Principle Components Analysis (PCA)

Suppose we have a set of data \(x_1, \ldots, x_n \in \mathbb{R}^D\). For example, we could have each \(x_i\) be a 500 \times 500\ pixel image. In this case our \(D\) would be \(500^2 = 250,000\).

We would like to find a linear projection from \(\mathbb{R}^D\) to \(\mathbb{R}^M\) to maximize the variance

\[
\sum_{i=1}^{n} \| y_i - \bar{y} \|^2
\]

where \(y_i\) is the projection of \(x_i\) onto some subspace \(l\) and \(\bar{y}\) is the mean \(\frac{1}{n} \sum_{i=1}^{n} y_i\).

Last time we projected two dimensional data down onto a line (one dimensional space) defined by the unit vector \(v\). The projections were defined by \(y_i = v^T x_i\), and the sample covariance was given by

\[
S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T
\]

We defined the matrix \(A \in \mathbb{R}^{M \times D}\) using the \(M\) eigenvectors of \(S\) such that

\[
y_i = Ax_i
\]

A translation of \(l\) (space defined by \(A\)) gives the best affine subspace approximation to the data. The objective is

\[
\min \sum_{i=1}^{n} \| \xi_i - x_i \|^2
\]

where \(\xi_i\) is the projection of \(x\), i.e.

\[
\xi_i = \sum_{j=1}^{M} v_j^T (x_i - \bar{x})v_j + \bar{x}
\]  \hspace{1cm} (1)
Projection onto a Subspace

To see where expression (1) comes from, we first try to write down the expression for a projection of a vector \( x \) onto a subspace spanned by the vectors \( \{v_1, \ldots, v_n\} \).

The projection of a vector \( x \) onto the subspace spanned by the single vector \( v \) is given by \((vv^T)x\) (the outer product of \( v \) applied to \( x \)). Since matrix multiplication is associative, this expression is equal to

\[
v(v^Tx) = (v^Tx)v
\]
since \( v^Tx \in \mathbb{R} \) (the inner product of \( v \) and \( x \)).

In class we represented the projection \( u = v^Txv \) as just \( v^Tx \in \mathbb{R} \), the coefficient of \( v \) that uniquely determines \( u \). This was in order to represent the vectors that live in \( \mathbb{R}^2 \) using only one dimension of data. This reduced representation makes use of the fact that the vectors we are trying to represent all lie on the space spanned by a single vector.

In general, to project a vector \( x \) onto the subspace spanned by the vectors \( v_1, \ldots, v_M \), the projection is given by

\[
\sum_{j=1}^{M} (v_jv_j^T)x = \sum_{j=1}^{M} (v_j^Tx)v_j = (v_1^Tx)v_1 + \cdots + (v_M^Tx)v_M
\]

In expression (1), we are projecting onto the space spanned by the \( M \) eigenvectors \( \{v_1, \ldots, v_M\} \) corresponding to the largest \( M \) eigenvalues of the sample covariance matrix. We subtract \( \bar{x} \) in order to center the data points about the origin, since the eigenvectors were obtained from the centered data in the sample covariance matrix calculation \( S = \frac{1}{n} \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \).

**Question:** What if the data does not lie in a linear subspace?

Then the PCA approximation will be bad. For example, a set of data that lies on the circumference of the unit circle in \( \mathbb{R}^2 \) does not fit nicely onto any linear subspace of \( \mathbb{R}^2 \).

**Johnson Lindestrauss Lemma**

Let \( x_1, \ldots, x_n \in \mathbb{R}^D \).

Let \( \varepsilon \in (0, 1) \).

**Lemma 1:** (Johnson Lindestrauss.) There exist \( M \in \mathcal{O}(\frac{\log n}{\varepsilon^2}) \) and a linear transformation \( L: \mathbb{R}^D \rightarrow \mathbb{R}^M \) such that for all \( i, j \),

\[
(1 - \varepsilon)\|x_i - x_j\| \leq \|L(x_i) - L(x_j)\| \leq (1 + \varepsilon)\|x_i - x_j\| \quad (2)
\]
Lemma 2: Suppose $Z \sim N(0, \sigma^2_z)$ and $W \sim N(0, \sigma^2_w)$ are independent. Then we have

$$Z + W \sim N(0, \sigma^2_z + \sigma^2_w)$$

$$\alpha Z \sim N(0, \alpha^2 \sigma^2_z)$$

Proof. This can be done using moment generating functions, which is beyond the scope of this class, but can be easily looked up online for those who are interested.

Instead we present heuristic (non rigorous) argument.

If $Z$ and $W$ are both normal random variables, then it is reasonable to believe that their sum is also normally distributed. To find the parameters which define the distribution of $Z + W$, we use linearity of expectations to find

$$E(Z + W) = EZ + EW = 0 + 0 = 0$$

Since $Z$ and $W$ are independent,

$$Var(Z + W) = Var(Z) + Var(W) = \sigma^2_z + \sigma^2_w$$

So it reasonable to believe that

$$Z + W \sim N(0, \sigma^2_z + \sigma^2_w)$$

The argument for why $\alpha Z \sim N(\alpha^2 \sigma^2_z)$ is similar and uses the fact that

$$Var(\alpha Z) = \alpha^2 Var(Z)$$

Randomized construction (Johnson Lindestrauss)

Let $L(x) = \frac{1}{\sqrt{M}}Ax$ where $A$ is $M \times D$ iid entries from $N(0, 1)$. $L$ satisfies (2) with high probability.

Claim: Random projections preserve lengths. For all $x$ in our input space, we have $\|L(x)\| \approx \|x\|$.

If we preserve lengths, then we preserve lengths of differences, i.e. distances between points. We will make use of the following lemma in the proof.

Fix $x \in \mathbb{R}^D$. $L(x) = \frac{1}{\sqrt{M}}Ax$ is a random variable.

Let $A_i$ denote the $i^{th}$ row of $A$. Each $A_ix$ is the sum

$$A_i x = \sum_j A_{ij}x$$
where $A_{ij} \sim N(0, 1)$. Thus by lemma 2, $A_i x \sim N(0, \sum_j x_j^2)$. So we have

$$E[\|Ax\|^2] = E[\sum_{i=1}^M (A_i x)^2]$$

$$= \sum_{i=1}^M E[(A_i x)^2]$$

$$= \sum_{i=1}^M \sum_{j=1}^D x_j^2 = M\|x\|^2$$

The third equality follows from the fact that $A_i x \sim N(0, \sum_1^D x_j^2)$ and that $Var(A_i x) = E[(A_i x)^2] - (EA_i x)^2 = E[(A_i x)^2]$. It then follows that

$$E[\|L(x)\|^2] = \frac{1}{M} E[\|Ax\|^2] = \|x\|^2$$

So we have shown that on average, the length is preserved.