Machine Learning I

Lecture 5.5: Dimensional Reduction

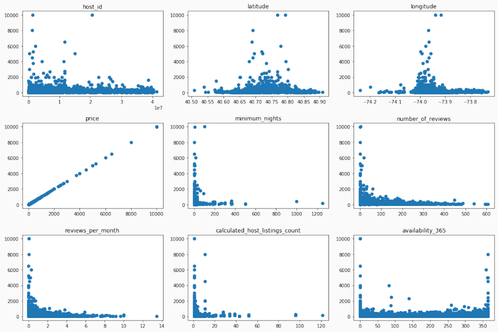
Nathaniel Bade

Northeastern University Department of Mathematics

Table of contents

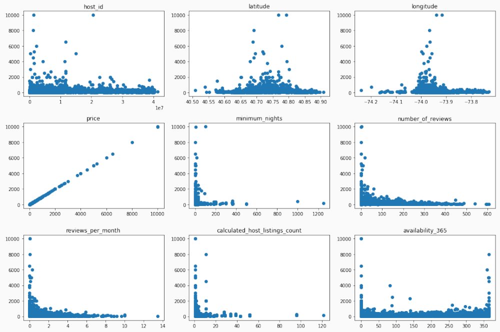
Feature Construction

Low Dimensional Projection and Meaning



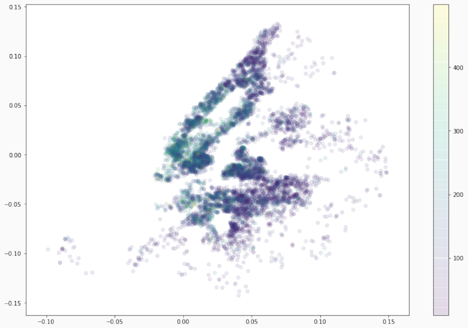
Sometimes the feature you have been handed are not actually the most salient features. As a simple examples, recall the AirBnB New York dataset from the homework. A quick analysis of the numeric features show that number of reviews and reviews per month seem to show the biggest correlation.

Low Dimensional Projection and Meaning



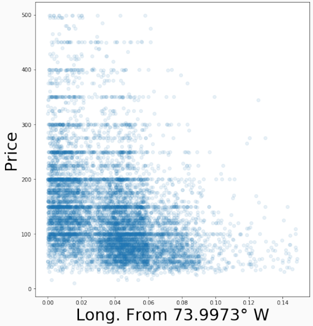
In particular though, latitude and longitude are effectively uncorrelated with price. Thinking for a second that cannon be right; geography has a everything to due to real estate. Actually, we notice two thing: latitude and longitude only don't look correlated because they're peaked in the middle.

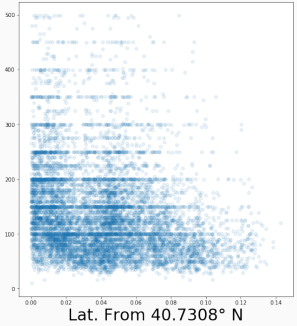
Low Dimensional Projection and Meaning



In fact, plotting latitude against longitude and coloring by the price we first see the New York land mass appear, but we also see that high prices are heavily centered around Lower and Mid Manhattan.

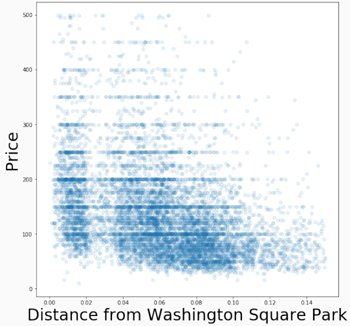
Low Dimensional Projection and Meaning





We can construct a features by measuring the distance in latitude from Washington Square Park (40.7308° N, 73.9973° W). In the image above I have thrown away some outliers, and we see a much stronger correlation than in for latitude and longitude.

Low Dimensional Projection and Meaning



Combining the two to find the distance gives and even better correlation. What we are seeing here is an example of a constructed feature. With a bit of understand of our data, we can often construct much more useful features than we are given to start out with. We'll see later that there is an interplay between feature construction and model selection.

