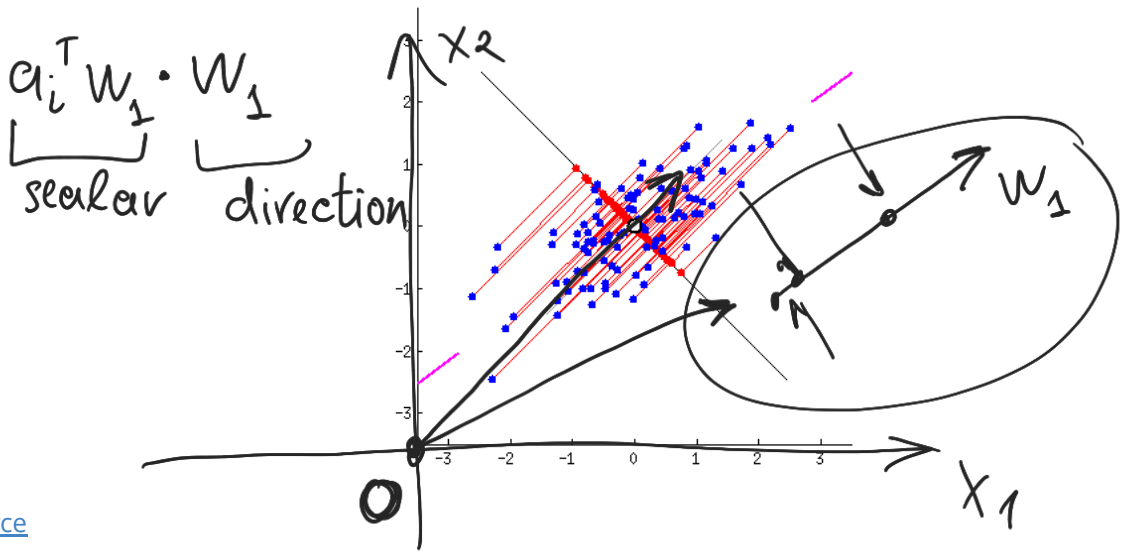
Principal component analysis

Intuition

Imagine, that you have a dataset of points. Your goal is to choose orthogonal axes, that describe your data the most informative way. To be precise, we choose first axis in such a way, that maximize the variance (expressiveness) of the projected data. All the following axes have to be orthogonal to the previously chosen ones, while satisfy largest possible variance of the projections.

Let's take a look at the simple 2d data. We have a set of blue points on the plane. We can easily see that the projections on the first axis (red dots) have maximum variance at the final position of the animation. The second (and the last) axis should be orthogonal to the previous one.



source

This idea could be used in a variety ways. For example, it might happen, that projection of complex data on the principal plane (only 2 components) bring you enough intuition for clustering. The picture below plots projection of the labeled dataset onto the first to principal components (PC's), we can clearly see, that only two vectors (these PC's) would be enogh to differ Finnish people from Italian in particular dataset (celiac disease (Dubois et al. 2010))

