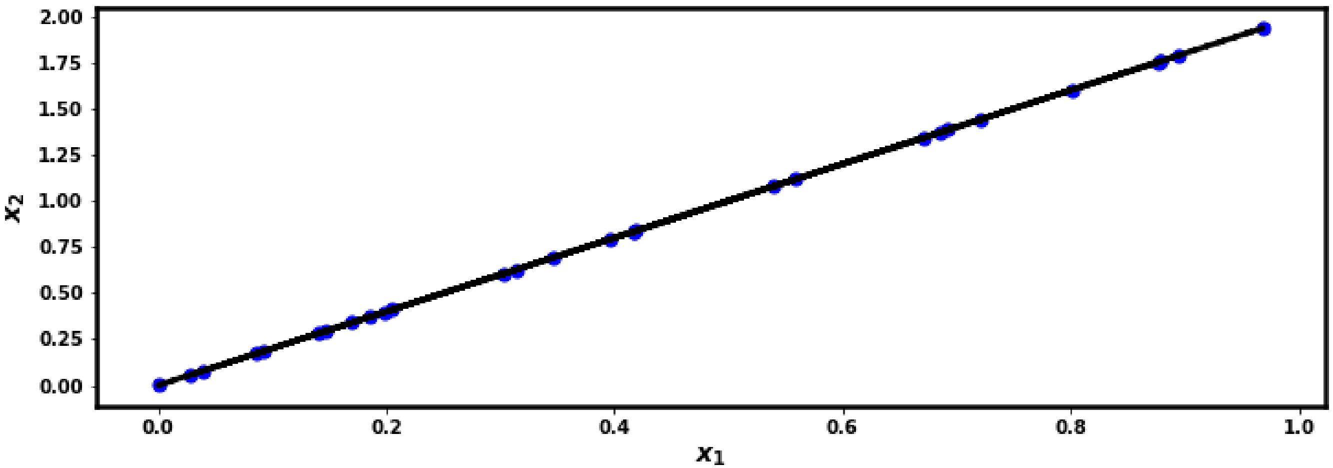
Correlated features

Consider the following set of examples with 2 features

Two features: perfect correlation



As you can see

- $\mathbf{x}_2$ is perfectly correlated with $\mathbf{x}_1$
$$\mathbf{x}_2^{(\mathbf{i})} = 2 * \mathbf{x}_1^{(\mathbf{i})}$$

Linear algebra

A way to conceptualize $\mathbf{x^{(i)}}$

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

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

- With $\mathbf{x^{(i)}}$ having exposure

$$
\begin{aligned}
\mathbf{x}^{(\mathbf{i})}_1 & \text{ to } \mathbf{u}_{(1)} \\
\mathbf{x}^{(\mathbf{i})}_2 & \text{ to } \mathbf{u}_{(2)}
\end{aligned}
$$

So example $\mathbf{x^{(i)}}$ is

For example

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

That is:

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

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

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

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

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

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

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

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

In the new basis space, example $\mathbf{x}^{\mathbf{(i)}}$ has co-ordinates $\tilde{\mathbf{x}}^{\mathbf{(i)}}$

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

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.

$$\mathbf{x}_2^{(\mathbf{i})} = 2 * \mathbf{x}_1^{(\mathbf{i})}$$

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

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

such that example $\mathbf{x}^{(\mathbf{i})}$ has coordinates $\tilde{\mathbf{x}}^{(\mathbf{i})}$

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

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

For example

$$
\begin{aligned}
\mathbf{x}^{(\mathbf{i})} &= \begin{pmatrix} 3 \\ 6 \end{pmatrix} \\
\\
\tilde{\mathbf{x}}^{(\mathbf{i})} &= (3) && \text{co-ordinates in alternate basis } \tilde{\mathbf{v}}_{(1)} \\
&= 3 * \tilde{\mathbf{v}}_{(1)} \\
&= 3 * \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\
\\
&= \begin{pmatrix} 3 \\ 6 \end{pmatrix}
\end{aligned}
$$

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

So

- Rather than representing $\mathbf{x}^{(\mathbf{i})}$ as a vector with 2 features (in the original basis)
- We can represent it as $\tilde{\mathbf{x}}^{(\mathbf{i})}$, a vector with 1 feature (in the new basis)