Low Dimensional Projection and Meaning

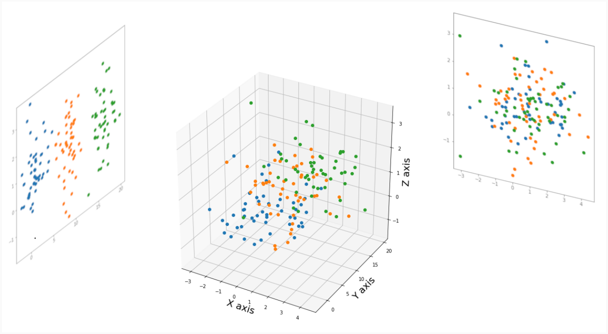
Constructed features play an important role in terms of subset selection. Using our naive subset selection for a linear model, we would quickly throw away both latitude and longitude as being unpredictive. However, when we pick the right combination of the features suddenly we're given one our best predictors.

The process of constructing meaningful features is called **feature engineering**. Feature engineering is useful both in terms of constructing better models, and communicating results in a human readable format. One of the project of statistics, and by extension machine learning, is to construct a minimal set of mean rich features that explain all of the variance in the data.

To this end, feature engineering often uses the techniques of dimensional reduction, and visa versa.

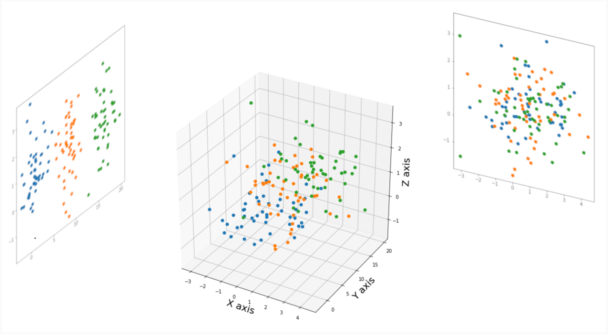
Dimensional Reduction

Low Dimensional Projection and Meaning



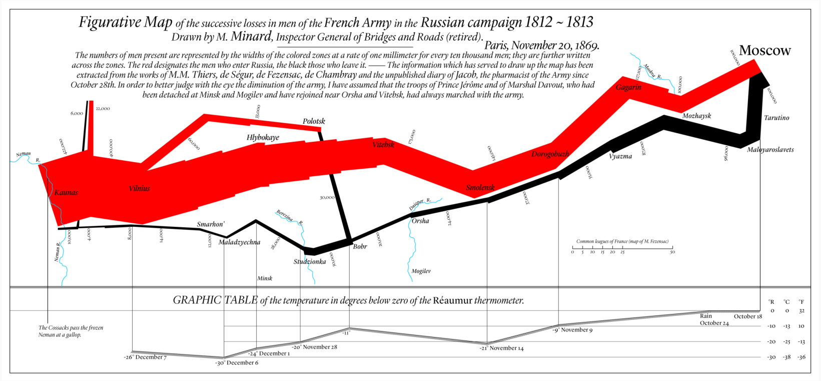
Dimensional Reduction is the process of combining high dimensional information into low dimensional representations. Low dimensional representations can be computationally easier to work with, while hopefully throwing away only extraneous information.

Low Dimensional Projection and Meaning



Low dimensional projections are *meaningful*. In fact, all of data visualization can be said to be the result careful low dimensional projections. One goal is to discover the small set of variables that controls the larger phenomena.

Low Dimensional Projection and Meaning



Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813
Drawn by M. Minard, Inspector General of Bridges and Roads (retired).
Paris, November 20, 1869.

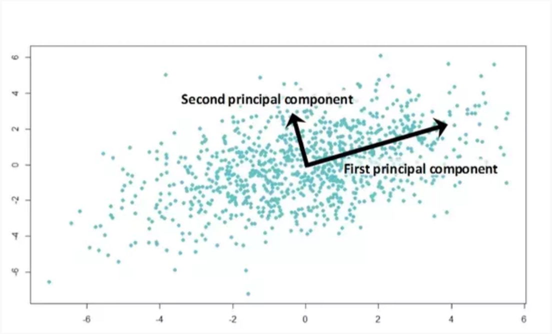
The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. —— The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.

GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.

The Cossacks pass the frozen Neman at a gallop.