$$f(w) = w^T A^T A w \longrightarrow \max$$

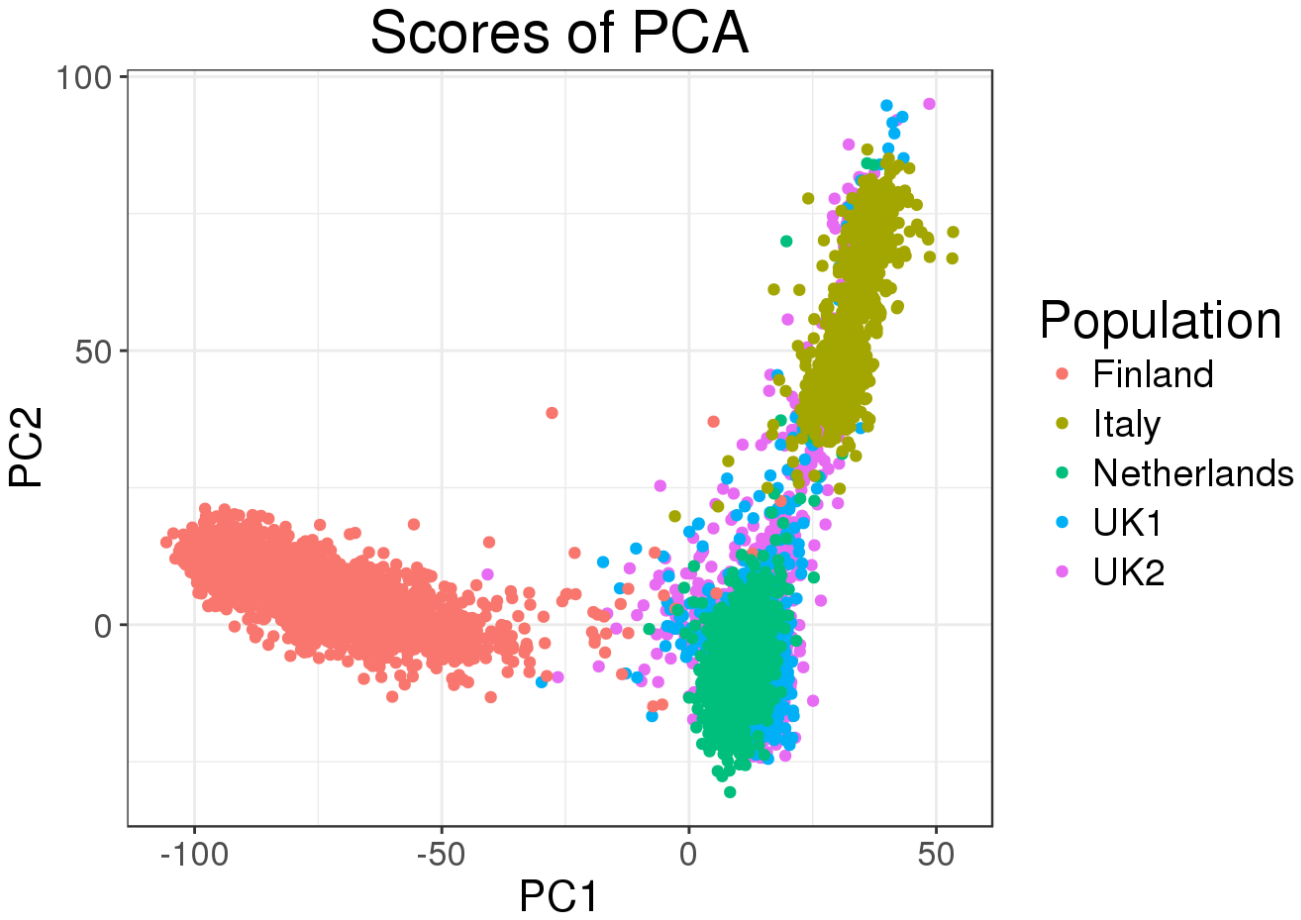
$$\|w\| = 1$$

$$L(w, \lambda) = w^T A^* A w + \lambda w^T w$$

$$\frac{dL}{dw} = 2A^* A w + 2\lambda w = 0$$

$$A^* A w = -\lambda w$$

eigenvector of $A^* A$



source

Problem

The first component should be defined in order to maximize variance. Suppose, we've already normalized the data, i.e. $\sum\limits_i a_i = 0$, then sample variance will become the sum of all squared projections of data points to our vector $\mathbf{w}_{(1)}$, which implies the following optimization problem:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{a}_{(i)} \cdot \mathbf{w} \right)^2 \right\}$$

or

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{ \|\mathbf{A}\mathbf{w}\|^2 \} = \arg\max_{\|\mathbf{w}\|=1} \{ \mathbf{w}^\top \mathbf{A}^\top \mathbf{A}\mathbf{w} \}$$

since we are looking for the unit vector, we can reformulate the problem:

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^\top \mathbf{A}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right\}$$

It is known, that for positive semidefinite matrix $A^\top A$ such vector is nothing else, but eigenvector of $A^\top A$, which corresponds to the largest eigenvalue. The following components will give you the same results (eigenvectors).

So, we can conclude, that the following mapping:

$$\underset{n \times k}{\Pi} = \underset{n \times d}{A} \cdot \underset{d \times k}{W}$$

describes the projection of data onto the k principal components, where W contains first (by the size of eigenvalues) k eigenvectors of $A^\top A$.

Now we'll briefly derive how SVD decomposition could lead us to the PCA.

