Principal component analysis

Principal component analysis is essentially boiling down multidimensional data with a lot of dimensions (aka columns) into a few dimensions while keeping **most** of the information.

Given n m-dimensional vectors, steps to find the top k principal components:

1. Calculate the component-wise average of all of the vectors $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i $
2. Form $m \times m$ matrix $S = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T $
3. Calculate the m-dimensional eigenvectors associated with the largest k eigenvalues of S: $v_1, \ldots v_k$ associated with $\lambda_1, \ldots, \lambda_k$
4. The k dimensional representation of x_i is then $\hat{x}_i = (x_i^Tv_1, \ldots x_i^Tv_k)$

Another way to state the objective:

$$ \min \sum_{i=1}^n || x_i - \hat{x}_i ||^2 $$

$$ \max \sum_{i=1}^n || \hat{x}_i ||^2 $$