Principal Components

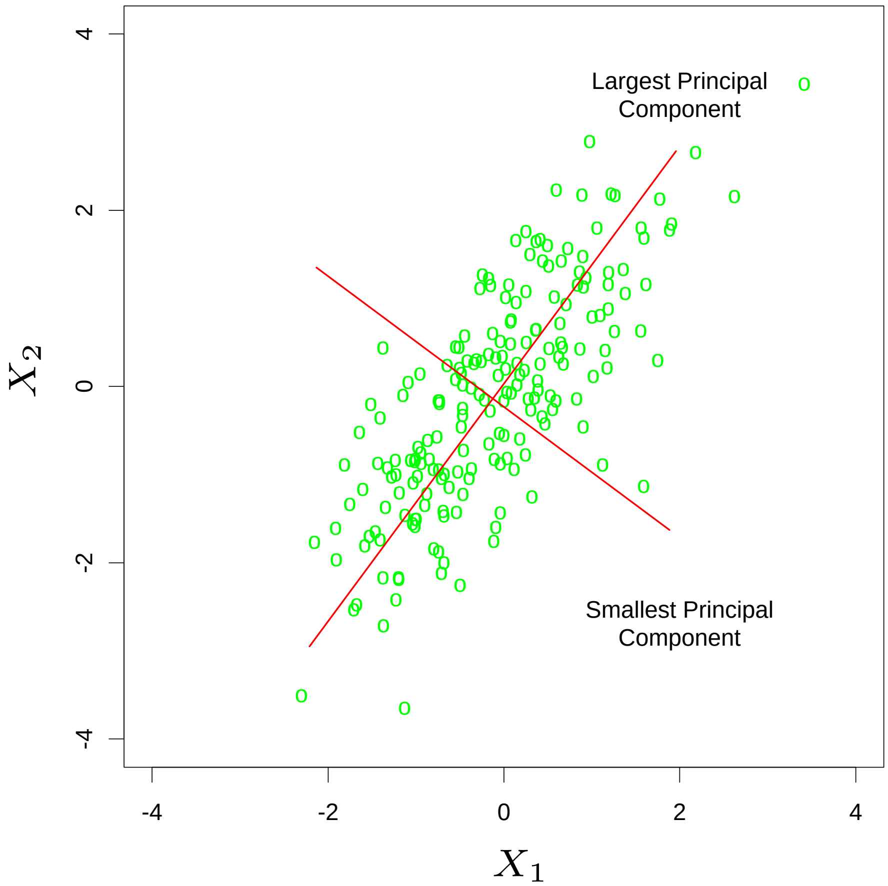
Suppose we have N measurements on each of p variables $X_j,\ j = 1, \ldots, p$. There are several equivalent approaches to principal components:

- Given $X = (X_1, \ldots X_p)$, produce a derived (and small) set of uncorrelated variables $Z_k = X\alpha_k,\ k = 1, \ldots, q < p$ that are linear combinations of the original variables, and that explain most of the variation in the original set.

- Approximate the original set of N points in $\mathbb{R}^p$ by a least-squares optimal linear manifold of co-dimension $q < p$.

- Approximate the $N \times p$ data matrix $\mathbf{X}$ by the best rank-q matrix $\hat{\mathbf{X}}_{(q)}$. This is the usual motivation for the SVD.

