CS724: Topics in Algorithms
Dimensionality Reduction
Slide Set 10

Prof. Dan A. Simovici

UMASS BOSTON

UMASS BOSTON

The Euler Functions

The integrals

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx \text{ and } \Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$$

are known as *Euler's integral of the first type* and *Euler's integral of the second type*, respectively. We assume here that a and b are positive numbers to ensure that the integrals are convergent.

A few facts on Euler's integrals:

- $B(a,b) = B(b,a)$;
- $$B(a,b) = \frac{b-1}{a+b-1} B(a,b-1);$$
- $$B(a,b) = \frac{a-1}{a+b-1} \cdot B(a-1,b);$$
- $$B(a,n) = B(n,a) = \frac{1 \cdot 2 \cdot \cdots \cdot (n-1)}{a \cdot (a+1) \cdot \cdots \cdot (a+n-1)}.$$

- $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
- The Γ function is a generalization of the factorial because $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.
- $$\Gamma'(a) = \int_0^\infty x^{a-1}(\ln x)e^{-x}dx,$$

  and, in general, $\Gamma^{(n)}(a) = \int_0^\infty x^{a-1}(\ln x)^n e^{-x}dx$. Thus, $\Gamma^{(2)}(a) > 0$, which shows that the first derivative is increasing.
- An integral that is useful for a variety of applications is

$$I = \int_{\mathbb{R}} e^{-\frac{1}{2}t^2}dt = \sqrt{2\pi}.$$

Using this integral, we can compute the value of $\Gamma\left(\frac{1}{2}\right)$.
Since $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-x}}{\sqrt{x}} dx$, by applying the change of variable $x = \frac{t^2}{2}$, we have

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{2} \cdot \int_0^\infty e^{-\frac{1}{2}t^2} dt = \sqrt{\pi}. \tag{1}$$

The relationship between B and Γ is

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Note that

$$\Gamma\left(p+\frac{1}{2}\right)=\left(p-\frac{1}{2}\right)\Gamma\left(p-\frac{1}{2}\right)=\left(p-\frac{1}{2}\right)\left(p-\frac{3}{2}\right)\Gamma\left(p-\frac{3}{2}\right)=\cdots$$

The last equality allows us to compute the values of the form $\Gamma\left(\frac{2p+1}{2}\right)$. It is easy to see that

$$\Gamma\left(\frac{2p+1}{2}\right)=\frac{(2p-1)\cdot(2p-3)\cdots 3\cdot 1}{2^p}\sqrt{\pi}=\frac{(2p)!}{p!2^{2p}}\sqrt{\pi}. \qquad (2)$$

Details on Euler's Integrals

Replacing x by $1 - x$ yields the equality

$$B(a, b) = -\int_1^0 (1 - x)^{a-1}(x)^{b-1} dx = B(b, a),$$

which shows that B is symmetric.

Integrating $B(a,b)$ by parts, we obtain

$$\begin{aligned} B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1}dx = \int_0^1 (1-x)^{b-1}d\frac{x^a}{a} \\ &= \frac{x^a(1-x)^{1-b)}}{a}\Bigg|_0^1 + \frac{b-1}{a}\int_0^1 x^a(1-x)^{b-2}dx \\ &= \frac{b-1}{a}\int_0^1 x^{a-1}(1-x)^{b-2}dx - \frac{b-1}{a}\int_0^1 x^{a-1}(1-x)^{b-1}dx \\ &= \frac{b-1}{a}B(a,b-1) - \frac{b-1}{a}B(a,b), \end{aligned}$$

which yields

$$B(a,b) = \frac{b-1}{a+b-1}B(a,b-1). \tag{3}$$

The symmetry of the function B allows us to infer the formula

$$B(a,b) = \frac{a-1}{a+b-1}\cdot B(a-1,b).$$

If b is a natural number n, a repeated application of Equality (3) allows us to write

$$B(a,n) = \frac{n-1}{a+n-1} \cdot \frac{n-2}{a+n-2} \cdots \frac{1}{a+1} \cdot B(a,1).$$

The last factor of this equality, $B(a,1)$, is easily seen to equal $\frac{1}{a}$. Thus,

$$B(a,n) = B(n,a) = \frac{1 \cdot 2 \cdot \cdots \cdot (n-1)}{a \cdot (a+1) \cdot \cdots \cdot (a+n-1)}.$$

If a is also a natural number, $a = m \in \mathbb{N}$, then

$$B(m,n) = \frac{(n-1)!(m-1)!}{(m+n-1)!}.$$

Next, we show the connection between Euler's integral functions:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{4}$$

Replacing x in the integral

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

by $x = ry$ with $r > 0$ gives $\Gamma(a) = r^a \int_0^\infty y^{a-1}e^{-ry}dy$.
Replacing a by $a+b$ and r by $r+1$ yields the equality

$$\Gamma(a+b)(r+1)^{-(a+b)} = \int_0^\infty y^{a+b-1}e^{-(r+1)y}dy.$$

By multiplying both sides by r^{a-1} and integrating, we have

$$\Gamma(a+b)\int_0^\infty r^{a-1}(r+1)^{-(a+b)}dr = \int_0^\infty r^{a-1}\left(\int_0^\infty y^{a+b-1}e^{-(r+1)y}dy\right)dr.$$

By the definition of B, the last equality can be written

$$\Gamma(a+b)B(a,b) = \int_0^\infty r^{a-1}\left(\int_0^\infty y^{a+b-1}e^{-(r+1)y}dy\right)dr.$$

By permuting the integrals from the right member (we omit the justification of this manipulation), the last equality can be written as

$$\Gamma(a+b)B(a,b) = \int_0^\infty y^{a+b-1}e^{-y}\left(\int_0^\infty r^{a-1}e^{-ry}dr\right)dy.$$

Note that $\int_0^\infty r^{a-1}e^{-ry}dr = \frac{\Gamma(a)}{y^a}$. Therefore,

$$\Gamma(a+b)B(a,b) = \int_0^\infty y^{a+b-1}e^{-y}\frac{\Gamma(a)}{y^a}dy = \int_0^\infty y^{b-1}e^{-y}\Gamma(a)dy = \Gamma(a)\Gamma(b),$$

which is Formula (4).

The Γ function is a generalization of the factorial. Starting from the definition of Γ and integrating by parts, we obtain

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx = \frac{x^a}{a} e^{-x} \Big|_0^\infty + \frac{1}{a} \int_0^\infty x^a e^{-x} dx = \frac{1}{a} \Gamma(a+1).$$

Thus, $\Gamma(a+1) = a\Gamma(a)$. Since $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$, it is easy to see that $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

It is possible to show that Γ has derivatives of arbitrary order and that we can compute these derivatives by deriving the function under the integral sign. Namely, we can write:

$$\Gamma'(a) = \int_0^\infty x^{a-1}(\ln x)e^{-x}dx,$$

and, in general, $\Gamma^{(n)}(a) = \int_0^\infty x^{a-1}(\ln x)^n e^{-x}dx$. Thus, $\Gamma^{(2)}(a) > 0$, which shows that the first derivative is increasing.
Since $\Gamma(1) = \Gamma(2) = 1$, there exists $a \in [1, 2]$ such that $\Gamma'(a) = 0$. For $0 < x < a$, we have $\Gamma'(x) \leqslant 0$, so Γ is decreasing. For $x > a$, $\Gamma'(x) \geqslant 0$, so Γ is increasing. It is easy to see that

$$\lim_{x \to 0+} \Gamma(x) = \frac{\Gamma(x+1)}{x} = \infty,$$

and $\lim_{x\to\infty} \Gamma(x) = \infty$.