Firstly, we write down SVD decomposition of our matrix:

$$A = U \Sigma W^\top$$

and to its transpose:

$$\begin{aligned} A^\top &= (U \Sigma W^\top)^\top \\ &= (W^\top)^\top \Sigma^\top U^\top \\ &= W \Sigma^\top U^\top \\ &= W \Sigma U^\top \end{aligned}$$

Then, consider matrix $$A A^\top$$:

$$\begin{aligned} A^\top A &= (W \Sigma U^\top)(U \Sigma V^\top) \\ &= W \Sigma I \Sigma W^\top \\ &= W \Sigma \Sigma W^\top \\ &= W \Sigma^2 W^\top \end{aligned}$$

Which corresponds to the eigendecomposition of matrix $A^\top A$, where W stands for the matrix of eigenvectors of $A^\top A$, while Σ^2 contains eigenvalues of $A^\top A$.

At the end:

$$\begin{aligned} \Pi &= A \cdot W = \\ &= U \Sigma W^\top W = U \Sigma \end{aligned}$$

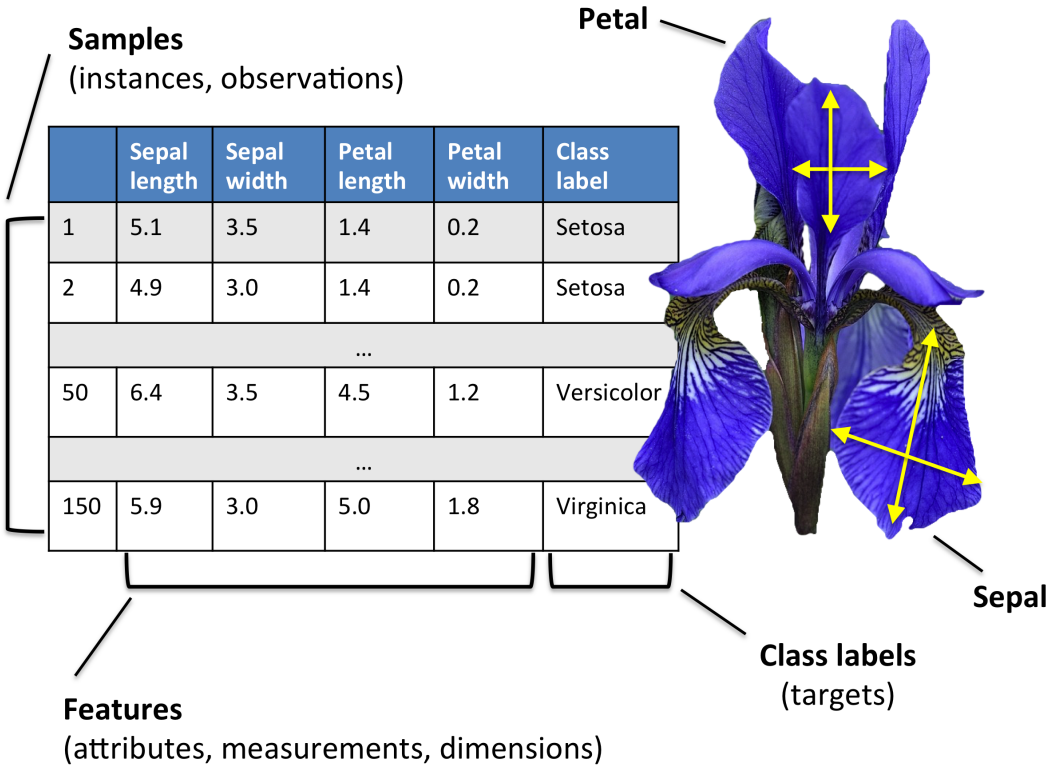
The latter formula provide us with easy way to compute PCA via SVD with any number of principal components:

