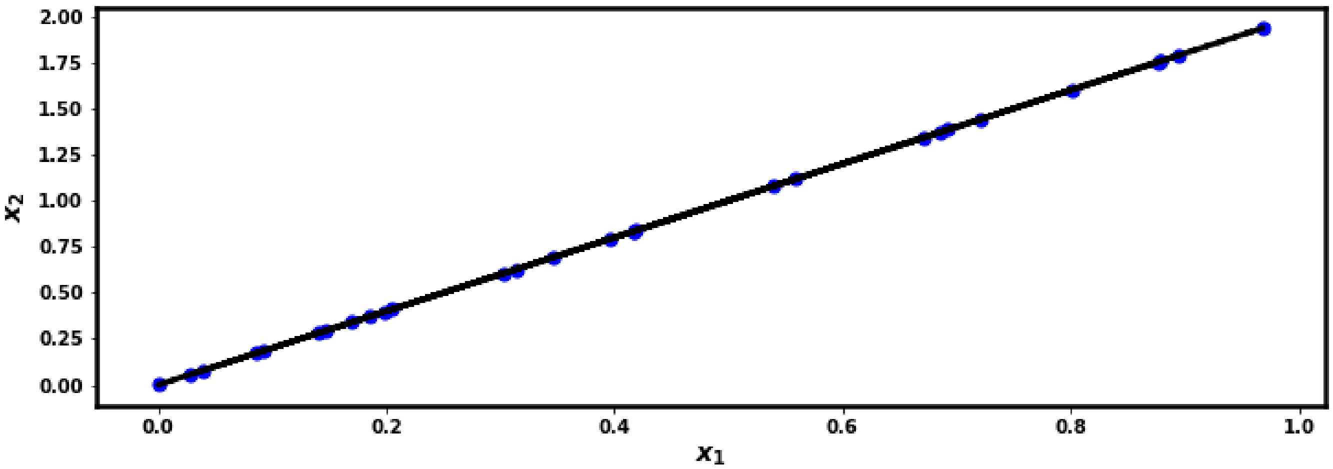
# Correlated features

Consider the following set of examples with $2$ features

## Two features: perfect correlation



As you can see

- $\backslash\mathbf{x}_2$ is perfectly correlated with $\backslash\mathbf{x}_1$

$$\backslash\mathbf{x}_2^{\backslash\mathrm{ip}} = 2 * \backslash\mathbf{x}_1^{\backslash\mathrm{ip}}$$

**Linear algebra**

A way to conceptualize $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$

- As a point in the space spanned by unit basis vectors parallel to the horizontal and vertical axes.

$$\backslash\mathbf{u}_{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$
$$\backslash\mathbf{u}_{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- With $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ having exposure

$$\backslash\mathbf{x}^{\backslash\mathrm{ip}}_1 \text{ to } \backslash\mathbf{u}_{(1)}$$
$$\backslash\mathbf{x}^{\backslash\mathrm{ip}}_2 \text{ to } \backslash\mathbf{u}_{(2)}$$

So example $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ is

For example

$$
\begin{aligned}
\backslash\mathbf{x}^{\backslash\mathrm{ip}} &= \begin{pmatrix} 3 \\ 6 \end{pmatrix} \\
&= 3 * \backslash\mathbf{u}_{(1)} + 6 * \backslash\mathbf{u}_{(2)} \\
&= 3 * \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 6 * \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} 3 \\ 6 \end{pmatrix}
\end{aligned}
$$

That is:

- Our feature space is defined by the basis vectors ("axes")

$$\backslash\mathbf{u}_{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\backslash\mathbf{u}_{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ describes a point in the span of the basis vectors
  - $\backslash\mathbf{x}_1^{\backslash\mathrm{ip}}$ is the displacement of observation $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ along basis vector $\backslash\mathbf{u}_{(1)}$
  - $\backslash\mathbf{x}_2^{\backslash\mathrm{ip}}$ is the displacement of observation $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ along basis vector $\backslash\mathbf{u}_{(2)}$
- In general, for any length $n$ vector of features

$$\backslash\mathbf{x}^{\backslash\mathrm{ip}} = \sum_{j'=1}^{n} \backslash\mathbf{x}_{j'}^{\backslash\mathrm{ip}} * \backslash\mathbf{u}_{(j')}$$