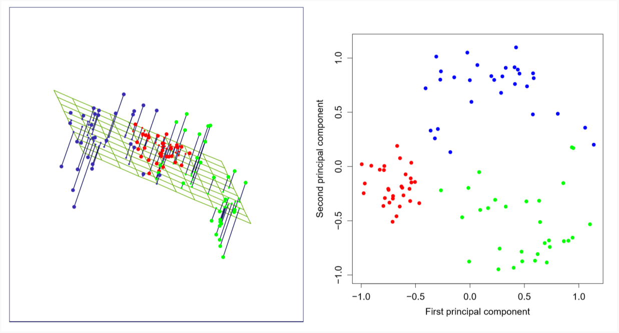
For example in *The Visual Display of Quantitative Information*, Edward Tufte points out that Charles Minard's map of Napoleon's march into Russia is one of the best examples of dimensional reduction, with 8 variables being represented using 2 dimensions, and effectually along 1.

Projections



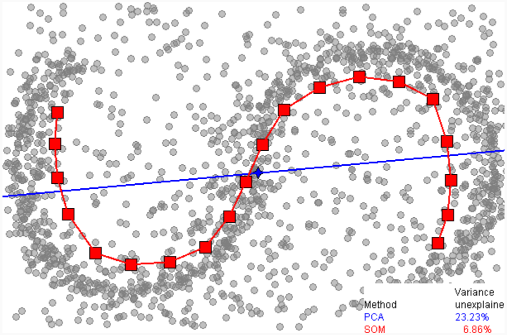
There are two main techniques in dimensional reduction: **projection onto linear subspaces** and **manifold learning**. In projection onto linear subspaces, we try to discover the linear combination of features the provide the clearest coordinate system for the dataset, ie the **component factors**.

Projections



In an alternative view, factor analysis tries to fit a linear subspace to a data set in such a way as to maximize the **projected variance**, that is maximize the amount of data variance captured in orthogonal project to the new space.

Projections



In **manifold learning**, we try to fit a more complicated manifold to the underlying data. Manifold learning is of course much more complicated than simple linear dimensional reduction, but there do exist good algorithms which we will talk more about later in the semester.

Factor Analysis

The Formal Problem

If we have some insight into the features, we can construct new features using previously known relationships. However, often in machine learning problems we may not have a road map for the construction of new features.

One project of machine machine learning is finding effective and meaningful projections of data into lower dimensions when we *don't* have any information about the features. This means finding a low dimensional representation that preserves the important structures of the high dimensional data.

Factor Analysis tries to find low dimensional projections that preserve certain quantities, for example the distance between data points, the variance of the dataset as a whole, or probability that two labels will co-occur. We take a moment to go over some of the common ideas in factor analysis.

The Formal Problem

Given N data points $x_i \in \mathbb{R}^p$, we want to find a linear transform U that reduces x_i from a p dimensional feature space to a $k < p$ dimensional feature space without distorting the relationship between points.

Practically, this means we want to preserve the inner product:

$$\langle Ux_i, Ux_j \rangle = \langle x_i, x_j \rangle .$$

Therefore U is a $k \times p$ isometry. For the Cartesian inner product the isometries are the orthogonal matrices $U^T U = I$. Note that $UU^T \neq I$, since this would imply that lowering dimension lost no information.

The Formal Problem

The goal then is to find a new set of vectors $z_i = Ux_i$ that best represent the data, that is find the k directions in $\mathbb{R}^p$ the best represent x_i.

We will proceed by optimizing two objectives:

Minimize the **reconstruction error.**

Maximize the **projected variance.**

We will see that these are in fact equivalent conditions.

The Reconstruction Error

We can think about $z_i = Ux_i$ as an **encoding** or **compression** of x_i. Undoing the linear transformation then amounts to a decoding

$$\widetilde{x}_i = U^T z_i = U^T U x_i \,.$$

The **reconstruction error** is the difference between the original data and the decoding:

$$\sum_{i=1}^{N} ||x_i - \widetilde{x}_i|| = \sum_{i=1}^{N} ||x_i - U^T U x_i|| \,,$$

and the problem can be stated as finding

$$\min_{U \in \mathbb{R}^{k \times p}} \sum_{i=1}^{N} ||x_i - U^T U x_i||^2 \,, \qquad UU^T = I_k \,.$$

The Projected Variance

Similarly, we may try to minimize the **projected sample variance** $\widehat{\text{var}}[Ux]$. The sample variance is given as usual by

$$\widehat{\text{var}}\left[\, Ux \,\right] = \widehat{E}\left[\, (Ux)^2 \,\right] - \widehat{E}\left[\, (Ux) \,\right]^2 .$$

