Informative projections

Dimensionality reduction

Why reduce the number of features in a data set?

1. It reduces storage and computation time.
2. High-dimensional data often has a lot of redundancy.
3. Remove noisy or irrelevant features.

Example: are all the pixels in an image equally informative?

If we were to choose a few pixels to discard, which would be the prime candidates?
Eliminating low variance coordinates

MNIST: what fraction of the total variance lies in the 100 (or 200, or 300) coordinates with lowest variance?

The effect of correlation

Suppose we wanted just one feature for the following data.

This is the direction of maximum variance.
Comparing projections

**Projection: formally**

What is the projection of $x \in \mathbb{R}^d$ in the **direction** $u \in \mathbb{R}^d$? Assume $u$ is a unit vector (i.e. $\|u\| = 1$).

Projection is

$$x \cdot u = u \cdot x = u^T x = \sum_{i=1}^{d} u_i x_i.$$
Examples

What is the projection of \( x = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \) along the following directions?

1. The \( x_1 \)-axis?
2. The direction of \( \begin{pmatrix} 1 \\ -1 \end{pmatrix} \)?

The best direction

Suppose we need to map our data \( x \in \mathbb{R}^d \) into just one dimension:

\[
x \mapsto u \cdot x \quad \text{for some unit direction } u \in \mathbb{R}^d
\]

What is the direction \( u \) of maximum variance?

Useful fact 1:
- Let \( \Sigma \) be the \( d \times d \) covariance matrix of \( X \).
- The variance of \( X \) in direction \( u \) (the variance of \( X \cdot u \)) is:

\[
u^T \Sigma u.
\]
Best direction: example

Here covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.85 \\ 0.85 & 1 \end{pmatrix}$

The best direction

Suppose we need to map our data $x \in \mathbb{R}^d$ into just one dimension:

$$x \mapsto u \cdot x \quad \text{for some unit direction } u \in \mathbb{R}^d$$

What is the direction $u$ of maximum variance?

Useful fact 1:
- Let $\Sigma$ be the $d \times d$ covariance matrix of $X$.
- The variance of $X$ in direction $u$ is given by $u^T \Sigma u$.

Useful fact 2:
- $u^T \Sigma u$ is maximized by setting $u$ to the first eigenvector of $\Sigma$.
- The maximum value is the corresponding eigenvalue.
Best direction: example

Direction: first eigenvector of the $2 \times 2$ covariance matrix of the data.

Projection onto this direction:
the top principal component of the data

Projection onto multiple directions

Projecting $x \in \mathbb{R}^d$ into the $k$-dimensional subspace defined by vectors $u_1, \ldots, u_k \in \mathbb{R}^d$.

This is easiest when the $u_i$’s are orthonormal:

- They have length one.
- They are at right angles to each other: $u_i \cdot u_j = 0$ when $i \neq j$

The projection is a $k$-dimensional vector:

$$
(x \cdot u_1, x \cdot u_2, \ldots, x \cdot u_k) = \begin{pmatrix}
\langle u_1 \rangle \\
\langle u_2 \rangle \\
\vdots \\
\langle u_k \rangle \\
\end{pmatrix}
\begin{pmatrix}
x
\end{pmatrix}
$$

U is the $d \times k$ matrix with columns $u_1, \ldots, u_k$. 

Projection onto multiple directions: example

E.g. project data in $\mathbb{R}^4$ onto the first two coordinates.

Take vectors $u_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, $u_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ (notice: orthonormal)

Write $U^T = \begin{pmatrix} u_1 & u_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$

The projection of $x \in \mathbb{R}^4$ is $U^T x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

The best $k$-dimensional projection

Let $\Sigma$ be the $d \times d$ covariance matrix of $X$. In $O(d^3)$ time, we can compute its eigendecomposition, consisting of

- real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
- corresponding eigenvectors $u_1, \ldots, u_d \in \mathbb{R}^d$ that are orthonormal (unit length and at right angles to each other)

Fact: Suppose we want to map data $X \in \mathbb{R}^d$ to just $k$ dimensions, while capturing as much of the variance of $X$ as possible. The best choice of projection is:

$$x \mapsto (u_1 \cdot x, u_2 \cdot x, \ldots, u_k \cdot x),$$

where $u_i$ are the eigenvectors described above.

