Figure 1: A collection of points in $\mathbb{R}^d$ for $d = 3$ that live in a $k = 2$ dimensional subspace

Assume we have a collection of $n$ high-dimensional points $x_i \in \mathbb{R}^d$. For example, if these represent vectorized images of size $200 \times 200$, $d = 40000$. For computational convenience, we would like to reduce their dimensionality to $k \ll d$. We can do this with small error if the points lie on a lower-dimensional subspace.

**Centering:** First, let us center these points

$$x'_i = x_i - \mu \quad (1)$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (2)$$

**Maximize variance:** Let us find direction $u$ such that when $x'_i$ is projected into $u$, the variance of the data is maximized. Let us define $u$ to be of length 1. We can then write the projection as $u^T x'_i$.

$$\max_u \text{var}(u^T x'_i), \quad ||u|| = 1 \quad (3)$$

Interpreting the projected values as samples of a random variable, we can estimate its variance by computing the average squared deviation from its mean. Since $x'_i$ are centered, the mean $\frac{1}{n} \sum_i u^T x'_i = u^T (\frac{1}{n} \sum_i x'_i) = u^T 0 = 0$ is zero and so we can
Figure 2: Finding a direction that preserves variance

directly optimize

\[
\max_u \frac{1}{n} \sum_i (u^T x'_i - 0)^2, \quad ||u|| = 1
\]  
(4)

\[
\max_u \frac{1}{n} \sum_i (u^T x'_i)(u^T x'_i)^T, \quad ||u|| = 1
\]  
(5)

\[
\max_u \frac{1}{n} \sum_i u^T x'_i x'_i^T u, \quad ||u|| = 1
\]  
(6)

\[
\max_u u^T \left( \frac{1}{n} \sum_i x'_i x'_i^T \right) u, \quad ||u|| = 1
\]  
(7)

\[
\max_u u^T \Sigma u, \quad u^T u = 1, \quad \Sigma = \frac{1}{n} \sum_i x'_i x'_i^T
\]  
(8)

To solve (8), we use lagrange multipliers. We know that the gradient of the objective function (with respect to \(u\)) and the constraint are in the same direction. We can equivalently state they are equal up to a scale factor

\[
\nabla_u u^T \Sigma u = \lambda \nabla_u u^T u
\]  
(9)

Using rules of matrix derivatives

\[
2\Sigma u = 2\lambda u
\]  
(10)

\[
\Sigma u = \lambda u
\]  
(11)

We recognize (11) as an eigenvalue equation. We call the direction \(u\) a principle component of the data \(x_i\). Note that an alternate derivation of \(u\) is the direction that minimizes the reconstruction error between a point and its projection

In order to project to two directions, one can first find the optimal single direction from (8) and call this direction \(u_1\). To find \(u_2\), one can optimize

\[
\max_{u_2} u_2^T \Sigma u_2, \quad u_2^T v_2 = 1, u_2^T u_1 = 0
\]  
(12)
1 Summary

PCA: In PCA, one would like to find the vectors $U = [u_1, u_2, \ldots, u_k]$ that preserve the variance of the projected points. One does this by setting $u_1, u_2, \ldots, u_k$ to be the top $k$ eigenvectors (associated with the $k$ largest eigenvalues) of the covariance matrix of centered points $\Sigma = \frac{1}{n} \sum x'_i x'_i^T$.

Projection: Let $U_{d \times k} = [u_1, u_2, \ldots, u_k]$ be the matrix of the top $k$ eigenvectors. Let $X_{d \times n} = [x'_1, \ldots, x'_n]$ be the design matrix. Note that $\Sigma = \frac{1}{n} XX^T$. Let $c_i$ be the $k$ coefficients specifying the coordinates of $x_i$ when projected onto $U$. We can write $c_i = U^T x'_i$.

Reconstruction: We can reconstruct the point $x_i$ by computing the linear combination $\hat{x}_i = \mu + U c_i$.

2 Geometric interpretation

We use the following fact from linear algebra:

Let $A_{d \times d}$ be a square, symmetric matrix ($A = A^T$). Let $u_i$ be an eigenvector $A$ with eigenvalue $\lambda_i$ such that $Au_i = \lambda_i u_i$. $A$ can be decomposed into $A = U \Lambda U^T$ where $U = [u_1, u_2, \ldots, u_k]$ is an orthonormal matrix $U^T U = I$ and $\Lambda$ is a diagonal matrix with entries $\Lambda_{i,i} = \lambda_i$.

We can interpret the decomposition within the context of the linear transformation $A z = U \Lambda U^T z$ for some vector $z$. First, note that the eigenvectors $u_i$ form an orthonormal basis for $R^d$.

1. $c = U^T z$ are the coordinates of the vector $z$ when projected onto the orthonormal basis $U$.
2. $\tilde{c} = \Lambda c$ scales each coordinate $c_i$ by $\lambda_i$.
3. $U \tilde{c}$ is the “reconstruction” of $z$ with the linear combination $\sum \tilde{c}_i u_i$

![Figure 3: We can interpret the linear transformation $Az$ as a sequence of projection, scaling, and reconstruction steps.](image)
3 SVD

If \( d \) is large, it maybe difficult to compute and store \( \Sigma_{d \times d} \). We show how one can use the SVD of \( X \) to compute \( U \) without computing \( \Sigma \). Let us compute the SVD of \( X = U'\Sigma'V'^T \). We use the accents to differentiate these matrices from \( \Sigma \) and \( U \) computed earlier.

Recall the following fact:

Any (possibly non-square matrix) \( A_{m \times n} \) can be factored \( A = U\Sigma V^T \), where \( U_{m \times m} \) an orthonormal matrix \((UU^T = I)\) and \( V_{n \times n} \) is also an orthonormal matrix \( (VV^T = I) \). \( \Sigma_{m \times n} \) is a diagonal matrix with entries \( \sigma_i \geq 0 \). These \( \sigma_i \) are called singular values of \( A \).

We know that the sample covariance matrix \( \Sigma \) can be written as

\[
\Sigma = \frac{1}{n} \sum_i x_i'x_i^T 
\]

(13)

\[
= \frac{1}{n} XX^T 
\]

(14)

\[
= \frac{1}{n} (U'\Sigma'V'^T)(U'\Sigma'V'^T)^T 
\]

(15)

\[
= \frac{1}{n} (U'\Sigma'(V'^T V')\Sigma'U'^T) 
\]

(16)

\[
= \frac{1}{n} (U'\Sigma'^2U'^T) 
\]

(17)

\[
= \frac{1}{n} (U'\Sigma'^2U'^T) 
\]

(18)

**PCA-via-the-SVD:** Because we know that \( \Sigma = U\Lambda U'^T \), we conclude that \( \Lambda = \frac{1}{n} \Sigma'^2 \) and \( U = U' \). From Sec. 1, we know that we can the optimal projection directions \( u_1, u_2, \ldots, u_k \) are equal to the top \( k \) eigenvectors of \( \Sigma \). To avoid computing \( \Sigma \), which often will not fit in memory, we can set the directions to be \( u_1', u_2', \ldots, u_k' \), the top \( k \) left singular vectors of the design matrix \( X \), where \( X = U'\Sigma'V'^T \).