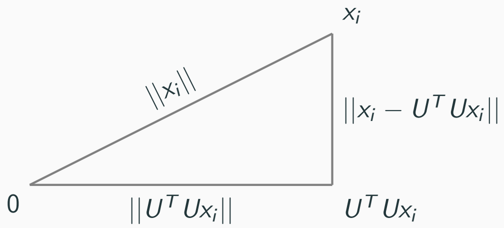
Assume our training data has been centered, that is $\widehat{E}[\,(x_i)_j\,]$ for each feature $j = 1, \ldots, p$. Then the sample variance is

$$\widehat{\text{var}}\left[\, Ux \,\right] = \widehat{E}\left[\, (Ux)^2 \,\right] .$$

Out goal then is to find

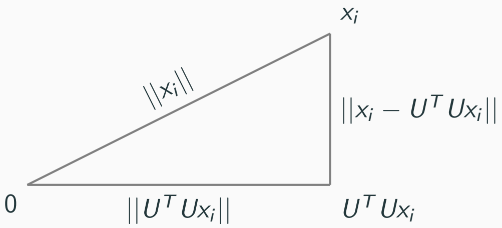
$$\max_{U \in \mathbb{R}^{k \times p}} \widehat{E}\left[\, (Ux)^2 \,\right] , \qquad UU^T = I_k .$$

Equivalence of Conditions



The diagram above show thatTo see that the two conditions are equivalent, we will show that the variance in the data (which is fixed) is the sum of the **projected sample variance** and the **reconstruction error**.

Equivalence of Conditions



Taking the expectation of the Pythagorean decomposition

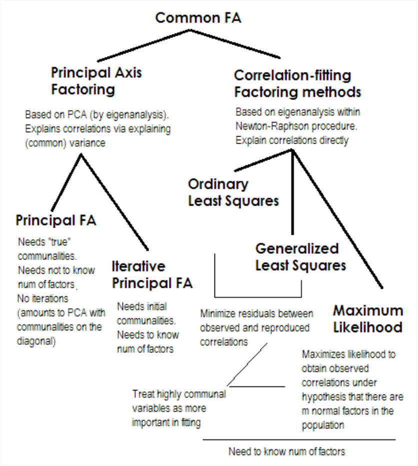
$$x_i = U^T U x_i + x_i - U^T U x_i$$

yields

$$\underbrace{\widehat{E}\left[\,||x_i||^2\,\right]}_{\text{Sample Variance}} = \underbrace{\widehat{E}\left[\,||U^T U x_i||^2\,\right]}_{\text{Projected Variance}} + \underbrace{\widehat{E}\left[\,||x_i - U^T U x_i||^2\,\right]}_{\text{Reconstruction Error}}$$

We see the minimizing reconstruction error is equivalent to maximizing projected variance.

Techniques in Factor Analysis


Common FA
Principal Axis Factoring
Based on PCA (by eigenanalysis). Explains correlations via explaining (common) variance
Correlation-fitting Factoring methods
Based on eigenanalysis within Newton-Raphson procedure. Explain correlations directly
Principal FA
Needs "true" communalities. Needs not to know num of factors. No iterations (amounts to PCA with communalities on the diagonal)
Iterative Principal FA
Needs initial communalities. Needs to know num of factors
Ordinary Least Squares
Generalized Least Squares
Minimize residuals between observed and reproduced correlations
Treat highly communal variables as more important in fitting
Maximum Likelihood
Maximizes likelihood to obtain observed correlations under hypothesis that there are m normal factors in the population
Need to know num of factors

There is no consensus on how to perform factor analysis. The problem mainly comes down to choosing the how to determine the number of factors to minimize on.

https://stats.stackexchange.com/questions/50745/

Principle Component Analysis

Principle Component Analysis

The most common technique for factor analysis is **Principle Component Analysis** or **PCA**. It is in fact so common that "factor analysis" usually means "factor analysis that isn't PCA."

PCA is a greedy algorithm with a beautiful mathematical interpretation. The idea is to proceed iteratively:

- Find principle component (direction) that accounts for the largest possible variance.
- Project onto the subspace orthogonal to that vector.
- Repeat, collecting the orthogonal vectors into the rows of U until U is a $p \times p$ orthogonal matrix.