This projection is called principal component analysis (PCA).
Example: MNIST

Contrast coordinate projections with PCA:

Applying PCA to MNIST: examples

Reconstruct this original image from its PCA projection to $k$ dimensions.

$k = 200$  $k = 150$  $k = 100$  $k = 50$

How do we get these reconstructions?
Reconstruction from a 1-d projection

Projection onto $\mathbb{R}$:

Reconstruction in $\mathbb{R}^2$:

Reconstruction from multiple projections

Projecting into the $k$-dimensional subspace defined by orthonormal $u_1, \ldots, u_k \in \mathbb{R}^d$.

The projection of $x$ is a $k$-dimensional vector:

\[
(x \cdot u_1, x \cdot u_2, \ldots, x \cdot u_k) = \begin{pmatrix}
\langle u_1 \rangle & \langle u_2 \rangle & \cdots & \langle u_k \rangle
\end{pmatrix}
\begin{pmatrix}
x
\end{pmatrix}
\]

The reconstruction from this projection is:

\[
(x \cdot u_1)u_1 + (x \cdot u_2)u_2 + \cdots + (x \cdot u_k)u_k = UU^T x.
\]
MNIST: image reconstruction

Reconstruct this original image $x$ from its PCA projection to $k$ dimensions.

$\begin{align*}
  k = 200 & \quad k = 150 & \quad k = 100 & \quad k = 50 \\
  \begin{array}{c}
  \includegraphics[width=0.2\textwidth]{image1}
  \\
  \includegraphics[width=0.2\textwidth]{image2}
  \\
  \includegraphics[width=0.2\textwidth]{image3}
  \\
  \includegraphics[width=0.2\textwidth]{image4}
  \\end{array}
\end{align*}$

Reconstruction $UU^T x$, where $U$'s columns are top $k$ eigenvectors of $\Sigma$.

Linear algebra: eigenvalues and eigenvectors
The linear function defined by a matrix

- Any matrix $M$ defines a linear function, $x \mapsto Mx$. If $M$ is a $d \times d$ matrix, this maps $\mathbb{R}^d$ to $\mathbb{R}^d$.

- This function is easy to understand when $M$ is diagonal:

$$
\begin{pmatrix}
2 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 10
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix}
= 
\begin{pmatrix}
2x_1 \\
-x_2 \\
10x_3
\end{pmatrix}
$$

In this case, $M$ simply scales each coordinate separately.

- General symmetric matrices also just scale coordinates separately... but in a different coordinate system!

Eigenvector and eigenvalue: definition

Let $M$ be a $d \times d$ matrix. We say $u \in \mathbb{R}^d$ is an eigenvector of $M$ if

$$
Mu = \lambda u
$$

for some scaling constant $\lambda$. This $\lambda$ is the eigenvalue associated with $u$.

Key point: $M$ maps eigenvector $u$ onto the same direction.
Question: What are the eigenvectors and eigenvalues of:

\[
M = \begin{pmatrix}
2 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 10
\end{pmatrix}
\]

**Eigenvectors of a real symmetric matrix**

**Fact:** Let \( M \) be any real symmetric \( d \times d \) matrix. Then \( M \) has

- \( d \) eigenvalues \( \lambda_1, \ldots, \lambda_d \)
- corresponding eigenvectors \( u_1, \ldots, u_d \in \mathbb{R}^d \) that are orthonormal

Can think of \( u_1, \ldots, u_d \) as the axes of the natural coordinate system for \( M \).
Example

\[ M = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \] has eigenvectors \( u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \)

1. Are these orthonormal?
2. What are the corresponding eigenvalues?

Spectral decomposition

**Fact:** Let \( M \) be any real symmetric \( d \times d \) matrix. Then \( M \) has orthonormal eigenvectors \( u_1, \ldots, u_d \in \mathbb{R}^d \) and corresponding eigenvalues \( \lambda_1, \ldots, \lambda_d \).