An integral that is useful for a variety of applications is

$$I = \int_{\mathbb{R}} e^{-\frac{1}{2}t^2} dt.$$

We prove that $I = \sqrt{2\pi}$.
We can write

$$I^2 = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} dx \cdot \int_{\mathbb{R}} e^{-\frac{1}{2}y^2} dy = \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dxdy.$$

Changing to polar coordinates by using the transformation $x = \rho\cos\theta$ and $y = \rho\sin\theta$ whose Jacobian is

$$\begin{vmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{vmatrix} = \rho,$$

we have

$$I^2 = \int_{\mathbb{R}^2} e^{-\frac{\rho^2}{2}} \rho d\rho d\theta = \int_0^{2\pi} d\theta \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho = 2\pi.$$

Thus, $I = \sqrt{2\pi}$. Since $e^{-\frac{1}{2}t^2}$ is an even function, it follows that

$$\int_0^{\infty} e^{-\frac{1}{2}t^2} dt = \sqrt{\frac{\pi}{2}}.$$

Volume of Spheres

A closed sphere centered in $(0, \ldots, 0)$ and having the radius R in $\mathbb{R}^n$ is defined as the set of points:

$$S_n(R) = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 = 1 \right\}.$$

The volume of this sphere is denoted by $V_n(R)$.
We approximate the volume of an n-dimensional sphere of radius R as a sequence of $n-1$-dimensional spheres of radius $r(u) = \sqrt{R^2 - u^2}$, where u varies between $-R$ and R. This allows us to write

$$V_{n+1}(R) = \int_{-R}^{R} V_n(r(u))du.$$

We seek $V_n(R)$ as a number of the form $V_n(R) = k_n R^n$. Thus, we have

$$\begin{aligned} V_{n+1}(R) &= k_n \int_{-R}^{R} (r(u))^n du = k_n \int_{-R}^{R} (R^2 - u^2)^{\frac{n}{2}} du \\ &= k_n R^n \int_{-R}^{R} \left(1 - \left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}} du \\ &= V_n(R) \int_{-R}^{R} \left(1 - \left(\frac{u}{R}\right)^2\right)^{\frac{n}{2}} du = R V_n(R) \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx. \end{aligned}$$

In turn, this yields the recurrence

$$k_{n+1} = k_n \int_{-1}^{1} (1 - x^2)^{\frac{n}{2}} dx.$$

Note that

$$\int_{-1}^{1}(1-x^2)^{\frac{n}{2}}dx = 2\cdot\int_{0}^{1}(1-x^2)^{\frac{n}{2}}dx$$

because the function $(1-x^2)^{\frac{n}{2}}$ is even. To compute the latest integral, substitute $u = x^2$. We obtain

$$\int_{0}^{1}(1-x^2)^{\frac{n}{2}}dx = \frac{1}{2}\int_{0}^{1}u^{-\frac{1}{2}}(1-u)^{\frac{n}{2}}du,$$

which equals $\frac{1}{2}\cdot B(\frac{1}{2},\frac{n}{2}+1)$. Using the Γ function, the integral can be written as

$$\int_{0}^{1}(1-x^2)^{\frac{n}{2}}dx = \frac{1}{2}\cdot\frac{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2}+1)}{\Gamma\left(\frac{n}{2}+\frac{3}{2}\right)}.$$

Thus,

$$k_{n+1} = k_n \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\left(\frac{n+1}{2}+1\right)}.$$

Since $k_1 = 2$, this implies

$$k_n = 2\left(\Gamma\left(\frac{1}{2}\right)\right)^{n-1}\frac{\Gamma\left(\frac{1}{2}+1\right)}{\Gamma\left(\frac{n}{2}+1\right)} = \left(\Gamma\left(\frac{1}{2}\right)\right)^{n}\frac{1}{\Gamma\left(\frac{n}{2}+1\right)} = \pi^{\frac{n}{2}}\frac{1}{\Gamma\left(\frac{n}{2}+1\right)}.$$

Thus, the volume of the n-dimensional sphere of radius R equals

$$\frac{\pi^{\frac{n}{2}}R^n}{\Gamma\left(\frac{n}{2}+1\right)}.$$

For $n = 1, 2, 3$, by applying Formula (2), we obtain the well-known values $2R, \pi R^2$, and $\frac{4\pi R^3}{3}$, respectively. For $n = 4$, the volume of the sphere is $\frac{\pi^2 R^4}{2}$.

The Dimensionality Curse

The term "dimensionality curse," invented by Richard Bellman is used to describe the difficulties of exhaustively searching a space of high dimensionality for an optimum value of a function defined on such a space. These difficulties stem from the fact that the size of the sets that must be searched increases exponentially with the number of dimensions. Moreover, phenomena that are at variance with the common human intuition acquired in two- or three-dimensional spaces become more significant. This section is dedicated to a study of these phenomena.

The dimensionality curse impacts many data mining tasks, including classification and clustering. Thus, it is important to realize the limitations that working with high-dimensional data impose on designing data mining algorithms.

Let $Q_n(\ell)$ be an n-dimensional cube in $\mathbb{R}^n$. The volume of this cube is ℓ^n. Consider the n-dimensional closed sphere of radius R that is centered in the center of the cube $Q_n(2R)$ and is tangent to the opposite faces of this cube. We have:

$$\lim_{n\to\infty} \frac{V_n(R)}{2^n R^n} = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma\left(\frac{n}{2}+1\right)} = 0.$$

In other words, as the dimensionality of the space grows, the fraction of the cube volume that is located inside the sphere decreases and tends to become negligible for very large values of n.

It is interesting to compare the volumes of two concentric spheres of radii R and $R(1-\epsilon)$, where $\epsilon \in (0,1)$. The volume located between these spheres relative to the volume of the larger sphere is

$$\frac{V_n(R) - V_n(R(1-\epsilon))}{V_n(R)} = 1 - (1-\epsilon)^n,$$

and we have

$$\lim_{n\to\infty} \frac{V_n(R) - V_n(R(1-\epsilon))}{V_n(R)} = 1.$$

Thus, for large values of n, the volume of the sphere of radius R is concentrated mainly near the surface of this sphere.

Let $Q_n(1)$ be a unit side-length n-dimensional cube, $Q_n(1) = [0,1]^n$, centered in $\boldsymbol{c}_n = (0.5, \ldots, 0.5) \in R^n$. The d_2-distance between the center of the cube $\boldsymbol{c}_n$ and any of its vertices is $\sqrt{0.5^2 + \cdots 0.5^2} = 0.5\sqrt{n}$, and this value tends to infinity with the number of dimensions n despite the fact that the volume of the cube remains equal to 1. On the other hand, the distance from the center of the cube to any of its faces remains equal to 0.5. Thus, the n-dimensional cube is exhibits very different properties in different directions; in other words the n-dimensional cube is an anisotropic object.

An interesting property of the unit cube $Q_n(1)$:
Let $P = (p, \ldots, p) \in \mathbb{R}^n$ be a point located on the main diagonal of $Q_n(1)$ and let K be the subcube of $Q_n(1)$ that includes $(0, \ldots, 0)$ and P and has a side of length p; similarly, let K' be the subcube of $Q_n(1)$ that includes P and $(1, \ldots, 1)$ and has side of length $1 - p$. The ratio of the volumes V and V' of the cubes K and K' is

$$r(p) = \left(\frac{p}{1-p}\right)^n.$$

To determine the increase δ of p needed to double this ratio, we must find δ such that $\frac{r(p+\delta)}{r(p)} = 2$.

Thus, we must have

$$\frac{\frac{p+\delta}{1-p-\delta}}{\frac{p}{1-p}} = \frac{p(1-p)+\delta(1-p)}{p(1-p)-\delta p} = \sqrt[n]{2}.$$

Equivalently, we have

$$\delta = \frac{p(1-p)(\sqrt[n]{2}-1)}{1-p+p\sqrt[n]{2}}.$$

For large dimensionality n smaller and smaller moves of the point p are needed to double the ratio of the volumes of the cubes K and K'. This suggests that the division of $Q_n(1)$ into subcubes is very unstable. If data classifications are attempted based on the location of data vectors in subcubes, this shows in turn the instability of such classification schemes.