Principle Component Analysis

Let x_i be demeaned data points and let $\mathbf{X}$ be the matrix whose rows are x_i^T. We want to find a vector $w_{(1)} = (w_1, \ldots, w_p)_{(1)}$ with $||w_{(1)}|| = 1$ that maximizes the projected variance. Projection of x_i onto $w_{(1)}$ is given by

$$\mathbf{proj}_{w_{(1)}}(x_i) = \frac{\langle x_1, w_{(1)} \rangle}{||w_{(1)}||^2} w_{(1)} = \langle x_1, w_{(1)} \rangle w_{(1)} \,.$$

Since x_i have mean 0, the variance along $w_{(1)}$ is just

$$\widehat{E}\left[\, ||\langle x_1, w_{(1)} \rangle w_{(1)}||^2 \,\right] = \widehat{E}\left[\, \langle x_1, w_{(1)} \rangle^2 \,\right].$$

Writing $\langle x_1, w_{(1)} \rangle = x_1^T w_{(1)}$ for the Cartesian inner product, our goal is to find

$$w_{(1)} = \underset{||w_{(1)}||=1}{\operatorname{argmax}} \left(\sum_{i=1}^{N} (x_i \cdot w_{(1)})^2 \right) = \underset{||w_{(1)}||=1}{\operatorname{argmax}} \left(\, ||\mathbf{X} w_{(1)}||^2 \, \right).$$

Principle Component Analysis

Writing the Euclidean norm as matrix multiplication, we see that it is equivalent to find

$$w_{(1)} = \underset{||w_{(1)}||=1}{\operatorname{argmax}} \left(||\mathbf{X} w_{(1)}||^2 \right) = \underset{||w_{(1)}||=1}{\operatorname{argmax}} \left(w_{(1)}^T \mathbf{X}^T \mathbf{X} w_{(1)} \right).$$

The maximum is not so clear using our usual matrix differentiation. Indeed, the condition that $||w_{(1)}|| = 1$ plays an important role in the expression even having a maximum. We can use a Lagrangian procedure to rewrite the above as a single maximization problem:

$$\mathcal{L} = w_{(1)}^T \mathbf{X}^T \mathbf{X} w_{(1)} + \lambda(1 - w_{(1)}^T w_{(1)}).$$

(Homework) Show that the stationary points of $\mathcal{L}$ are the solutions to the eigenvalue equation $\mathbf{X}^T \mathbf{X} w_{(1)} = \lambda w_{(1)}$.

Principle Component Analysis

Since the unit eigenvectors of $\mathbf{X}^t\mathbf{X}$ extremize $w_{(1)}^T\mathbf{X}^T\mathbf{X}w_{(1)}$,

$$\underset{||w_{(1)}||=1}{\operatorname{argmax}} \left(w_{(1)}^T\mathbf{X}^T\mathbf{X}w_{(1)} \right)$$

is the eigenvector of $\mathbf{X}^T\mathbf{X}$ with the largest eigenvalue λ. Furthermore,

$$\widehat{\mathrm{var}}(\mathbf{X}x_{(1)}) = \frac{1}{N}w_{(1)}^T\mathbf{X}^T\mathbf{X}w_{(1)} = \frac{1}{N}w_{(1)}^T\left(\lambda w_{(1)}\right) = \frac{\lambda}{N}\,.$$

But recall that the covariance matrix of the data

$$\widehat{\mathrm{Cov}}(\mathbf{X}) = \frac{1}{N}\mathbf{X}^T\mathbf{X}\,.$$

We have shown that the first principle component is the eigenvector of the covariance matrix with the highest eigenvalue.

Principle Component Analysis

The result extends to the subsequent principle components. The proof follow exactly ass for the first principle component and in fact constructs the orthogonal decomposition

$$\widehat{\mathrm{Cov}}(\mathbf{X}) = W\Lambda^{(p)}W^T \,,$$

where $\lambda^{(p)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, with $\lambda_i > \lambda_{i+1}$.

Computationally, the complexity is $O(Np^2 + p^3)$, the sum computing the covariance and eigenbasis respectively.

Truncated Principle Component Analysis

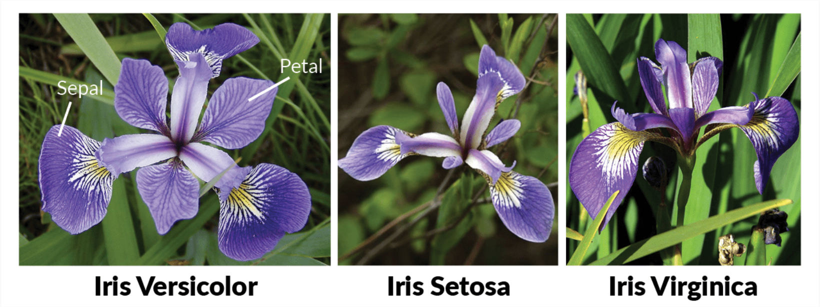
For dimensional reduction, we may stop this process at any step, revealing a **truncated** PCA matrix. Keeping only the first k principle components yields a orthogonal matrix W_k which projects x_i to the space spanned by the first k principle vectors.

