Dimensionality Reduction

Let \( x_1, \ldots, x_n \in \mathbb{R}^D \). We tried to estimate a density \( p(x) \) using histograms, but the number of bins grows on the order of \( C^D \) for some constant \( C \). For parametric estimation, if we use a multivariate gaussian model \( N(x \mid \mu, \Sigma) \) then we are estimating parameters \( \mu \in \mathbb{R}^D \) and \( \Sigma \in \mathbb{R}^{D \times D} \). Thus, the number of parameters we need to estimate gets very large as \( D \) grows.

Linear Projections

Let \( \phi : \mathbb{R}^D \to \mathbb{R}^m \) where \( m \ll D \).

1. We would like \( \phi \) to “preserve information”.
2. If \( y_i = \phi(x_i) \) we want \( y_i \) to approximate \( x_i \).

Affine Subspaces

Let \( u_1, \ldots, u_M \in \mathbb{R}^D \) be orthonormal vectors, i.e. \( \|u_i\| = 1 \) and \( u^T_i u_j = 0 \) for \( i \neq j \).

An affine space \( A \) is a space generated by linear combinations of the set of vectors \( \{u_i\}_{i=1}^M \) plus an offset \( u_0 \in \mathbb{R}^D \), i.e. a hyperplane \( A \) defined by

\[
A = \{ u_0 + a_1 u_1 + \cdots + a_M u_M : a_i \in \mathbb{R} \}
\]

An equivalent definition is the set of points \( x \in \mathbb{R}^D \) satisfying the equation

\[
x^T u_0 = C
\]

for some constant \( C \in \mathbb{R} \).

Example: Affine subspaces in \( \mathbb{R}^3 \) include lines and planes.

An orthogonal projection takes \( x \) to the closest point in \( l \), an affine subspace. The vector difference \( x - y \) is orthogonal to \( l \), where \( y \) is the projection of \( x \) onto \( l \).
Let \( x_1, \ldots, x_n \in \mathbb{R}^D \). \( x_i \) projects to \( y_i \). The goal is to find an affine subspace minimizing the error,

\[
E = \sum_{i=1}^{n} ||y_i - x_i||^2
\]

This is called the “least squares fit”.

Equivalently, we look for the affine subspaces which maximize the variance

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

Intuitively, we are just looking for a lower dimensional approximation of our data. For example, if we had a training set in \( \mathbb{R}^2 \) such that all points were contained on a line, we wouldn’t need the coordinate representation of each data point. Instead we could take the unit vector for the line and record the magnitude of each data point. This would retain all the information of the data set while reducing the number of dimensions required to represent it.

Now suppose the data doesn’t fit perfectly in a subspace. If we wanted to project down a dimension while preserving as much information as possible, then the least squares objective is the most immediately intuitive one. But the objective we choose to pursue in the following example is the one which maximizes the variance in the affine subspace. This is still a sensible approach, since we want to mimic the structure of the original data in a lower dimensional subspace, so we want the projected data points to be as spread out as possible in order to mimic the true spread.

**Example: Finding a 1 Dimensional Subspace**

Suppose we are given \( x_1, \ldots, x_n \in \mathbb{R}^D \). Find the “best” linear subspace of dimension 1. This is a line that passes through the origin, defined by a single vector \( u \in \mathbb{R}^D \) such that \( ||u|| = 1 \). So this subspace \( l \) is

\[
l_u = \{ au : a \in \mathbb{R} \}
\]

To find “the best line”, we look for a line that maximizes \( \sigma^2 \).

We represent \( y_i \in \mathbb{R}^2 \) (the projection of \( x_i \) onto the line \( l_u \)) by its magnitude since all \( y_i \)'s are on the line \( l_u \). So we can represent them as \( y_i = u^T x_i = ||u|| ||x_i|| \cos \theta = ||x_i|| \cos \theta \), where \( \theta \) is the angle between \( u \) and \( x_i \). (Draw a picture to see that this is the \( \pm \) magnitude of the projection of \( x_i \) onto \( l_u \), i.e. the \( \pm \) magnitude of \( y_i \)).

The mean of the projected data is

\[
\bar{y} = \frac{1}{n} \sum_i y_i = \frac{1}{n} \sum_i u^T x_i = u^T \frac{1}{n} \sum_i x_i = u^T \bar{x}
\]
And the variance is

\[ \sigma^2 = \frac{1}{n} \sum_i (y_i - \mu)^2 \]

\[ = \frac{1}{n} \sum_i (u^T x_i - u^T \bar{x})^2 \]

\[ = \frac{1}{n} \sum_i (u^T (x_i - \bar{x}))^2 \]

\[ = \frac{1}{n} \sum_i u^T (x - \bar{x})(x_i - \bar{x})^T u \]

\[ = u^T Su \]

where \( S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \) (this is the sample data covariance matrix). So we want to find \( u \) maximizing \( u^T Su \) with \( ||u|| = 1 \).

This problem can be restated as

maximize \( f(u) \) subject to \( g(u) = 1 \)

where

\[ f(u) \doteq u^T Su \]

\[ g(u) \doteq u^T u \]

Using the method of Lagrange Multipliers.

\[ \nabla f = \lambda \nabla g \Rightarrow 2Su = \lambda 2u \Rightarrow Su = \lambda u \]

for \( \lambda \in \mathbb{R} \). Notice that \( u \) is an eigenvector of \( S \) with corresponding eigenvalue \( \lambda \).

Recall that our objective was to choose \( u \) to maximize variance. The variance \( \sigma^2 \) is

\[ \sigma^2 = u^T Su = u^T \lambda u = \lambda u^T u = \lambda \]

Thus we should choose the eigenvector with the largest corresponding eigenvalue. Note that \( \sigma^2 \) doesn’t change if we translate \( l \).

**Generalization**

To find the best subspace of dimension \( M \), take the \( M \) eigenvectors of \( S \) with largest eigenvalues, where \( S \) is the data covariance matrix.

This method of data reduction is called **PCA**, or **Principle Components Analysis**.