**Spectral decomposition:** Another way to write \( M \):

\[
M = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix} \quad \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix}
\]

Thus \( Mx = U\Lambda U^T x \):

- \( U^T \) rewrites \( x \) in the \( \{u_i\} \) coordinate system
- \( \Lambda \) is a simple coordinate scaling in that basis
- \( U \) sends the scaled vector back into the usual coordinate basis
Apply spectral decomposition to the matrix we saw earlier:

\[ M = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \]

- Eigenvectors \( u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \)

- Eigenvalues \( \lambda_1 = -1, \quad \lambda_2 = 3. \)

\[ M = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \frac{-1}{0} \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \]

\[ M \begin{pmatrix} 1 \\ 2 \end{pmatrix} = U \Lambda U^T \begin{pmatrix} 1 \\ 2 \end{pmatrix} \]
Principal component analysis revisited

Data vectors $X \in \mathbb{R}^d$

- $d \times d$ covariance matrix $\Sigma$ is symmetric.
- Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
- Eigenvectors $u_1, \ldots, u_d$.
- $u_1, \ldots, u_d$: another basis for data.
- Variance of $X$ in direction $u_i$ is $\lambda_i$.
- Projection to $k$ dimensions:
  $x \mapsto (x \cdot u_1, \ldots, x \cdot u_k)$.

What is the covariance of the projected data?

Case study: Quantifying personality

What are the dimensions along which personalities differ?

- **Lexical hypothesis**: most important personality characteristics have become encoded in natural language.
- Allport and Odbert (1936): identified 4500 words describing personality traits.
- Group these words into (approximate) synonyms, by manual clustering.
  E.g. Norman (1967):

<table>
<thead>
<tr>
<th>Spirit</th>
<th>Jolly, merry, witty, lively, peppery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talkativeness</td>
<td>Talkative, articulate, verbose, gossipy</td>
</tr>
<tr>
<td>Sociability</td>
<td>Companionable, social, outgoing</td>
</tr>
<tr>
<td>Spontaneity</td>
<td>Impulsive, carefree, playful, zany</td>
</tr>
<tr>
<td>Boisterousness</td>
<td>Mischiefous, rowdy, loud, prankish</td>
</tr>
<tr>
<td>Adventure</td>
<td>Brave, venturesome, fearless, reckless</td>
</tr>
<tr>
<td>Energy</td>
<td>Active, assertive, dominant, energetic</td>
</tr>
<tr>
<td>Conceit</td>
<td>Boastful, conceited, egotistical</td>
</tr>
<tr>
<td>Vanity</td>
<td>Affectcd, vain, chic, dapper, jaunty</td>
</tr>
<tr>
<td>Indiscipline</td>
<td>Noisy, snoopy, indecent, meddlesome</td>
</tr>
<tr>
<td>Sensuality</td>
<td>Sexy, passionate, sensual, flirtatious</td>
</tr>
</tbody>
</table>

- Data collection: subjects whether these words describe them.
Personality assessment: the data

Matrix of data (1 = strongly disagree, 5 = strongly agree)

<table>
<thead>
<tr>
<th></th>
<th>shy</th>
<th>merry</th>
<th>tense</th>
<th>boastful</th>
<th>forgiving</th>
<th>quiet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Person 2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Person 3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How to extract important directions?

- Treat each column as a data point, find tight clusters
- Treat each row as a data point, apply PCA
- Or factor analysis, independent component analysis, etc.

What would PCA accomplish?

E.g.: Suppose two traits (generosity, trust) are so highly correlated that each person either answers “1” to both or “5” to both.

A single PCA dimension would entirely account for both traits.
Personality assessment: the data

Matrix of data (1 = strongly disagree, 5 = strongly agree)

<table>
<thead>
<tr>
<th></th>
<th>shy</th>
<th>merry</th>
<th>tense</th>
<th>boastful</th>
<th>forgiving</th>
<th>quiet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Person 2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Person 3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Methodology: apply PCA to the rows of this matrix.

The “Big Five” taxonomy

**Extraversion**
- quiet (-.83), reserved (-.80), shy (-.75), silent (-.71)
+ talkative (.85), assertive (.83), active (.82), energetic (.82)

**Agreeableness**
- fault-finding (-.52), cold (-.48), unfriendly (-.45), quarrelsome (-.45)
+ sympathetic (.87), kind (.85), appreciative (.85), affectionate (.84)

**Conscientiousness**
- careless (-.58), disorderly (-.53), frivolous (-.50), irresponsible (-.49)
+ organized (.80), thorough (.80), efficient (.78), responsible (.73)

**Neuroticism**
- stable (-.39), calm (-.35), contented (-.21)
+ tense (.73), anxious (.72), nervous (.72), moody (.71)

**Openness**
- commonplace (-.74), narrow (-.73), simple (-.67), shallow (-.55)
+ imaginative (.76), intelligent (.72), original (.73), insightful (.68)