# Dimensionality Reduction

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

# Mathematical model

- Recall

$$t_n = f(\mathbf{x}_n) + \epsilon_n = \sum_{m=1}^{M} c_m d_m(\mathbf{x}_n) + \epsilon_n,$$

- The *design matrix* $\mathbf{D}$ has entries

$$D_{n,m} = d_m(\mathbf{x}_n).$$

With $\mathbf{t}^T = (t_1, \ldots, t_N)$, $\mathbf{c}^T = (c_1, \ldots, c_M)$ and $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_N)$, then

$$\mathbf{t} = \mathbf{D}\mathbf{c} + \boldsymbol{\epsilon}.$$

# Multicollinearity

- *Multicollinearity* happens when one or more predictor variables are highly correlated.
- In this case one of the predictors can be modeled by the others.
- A high degree of correlation increases the variance in the coefficients, since different models are equivalent.
- Small changes in the input data can lead to large changes in the model.
- *Perfect multicollinearity* means the regressors are linearly dependent, that is one can be exactly expressed by the others.
- In this case $(\mathbf{D}^T\mathbf{D})^{-1}$ does not exist.

# Multicollinearity

- In a good regression model, the regressors correlate minimally with each other, but are each highly correlated with the regressand.
- The aim is to find a linear combination of few regressors which summarize and explain the data without too much loss of information.
- *Principal component regression* uses the principal components of $\mathbf{D}$ as regressors instead of the columns of $\mathbf{D}$.

# Principal Component Regression

- Let $\mathbf{d}_1, \ldots, \mathbf{d}_M$ denote the columns of $\mathbf{D}$.
- We assume that the columns are standardized that is they have mean $0$ and length $1$.
- The first assumption is valid, since the measurements are mean centred, and the second assumption is valid, since regressors are invariant to scaling.
- The correlation between the $i$-th and $j$-th regressor is then

$$\mathbf{d}_i^T \mathbf{d}_j.$$

- The correlation matrix is $\mathbf{D}^T \mathbf{D}$.

# Principal Component Regression

- A new set of regressors is generated as linear combinations of regressors, say
$$v_1 d_1(\mathbf{x}) + \ldots + v_M d_M(\mathbf{x}).$$
- Evaluating this at the $N$ different points we arrive at a linear combination of the columns of $\mathbf{D}$,
$$v_1 \mathbf{d}_1 + \ldots + v_M \mathbf{d}_M = \mathbf{D}\mathbf{v}.$$
- The correlation between two such linear combinations is
$$\frac{\mathbf{v}_1^T \mathbf{D}^T \mathbf{D}\mathbf{v}_2}{\|\mathbf{D}\mathbf{v}_1\| \|\mathbf{D}\mathbf{v}_2\|},$$
where $\mathbf{v}_1$ and $\mathbf{v}_2$ are such that $\mathbf{D}\mathbf{v}_1 \neq 0$ and $\mathbf{D}\mathbf{v}_2 \neq 0$.

# Principal Component Regression

- The matrix $\mathbf{D}^T\mathbf{D}$ is symmetric and positive semidefinite.
- It has $M$ non-negative eigenvalues and corresponding orthonormal eigenvectors.
- Thus choosing $\mathbf{v}_1$ and $\mathbf{v}_2$ to be eigenvectors, the new regressors are uncorrelated.
- The eigenvectors corresponding to the $K$ nonzero eigenvalues are chosen. The prediction $\hat{\mathbf{t}}$ is the projection of $\mathbf{t}$ onto the space spanned by $\mathbf{Dv}_1, \ldots, \mathbf{Dv}_K$.
- The distance is

$$\|\mathbf{t} - \hat{\mathbf{t}}\|^2 = \mathbf{t}^T\mathbf{t} - \sum_{i=1}^{K} \frac{\left[(\mathbf{Dv}_i)^T\mathbf{t}\right]^2}{\lambda_k}.$$

# Principal Component Regression

- We decompose the regressand $\mathbf{t}$ into one portion lying in the subspace spanned by $\mathbf{Dv}_1, \ldots, \mathbf{Dv}_K$ and a remainder $\mathbf{a}$ which is orthogonal to $\mathbf{Dv}_1, \ldots, \mathbf{Dv}_K$, $\mathbf{t} = \sum_{k=1}^{K} a_k \mathbf{Dv}_k + \mathbf{a}$.
- Then $(\mathbf{Dv}_i)^T \mathbf{t} = \sum_{k=1}^{K} a_k \mathbf{v}_i^T \mathbf{D}^T \mathbf{Dv}_k + \mathbf{v}_i^T \mathbf{D}^T \mathbf{a} = a_i \lambda_i$,
- $\|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 = \mathbf{t}^T \mathbf{t} - \sum_{i=1}^{K} a_i^2 \lambda_i = \|\mathbf{a}\|_2^2$.
- If sparsity is required and not all eigenvectors can be used, those for which $a_i^2 \lambda_i$ is largest should be chosen. However, this can cause the model to not generalize well to unseen data. To avoid this, it is customary to choose the eigenvectors with the largest eigenvalues.

# Partial Least Squares (PLS)

- Principal component regression does not address the correlation with the regressand.
- *Partial Least Squares (PLS)* aims to maximize the correlation between regressors and regressand.
- This time a new set of regressors are generated iteratively as linear combinations of regressors, in the first iteration say

$$z_1\mathbf{d}_1 + \ldots + z_M\mathbf{d}_M = \mathbf{Dz}.$$

- Wlog $\|\mathbf{Dz}\| = 1$, since regressors are invariant to scaling.

# PLS

- The matrix $\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D}$ is a symmetric, positive semidefinite $M \times M$ matrix, since

$$(\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D})^T = \mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D} \text{ and } \mathbf{v}^T\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{D}\mathbf{v} = (\mathbf{v}^T\mathbf{D}^T\mathbf{t})^2 \geq 0.$$

- It has $M$ non-negative eigenvalues and corresponding orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_M$ where the eigenvectors are ordered with regards to the corresponding eigenvalues from largest to smallest.
- $\mathbf{z} \in \mathbb{R}^M$, and thus can be expressed as a linear combination of these eigenvectors:

$$\mathbf{z} = \hat{z}_1\mathbf{v}_1 + \ldots + \hat{z}_M\mathbf{v}_M.$$

# PLS

- The square of the correlation between $\mathbf{t}$ and the new regressor $\mathbf{Dz}$ is

$$\left(\frac{\mathbf{t}^T\mathbf{Dz}}{\|\mathbf{t}\|_2\|\mathbf{Dz}\|_2}\right)^2 = \frac{1}{\|\mathbf{t}\|_2^2}\mathbf{z}^T\mathbf{D}^T\mathbf{t}\mathbf{t}^T\mathbf{Dz} = \frac{1}{\|\mathbf{t}\|_2^2}\left(\lambda_1\hat{z}_1^2 + \ldots \lambda_M\hat{z}_M^2\right).$$

- This is maximal for $\hat{z}_2 = \ldots = \hat{z}_M = 0$.
- Thus the first new regressor is $\mathbf{t}_1 = \mathbf{Dv}_1/\|\mathbf{Dv}_1\|_2$.

# PLS

- Having generated $\mathbf{t}_1$, we calculate

$$\mathbf{D}_1 = \left(\mathbf{I} - \mathbf{t}_1\mathbf{t}_1^T\right)\mathbf{D}.$$

- Note

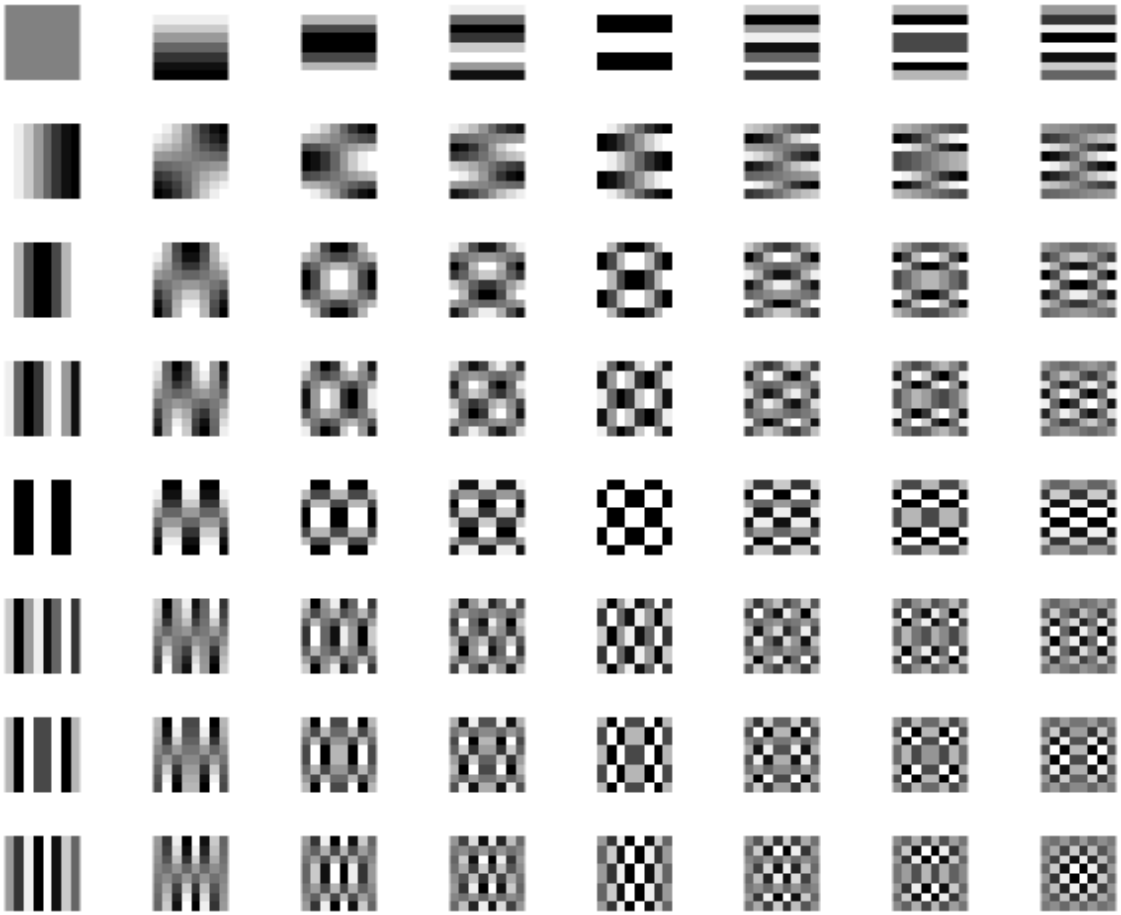
$$\mathbf{D}_1\mathbf{v}_1 = \left(\mathbf{I} - \mathbf{t}_1\mathbf{t}_1^T\right)\mathbf{D}\mathbf{v}_1 = \|\mathbf{D}\mathbf{v}_1\|_2(1 - \|\mathbf{t}_1\|_2^2)\mathbf{t}_1 = 0.$$