Another interesting example of the counterintuitive behavior of spaces of high dimensionality:
Let $Q_n(1)$ be the unit cube centered in the point $\boldsymbol{c}_n \in \mathbb{R}^n$, where $\boldsymbol{c}_n = (0.5, \ldots, 0.5)$. For $n = 2$ or $n = 3$, it is easy to see that every sphere that intersects the sides of $Q_2(1)$ or all faces of $Q_3(1)$ must contain the center of the cube $\boldsymbol{c}_n$. We shall see that, for sufficiently high values of n a sphere that intersects all $(n-1)$-dimensional faces of $Q_n(1)$ does not necessarily contain the center of $Q_n(1)$.

Consider the closed sphere $B(\boldsymbol{q}_n, r)$, whose center is the point $\boldsymbol{q}_n = (q, \ldots, q)$, where $q \in [0, 1]$. Clearly, we have $\boldsymbol{q}_n \in Q_n(1)$ and $d_2(\boldsymbol{c}_n, \boldsymbol{q}_n) = \sqrt{n(q^2 - q + 0.25)}$.
If the radius r of the sphere $B(\boldsymbol{q}_n, r)$ is sufficiently large, then $B(\boldsymbol{q}_n, r)$ intersects all faces of Q_n. Indeed, the distance from $\boldsymbol{q}_n$ to an $(n-1)$-dimensional face is no more than $\max\{q, 1-q\}$, which shows that $r \geqslant \max\{q, 1-q\}$ ensures the nonemptiness of all these intersections. Thus, the inequalities

$$n\,(q - 0.5)^2 > r^2 > \max\{q^2, (1-q)^2\} \tag{5}$$

ensure that $B(\boldsymbol{q}_n, r)$ intersects every $(n-1)$-dimensional face of Q_n, while leaving $\boldsymbol{c}_n$ outside $B(\boldsymbol{q}_n, r)$. This is equivalent to requiring

$$n > \frac{\max\{q^2, (1-q)^2\}}{(q - 0.5)^2}.$$

For example, if we choose $q = 0.3$, then $n > \frac{0.7^2}{0.2^2} = 12.25$. Thus, in the case of R^{13}, Inequality (5) amounts to $0.52 > r^2 > 0.49$. Choosing $r = \frac{\sqrt{2}}{2}$ gives the sphere with the desired "paradoxical" property.

The examples discussed suggest that precautions and sound arguments are needed when trying to extrapolate familiar properties of two- or three-dimensional spaces to spaces of higher dimensionality.
Physical and biological data as well as economic and demographic data have often high dimensionality. Intelligent data-mining algorithms work best in interpretation and decision making based on this data when we are able to simplify their tasks by reducing the high-dimensionality of the data. Dimensionality reduction refers to the extraction of the relevant information for a specific objective, while ignoring the unnecessary information and is a key concept in the pattern recognition, data mining, feature processing and machine learning. Dimensionality reduction requires tuning in terms of the expected number of dimensions, or the parameters of the learning algorithms.

Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique that aims to create a few new, uncorrelated linear combinations of the variables of an experiments that "explain" the major parts of the data variability.

Definition

Let $\boldsymbol{w} \in \mathbb{R}^n$ be a unit vector and let $\boldsymbol{u}$ be a vector in $\mathbb{R}^n$.
The *residual* of $\boldsymbol{u}$ relative to $\boldsymbol{w}$ is the number $r(\boldsymbol{u}) = \| \boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w} \|^2$ and it represents the error committed when the vector $\boldsymbol{u}$ is replaced by its projection on $\boldsymbol{w}$.

Theorem

If $r(\boldsymbol{u})$ is the residual of the vector $\boldsymbol{u}$ of a data matrix X relative to the vector $\boldsymbol{w}$ with $\| \boldsymbol{w} \| = 1$, then

$$r(\boldsymbol{u}) = \| \boldsymbol{u} \|^2 - (\boldsymbol{u}'\boldsymbol{w})^2.$$

Proof.

We have

$$\begin{aligned} r(\boldsymbol{u}) &= \| \boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w} \|^2 \\ &= (\boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w})'(\boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w}) \\ &= (\boldsymbol{u}' - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w}')(\boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w}) \\ &= \boldsymbol{u}'\boldsymbol{u} - (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w}'\boldsymbol{u} - \boldsymbol{u}'(\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w} + (\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w}'(\boldsymbol{u}'\boldsymbol{w})\boldsymbol{w} \\ &= \| \boldsymbol{u} \|^2 - 2(\boldsymbol{u}'\boldsymbol{w})^2 + (\boldsymbol{u}'\boldsymbol{w})^2 \\ &= \| \boldsymbol{u} \|^2 - (\boldsymbol{u}'\boldsymbol{w})^2 \end{aligned}$$

because $\boldsymbol{w}'\boldsymbol{w} = 1$. □

Let $X \in \mathbb{R}^{m \times n}$ be a data matrix whose rows are $\boldsymbol{u}_1', \ldots, \boldsymbol{u}_m'$.

Definition

The *mean square error* $\text{MSE}(X, \boldsymbol{w})$ of the projections of the experiments $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of the data matrix $X \in \mathbb{R}^{m \times n}$ on the unit vector $\boldsymbol{w} \in \mathbb{R}^n$ is the sum of the residuals, $\text{MSE}(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} r(\boldsymbol{u}_i)$.
The *average* of the projections of the experiment vectors on the unit vector $\boldsymbol{w}$ is the scalar:

$$u_{\boldsymbol{w}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{u}_i' \boldsymbol{w}.$$

The data is *centered* if and only if $u_{\boldsymbol{w}} = 0$.
The *variance* of the projections of the experiment vectors on $\boldsymbol{w}$ is

$$V(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{u}_i' \boldsymbol{w} - u_{\boldsymbol{w}})^2.$$

Note that

- if X is centered we have $u_{\boldsymbol{w}} = 0$;
- we choose $\boldsymbol{w}$ such that the variance $V(X, \boldsymbol{w})$ of the projections of the experiment vectors of $\boldsymbol{w}$ to be maximal.

Theorem

For the variance $V(X, \boldsymbol{w})$ *and the mean square error* $MSE(X, \boldsymbol{w})$ *we have the equalities:*

$$V(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{u}_i' \boldsymbol{w})^2 - u_{\boldsymbol{w}}^2,$$

and

$$MSE(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \| \boldsymbol{u}_i \|^2 - u_{\boldsymbol{w}}^2 - V(X, \boldsymbol{w}).$$

Proof

For the first equality we have

$$\begin{aligned} V(X, \boldsymbol{w}) &= \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}_i'\boldsymbol{w} - u_{\boldsymbol{w}})^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left((\boldsymbol{u}_i'\boldsymbol{w})^2 + u_{\boldsymbol{w}}^2 - 2(\boldsymbol{u}_i'\boldsymbol{w})u_{\boldsymbol{w}}\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}_i'\boldsymbol{w})^2 + u_{\boldsymbol{w}}^2 - 2u_{\boldsymbol{w}}^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}_i'\boldsymbol{w})^2 - u_{\boldsymbol{w}}^2. \end{aligned}$$

Proof cont'd

For the second equality we can write:

$$\begin{aligned}
\text{MSE}(X, \boldsymbol{w}) &= \frac{1}{m}\sum_{i=1}^{m} r(\boldsymbol{u}_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(\| \boldsymbol{u}_i \|^2 - (\boldsymbol{u}_i'\boldsymbol{w})^2 \right) \\
&= \frac{1}{m}\sum_{i=1}^{m} \| \boldsymbol{u}_i \|^2 - \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{u}_i'\boldsymbol{w})^2 \\
&= \frac{1}{m}\sum_{i=1}^{m} \| \boldsymbol{u}_i \|^2 - u_{\boldsymbol{w}}^2 - V(X, \boldsymbol{w}).
\end{aligned}$$

Corollary

If X is a centered data matrix then minimizing $MSE(X, \boldsymbol{w})$ amounts to maximizing the variance of the projections of the vectors of the experiments.

Proof.