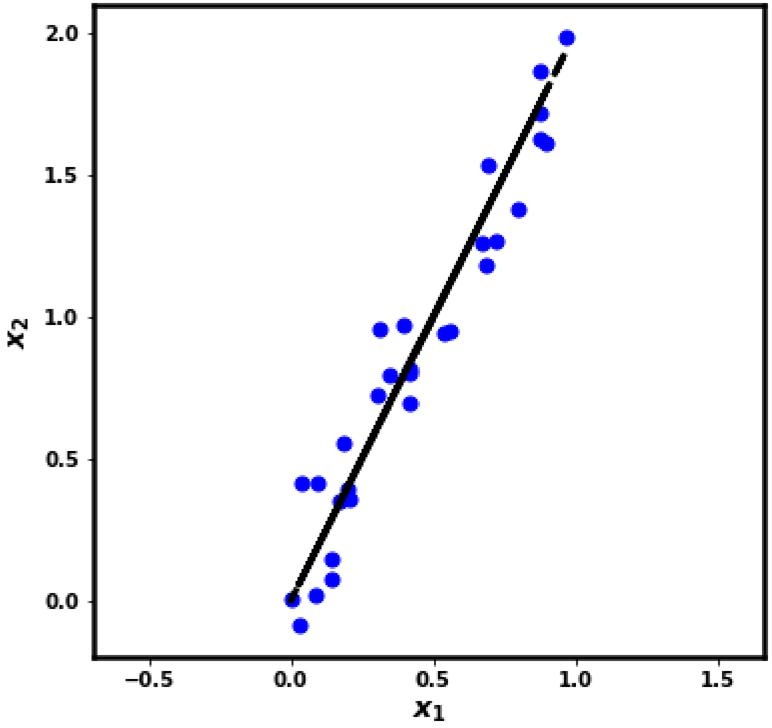
This is the essence of dimensionality reduction

- 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



We can still find an alternate basis of 2 vectors to perfectly describe the set of examples.

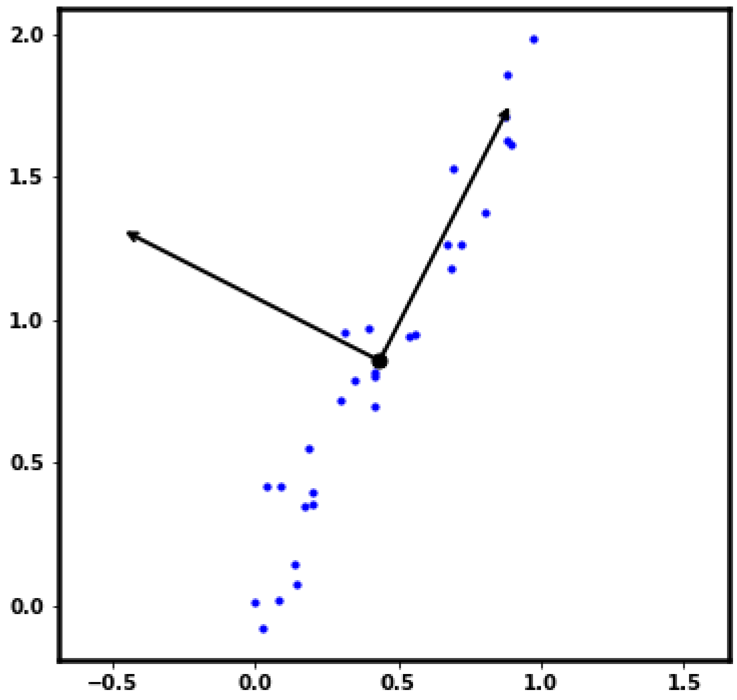
$$\tilde{\mathbf{x}}^{\mathbf{(i)}} = \sum_{j'=1}^{2} \tilde{\mathbf{x}}^{\mathbf{(i)}}_{j'} * \tilde{\mathbf{v}}_{(j')}$$

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

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

- orthogonal to the first

Two features: imperfect correlation, alternate basis



As you can see:

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

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

- The examples would be projected onto the line $\tilde{\mathbf{v}}_{(1)}$
- 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 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
```