- Let $\mathbf{v}_2$ with $\|\mathbf{v}_2\|_2 = 1$ be the eigenvector corresponding to the largest eigenvalue of $\mathbf{D}_1^T\mathbf{t}\mathbf{t}^T\mathbf{D}_1$. The second new regressor is $\mathbf{t}_2 = \mathbf{D}_1\mathbf{v}_2$ normalized such that $\|\mathbf{t}_2\|_2 = 1$. Again, the correlation is maximal.
- The process continues until $\mathbf{D}_r$ is a null matrix, i.e. its rank is zero.
- We have $\mathrm{rank}\mathbf{D}_j \leq \mathrm{rank}\mathbf{D}_{j-1} - 1$ since a vector $(\mathbf{v}_j)$ which previously was not mapped to zero, now is mapped to zero.
- (PLS can be used for multivariate regression, that is the regressand is a matrix $\mathbf{t}$ of size $N \times q$.)

# Dimensionality Reduction

- Data lies in a high dimensional space.
- Finding a viewpoint along a small number of dimensions where the most relevant information is visible.
- When driving, the most important viewpoints are the front and mirrors, and looking over ones shoulder when necessary.
- An $8$-bit gray scale image, each pixel has a value between $0$ (black) and $255$ (white). A standard sized image of $1280$ by $720$ pixels therefore is encoded in $921,600$ bytes or approaching one megabyte.
- JPEG 1992: $8 \times 8$ pixels $\Rightarrow 14400$ blocks.

# Dimensionality Reduction



# Principal Component Analysis

- *Principal Component Analysis (PCA)* seeks a subspace (*principal subspace*) of a given dimension $K$ of the feature space such that projections of the data samples $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ onto that subspace are as spread out as possible.
- Start with $K = 1$.
- Projection onto a one-dimensional subspace defined by the vector $\mathbf{w}$.
- Length is chosen to be $\|\mathbf{w}\| = 1$, so that the projection of $\mathbf{v}_n$ is calculated as $\mathbf{w}^T\mathbf{v}_n$.

# Principal Component Analysis

- Sample mean: $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{v}_n$.
- Sample covariance matrix: $\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu})^T$.
- Variance of the projected samples is

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{v}_n - \mathbf{w}^T\boldsymbol{\mu})^2 &= \mathbf{w}^T\left[\frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu})^T\right]\mathbf{w} \\ &= \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}.\end{aligned}$$

- Maximize subject to the constraint $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = 1$.

# Principal Component Analysis

- The *Lagrangian function* is

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1).$$

  A stationary point of $L(\mathbf{w}, \lambda)$ is a maximum of the constraint optimization.
- The derivative of $L(\mathbf{w}, \lambda)$ with respect to $\mathbf{w}$ is

$$\frac{d}{d\mathbf{w}} L(\mathbf{w}, \lambda) = \mathbf{\Sigma} \mathbf{w} - \lambda \mathbf{w}.$$

- Setting this to zero, gives

$$\mathbf{\Sigma} \mathbf{w} = \lambda \mathbf{w},$$

$\Rightarrow \mathbf{w}$ is an eigenvector of $\mathbf{\Sigma}$. Using $\mathbf{w}^T \mathbf{w} = 1$, the eigenvalue is

$$\lambda = \lambda \mathbf{w}^T \mathbf{w} = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w},$$

which is the variance of the projected data.

# Principal Component Analysis

- Being a covariance matrix, $\mathbf{\Sigma}$ has non-negative eigenvalues.
- Since $\mathbf{\Sigma}$ is symmetric, its eigenvectors are orthogonal to each other.
- The principal space of dimension $K$ is the subspace spanned by the eigenvectors of the $K$ largest eigenvalues.
- These are known as *principal components*.

# Principal Component Analysis

- Sample $\mathbf{v}_n$ expressed in the eigenvector basis is

$$\mathbf{v}_n = \sum_{d=1}^{D} (\mathbf{v}_n^T \mathbf{w}_d) \mathbf{w}_d,$$

- The squared distance between a sample and its projection is

$$\| \sum_{d=K+1}^{D} (\mathbf{v}_n^T \mathbf{w}_d) \mathbf{w}_d \|^2 = \sum_{d=K+1}^{D} (\mathbf{v}_n^T \mathbf{w}_d)^2.$$

- The average squared distance is

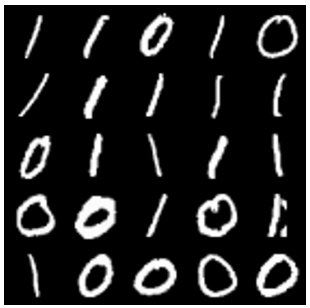
$$\frac{1}{N} \sum_{n=1}^{N} \sum_{d=K+1}^{D} (\mathbf{v}_n^T \mathbf{w}_d)^2 = \sum_{d=K+1}^{D} \lambda_d + (\boldsymbol{\mu}^T \mathbf{w}_d)^2.$$

- As more eigenvectors are used, the average squared distance becomes smaller. $\Rightarrow$ Reconstruction gets better.

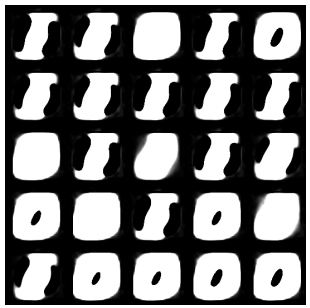
# MNIST

- MNIST data set of handwritten digits.
- $60,000$ images of handwritten digits of size $28 \times 28 = 784$ pixels.
- Images are data points in a $784$ dimensional space.

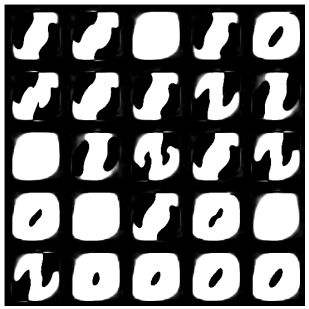
# MNIST



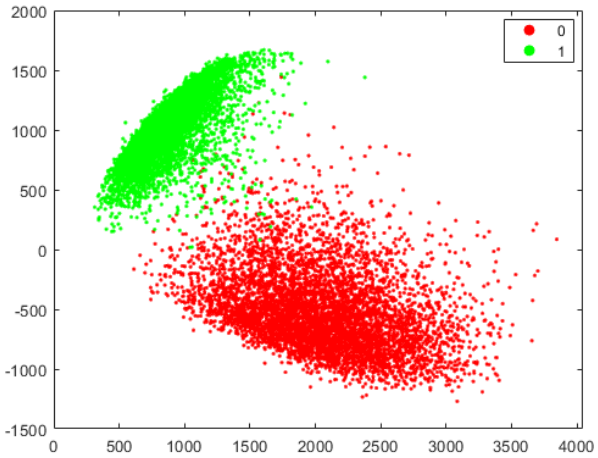
Original.



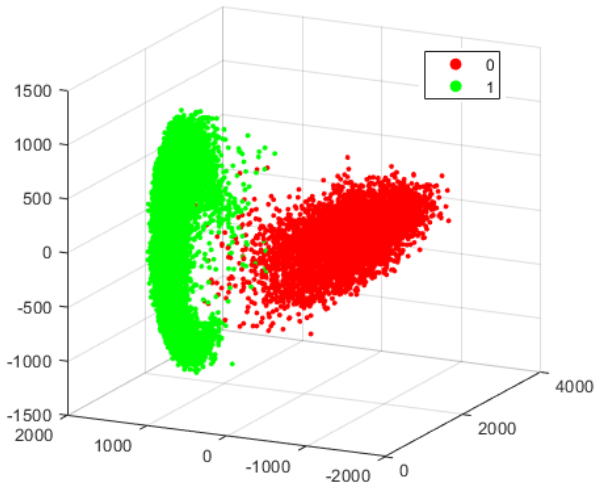
Reconstruction with two principal components.



Reconstruction with three principal components.

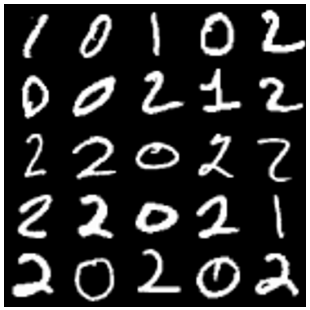



2-D principal subspace.

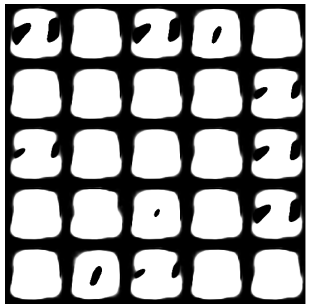



3-D principal subspace.