Since X is centered we have $u_{\boldsymbol{w}} = 0$. Therefore, the equality involving $\text{MSE}(X, \boldsymbol{w})$ from Slide 41 becomes:

$$\text{MSE}(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \| \boldsymbol{u}_i \|^2 - V(X, \boldsymbol{w}).$$

The first term does not depend on $\boldsymbol{w}$. Therefore, to minimize the mean square error we need to maximize the variance of the projections of the vectors of the experiments. □

If X is a centered data matrix we have $u_{\boldsymbol{w}} = 0$ and the variance of the data matrix reduces to:

$$V(X, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{u}_i' \boldsymbol{w})^2.$$

This expression can be transformed as

$$V(X, \boldsymbol{w}) = \frac{1}{m} (X\boldsymbol{w})'(X\boldsymbol{w}) = \frac{1}{m} \boldsymbol{w}' X' X \boldsymbol{w} = \boldsymbol{w}' Z \boldsymbol{w},$$

where $Z = \frac{1}{m} X' X$.

We need to choose the unit vector $\boldsymbol{w}$ to maximize $V(X, \boldsymbol{w})$. In other words we need to maximize $V(X, \boldsymbol{w})$ subjected to the restriction $\boldsymbol{w}'\boldsymbol{w} - 1 = 0$. This can be resolved using a Lagrange multiplier λ to optimize the function

$$L(\boldsymbol{w}, \lambda) = \frac{1}{m}\boldsymbol{w}'X'X\boldsymbol{w} - \lambda(\boldsymbol{w}'\boldsymbol{w} - 1).$$

Since

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= -\boldsymbol{w}'\boldsymbol{w} + 1,\\ \frac{\partial L}{\partial \boldsymbol{w}} &= 2Z'\boldsymbol{w} - 2\lambda\boldsymbol{w},\end{aligned}$$

which implies $\boldsymbol{w}'\boldsymbol{w} = 1$ and $Z\boldsymbol{w} = \lambda\boldsymbol{w}$. The last equality amounts to

$$\frac{1}{m}X'X\boldsymbol{w} = \lambda\boldsymbol{w},$$

which means that $\boldsymbol{w}$ must be an eigenvector of the covariance matrix $cov(X)$. This is an $n \times n$ symmetric matrix, so its eigenvectors are mutually orthogonal and all its eigenvalues are non-negative. These eigenvectors are the *principal components* of the data.

Definition

Let

$$X = \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_m' \end{pmatrix} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{m \times n}$$

be a data sample matrix.

The *centered data matrix* is $\hat{X} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_n) = H_m X \in \mathbb{R}^{m \times n}$.

Definition

The *principal directions* of X are the eigenvectors of the covariance matrix

$$cov(X) = \frac{1}{m-1}\hat{X}'\hat{X} = \frac{1}{m-1}X'H_m'H_mX = \frac{1}{m-1}X'H_mX \in \mathbb{R}^{n\times n}.$$

If $R \in \mathbb{R}^{n\times n}$ is the orthogonal matrix that diagonalizes $cov(X)$, then the *principal directions* of X are the columns of R because $R'cov(X)R = D$, or equivalently, $cov(X)R = RD$.

Note that:

- the covariance matrix $cov(X)$ is a scalar multiple of the Gram matrix $\hat{X}'\hat{X}$ of the columns $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_n$ of the centered data matrix $\hat{X}$;
- If $R'cov(X)R = D = \text{diag}(d_1, d_2, \ldots, d_n)$ and $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n$ then the first eigenvector (which corresponds to d_1) is the *first principal direction* of $cov(X)$; in general, the k^{th} eigenvector $\boldsymbol{r}_k$ is called the k^{th} *principal direction* of X.

The sum of the elements of D's main diagonal equals the total variance $\text{tvar}(X)$. The principal directions "explain" the sources of the total variance: sample vectors grouped around $\boldsymbol{r}_1$ explain the largest portion of the variance; sample vectors grouped around $\boldsymbol{r}_2$ explain the second largest portion of the variance, etc.

Let $Q \in \mathbb{R}^{n \times \ell}$ be a matrix having orthogonal columns.
Starting from a sample matrix $X \in \mathbb{R}^{m \times n}$ we can construct a new sample matrix $W \in \mathbb{R}^{m \times \ell}$ having ℓ variables as $W = XQ$
Each experiment E_i is represented now by a row $\boldsymbol{w}'_i$ that is linked by $\boldsymbol{u}'_i$ by the equality $\boldsymbol{w}'_i = \boldsymbol{u}'_i Q$. This means that the component $(\boldsymbol{w}'_i)_k$ that corresponds to the new variable $\mathcal{W}_k$ is obtained as $(\boldsymbol{w}'_i)_k = \sum_{p=1}^{n} (u'_i)_p q_{pk}$, a linear combination of the values that correspond to the previous variables. ‘

Theorem

Let $W \in \mathbb{R}^{m \times n}$ be a centered sample matrix and let $R \in \mathbb{R}^{n \times n}$ be an orthogonal matrix such that $R' cov(W) R = D$, where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix $D = diag(d_1, \ldots, d_n)$ and $d_1 \geqslant \cdots \geqslant d_n$.
Let $Q \in \mathbb{R}^{n \times \ell}$ be a matrix having orthogonal columns and let $X = WQ \in \mathbb{R}^{m \times \ell}$. Then, $trace(cov(X))$ is maximized when Q consists of the first ℓ columns of R and is minimized when Q consists of the last ℓ columns of R.

Proof: This result follows from Ky Fan's Theorem applied to the symmetric covariance matrix of the transformed data set.

Let $\hat{X} = UDV'$ be the thin SVD of the centered data matrix $\hat{X} \in \mathbb{R}^{m \times n}$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ are matrices having orthogonal columns and

$$D = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{pmatrix},$$

where $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$ are the singular values of $\hat{X}$. For the covariance matrix $cov(X)$ we have

$$\begin{aligned} cov(X) &= \frac{1}{m-1}\hat{X}'\hat{X} = \frac{1}{m-1}VD'U'UDV' \\ &= \frac{1}{m-1}VD'DV' = \frac{1}{m-1}VD^2V', \end{aligned}$$

due to the orthogonality of the columns of U.

It is clear that $\sigma_1^2, \ldots, \sigma_r^2$ coincide with the eigenvalues of $\hat{X}'\hat{X}$.
Starting with the thin SVD of the centered data matrix: $\hat{X} = UDV'$ we have the following definitions:

- The columns of V are the eigenvectors of $cov(X)$. The matrix V is known as the *matrix of loadings*.
- The matrix $S = UD \in \mathbb{R}^{m \times r}$ is known as the *matrix of scores*.
- Observe that $\hat{X} = SV'$, where S is the scores matrix and V is the loadings matrix. Since the columns of V are orthogonal we also have $S = \hat{X}V$.

The SVD of $\hat{X}$ can be written as:

$$\hat{X} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i'.$$

This implies $\hat{X}'\hat{X}\boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v}_i$. Since $\boldsymbol{u}_i'\hat{X} = \sigma_i \boldsymbol{v}_i'$, it follows that $\boldsymbol{v}_i'$ is a weighted sum of the rows of the matrix $\hat{X}$. Similarly, $\boldsymbol{u}_i$ are weighted sums of the columns of $\hat{X}$.

If

$$\hat{X} = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r)(\sigma_1 \boldsymbol{v}_1' \ \ldots \ \sigma_r \boldsymbol{v}_r)',$$

then

$$\begin{aligned} I_r &= (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r)'(\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r) \\ (n-1)\hat{X}'\hat{X} &= (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r)(\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r)' \end{aligned}$$

The data set that we analyze was published by FAO (Food and Agricultural Organization of UN) and shows the protein and fat consumption for 37 European countries in grams per person per day. The countries are identified by a two-letter code.

code	prot	fat	code	prot	fat
AL	97	87	IT	113	158
AT	107	155	LV	87	116
BY	88	97	LT	112	105
BE	97	164	LU	124	164
BA	86	67	MK	72	102
BG	79	101	MT	116	110
HR	74	97	MD	73	59
CY	99	133	NL	103	135
CZ	95	121	NO	104	144
DK	108	135	PL	100	113
EE	88	96	PT	114	137
FI	105	127	RO	110	107
FR	117	164	RU	92	87
GE	77	58	YU	75	116
DE	99	142	SK	72	108
GR	117	146	SI	102	131
HU	90	145	ES	109	152
IS	128	143	CH	91	152
IE	115	135			

- The sample matrix $X \in \mathbb{R}^{37\times 2}$ is obtained from the second and third columns of this table that correspond to the variables *prot* and *fat*.
- The vector of the sample variances of the two columns is $\boldsymbol{s} = (15.5213\ 28.9541)$. Since the magnitudes of the sample variances are substantial and quite distinct we normalize the data by dividing the columns of X be their respective sample variances.
- The normalization is done by using the function `zscore`; namely, `zscore(X)` returns a centered and scaled version of X having the same format as X such that the columns of the result have sample mean 0 and sample variance 1.