Mathematically, this performs decomposition of the covariance matrix into

$$\widehat{\mathrm{Cov}}(\mathbf{X}) = W \Lambda^{(k)} W^T ,$$

where $\lambda^{(k)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, with $\lambda_i > \lambda_{i+1}$.

Machine Learning Examples: Iris Classification

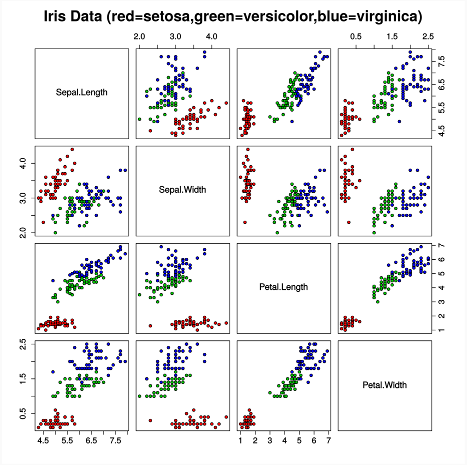


Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	I. Setosa
7.0	3.2	4.7	1.4	I. Versicolor

$\vdots$

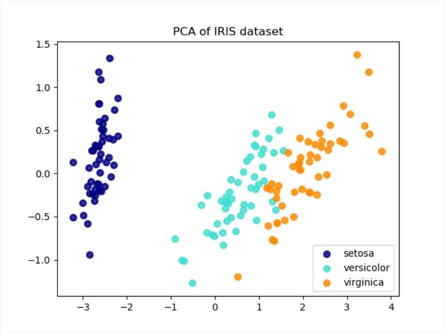
Take as an example projection onto the first two principle component in the Iris data set from Lecture 1.

Example: PCA for Iris



There are four variables, with correlations given above.

Example: PCA for Iris

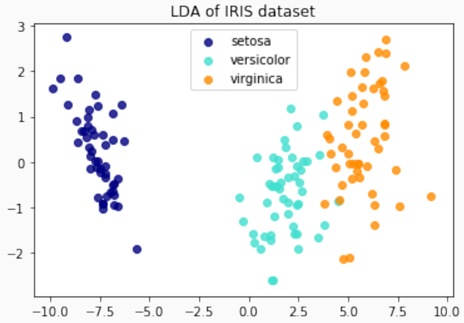


Using PCA to project onto the first two principle components yields a projection onto

$$w_{(1)} = (0.36, -0.08, 0.86, 0.36)\,, \qquad w_{(2)} = (0.66, 0.73, -0.18, -0.07)$$

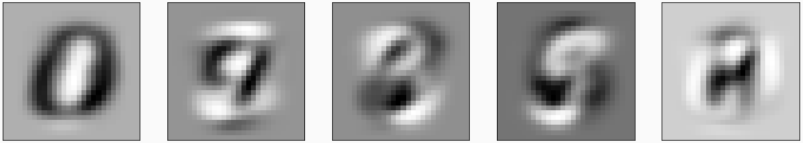
We see the species are separable knowing only their features, and furthermore we have concrete measurement ratios to determine species.

Example: PCA for Iris



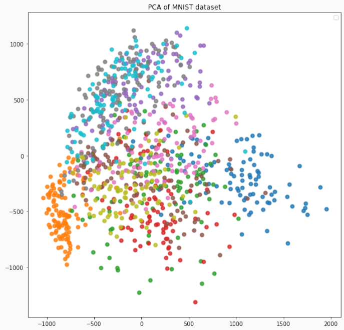
Compare for a moment with the LDA projection. LDA requires knowing the labels and requires significantly more computational time for a similar fit.

Example: PCA MNIST



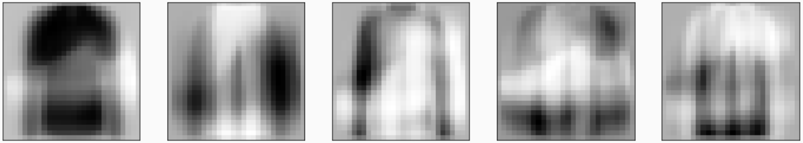
The MNIST data set may provide one of the most startling examples. A simple PCA can be computed quite quickly and yields a wealth of information about the structure of the data set.