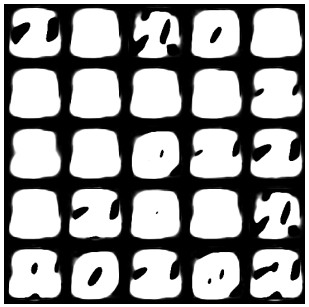
# MNIST



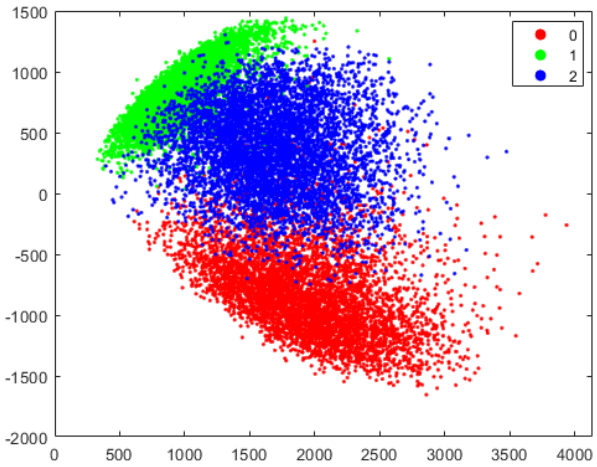
Original.



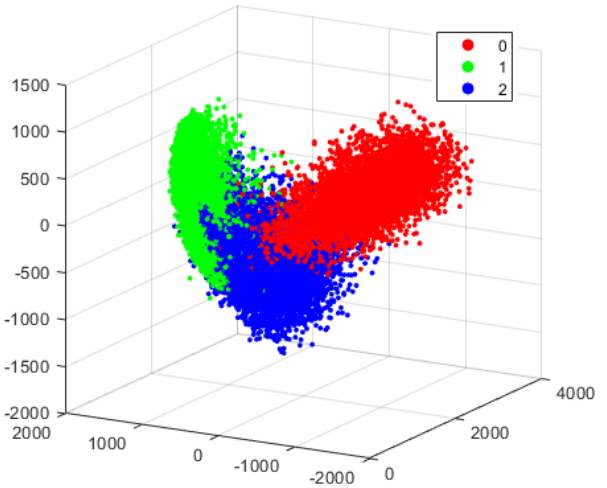
Reconstruction with two principal components.



Reconstruction with three principal components.



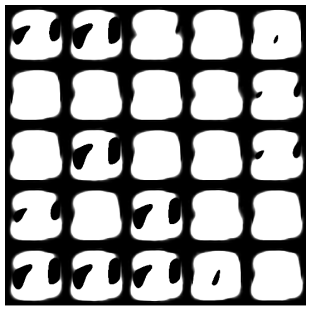

2-D principal subspace.



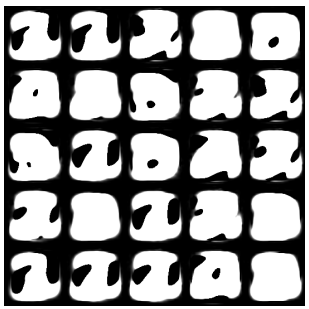

3-D principal subspace.

# MNIST

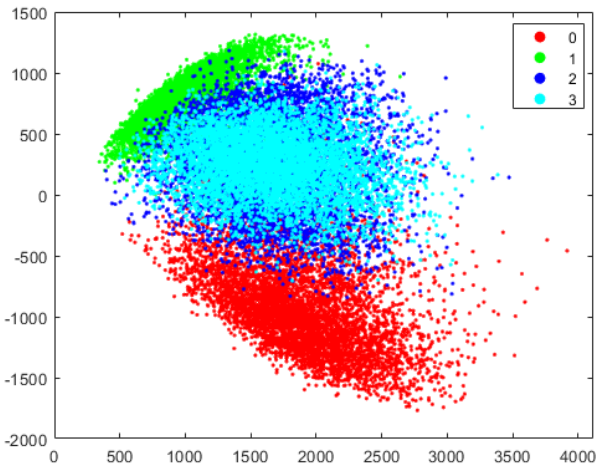
Original.



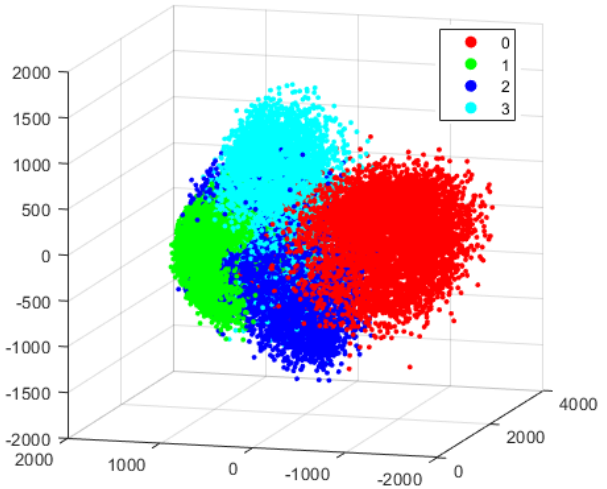
Reconstruction with two principal components.



Reconstruction with three principal components.




2-D principal subspace.




3-D principal subspace.

# Whitening

- Eigenvectors and eigenvalues are also used for *whitening* or *sphereing* the data.
- Let $\mathbf{W}$ be the orthogonal matrix formed from the eigenvectors $\mathbf{w}_d$ as columns and $\mathbf{\Lambda}^{-1/2}$ the diagonal matrix with $\lambda_d^{-1/2}$ on the diagonal.
- $\mathbf{v}_n$ is transformed to $\hat{\mathbf{v}}_n = \mathbf{\Lambda}^{-1/2}\mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu})$.
- Transformed mean is zero.
- The transformed covariance matrix is

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{v}}_n\hat{\mathbf{v}}_n^T &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{\Lambda}^{-1/2}\mathbf{W}^T(\mathbf{v}_n-\boldsymbol{\mu})(\mathbf{v}_n-\boldsymbol{\mu})^T\mathbf{W}\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{\Lambda}^{-1/2}\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbf{I}.\end{aligned}$$

- *Do not use, if the technique to be used relies on distinct eigenvalues of the sample covariance matrix.*

# Different eigenvalue problem

- Let $\mathbf{V}$ be the matrix where the $n^{\text{th}}$ row is $\mathbf{v}_n - \boldsymbol{\mu}$.
- The sample covariance matrix is $\boldsymbol{\Sigma} = \frac{1}{N}\mathbf{V}^T\mathbf{V}$.
- The eigenvalue equation is

$$\begin{aligned} \frac{1}{N} \overbrace{\mathbf{V}^T\mathbf{V}}^{D \times D} \mathbf{w}_d &= \lambda_d \mathbf{w}_d. \\ \frac{1}{N} \underbrace{(\mathbf{V}\mathbf{V}^T)}_{N \times N} \mathbf{V}\mathbf{w}_d &= \lambda_d \mathbf{V}\mathbf{w}_d. \end{aligned}$$

- So $\mathbf{V}\mathbf{w}_d$ is an eigenvector of $\mathbf{V}\mathbf{V}^T/N$.

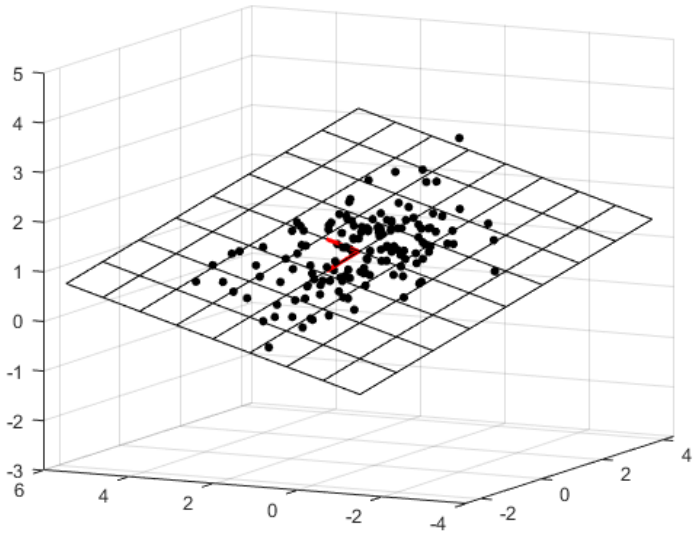
# Data generation

How should $K$ be chosen?

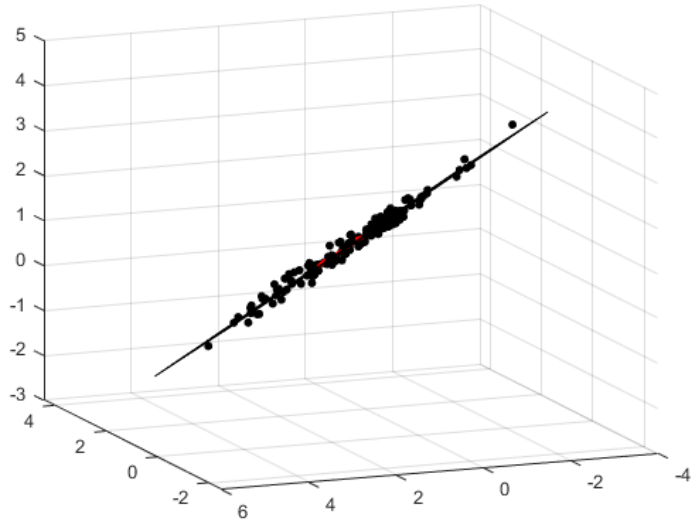
Assumptions:

- $\mathbf{v} = \mathbf{W}\mathbf{u} + \mathbf{m} + \boldsymbol{\epsilon}$, with $D \times K$ matrix $\mathbf{W}$, which has full rank, and $\mathbf{m} \in \mathbb{R}^D$.
- Columns of $\mathbf{W}$ are orthogonal, since they can be made orthogonal which corresponds to a change of basis in the representation of $\mathbf{u}$.
- $\boldsymbol{\epsilon}$ is normally distributed noise with zero mean and covariance matrix $\sigma^2\mathbf{I}$, known as *isotropic* or *spherical covariance matrix*.
- The noise explains how much the data protrudes from a $K$-dimensional subspace.