The fao.csv file

```
code, prot, fat
AL, 97, 87
AT, 107, 155
BY,88,97
BE,97,164
BA,86,67
BG,79,101
HR,74,97
CY,99,133
CZ,95,121
DK,108,135
EE,88,96
FI,105,127
FR,117,164
GE,77,58
DE,99,142
GR,117,146
HU,90,145
IS,128,143
```

The fao.csv file cont.

```
code, prot, fat
IE,115,135
IT,113,158
LV,87,116
LT,112,105
LU,124,164
MK,72,102
MT,116,110
MD,73,59
NL,103,135
NO,104,144
PL,100,113
PT,114,137
RO,110,107
RU,92,87
YU,75,116
SK,72,108
SI,102,131
ES,109,152
```

```
>> opts=detectImportOptions('fao.csv');
>> preview('fao.csv',opts)

ans =
     code      prot    fat
    {'AL'}      97      87
    {'AT'}     107     155
    {'BY'}      88      97
    {'BE'}      97     164
    {'BA'}      86      67
    {'BG'}      79     101
    {'HR'}      74      97
    {'CY'}      99     133


>> opts.SelectedVariableNames=[2:3]
```

```
>> M=readmatrix('fao.csv',opts)

M =

    97    87
   107   155
    88    97
    97   164
    86    67
    79   101
    74    97
    99   133
    95   121
   108   135
    88    96
   105   127
   117   164
    77    58
    99   142
   117   146
    90   145
```

```
>> Z=zscore(M)
Z =

   -0.0801   -1.2041
    0.5642    1.1444
   -0.6599   -0.8588
   -0.0801    1.4552
   -0.7888   -1.8949
   -1.2398   -0.7206
   -1.5619   -0.8588
    0.0488    0.3846
   -0.2090   -0.0299
    0.6286    0.4537
   -0.6599   -0.8933
    0.4353    0.1774
    1.2085    1.4552
   -1.3686   -2.2057
    0.0488    0.6954
    1.2085    0.8336
   -0.5311    0.7990
    1.9172    0.7300
```

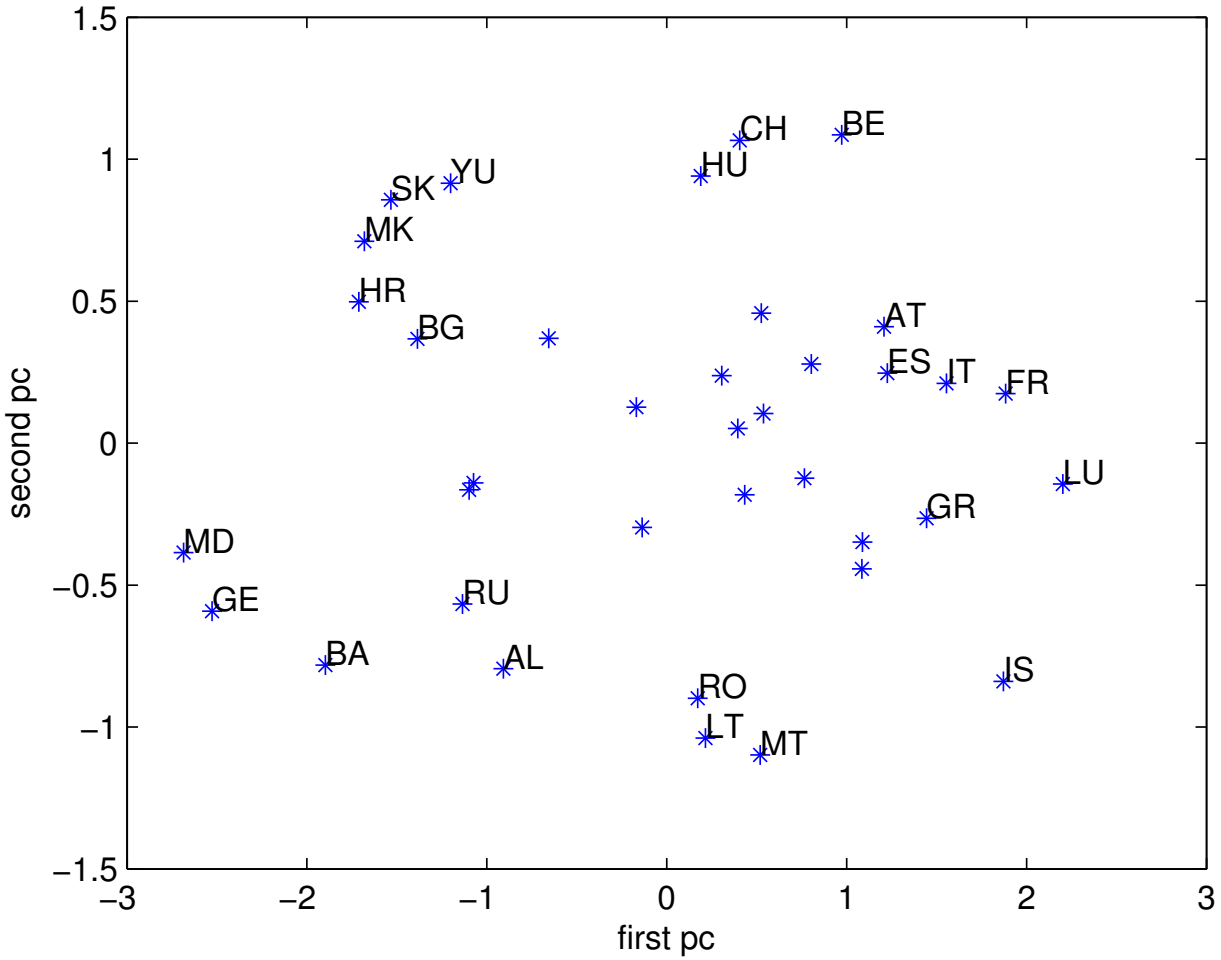
```
>> pca(zscore(M))

ans =

    0.7071    0.7071
    0.7071   -0.7071

>>
```

The first two principal components of the FAO dataset:



- Principal component analysis in MATLAB is done using the function `pca` of the statistics toolbox. There are several signatures of this function which we review next.
- The statement `coeff = pca(A)` performs principal components analysis (PCA) on the matrix $A \in \mathbb{R}^{m \times n}$, and returns the principal component coefficients, also known as *loadings*.
- Rows of A correspond to observations, and columns to variables. The columns of the matrix `coeff` (a $n \times n$ matrix) contain coefficients for one principal component and these columns are in order of decreasing component variance.

The function `pca` computes the principal components of a sample matrix X. The are several incarnations of the function `pca` described below.

- `[coeff,score] = pca(X)` returns the matrix `score`, the principal component scores, that is, the representation of X in the principal component space. The rows of `score` correspond to observations, columns to components.
- `[coeff,score,latent] = pca(X)` returns the vector `latent` which contains the eigenvalues of the covariance matrix of X.

- The matrix `score` contains the data formed by transforming the original data into the space of the principal components. The values of the vector `latent` are the variance of the columns of `score`.
- The function `pca` centers X by subtracting off column variance means, but does not rescale the columns of X. To perform principal components analysis with standardized variables we need to use `pca(zscore(X))`.

The loading matrix or the coefficient matrix is given by

$$\begin{pmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}$$

Both coefficients in the first column (which represents the first principal component) are equal and positive, which means that the first principal component is a weighted average of the two variables. The second principal component corresponds to a weighted difference of the original variables. The coordinates of the data in the new coordinate system is defined by the matrix scores.