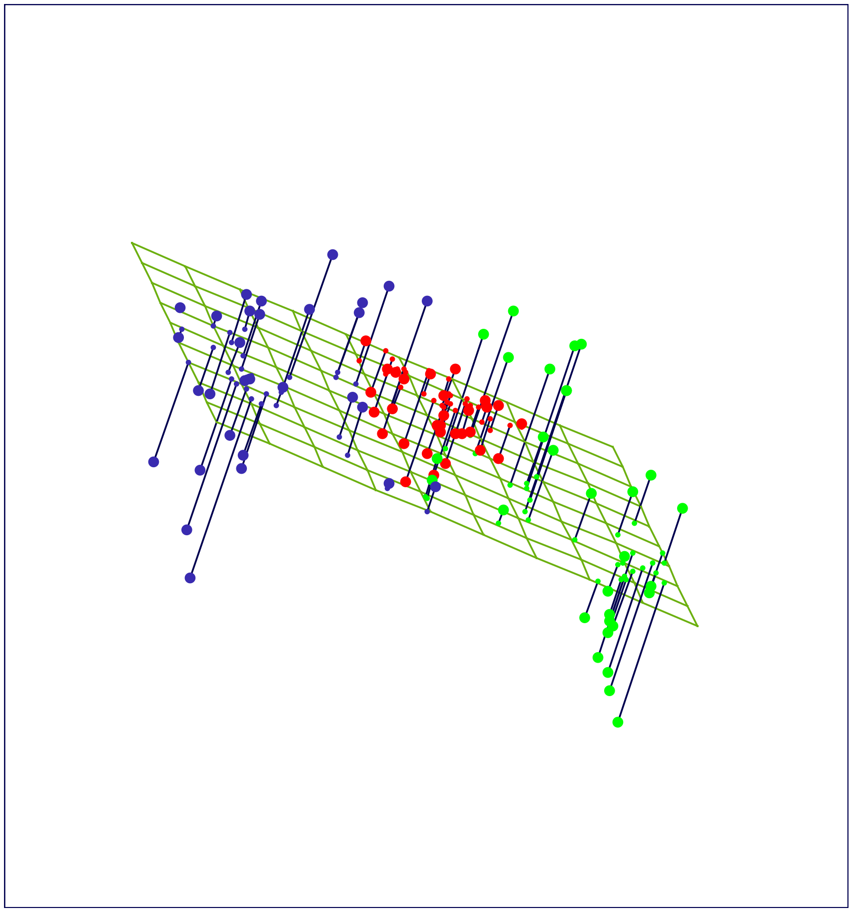
PC: Derived Variables



$Z_1 = X\alpha_1$ is the projection of the data onto the longest direction, and has the largest variance amongst all such normalized projections.

α_1 is the eigenvector corresponding to the largest eigenvalue of $\hat{\mathbf{\Sigma}}$, the sample covariance matrix of X. Z_2 and α_2 correspond to the second-largest eigenvector.

PC: Least Squares Approximation



Find the linear manifold $f(\lambda) = \mu + \mathbf{V}_q\lambda$ that best approximates the data in a least-squares sense:

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \mathbf{V}_q\lambda_i\|^2.$$

Solution: $\mu = \bar{x}$, $v_k = \alpha_k$, $\lambda_k = \mathbf{V}_q^T(x_i - \bar{x})$.

PC: Singular Value Decomposition

Let $\tilde{\mathbf{X}}$ be the $N \times p$ data matrix with centered columns (assume $N > p$).