# Noise




Viewing the plane spanned by two principal components.




Viewing the data along one of the principal components.

# Data Generation

- Assumption $\mathbf{u}$ follows a normal distribution with mean $\hat{\mathbf{m}}$ and covariance matrix $\mathbf{S}$.
- We can further assume that $\hat{\mathbf{m}} = 0$ and $\mathbf{S} = \mathbf{I}$, since any other normal distribution can be generated from the standard normal distribution by a linear transformation and shift, both of which can be absorbed in $\mathbf{W}$ and $\mathbf{m}$.

$$\mathbf{v} \sim \mathcal{N}(\mathbf{m}, \underbrace{\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}}_{\mathbf{C}}).$$

# Independent Component Analysis

- If a more general distribution for $\mathbf{u}$ is assumed, it is known as *Independent Component Analysis (ICA)*.
- For example,

$$p(\mathbf{u}) = \prod_{k=1}^{K} p(u_k),$$

where each $u_k$ follows the distribution given by

$$p(u_k) = \frac{2}{\pi(\exp(u_k) + \exp(-u_k))}.$$

- Larger *kurtosis* than the normal distribution, meaning more suitable to model heavy l=tails or outliers.

# Maximizing the Likelihood

- The likelihood of the data set $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ is

$$\prod_{n=1}^{N} \frac{1}{\sqrt{|2\pi\mathbf{C}|}} \exp\left(-\frac{1}{2}(\mathbf{v}_n - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{v}_n - \mathbf{m})\right).$$

- Maximize its logarithm, the *log likelihood*,

$$\mathcal{L} = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{v}_n - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{v}_n - \mathbf{m}).$$

# Maximizing the Likelihood

- The derivative with respect to $\mathbf{m}$ is

$$\frac{\partial}{\partial \mathbf{m}} \mathcal{L} = \sum_{n=1}^{N} \mathbf{C}^{-1}(\mathbf{v}_n - \mathbf{m}) = \mathbf{C}^{-1} \left[ \sum_{n=1}^{N} (\mathbf{v}_n - \mathbf{m}) \right].$$

- Zero, if

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{v}_n = \boldsymbol{\mu}.$$

# Maximizing the Likelihood

- Rewriting the log likelihood.
- $(\mathbf{v}_n - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{v}_n - \boldsymbol{\mu})$ is scalar and therefore equal to its trace.
- Trace of a three term product is invariant to cyclic permutations.

$$\begin{aligned}
\mathcal{L} &= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}\mathrm{tr}(\mathbf{C}^{-1}(\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu})^T) \\
&= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu})^T\right) \\
&= -\frac{N}{2}\left[D\log(2\pi) + \log|\mathbf{C}| + \mathrm{tr}\left(\mathbf{C}^{-1}\boldsymbol{\Sigma}\right)\right].
\end{aligned}$$

# Maximizing the Likelihood

- Let $\mathbf{w}_1, \ldots, \mathbf{w}_K$ be the columns of $\mathbf{W}$, which are mutually orthogonal.
- Then

$$\mathbf{C} = \sigma^2 \mathbf{I} + \sum_{k=1}^{K} \mathbf{w}_k \mathbf{w}_k^T .$$

- The inverse of $\mathbf{C}$ is

$$\mathbf{C}^{-1} = \sigma^{-2} \left( I - \sum_{k=1}^{K} \frac{\mathbf{w}_k \mathbf{w}_k^T}{\sigma^2 + \mathbf{w}_k^T \mathbf{w}_k} \right) .$$

# Maximizing the Likelihood

- Defining $\mathbf{C}_{-j} = \sigma^2\mathbf{I} + \sum_{\substack{k=1 \\ k\neq j}}^{K} \mathbf{w}_k\mathbf{w}_k^T$.
- Then $\mathbf{C} = \mathbf{C}_{-j} + \mathbf{w}_j\mathbf{w}_j^T$ and

$$\mathbf{C}^{-1} = \mathbf{C}_{-j}^{-1} - \frac{\sigma^{-2}}{\sigma^2 + \mathbf{w}_j^T\mathbf{w}_j}\mathbf{w}_j\mathbf{w}_j^T.$$

- Using the *matrix determinant lemma*, the determinant of $\mathbf{C}$ is

$$|\mathbf{C}| = |\mathbf{C}_{-j}|(1 + \mathbf{w}_j^T\mathbf{C}_{-j}^{-1}\mathbf{w}_j) = |\mathbf{C}_{-j}|(1 + \sigma^{-2}\mathbf{w}_j^T\mathbf{w}_j).$$

# Maximizing the Likelihood

- The log likelihood becomes

$$\begin{aligned}\mathcal{L} &= -\frac{N}{2}\left[D\log(2\pi)+\log|\mathbf{C}_{-j}|+\log(1+\sigma^{-2}\mathbf{w}_j^T\mathbf{w}_j)\right.\\ &\left.+\mathrm{tr}(\mathbf{C}_{-j}^{-1}\mathbf{\Sigma})-\frac{\sigma^{-2}\mathbf{w}_j^T\mathbf{\Sigma}\mathbf{w}_j}{\sigma^2+\mathbf{w}_j^T\mathbf{w}_j}\right]\end{aligned}$$

- Differentiating with respect to $\mathbf{w}_j$ gives

$$\frac{\partial}{\partial\mathbf{w}_j}\mathcal{L}=-\frac{\sigma^{-2}N}{\sigma^2+\mathbf{w}_j^T\mathbf{w}_j}\left[\left(\sigma^2+\frac{\mathbf{w}_j^T\mathbf{\Sigma}\mathbf{w}_j}{\sigma^2+\mathbf{w}_j^T\mathbf{w}_j}\right)\mathbf{w}_j-\mathbf{\Sigma}\mathbf{w}_j\right].$$

# Maximizing the Likelihood

- The derivative with respect to $\mathbf{w}_j$ vanishes, if $\mathbf{w}_j$ is an eigenvector of $\mathbf{\Sigma}$ with eigenvalue $\lambda_j$ and the length of $\mathbf{w}_j$ is such that

$$\sigma^2 + \frac{\mathbf{w}_j^T \mathbf{\Sigma} \mathbf{w}_j}{\sigma^2 + \mathbf{w}_j^T \mathbf{w}_j} = \lambda_j.$$

- $\Rightarrow \mathbf{w}_j^T \mathbf{w}_j = \lambda_j - \sigma^2$.

# Maximizing the Likelihood

Inserting this gives

$$\begin{aligned} |\mathbf{C}| &= |\mathbf{C}_{-j}|(1+\sigma^{-2}(\lambda_j-\sigma^2)) = \sigma^{-2}\lambda_j|\mathbf{C}_{-j}| = \dots \\ &= (\sigma^{-2})^K \prod_{k=1}^{K} \lambda_k |\sigma^2\mathbf{I}| = (\sigma^2)^{D-K} \prod_{k=1}^{K} \lambda_k. \end{aligned}$$

and

$$\mathrm{tr}(\mathbf{C}^{-1}\mathbf{\Sigma}) = \sigma^{-2}\left(\mathrm{tr}(\mathbf{\Sigma}) - \sum_{k=1}^{K}\lambda_k\right) + K = \sigma^{-2}\left(\sum_{k=K+1}^{D}\lambda_k\right) + K.$$

# Maximizing the Likelihood

- The log likelihood becomes

$$\begin{aligned}\mathcal{L} &= -\frac{N}{2}\left[D\log(2\pi) + (D-K)\log\sigma^2 + \sum_{k=1}^{K}\log\lambda_k \right. \\ &\left. +\sigma^{-2}\left(\sum_{k=K+1}^{D}\lambda_k\right) + K\right],\end{aligned}$$

- Differentiating with respect to $\sigma^2$

$$\frac{\partial}{\partial\sigma^2}\mathcal{L} = -\frac{N}{2}\left((D-K)(\sigma^2)^{-1} - (\sigma^2)^{-2}\sum_{k=K+1}^{D}\lambda_k\right).$$

# Maximizing the Likelihood

- Zero for

$$\sigma^2 = \frac{1}{D-K} \sum_{k=K+1}^{D} \lambda_k.$$

- $\sigma^2$ is the average of all the other eigenvalues apart from $\lambda_1, \ldots, \lambda_K$.
- Calculating the covariance matrix is of complexity $O(ND^2)$.
- Finding the $K$ largest eigenvalues is $O(KD^2)$.
- Estimating $\sigma^2$ needs all $D$ eigenvalues.

# Expectation-Maximization

- The *Expectation-Maximization* algorithm calculates the expectation of the logarithm of the joint probability of the data and its latent variables with respect to the probability distribution of the latent variables and then maximizes this.
- For each $\mathbf{v}$, there is a latent variable $\mathbf{u}$ such that $p(\mathbf{v}|\mathbf{u})$ has mean $\mathbf{W}\mathbf{u} + \mathbf{m}$ and variance $\sigma^2\mathbf{I}$.
- We use $\mathbf{m} = \boldsymbol{\mu}$, the sample mean, since this is easily calculated.