Example

The data set that we are about to analyze originates in a study of the health condition of Boston neighborhoods produced by the Health Department of the City of Boston. The data includes incidence of various diseases and health events that occur in the 16 neighborhoods of the city identified as

Neighborhood	Code	Neighborhood	Code
Allston/Brighton	AB	North Dorchester	ND
Back Bay	BB	North End	NE
Charlestown	CH	Roslindale	RO
East Boston	EB	Roxbury	RX
Fenway	FW	South Boston	SB
Hyde Park	HP	South End	SE
Jamaica Plain	JP	South Dorchester	SD
Mattapan	MT	West Roxbury	WR

This is entered in MATLAB as

```
neighborhoods = ['AB';'BB';'CH';'EB';'FW';'HP';'JP';...
'MT';'ND';'NE';'RS';'RX';'SB';'SD';'SE';'WR']
```

The diseases and the health conditions are listed as the vector `categories`:

Category	Code	Category	Code
Hepatitis B	HepB	Tuberculosis	TBCD
Hepatitis C	HepC	Live Births	B154
HIV/AIDS	HIVA	Low weight at birth	LBWE
Chlamydia	CHLA	Infant Mortality	INFM
Syphilis	SYPH	Children with Elevated Lead	CELL
Gonorrhea	GONO	Subst. Abuse Treat. Admissions	SATA

This is entered in MATLAB as

```
categories = ['HepB';'HepC';'HIVA';'CHLA';'SYPH';'GONO';...
              'TBCD';'B154';'LBWE';'INFM';'CELL';'SATA']
```

The data itself is contained in the matrix `diseaseinc` in $\mathbb{R}^{16\times 12}$ and it has the form

38	57	13	168	5	22	20	607	40	7	13	624
17	24	16	179	11	52	8	306	25	0	0	497
10	13	0	46	0	8	0	284	25	0	0	489
12	46	11	150	10	16	17	718	43	10	43	1009
18	19	8	163	9	44	7	125	10	0	0	272
11	18	8	179	9	25	9	487	46	12	19	2781
6	32	18	213	10	46	6	420	35	5	10	1071
15	24	11	264	14	56	8	285	26	7	25	390
42	76	22	611	15	135	29	1350	168	28	88	1492
0	0	0	0	0	0	0	89	5	0	0	130
13	22	0	115	6	31	0	488	39	6	28	330
21	50	17	477	8	72	8	829	87	27	30	2075
9	52	0	85	7	25	5	403	25	0	18	1335
68	78	23	760	24	176	24	656	67	9	63	1464
51	35	35	124	31	61	11	439	34	0	0	6064
9	10	0	17	0	0	0	419	34	0	11	179

The array containing the sample variances of columns is computed by applying the function `std`:

```
stdinc = std(diseaseinc)
```

Next, by using the function `repmat` as in

```
si = diseaseinc./repmat(stdinc,16,1)
```

we create a 16×12 matrix consisting of 16 copies of `stdinc` and compute the normalized matrix `si` that is subjected to PCA in

```
[loadings,scores,variances]=pca(si)
```

This is one of several formats of the function `pca`. This function is applied to a data matrix and it centers the matrix by subtracting off column means.

This is one of several formats of the function `pca`. This function is applied to a data matrix and it centers the matrix by subtracting off column means. In the format that we use here, the function returns the matrices `loadings`, `scores`, and `variances` that contain the following data;
The columns of the matrix `loadings` contains the principal components. The entries of this matrix are known as *loadings*. In our case `loadings` is a 12×12 matrix, where each column contains represents one principal component. The columns are in order of decreasing component variance. We reproduce below the first three columns of this matrix

0.2914	0.2732	-0.2641
0.3207	-0.0568	-0.1593
0.2666	0.3848	0.1427
0.3267	-0.0800	-0.2671
0.2426	0.4650	-0.0270
0.3209	0.0668	-0.3463
0.3215	-0.0161	-0.1444
0.3055	-0.2594	0.3184
0.3061	-0.2659	0.2735
0.2655	-0.3215	0.3832
0.3026	-0.2903	-0.0815
0.1423	0.4702	0.5848

- The matrix `scores` in $\mathbb{R}^{16\times 12}$ contains the principal component scores, that is, the representation of `si` in the principal component space. Rows of `score` correspond to neighborhoods and columns to components.
- The matrix `variances` contains the principal component variances, that is, the eigenvalues of the covariance matrix of `si`.

The first two columns of the matrix `scores` contain the projections of data on the first two principal components. This is done by running

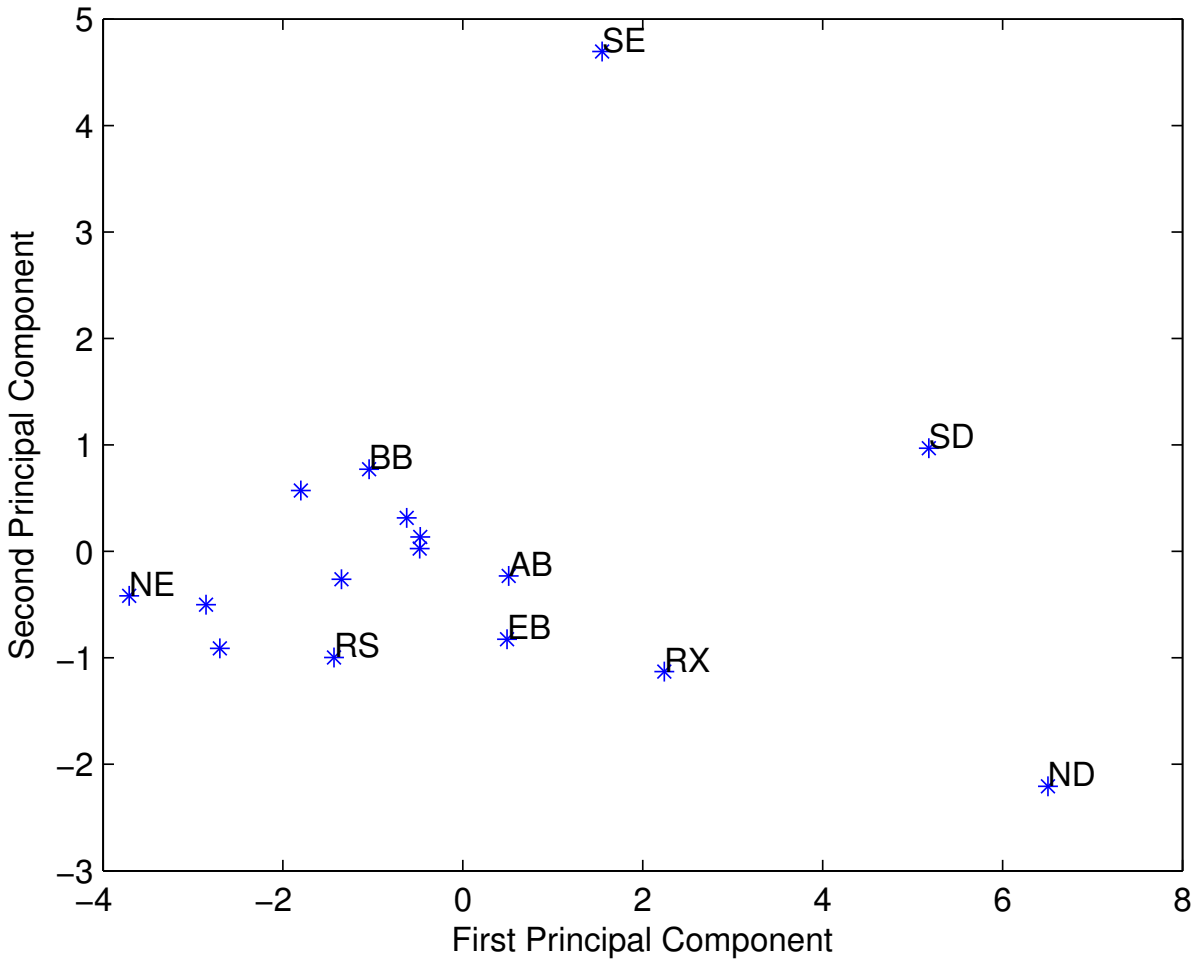
```
plot(scores(:,1),scores(:,2),'*')
```

After the plot is created labels can be added to the axes using

```
xlabel('First Principal Component')
ylabel('Second Principal Component')
```

The resulting plot is shown next.