One could easily imagine a *different* set of basis vectors to describe the feature space

- For example: a rotation of basis vectors $\backslash\mathbf{u}_{(1)}, \ldots, \backslash\mathbf{u}_{(n)}$
- Let this alternate set of basis vectors be denoted by $\backslash\tilde{\mathbf{v}}_{(1)}, \ldots, \backslash\tilde{\mathbf{v}}_{(n)}$
- The basis vectors are mutually orthogonal

$$\backslash\tilde{\mathbf{v}}_{(1)} \cdot \backslash\tilde{\mathbf{v}}_{(2)} = 0$$

In the new basis space, example $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ has co-ordinates $\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}$

$$\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}} = \sum_{j'=1}^{n} \backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}_{j'} * \backslash\tilde{\mathbf{v}}_{(j')}$$

PCA is a technique for finding particularly interesting alternate basis vectors.

The alternate basis is motivated by the fact that, for a given set of examples, there may be pairwise correlation among features.

- If the correlation is *perfect* for some pair of features, they are redundant
    - May drop one feature

Consider the set of examples above. Features 1 and 2 are perfectly correlated.

$$\backslash\mathbf{x}_2^{\backslash\mathrm{ip}} = 2 * \backslash\mathbf{x}_1^{\backslash\mathrm{ip}}$$

We can create an *alternate* basis vector (no longer parallel to the axes)

$$\backslash\tilde{\mathbf{v}}_{(1)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

such that example $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ has coordinates $\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}$

$$\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}} = \backslash\tilde{\mathbf{x}}_1^{\backslash\mathrm{ip}} * \backslash\tilde{\mathbf{v}}_{(1)}$$

Note that this alternate basis has only $1$ basis vector, rather than the $2$ basis vectors of the original representation.

For example

That is, $\x^{\ip}$ has exposure $\tilde{\x}_1^{\ip}$ to the new, single basis vector.

So

- Rather than representing $\x^{\ip}$ as a vector with 2 features
  - in the original basis
    - one basis vector per *raw* feature
    - mutually orthogonal basis vectors
- We can represent it as $\tilde{\x}^{\ip}$, a vector with 1 feature
  - in the new basis
  - which captures the correlation in 2 of the raw features when measured in the original basis

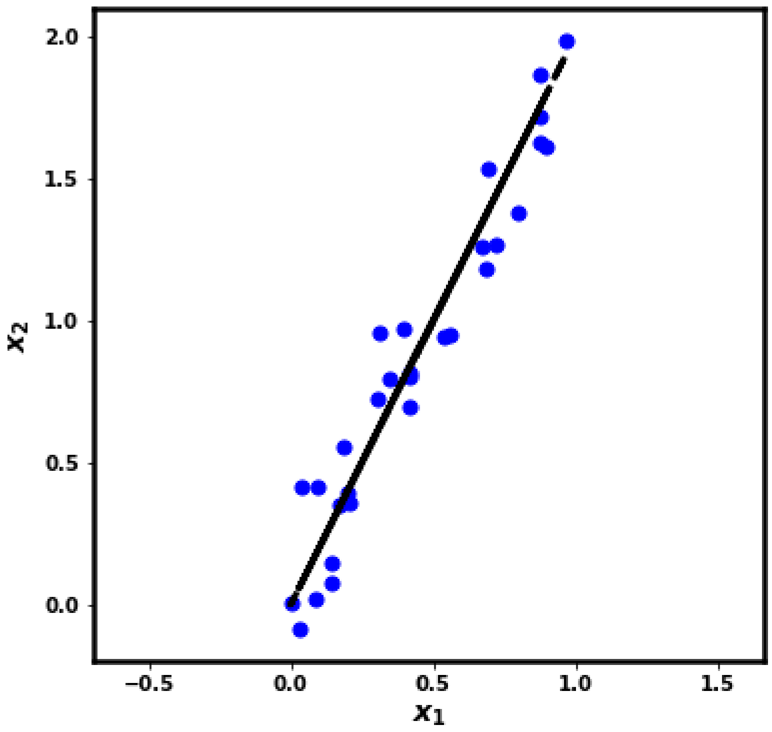
This is the essence of dimensionality reduction

- Changing bases to one with fewer basis vectors

It is rarely the case for features to be perfectly correlated

Let's modify the set of examples just a bit.