# Expectation-Maximization

- The joint distribution $p(\mathbf{v}, \mathbf{u})$ is

$$\begin{aligned} p(\mathbf{v}, \mathbf{u}) &= p(\mathbf{u})p(\mathbf{v}|\mathbf{u}) \\ &= (2\pi)^{-K/2} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{u}\right) \times \\ &\quad (2\pi\sigma^2)^{-D/2} \exp\left(-\frac{1}{2}\sigma^{-2}(\mathbf{v} - \mathbf{W}\mathbf{u} - \boldsymbol{\mu})^T(\mathbf{v} - \mathbf{W}\mathbf{u} - \boldsymbol{\mu})\right) \end{aligned}$$

- It has mean $\begin{pmatrix} \boldsymbol{\mu} & 0 \end{pmatrix}^T$
- and covariance matrix

$$\begin{pmatrix} \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T & \mathbf{W} \\ \mathbf{W}^T & \mathbf{I} \end{pmatrix}.$$

# Expectation-Maximization

The logarithm of the joint data likelihood is

$$\mathcal{L} = \log \prod_{n=1}^{N} p(\mathbf{v}_n, \mathbf{u}_n) = -\frac{1}{2} \sum_{n=1}^{N} \Big[ (K+D)\log(2\pi) + D\log(\sigma^2) + \mathbf{u}_n^T \mathbf{u}_n + \sigma^{-2}(\mathbf{v}_n - \boldsymbol{\mu})^T(\mathbf{v}_n - \boldsymbol{\mu}) - 2\sigma^{-2}\mathbf{u}_n^T \mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu}) + \sigma^{-2}\mathbf{u}_n^T \mathbf{W}^T \mathbf{W}\mathbf{u}_n \Big].$$

# Expectation-Maximization

The expectation of the logarithm of the joint data likelihood is therefore

$$\begin{aligned}
\mathbb{E}[\mathcal{L}] &= -\frac{1}{2}\sum_{n=1}^{N}\Big[(K+D)\log(2\pi) + D\log(\sigma^2) + \mathbb{E}[\mathbf{u}_n^T\mathbf{u}_n] \\
&\quad +\sigma^{-2}\|\mathbf{v}_n - \boldsymbol{\mu}\|^2 - 2\sigma^{-2}\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu}) \\
&\quad +\sigma^{-2}\mathbb{E}[\mathbf{u}_n^T\mathbf{W}^T\mathbf{W}\mathbf{u}_n]\Big] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\Big[(K+D)\log(2\pi) + D\log(\sigma^2) + \mathrm{tr}(\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]) \\
&\quad +\sigma^{-2}\|\mathbf{v}_n - \boldsymbol{\mu}\|^2 - 2\sigma^{-2}\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu}) \\
&\quad +\sigma^{-2}\mathrm{tr}(\mathbf{W}^T\mathbf{W}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T])\Big]
\end{aligned}$$

# Expectation-Maximization

- The derivative with respect to $\mathbf{W}$ is

$$\frac{\partial}{\partial \mathbf{W}} \mathbb{E}[\mathcal{L}] = \sum_{n=1}^{N} \left[ \sigma^{-2} (\mathbf{v}_n - \boldsymbol{\mu}) \mathbb{E}[\mathbf{u}_n]^T - \sigma^{-2} \mathbf{W} \mathbb{E}[\mathbf{u}_n \mathbf{u}_n^T] \right].$$

- Setting this to zero and solving for $\mathbf{W}$ results in the update formula

$$\mathbf{W}_{\text{new}} = \left[ \sum_{n=1}^{N} (\mathbf{v}_n - \boldsymbol{\mu}) \mathbb{E}[\mathbf{u}_n]^T \right] \left[ \sum_{n=1}^{N} \mathbb{E}[\mathbf{u}_n \mathbf{u}_n^T] \right]^{-1}.$$

# Expectation-Maximization

- The derivative with respect to $\sigma^2$ is

$$\begin{aligned}\frac{\partial}{\partial \sigma^2}\mathbb{E}[\mathcal{L}] &= -\frac{1}{2}\sum_{n=1}^{N}\Big[\frac{D}{\sigma^2} - (\sigma^2)^{-2}\|\mathbf{v}_n - \boldsymbol{\mu}\|^2 \\ &\quad +2(\sigma^2)^{-2}\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu}) \\ &\quad -(\sigma^2)^{-2}\mathrm{tr}(\mathbf{W}^T\mathbf{W}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T])\Big].\end{aligned}$$

- Setting this to zero and solving for $\sigma^2$ results in the update formula

$$\sigma^2_{\text{new}} = \frac{1}{ND}\sum_{n=1}^{N}\Big[\|\mathbf{v}_n - \boldsymbol{\mu}\|^2 - 2\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T(\mathbf{v}_n - \boldsymbol{\mu}) + \mathrm{tr}(\mathbf{W}^T\mathbf{W}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T])\Big].$$

- Usually $\mathbf{W} = \mathbf{W}_{\text{new}}$ is used when updating.

# Expectation-Maximization

- The covariance matrix of $p(\mathbf{u}|\mathbf{v})$ is independent of $\mathbf{v}$ and is

$$\mathbf{\Sigma_{u|v}} = \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}.$$

- Its mean depends on $\mathbf{v}$ and is

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{u}|\mathbf{v}} &= \sigma^{-2}\mathbf{\Sigma_{u|v}}\mathbf{W}^T(\mathbf{v} - \boldsymbol{\mu}) \\ &= (\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\mathbf{v} - \boldsymbol{\mu}). \end{aligned}$$

# Expectation-Maximization

*E-step*: Calculate

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{u}_n|\mathbf{v}_n} = \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} &= (\sigma_{\text{old}}^{-2}\mathbf{W}_{\text{old}}^T\mathbf{W}_{\text{old}} + \mathbf{I})^{-1}, \\
\mathbb{E}[\mathbf{u}_n] = \boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n} &= \sigma_{\text{old}}^{-2}\boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}}\mathbf{W}_{\text{old}}^T(\mathbf{v}_n - \boldsymbol{\mu}), \\
\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T] &= \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} + \boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n}\boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n}^T.
\end{aligned}$$

Interpretation of $\sigma^2_{\text{new}}$: Using

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]\right]^{-1},$$

$$\begin{aligned}
\sum_{n=1}^{N}\text{tr}(\mathbf{W}_{\text{new}}^T\mathbf{W}_{\text{new}}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]) &= \text{tr}\left(\mathbf{W}_{\text{new}}^T\mathbf{W}_{\text{new}}\sum_{n=1}^{N}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T])\right) \\
&= \text{tr}\left(\mathbf{W}_{\text{new}}^T\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\right) \\
&= \sum_{n=1}^{N}\text{tr}(\mathbf{W}_{\text{new}}^T(\mathbf{v}_n - \boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T) \\
&= \sum_{n=1}^{N}\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}_{\text{new}}^T(\mathbf{v}_n - \boldsymbol{\mu})
\end{aligned}$$

# Interpretation

- The update formula for $\sigma^2_{\text{new}}$ becomes

$$\begin{aligned}\sigma^2_{\text{new}} &= \frac{1}{ND}\sum_{n=1}^{N}\left[\|\mathbf{v}_n - \boldsymbol{\mu}\|^2 - \mathbb{E}[\mathbf{u}_n]^T\mathbf{W}_{\text{new}}^T(\mathbf{v}_n - \boldsymbol{\mu})\right] \\ &= \frac{1}{ND}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu} - \mathbf{W}_{\text{new}}\mathbb{E}[\mathbf{u}_n])^T(\mathbf{v}_n - \boldsymbol{\mu}).\end{aligned}$$

- $\sigma^2_{\text{new}}$ is the average inner product of $\mathbf{v}_n - \boldsymbol{\mu}$ with the difference of $\mathbf{v}_n - \boldsymbol{\mu}$ and the image under $\mathbf{W}_{\text{new}}$ of the expectation of $\mathbf{u}_n$ scaled by the number of dimensions $D$.
- It measures, how close on average $\mathbf{W}_{\text{new}}\mathbb{E}[\mathbf{u}_n]$ is to the orthogonal projection of $\mathbf{v}_n$ onto the subspace spanned by the columns of $\mathbf{W}_{\text{new}}$.

# Complexity

$$\mathbf{W}_{\text{new}} = [\sum_{n=1}^{N} \underbrace{(\mathbf{v}_n - \boldsymbol{\mu})}_{D\times 1} \underbrace{\mathbb{E}[\mathbf{u}_n]^T}_{1\times K}][\sum_{n=1}^{N} \underbrace{\mathbb{E}[\mathbf{u}_n \mathbf{u}_n^T]}_{K\times K}]^{-1}.$$

$$\sigma^2_{\text{new}} = \frac{1}{ND}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu} - \underbrace{\mathbf{W}_{\text{new}}}_{D\times K}\underbrace{\mathbb{E}[\mathbf{u}_n]}_{K\times 1})^T(\mathbf{v}_n - \boldsymbol{\mu}).$$

$$\boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} = (\sigma^{-2}_{\text{old}} \underbrace{\mathbf{W}^T_{\text{old}}}_{K\times D}\underbrace{\mathbf{W}_{\text{old}}}_{D\times K} + \mathbf{I})^{-1}$$

$$\mathbb{E}[\mathbf{u}_n] = \sigma^{-2}_{\text{old}} \underbrace{\boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}}}_{K\times K}\underbrace{\mathbf{W}^T_{\text{old}}}_{K\times D}\underbrace{(\mathbf{v}_n - \boldsymbol{\mu})}_{D\times 1}$$

$$\mathbb{E}[\mathbf{u}_n \mathbf{u}_n^T] = \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} + \underbrace{\boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n}}_{K\times 1}\underbrace{\boldsymbol{\mu}^T_{\mathbf{u}_n|\mathbf{v}_n}}_{1\times K}$$

# Factor Analysis

- The noise variance is modeled by a diagonal matrix $\mathbf{D}$ such that each direction is treated separately.
- Elements of $\mathbf{D}$ are known as *uniquenesses*.
- Columns of $\mathbf{W}$ are known as *factor loadings*
- The logarithm of the joint data likelihood is

$$\begin{aligned}\mathcal{L} = \log \prod_{n=1}^{N} p(\mathbf{v}_n, \mathbf{u}_n) \quad = \quad & -\frac{1}{2}\sum_{n=1}^{N}\Big[(K+D)\log(2\pi) + \log(|\mathbf{D}|) \\ & + \mathbf{u}_n^T\mathbf{u}_n + (\mathbf{v}_n - \boldsymbol{\mu})^T\mathbf{D}^{-1}(\mathbf{v}_n - \boldsymbol{\mu}) \\ & - 2\mathbf{u}_n^T\mathbf{W}^T\mathbf{D}^{-1}(\mathbf{v}_n - \boldsymbol{\mu}) \\ & + \mathbf{u}_n^T\mathbf{W}^T\mathbf{D}^{-1}\mathbf{W}\mathbf{u}_n\Big].\end{aligned}$$

- Update formula for $\mathbf{W}$ remains the same.