Projections on the first two principal components:



The neighborhood codes are applied to this plot by running `gname(neighborhoods)`. An inspection of the figure shows that the health issues are different for neighborhoods like South End (SE), South Dorchester (SD) and North Dorchester (ND).

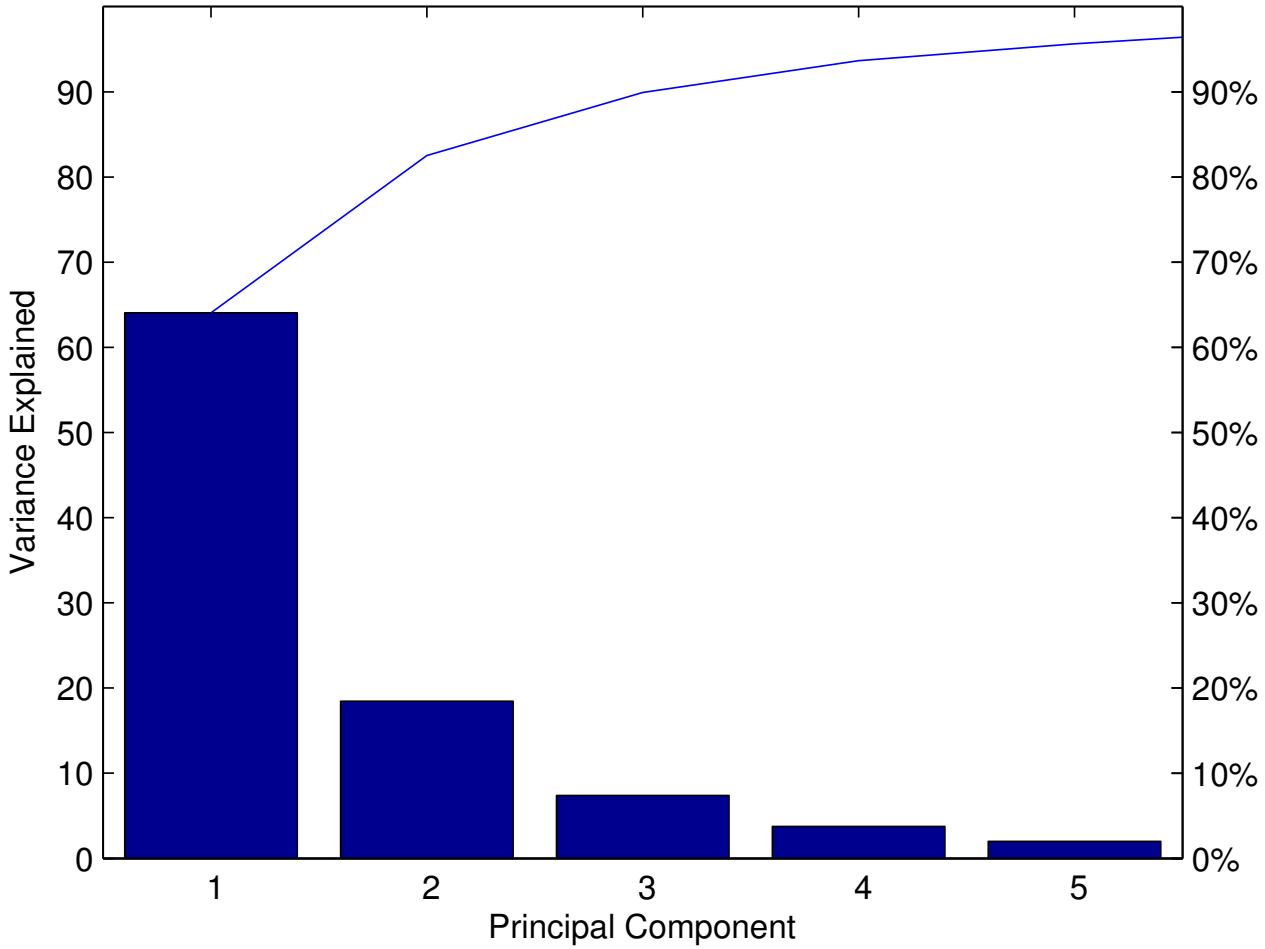
The matrix `variances` allows us to examine the percentage of the total variability explained by each principal component. Initially, we compute the matrix `percent_explained` as

```
percent_explained=100*variances/sum(variances)
```

and using the function `pareto` we write

```
pareto(percent_explained)
xlabel('Principal Component')
ylabel('Variance Explained')
```

This code produces the histogram shown below:

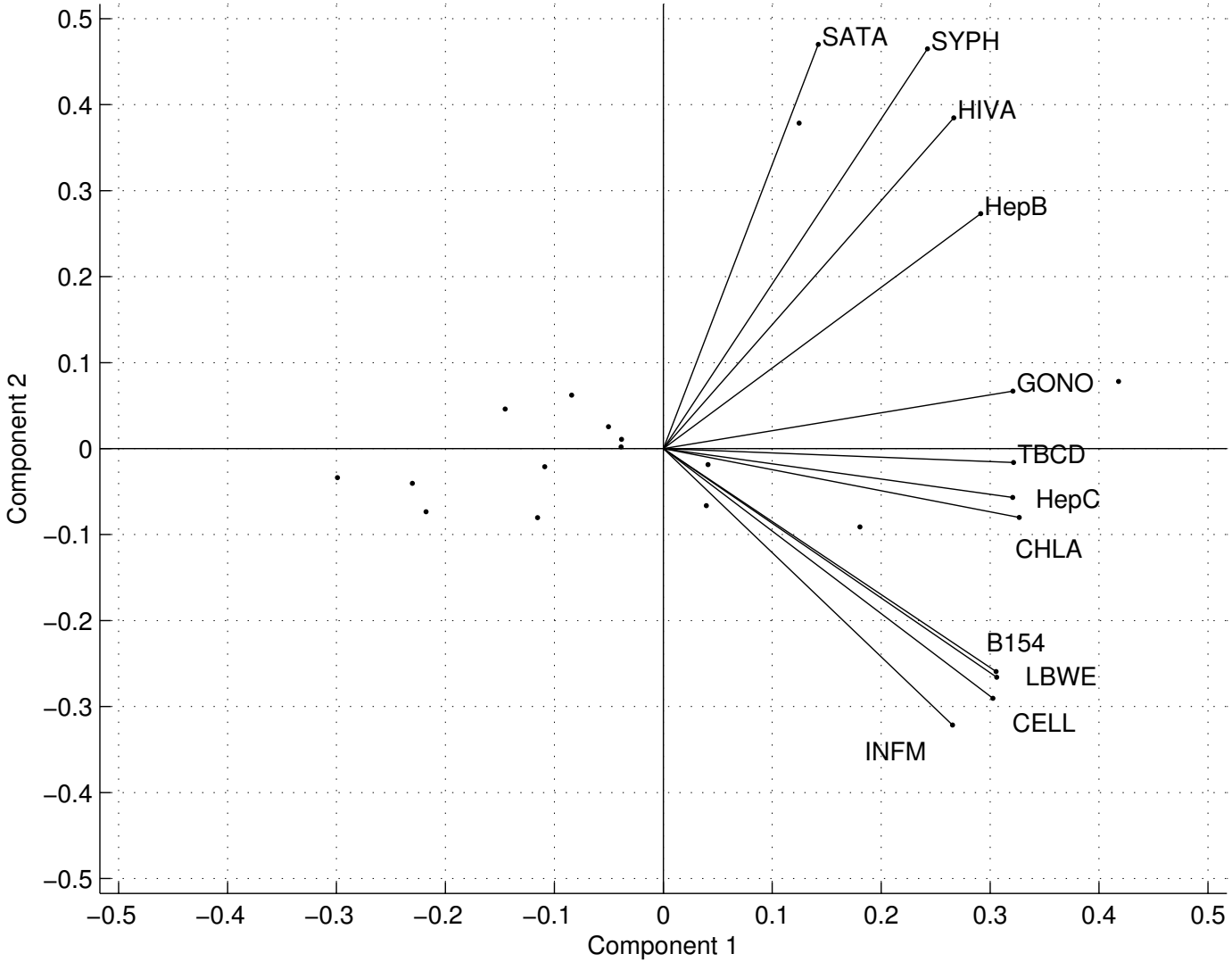


Percentage of variability explained by principal components

To visualize the results one can use the `biplot` function as in

```
biplot(loadings(:,1:2),'scores',scores(:,1:2),'varlabels',categories
```

resulting in



Representation of the 12 variables

Each of the 12 variable is represented by a vector in this figure.