$$\Pi_r = U_r \Sigma_r$$

Examples

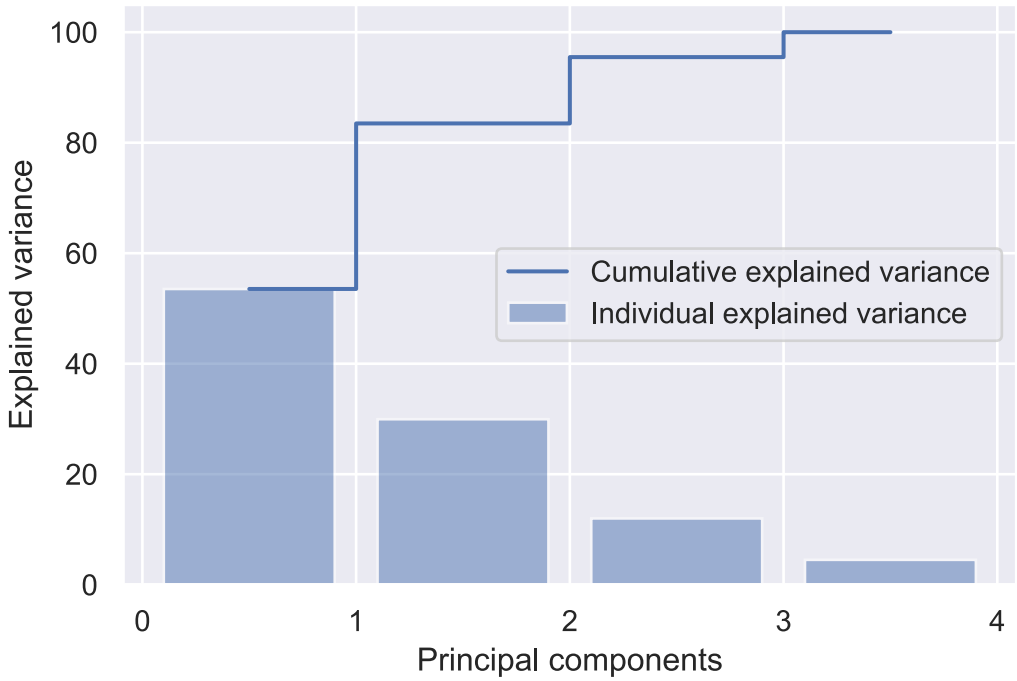
🏵 Iris dataset

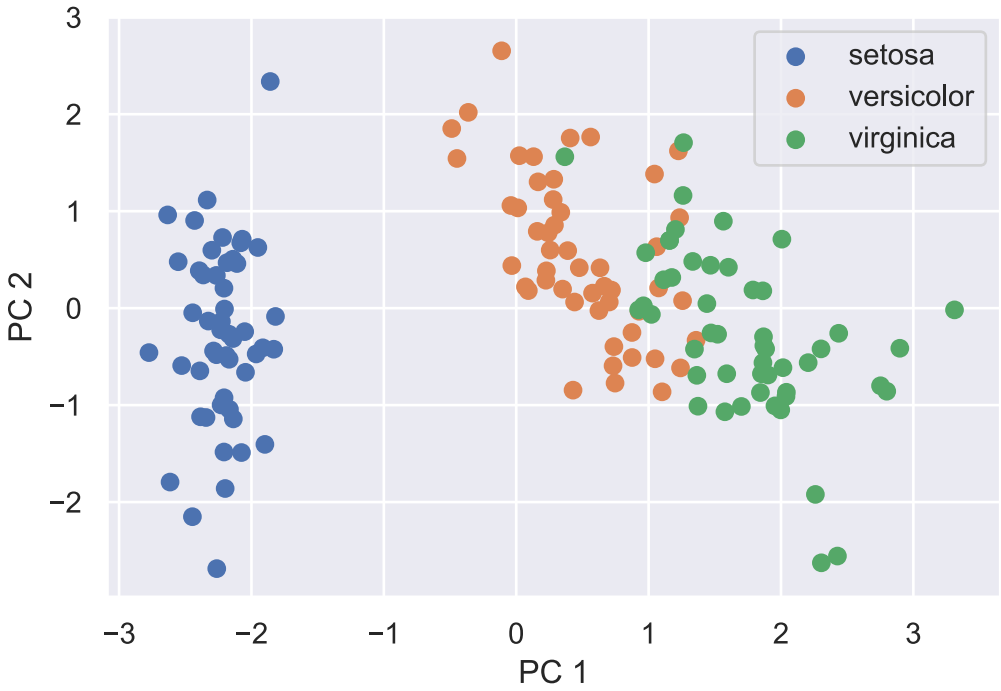
Consider the classical Iris dataset



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

source We have the dataset matrix $A \in \mathbb{R}^{150\times 4}$





Code

Open in Colab

Related materials

- Wikipedia
- Blog post
- Blog post

Matrix calculus

Useful definitions and notations

We will treat all vectors as column vectors by default.