# Factor Analysis

- Rewriting

$$\mathbb{E}[\mathcal{L}] = -\frac{1}{2}\sum_{n=1}^{N}\Big[(K+D)\log(2\pi) + \log(|\mathbf{D}|) + \mathrm{tr}(\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]) +$$
$$\mathrm{tr}\Big(\big[(\mathbf{v}_n-\boldsymbol{\mu})(\mathbf{v}_n-\boldsymbol{\mu})^T - 2(\mathbf{v}_n-\boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T + \mathbf{W}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]\mathbf{W}^T\big]\mathbf{D}^{-1}\Big)$$

- The derivative with respect to $\mathbf{D}$ is

$$\begin{aligned}\frac{\partial}{\partial\mathbf{D}}\mathbb{E}[\mathcal{L}] &= -\frac{1}{2}\sum_{n=1}^{N}\Big[\frac{1}{|\mathbf{D}|}|\mathbf{D}|\mathbf{D}^{-1} - \mathbf{D}^{-1}\mathrm{diag}\Big((\mathbf{v}_n-\boldsymbol{\mu})(\mathbf{v}_n-\boldsymbol{\mu})^T \\ &\quad -2(\mathbf{v}_n-\boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T + \mathbf{W}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]\mathbf{W}^T\Big)\mathbf{D}^{-1}\Big].\end{aligned}$$

# Factor Analysis

$$\mathbf{D}_{\text{new}} = \text{diag}\left( \frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu})^T - 2\frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\mathbf{W}^T + \frac{1}{N}\mathbf{W}\sum_{n=1}^{N}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]\mathbf{W}^T \right).$$

Using $\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{N}(\mathbf{v}_n - \boldsymbol{\mu})\mathbb{E}[\mathbf{u}_n]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{u}_n\mathbf{u}_n^T]\right]^{-1}$,

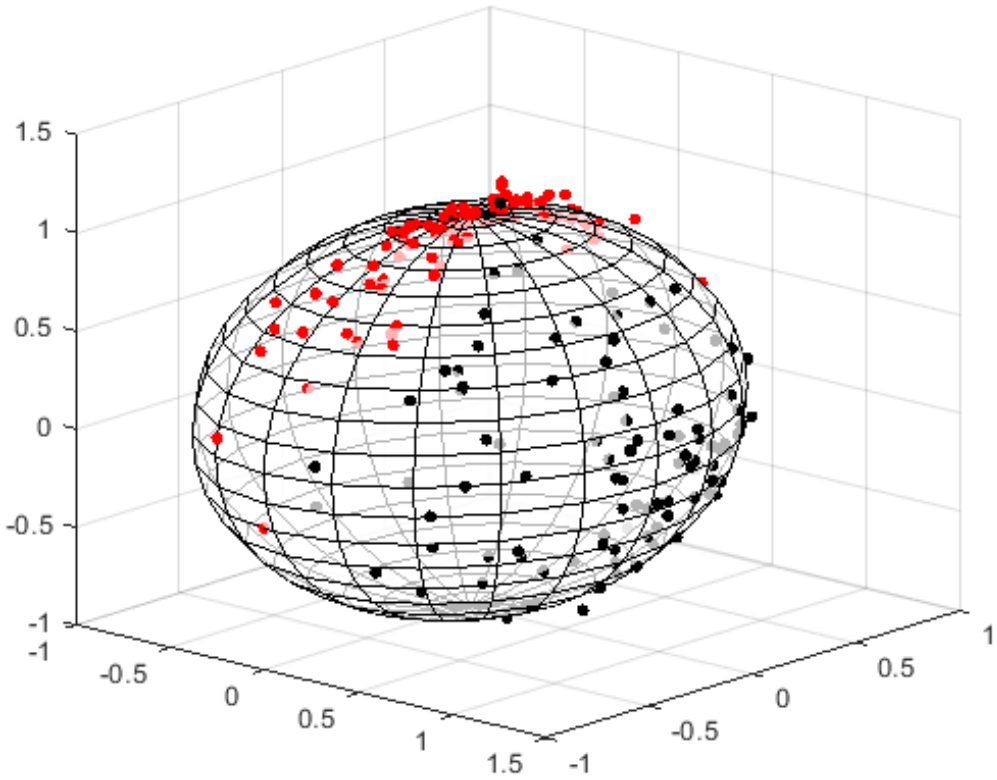
$$\mathbf{D}_{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\text{diag}\left((\mathbf{v}_n - \boldsymbol{\mu})(\mathbf{v}_n - \boldsymbol{\mu} - \mathbf{W}_{\text{new}}\mathbb{E}[\mathbf{u}_n])^T\right).$$

*E-step*: Calculate

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{u}_n|\mathbf{v}_n} = \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} &= (\mathbf{W}_{\text{old}}^T \mathbf{D}_{\text{old}}^{-1} \mathbf{W}_{\text{old}} + \mathbf{I})^{-1}, \\
\mathbb{E}[\mathbf{u}_n] = \boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n} &= \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} \mathbf{W}_{\text{old}}^T \mathbf{D}_{\text{old}}^{-1} (\mathbf{v}_n - \boldsymbol{\mu}), \\
\mathbb{E}[\mathbf{u}_n \mathbf{u}_n^T] &= \boldsymbol{\Sigma}_{\mathbf{u}|\mathbf{v}} + \boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n} \boldsymbol{\mu}_{\mathbf{u}_n|\mathbf{v}_n}^T.
\end{aligned}$$

# Kernel Principal Component Analysis

- Assumption so far that the data samples are related to latent variables in a lower dimensional space via a *linear* mapping.
- Data lie close to a line, plane or hyperplane.
- Data lie close to a $K$-dimensional *manifold*.



# Kernel Principal Component Analysis

- Mapping $\phi : \mathbb{R}^D \to \mathbb{R}^{\widehat{D}}$ from $D$ dimensions to $\widehat{D}$ dimensions.
- *Pre-images*: $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^D$.
- *Images*: $\phi(\mathbf{v}_1), \ldots, \phi(\mathbf{v}_N) \in \mathbb{R}^{\widehat{D}}$.
- *Kernel function* $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ defined as

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}).$$

- Let, $\mathbf{\Phi}$ be the $\widehat{D} \times N$ matrix whose columns are the vectors $\phi(\mathbf{v}_1), \ldots, \phi(\mathbf{v}_N)$.

# Kernel Principal Component Analysis

- Mean centering: Let $\mathbf{E} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$.
- $\mathbf{E}^2 = N\mathbf{E}$
- Then for $\mathbf{a}^T = (a_1, \ldots, a_N)$

$$\mathbf{a}^T(\mathbf{I} - \frac{1}{N}\mathbf{E}) = (a_1 - \frac{1}{N}\sum_{n=1}^{N} a_n, \cdots, a_N - \frac{1}{N}\sum_{n=1}^{N} a_n).$$

- In particular,

$$\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E}) = (\phi(\mathbf{v}_1) - \frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{v}_n), \cdots, \phi(\mathbf{v}_N) - \frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{v}_n)).$$

# Kernel Principal Component Analysis

- The covariance matrix of the shifted images is the average of the outer products of the $n^{\text{th}}$ column of $\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})$ with its transpose.

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{e}_n\mathbf{e}_n^T(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T,$$

  where $\mathbf{e}_n$ is the $n^{\text{th}}$ standard unit basis vector.

- $\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})$ and its transpose are independent of $n$ and $\sum_{n=1}^{N} \mathbf{e}_n\mathbf{e}_n^T = \mathbf{I}$.