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

is the SVD of $\tilde{\mathbf{X}}$, where

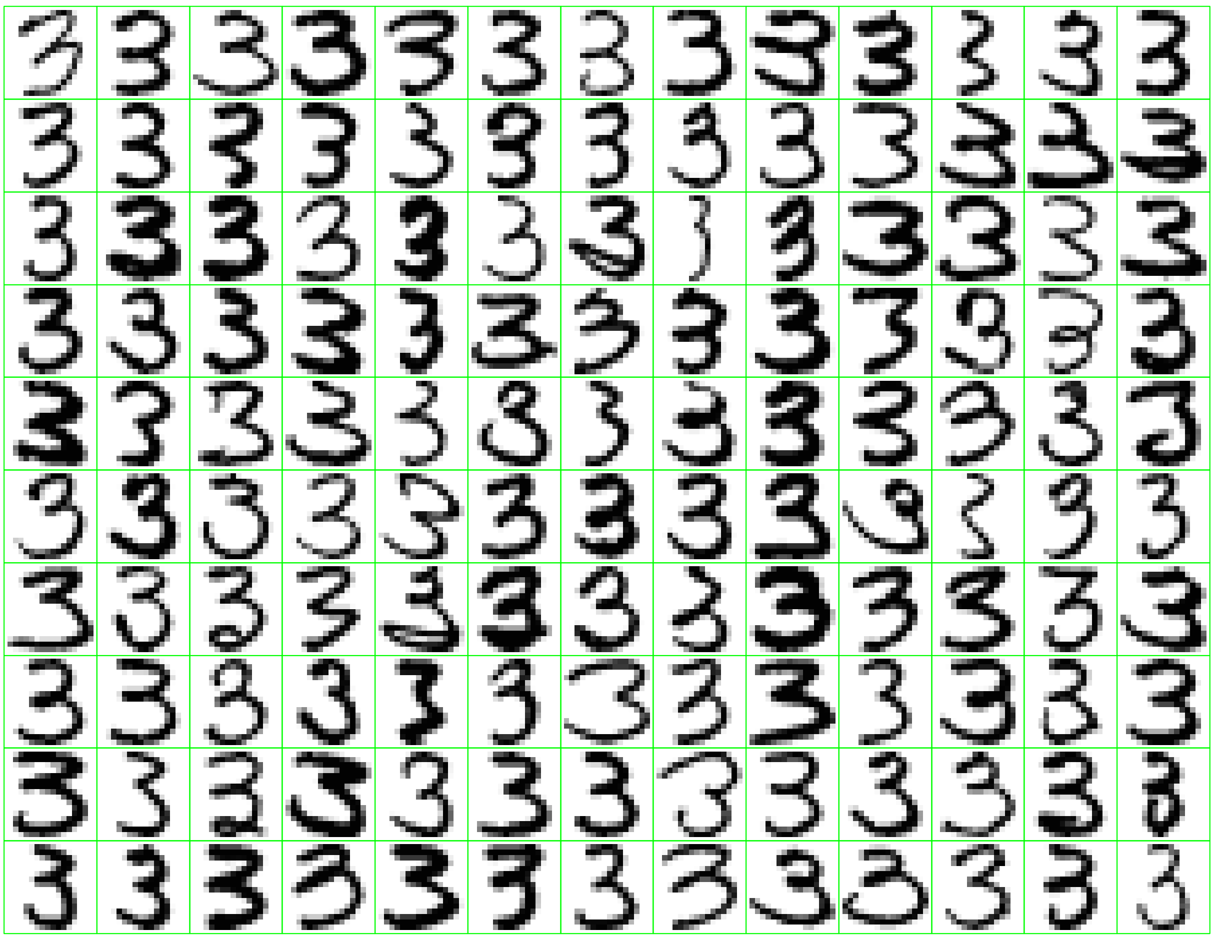
- $\mathbf{U}$ is $N \times p$ orthogonal, the left singular vectors.
- $\mathbf{V}$ is $p \times p$ orthogonal, the right singular vectors.
- $\mathbf{D}$ is diagonal, with $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$, the singular values.

The SVD always exists, and is unique up to signs. The columns of $\mathbf{V}$ are the principal components, and $Z_j = U_j d_j$.

Let $\mathbf{D_q}$ be $\mathbf{D}$, with all but the first q diagonal elements set to zero. Then $\hat{\mathbf{X}}_q = \mathbf{U}\mathbf{D}_q\mathbf{V}^T$ solves

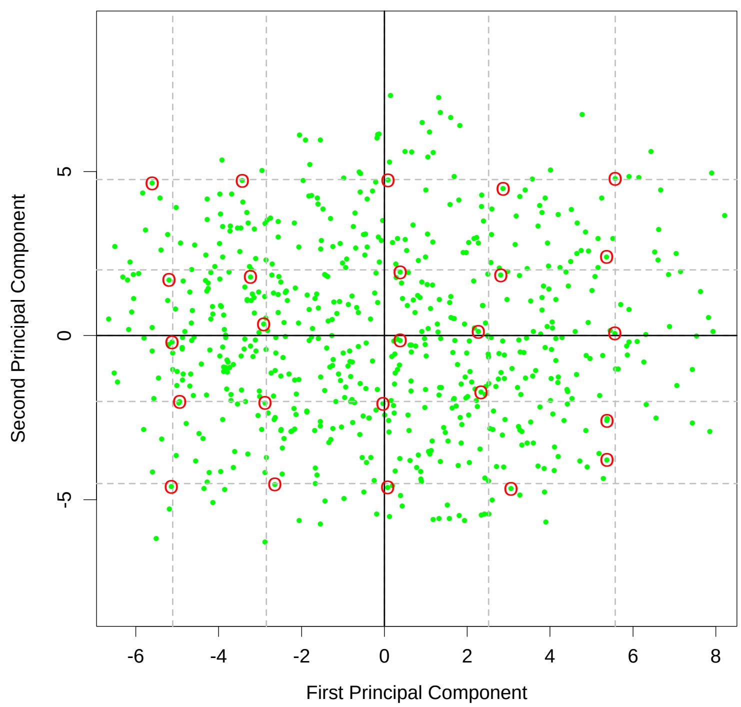
$$\min_{\operatorname{rank}(\hat{\mathbf{X}}_q)=q} ||\tilde{\mathbf{X}} - \hat{\mathbf{X}}_q||$$