Matrix and vector multiplication

Let A be $m \times n$, and B be $n \times p$, and let the product AB be

$$C = AB$$

then C is a $m \times p$ matrix, with element (i, j) given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Let A be $m \times n$, and x be $n \times 1$, then the typical element of the product

$$z = Ax$$

is given by

$$z_i = \sum_{k=1}^{n} a_{ik} x_k$$

Finally, just to remind:

- $C = AB \quad C^\top = B^\top A^\top$
- $AB \neq BA$
- $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$
- $e^{A+B} \neq e^A e^B$

Gradient

Gradient Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then vector, which contains all first order partial derivatives:

$$\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Hessian

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then matrix, containing all the second order partial derivatives:

$$f''(x) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial f}{\partial x_n \partial x_n} \end{pmatrix}$$

But actually, Hessian could be a tensor in such a way: $(f(x) : \mathbb{R}^n \to \mathbb{R}^m)$ is just 3d tensor, every slice is just hessian of corresponding scalar function $(H(f_1(x)), H(f_2(x)), \ldots, H(f_m(x)))$

Jacobian

The extension of the gradient of multidimensional $f(x) : \mathbb{R}^n \to \mathbb{R}^m$:

$$f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Summary

$$f(x) : X \to Y; \qquad \frac{\partial f(x)}{\partial x} \in G$$

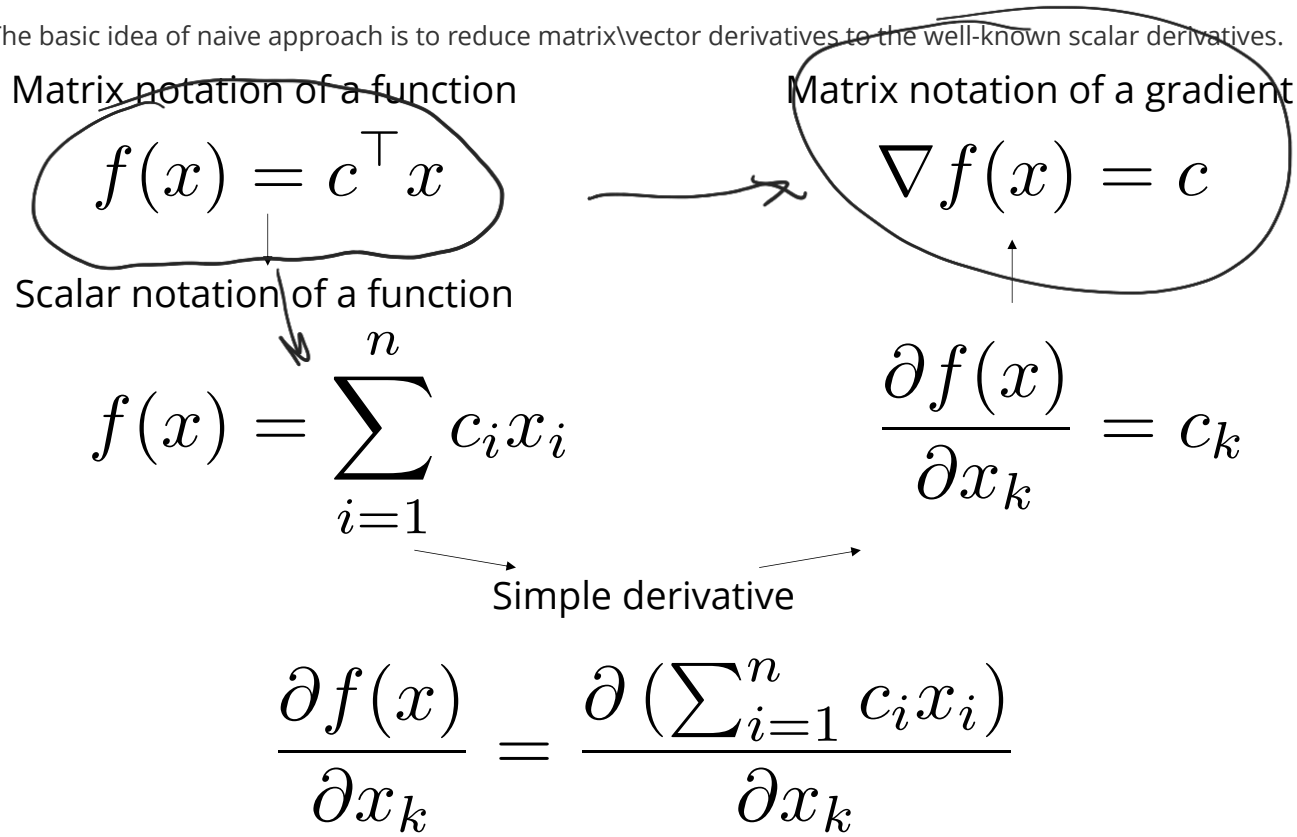
X	Y	G	Name
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$f'(x)$ (derivative)
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	$\frac{\partial f}{\partial x_i}$ (gradient)
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{n \times m}$	$\frac{\partial f_i}{\partial x_j}$ (jacobian)
$\mathbb{R}^{m \times n}$	$\mathbb{R}$	$\mathbb{R}^{m \times n}$	$\frac{\partial f}{\partial x_{ij}}$