- $\mathbf{I} - \frac{1}{N}\mathbf{E}$ is *idempotent*.

$$(\mathbf{I} - \frac{1}{N}\mathbf{E})(\mathbf{I} - \frac{1}{N}\mathbf{E}) = I - \frac{2}{N}\mathbf{E} + \frac{1}{N^2}\mathbf{E}^2 = \mathbf{I} - \frac{1}{N}\mathbf{E}.$$

$$\Rightarrow \mathbf{\Sigma} = \frac{1}{N}\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T.$$

# Eigenvalue equation

$$\mathbf{\Sigma} = \frac{1}{N}\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T.$$

- The eigenvalue equation is

$$\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T\mathbf{w}_k = N\lambda_k\mathbf{w}_k.$$

- Multiplying through with $(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T$ from the left and using the idempotence of $\mathbf{I} - \frac{1}{N}\mathbf{E}$,

$$(\mathbf{I} - \frac{1}{N}\mathbf{E})\underbrace{\mathbf{\Phi}^T\mathbf{\Phi}}_{\mathbf{K}}(\mathbf{I} - \frac{1}{N}\mathbf{E})\underbrace{(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T\mathbf{w}_k}_{\widehat{\mathbf{w}}_k} = N\lambda_k\underbrace{(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T\mathbf{w}_k}_{\widehat{\mathbf{w}}_k}.$$

# Kernel Principal Component Analysis

- Due to $\mathbf{I} - \frac{1}{N}\mathbf{E}$ being idempotent, we have $(\mathbf{I} - \frac{1}{N}\mathbf{E})\widehat{\mathbf{w}}_k = \widehat{\mathbf{w}}_k$.
- Thus, the components of $\widehat{\mathbf{w}}_k$ sum to zero.
- The equivalent equation is $\mathbf{K}\widehat{\mathbf{w}}_k = N\lambda_k\widehat{\mathbf{w}}_k$.
- $(i, j)$ entry of $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$ is the inner product

$$\phi(\mathbf{v}_i)^T\phi(\mathbf{v}_j) = k(\mathbf{v}_i, \mathbf{v}_j).$$

- $\mathbf{w}_k$ is a linear combination of the images, since

$$\begin{aligned}\frac{1}{N}\mathbf{\Phi}\widehat{\mathbf{w}}_k &= \frac{1}{N}\sum_{n=1}^{N}\widehat{w}_{kn}\phi(\mathbf{v}_n) \\ &= \frac{1}{N}\mathbf{\Phi}(\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{\Phi}^T\mathbf{w}_k = \mathbf{\Sigma}\mathbf{w}_k = \lambda_k\mathbf{w}_k.\end{aligned}$$

- *Not calculated*.

# Kernel Principal Component Analysis

- Only the values of the projections are calculated:

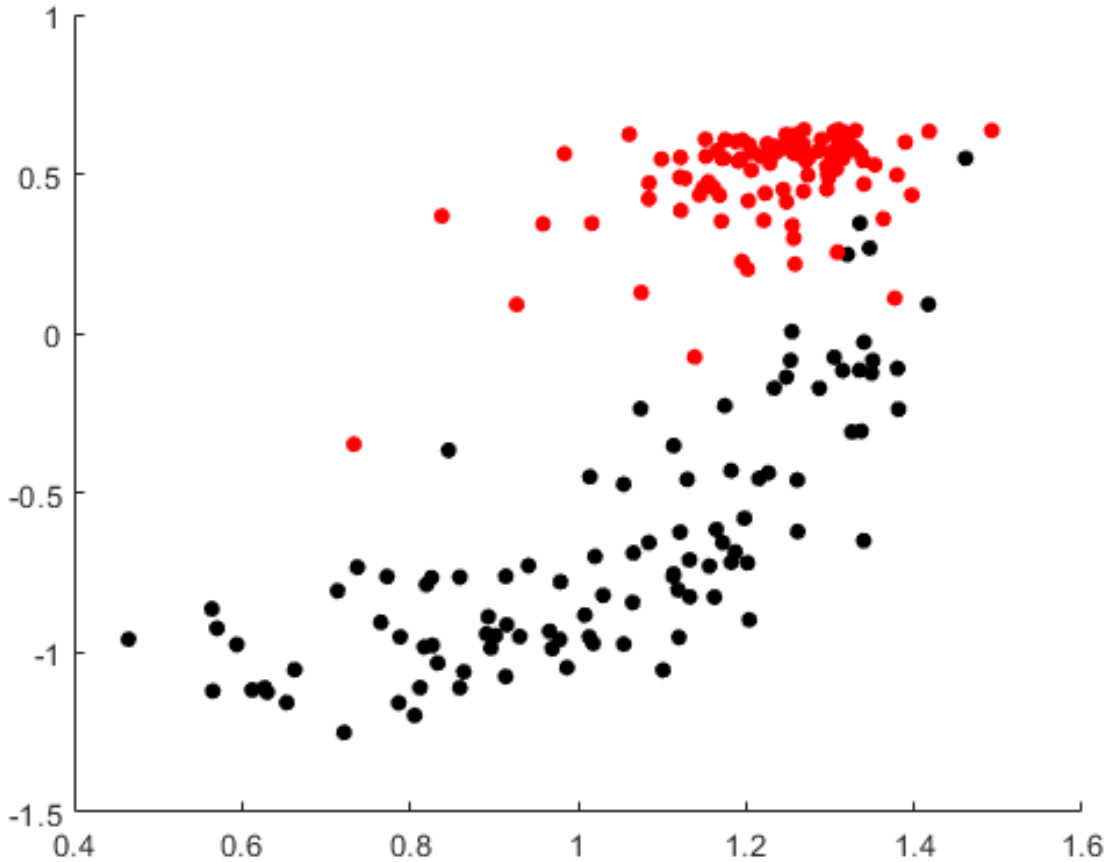
$$\begin{aligned}\mathbf{w}_k^T\phi(\mathbf{v})^T &= \frac{1}{N\lambda_k}\widehat{\mathbf{w}}_k^T\mathbf{\Phi}^T\phi(\mathbf{v}) = \frac{1}{N\lambda_k}\widehat{\mathbf{w}}_k^T\begin{pmatrix}\phi(\mathbf{v}_1)^T\phi(\mathbf{v})\\ \vdots \\ \phi(\mathbf{v}_N)^T\phi(\mathbf{v})\end{pmatrix}\\ &= \frac{1}{N\lambda_k}\widehat{\mathbf{w}}_k^T\begin{pmatrix}k(\mathbf{v}_1,\mathbf{v})\\ \vdots \\ k(\mathbf{v}_N,\mathbf{v})\end{pmatrix}.\end{aligned}$$

- Recall the condition on the principal components

$$\begin{aligned}1 &= \mathbf{w}_k^T\mathbf{w}_k = \frac{1}{(N\lambda_k)^2}\widehat{\mathbf{w}}_k^T\mathbf{\Phi}^T\mathbf{\Phi}\widehat{\mathbf{w}}_k\\ &= \frac{1}{(N\lambda_k)^2}\widehat{\mathbf{w}}_k^T(\mathbf{I}-\frac{1}{N}\mathbf{E})\mathbf{K}(\mathbf{I}-\frac{1}{N}\mathbf{E})\widehat{\mathbf{w}}_k = \frac{1}{N\lambda_k}\widehat{\mathbf{w}}_k^T\widehat{\mathbf{w}}_k,\end{aligned}$$

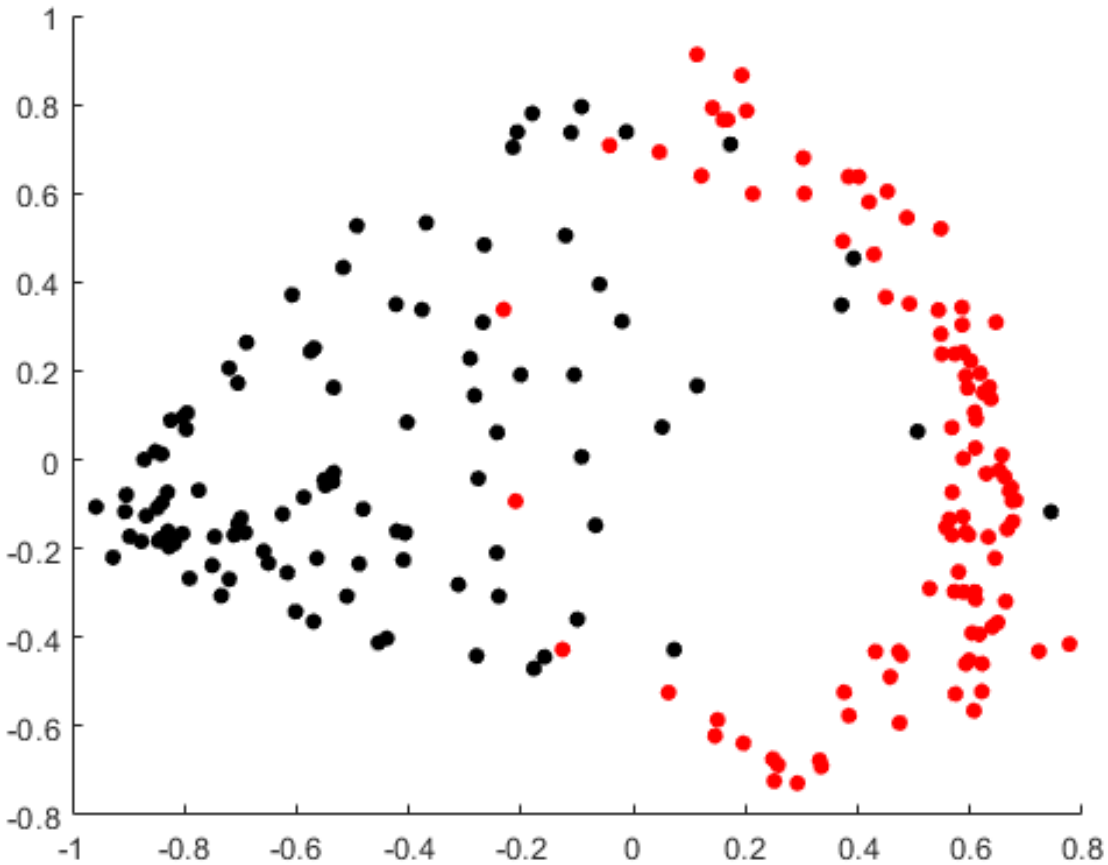
- $\widehat{\mathbf{w}}_k$ has to have length $\sqrt{N\lambda_k}$.