Example: PCA MNIST



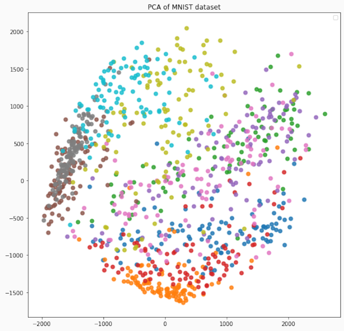
The MNIST data set may provide one of the most startling examples. A simple PCA can be computed quite quickly and yields a wealth of information about the structure of the data set. Projecting onto the first two components shows a lot of structure on the MNIST dataset.

Example: PCA MNIST



This is not always as useful. Performing PCA on the fashion MNIST data set doesn't yield as strong of forms as on the digits MNIST.

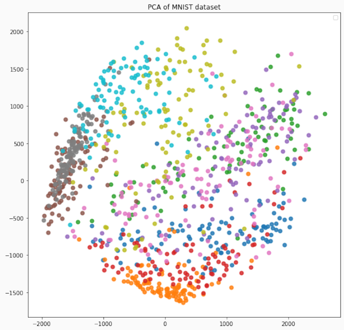
Example: PCA MNIST



This is not always as useful. Performing PCA on the fashion MNIST data set doesn't yield as strong of forms as on the digits MNIST. On the other hand, we see much more structure and organization in the projection onto eignevectors.

For an excellent article on visualization of MNIST, see `http://colah.github.io/posts/2014-10-Visualizing-MNIST/`

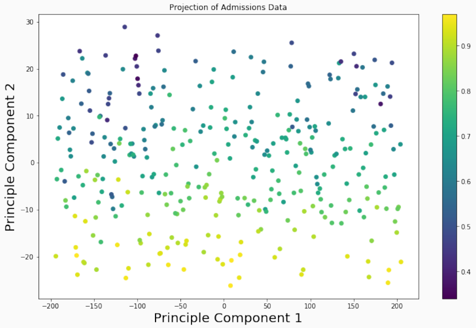
Example: PCA MNIST



This is not always as useful. Performing PCA on the fashion MNIST data set doesn't yield as strong of forms as on the digits MNIST. On the other hand, we see much more structure and organization in the projection onto eignevectors.

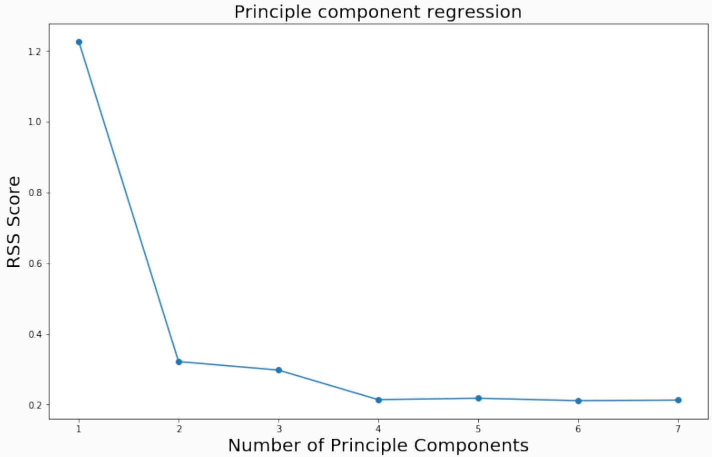
For an excellent article on visualization of MNIST, see `http://colah.github.io/posts/2014-10-Visualizing-MNIST/`

Nonlinear PCA



In the context of regression, PCA can perform subset selection for us. In the graph above, we used PCA to project the features of the Student Admission dataset onto the first two principle components. It's important to note though that PCA doesn't know anything about the labels Y. This can be an advantage, unlabeled data can import the PCA (an unlabeled data is often cheap).

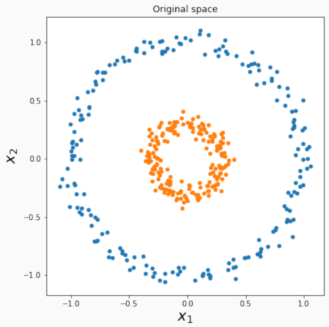
Nonlinear PCA



But that doesn't necessarily lead to a better fit.