## Two features: imperfect correlation



The single basis vector (black line)

- is insufficient to correctly capture each example
- *error*: displacement from black line

In order to eliminate the error, we add a *second* basis vector

- orthogonal to the first

So now $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ (measured in original basis) can be represented as $\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}$ (measured in new basis)

- where
  - $\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}_{(1)}$ measures the displacement along the first basis vector $\backslash\tilde{\mathbf{v}}_1$
  - $\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}_{(2)}$ measures the displacement along the second basis vector $\backslash\tilde{\mathbf{v}}_2$

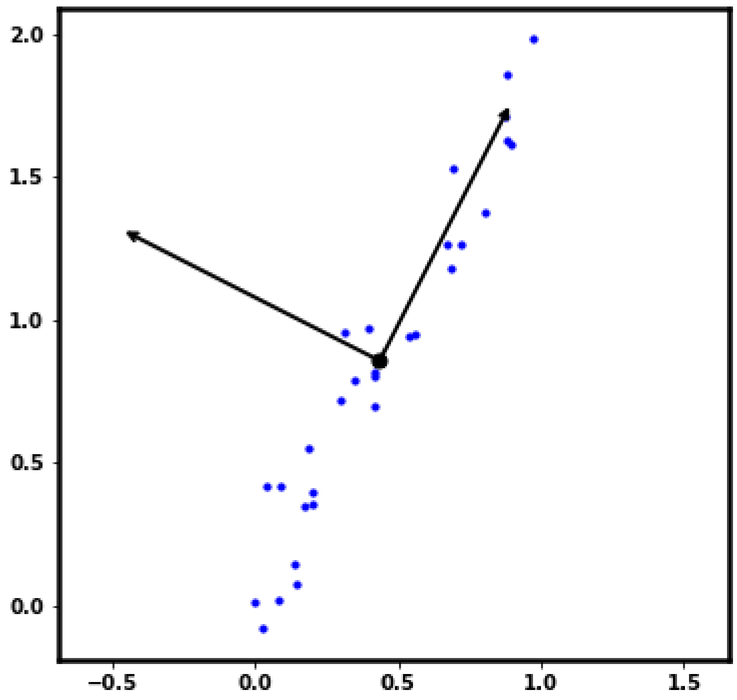
$$\backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}} = \sum_{j'=1}^{2} \backslash\tilde{\mathbf{x}}^{\backslash\mathrm{ip}}_{j'} * \backslash\tilde{\mathbf{v}}_{(j')}$$

- The dark black line in the diagram above is the first alternate basis vector $\backslash\tilde{\mathbf{v}}_{(1)}$

In the diagram below, we add a second basis vector $\backslash\tilde{\mathbf{v}}_{(2)}$

- orthogonal to the first

## Two features: imperfect correlation, alternate basis



As you can see:

- The variation along $\backslash\tilde{\mathbf{v}}_{(1)}$ is much greater than that around $\backslash\tilde{\mathbf{v}}_{(2)}$
- Capturing the notion that the "main" relationship is along $\backslash\tilde{\mathbf{u}}_{(1)}$

In fact, if we dropped $\backslash\tilde{\mathbf{v}}_{(2)}$ such that $||\backslash\tilde{\mathbf{x}}|| = 1$

- The examples would be projected onto the line $\backslash\tilde{\mathbf{v}}_{(1)}$
- With little information being lost

PCA finds alternate basis vectors and *orders them* in order of decreasing variation.

# Subsets of correlated features

It may not be the case that a group of features is correlated across *all* examples

Consider our "equity factor model"

- consider two subsets of examples: stocks in/not in the "tech" sector
- all stocks in the first/second subset have the same loading on the "tech" factor (1/0)
- so there is correlation *within* the subsets but *not between* the subsets

Consider the MNIST digits

- The subset of examples corresponding to the digit "1"
- Have a particular set of correlated features (forming a vertical column of pixels near the middle of the image)
- Which *may not* be correlated with the same features in examples corresponding to *other* digits

Thus, a synthetic feature encodes a "concept" that occurs in many but not all examples

We will present a method to *discover* "concepts"

- It may not necessarily be the pattern of features that corresponds to an entire digit
- It may be a partial pattern common to several digits
    - Vertical band (0, 1, 4, 7)
    - Horizontal band at top (5, 7, 9)

In [5]:

```
print("Done")
```

Done