# Kernel Principal Component Analysis



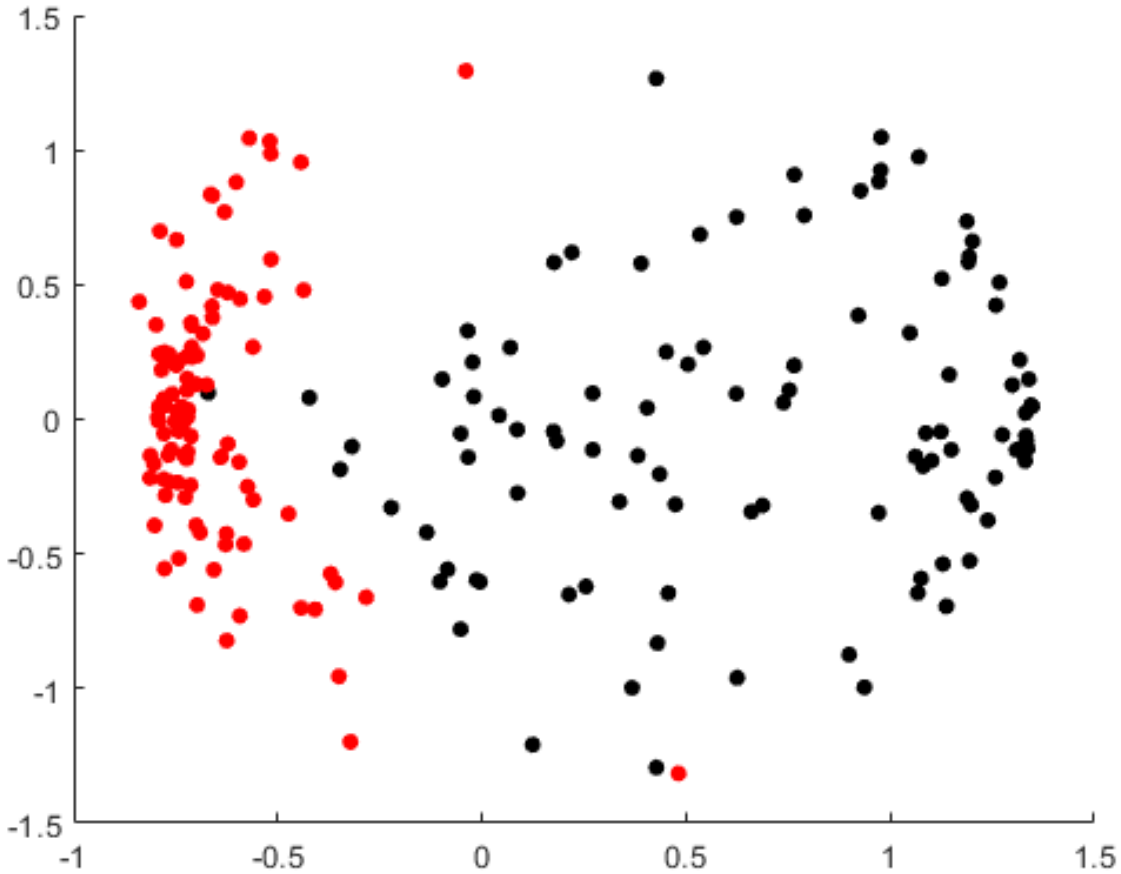

Projection onto two principal components after a transformation to a four-dimensional space.

# Kernel Principal Component Analysis



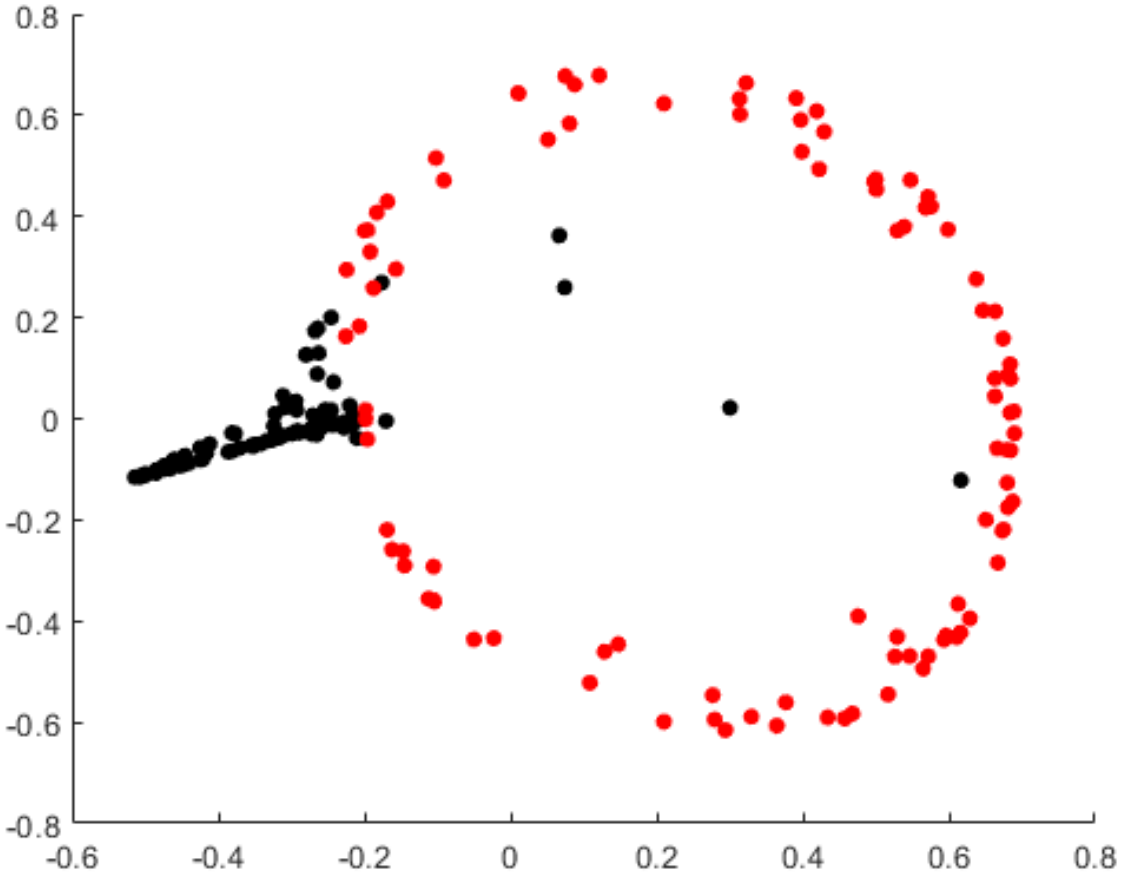

Quadratic kernel, $a = 1, c = 0.1$.

# Kernel Principal Component Analysis



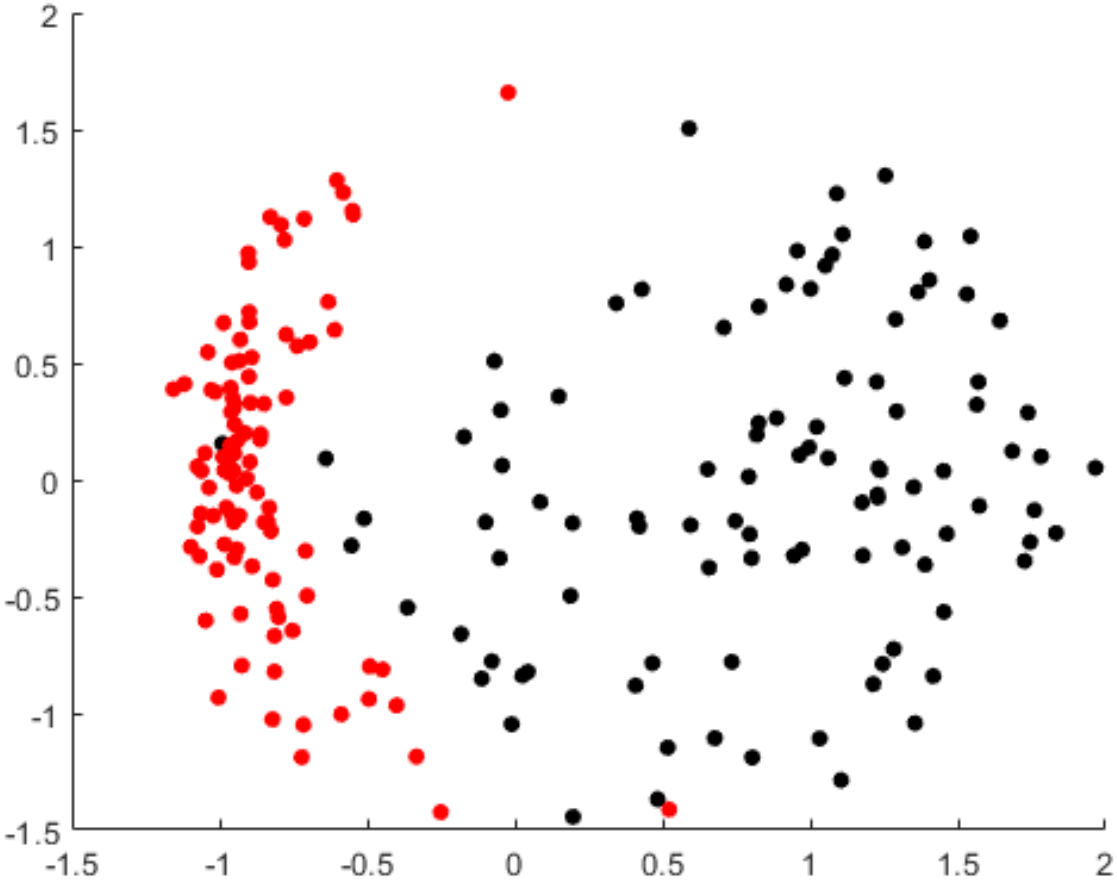

Hyperbolic tangent (Sigmoid) kernel, $a = 5, c = 0.5$.

# Kernel Principal Component Analysis




Gaussian kernel, $\sigma = 0.3$.

# Kernel Principal Component Analysis




Thin plate spline kernel, $c = 1$.