- Since the first principal component has positive coefficients, all vectors are located in the right half-plane.
  On the other hand the signs of the coefficients of the second principal component are varying.
- These components distinguish between neighborhoods where there is a high incidence of substance abuse treatment admissions, (SATA), syphilis (SYPH), HIV/Aids (HIVA), Hepatitis B (HepB), and Gonorrhea (GONO) and low incidence of the others and neighborhoods where the opposite situation occurs.
- The conclusions of a PCA analysis of data are mainly qualitative. The numerical precision (4 decimal digits) is not especially relevant for the PCA.

Next, we present a geometric point of view of principal component analysis.
Let $\boldsymbol{t} \in \mathbb{R}^n$ be a unit vector. The projection of a vector $\boldsymbol{w} \in \mathbb{R}^n$ on the subspace $\langle \boldsymbol{t} \rangle$ generated by $\boldsymbol{t}$ is given by

$$\text{proj}_{\langle \boldsymbol{t} \rangle}(\boldsymbol{w}) = \boldsymbol{t}\boldsymbol{t}'\boldsymbol{w}.$$

To simplify the notation we shall write $\text{proj}_{\boldsymbol{t}}$ instead of $\text{proj}_{\langle \boldsymbol{t} \rangle}$. Let $\hat{W} \in \mathbb{R}^{m \times n}$ be a centered sample matrix that corresponds to a sequence of experiments $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, that is

$$\hat{W} = \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_m' \end{pmatrix}.$$

We seek to evaluate the inertia $I_{\mathbf{0}}(\text{proj}_{\boldsymbol{t}}(\hat{W}'))$ on the subspace generated by the unit vector $\boldsymbol{t} \in \mathbb{R}^n$. Since $\hat{W}' = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_m)$, by the definition of the inertia, we have:

$$
\begin{aligned}
I_{\mathbf{0}}(\text{proj}_{\boldsymbol{t}}(\hat{W}')) &= \sum_{j=1}^{m} \| \ \boldsymbol{t}\boldsymbol{t}'\boldsymbol{u}_j \ \|_2^2 \\
&= \sum_{j=1}^{m} \boldsymbol{u}_j'\boldsymbol{t}\boldsymbol{t}'\boldsymbol{t}\boldsymbol{t}'\boldsymbol{u}_j \\
&= \sum_{j=1}^{m} \boldsymbol{u}_j'\boldsymbol{t}\boldsymbol{t}'\boldsymbol{u}_j \\
&\quad \text{(because } \boldsymbol{t}'\boldsymbol{t} = 1\text{)} \\
&= \sum_{j=1}^{m} \boldsymbol{t}'\boldsymbol{u}_j\boldsymbol{u}_j'\boldsymbol{t} \\
&\quad \text{(because both } \boldsymbol{u}_j'\boldsymbol{t} \text{ and } \boldsymbol{t}'\boldsymbol{u}_j \text{ are scalars)} \\
&= \boldsymbol{t}'X'X\boldsymbol{t}.
\end{aligned}
$$

The necessary condition for the existence of extreme values of this inertia as a function of $\boldsymbol{t}$ is

$$\begin{aligned} \text{grad } \left(I_{\mathbf{0}}(\text{proj}_{\boldsymbol{t}}(\hat{W}')) + \lambda(1 - \boldsymbol{t}'\boldsymbol{t}) \right) &= \text{grad } \left(\boldsymbol{t}'\hat{W}'\hat{W}\boldsymbol{t} + \lambda(1 - \boldsymbol{t}'\boldsymbol{t}) \right) \\ &= 2\hat{W}'\hat{W}\boldsymbol{u} - 2\lambda\boldsymbol{t} = \mathbf{0}, \end{aligned}$$

where λ is a Lagrange multiplier. This implies $\hat{W}'\hat{W}\boldsymbol{t} = \lambda\boldsymbol{t}$. In other words, to achieve extreme values of the inertia $I_{\mathbf{0}}(\text{proj}_{\boldsymbol{t}}(\hat{W}'))$, $\boldsymbol{t}$ must be chosen as a eigenvector of the covariance matrix of $\hat{W}$, that is, as a principal direction of $\hat{W}$.

Suppose that the eigenvalues of $\hat{W}'\hat{W}$ are the numbers $\lambda_1 \geqslant \cdots \geqslant \lambda_n$. The first principal direction $\boldsymbol{t}_1$ of W which corresponds to the largest eigenvalue of $\hat{W}'\hat{W}$ is

$$\begin{aligned} \boldsymbol{t}_1 &= \arg\max_{\boldsymbol{t}} \left\{ \boldsymbol{t}'\hat{W}\hat{W}'\boldsymbol{t} \mid \boldsymbol{t} \in \mathbb{R}^n, \| \boldsymbol{t} \|_2 = 1 \right\} \\ &= \arg\max_{\boldsymbol{t}} \left\{ \| \hat{W}'\boldsymbol{t} \|_2^2 \mid \| \boldsymbol{t} \|_2 = 1 \right\}. \end{aligned}$$

Suppose that we computed the principal directions $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k$ of $\hat{W}$. Then, $\boldsymbol{t}_{k+1} \in \mathbb{R}^n$ is a unit vector $\boldsymbol{t}$ that maximizes

$$\boldsymbol{t}'\hat{W}\hat{W}'\boldsymbol{t} = \| \hat{W}'\boldsymbol{t} \|_2^2$$

and belongs to the subspace orthogonal to the subspace generated by the first k principal directions of $\hat{W}$, that is,

$$\boldsymbol{t}_{k+1} = \arg\max_{\boldsymbol{t}} \left\{ \| \hat{W}'\boldsymbol{t} \|_2^2 \mid \boldsymbol{t} \in \mathbb{R}^n, \| \boldsymbol{t} \|_2 = 1, \boldsymbol{t} \in \langle \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k \rangle^{\perp} \right\}.$$

Note that for every vector $\boldsymbol{z} \in \mathbb{R}^n$ we have

$$(I - \sum_{j=1}^{k} \boldsymbol{t}_j \boldsymbol{t}_j')\boldsymbol{z} = \boldsymbol{z} - \text{proj}_{\langle \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k \rangle} \boldsymbol{z} \in \langle \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k \rangle^{\perp}.$$

Therefore, $\boldsymbol{x} \in \langle \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k \rangle^{\perp}$, is equivalent to $\boldsymbol{x} = (I - \sum_{j=1}^{k} \boldsymbol{t}_j \boldsymbol{t}_j')\boldsymbol{x}$. Thus, we can write

$$\boldsymbol{t}_{k+1} = \arg\max_{\boldsymbol{t}} \left\{ \left\| \hat{W}(I - \sum_{j=1}^{k} \boldsymbol{t}_j \boldsymbol{t}_j')\boldsymbol{t} \right\|_2 \mid \| \boldsymbol{t} \|_2 = 1 \right\},$$

for $0 \leq k \leq n-1$. This technique allows finding the principal directions of $\hat{W}$ by solving a sequence of optimization problems involving the matrix $\hat{W}$.