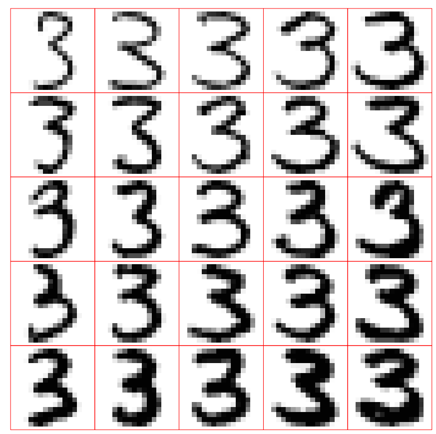
PC: Example — Digit Data



130 threes, a subset of 638 such threes and part of the handwritten digit dataset. Each three is a 16×16 greyscale image, and the variables $X_j,\ j = 1, \ldots, 256$ are the greyscale values for each pixel.

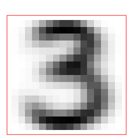
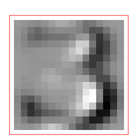
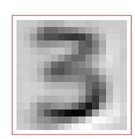
Rank-2 Model for Threes





Two-component model has the form

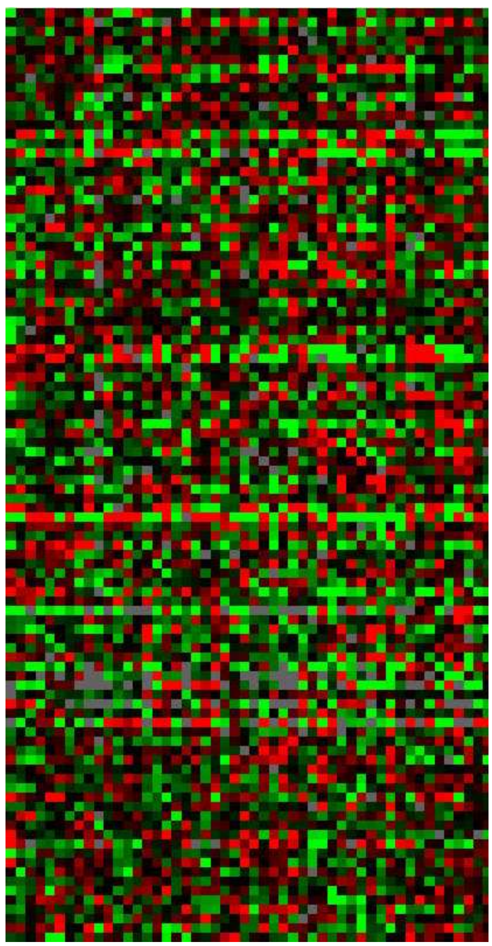
$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$=$  $+ \lambda_1 \cdot$  $+ \lambda_2 \cdot$  .

Here we have displayed the first two principal component directions, v_1 and v_2, as images.

SVD: Expression Arrays

The rows are genes (variables) and the columns are observations (samples, DNA arrays). Typically 6-10K genes, 50 samples.



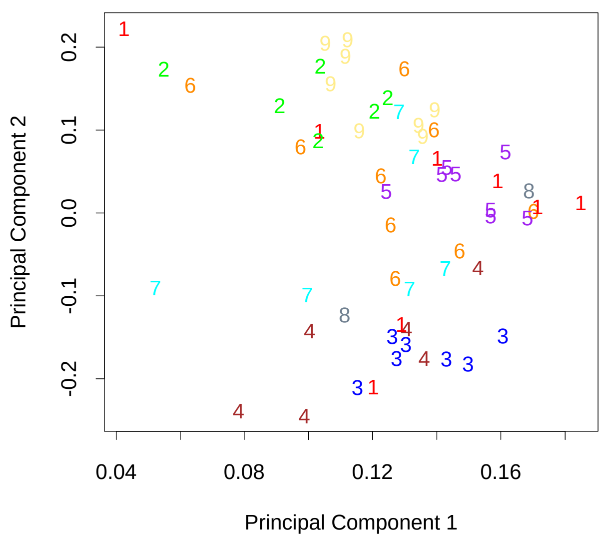
Eigengenes

- The first principal component or eigengene is the linear combination of the genes showing the most variation over the samples.
- The individual gene loadings for each eigengene or eigenarrays can have biological meaning.
- The sample values for the eigengenes show useful low-dimensional projections.

Example: NCI Cancer Data

First two eigengenes

Points are colored according to NCI cancer classes



First two eigenarrays