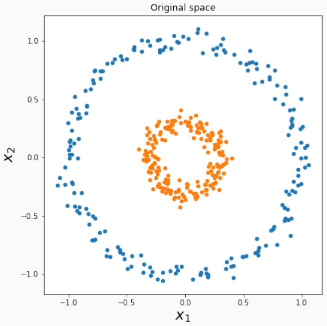
Nonlinear PCA

Nonlinear PCA



Of course, linear methods may be almost useless if the underlying structure isn't linear. For truly unknown structure we must construct smoothing maps using splines, random graphs, or other high complexity techniques. For example, in the distribution above linear PCA will do nothing to reduce the complexity of the data set.

Nonlinear PCA



However, whenever we have linear methods there is the hope that we can extend them in a computationally efficient manner using the kernel trick.

Nonlinear PCA

The kernel trick was to notice that if our optimization only refers to the higher dimensional feature space by way of the inner product, than we can often replace it with a much more computationally efficient kernel $k(x_i, x_j)$.

In keeping with the literature, let $\phi : \mathcal{X} \to \widetilde{\mathcal{X}}$ be an embedding of the feature space into a higher dimensional space (say, the space of degree d polynomials in the features). Assume that projection has 0 mean.

The covariance matrix in the embedded space can be written

$$\widetilde{C} = \widetilde{\mathrm{Cov}} = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)\phi(x_i)^T ,$$

with eigenvectors w_k and eigenvalues $\widetilde{C} w_k = \lambda_k w_k$.

Nonlinear PCA

Since

$$\widetilde{C} w_k = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)\big(\, \phi(x_i)^T \, w_k\big) = \lambda_k w_k \, ,$$

we can write

$$w_k = \sum_{i=1}^{N} a_{ki} \phi(x_i) \, .$$

If there exists $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, then for any $\phi(x_\ell)^T$ the first equation can be written

$$\phi(x_\ell)^T \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)\big(\, \phi(x_i)^T \, \sum_{j=1}^{N} a_{kj} \phi(x_j)\big) = \lambda_k \phi(x_\ell)^T \sum_{j=1}^{N} a_{kj} \phi(x_j) \, ,$$

or

$$\frac{1}{N} \sum_{i=1}^{N} K(x_\ell, x_i) \sum_{i=j}^{N} K(x_i, x_j) = \lambda_k \sum_{j=1}^{N} K(x_\ell, x_j) \, .$$

Nonlinear PCA

Writing $\mathbf{K}_{ij} = K(x_i, x_j)$, and $a_k = [a_{k1}, \ldots, a_{kN}]^T$

$$\frac{1}{N} \sum_{i=1}^{N} K(x_\ell, x_i) \sum_{i=j}^{N} K(x_i, x_j) = \lambda_k \sum_{j=1}^{N} K(x_\ell, x_j)\,,$$

can be written

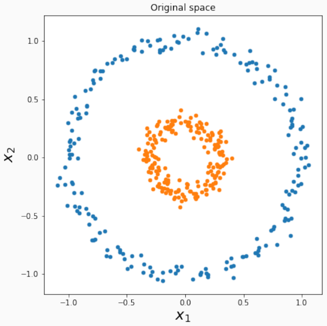
$$\mathbf{K}^2 a_k = \lambda_k N \mathbf{K} a_k\,.$$

So the higher dimensional fitting is solved by the kernel eigenvalue problem

$$\mathbf{K} a_k = \lambda_k N a_k\,.$$

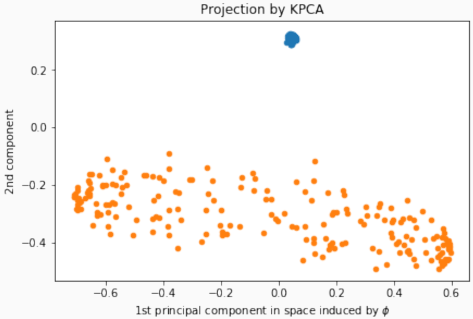
Again, this trick is restricted to transformations for which there exists a computable kernel, but if one exists this is a remarkable reduction in complexity.

Nonlinear PCA



Applying the radial basis function kernel PCA to the concentric circle data set results in a dramatic clustering.

Nonlinear PCA



Applying the radial basis function kernel PCA to the concentric circle data set results in a dramatic clustering.

References

References for this lecture: BS, Chapters 15, 16, 18 and ESII, Chapters 9, 12.

Charles Minard's Map:
`https://en.wikipedia.org/wiki/Charles_Joseph_Minard`