named gradient of $f(x)$. This vector indicates the direction of steepest ascent. Thus, vector $-\nabla f(x)$ means the direction of the steepest descent of the function in the point. Moreover, the gradient vector is always orthogonal to the contour line in the point.

General concept

Naive approach

The basic idea of naive approach is to reduce matrix\vector derivatives to the well-known scalar derivatives.



One of the most important practical trick here is to separate indicies of sum (i) and partial derivatives (k). Ignoring this simple rule tends to produce mistakes.

Guru approach

The guru approach implies formulating a set of simple rules, which allows you to calculate derivatives just like in a scalar case. It might be convinient to use the differential notation here.

Differentials

After obtaining the differential notaion of df we can retrieve the gradient using following formula:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Than, if we have differential of the above form and we need to calculate the second derivative of the matrix\vector function, we treat "old" dx as the constant dx_1, than calculate $d(df)$

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2 \rangle = \langle H_f(x) dx_1, dx_2 \rangle$$

Properties

Let A and B be the constant matrices, while X and Y are the variables (or matrix functions).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^\top) = (dX)^\top$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$

- $d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-\top}, dX \rangle$
- $d\mathrm{tr}\ X = \langle I, dX \rangle$
- $df(g(x)) = \frac{df}{dg} \cdot dg(x)$

References

- Good introduction
- The Matrix Cookbook
- MSU seminars (Rus.)
- Online tool for analytic expression of a derivative.

Example 1

Find $\nabla f(x)$, if $f(x) = \frac{1}{2} x^T A x + b^T x + c$.

1) Naive approach:

$$f(x) = \frac{1}{2} \sum_i x_i (Ax)_i + \sum_i b_i x_i + c$$

$$\frac{\partial f}{\partial x_k} = \frac{\partial \frac{1}{2} \sum_i x_i (Ax)_i}{\partial x_k} + \frac{\partial \left(\sum_i b_i x_i\right)}{\partial x_k} + \frac{\partial c}{\partial x_k} = 0 =$$

$$= \frac{1}{2} \sum_{ij} \frac{\partial (x_i a_{ij} x_j)}{\partial x_k} + \sum_i b_i \frac{\partial x_i}{\partial x_k} =$$

$$= \frac{1}{2} \sum_t \left(x_t a_{tk} + x_t a_{kt}\right) + b_k$$

$$\nabla f = \frac{1}{2} \left(A + A^T\right) x + b$$

$$\frac{\partial (x_i a_{ik} x_k)}{\partial x_k} = x_i a_{ik}$$

$$\frac{\partial (x_k a_{kj} x_j)}{\partial x_k} = a_{kj} x_j$$

$$\frac{\partial a_{kk} x_k^2}{\partial x_k} = 2 x_k a_{kk}$$

Example 2

2) Guru's approach

$$f(x) = \frac{1}{2} \langle x, Ax \rangle + \langle b, x \rangle + c$$

$$df = \frac{1}{2} d(\langle x, Ax \rangle) + d(\langle b, x \rangle) + dc =$$

$$= \frac{1}{2} \left(\langle dx, Ax \rangle + \langle x, A dx \rangle\right) + \langle b, dx \rangle + 0 =$$

$$= \frac{1}{2} \left(\langle Ax, dx \rangle + \langle A^T x, dx \rangle\right) + \langle b, dx \rangle =$$

$$= \left\langle \frac{1}{2} (A + A^T) + b, dx \right\rangle \quad \nabla f(x)$$

$x^T A dx = dx^T (x^T A)^T = dx^T (A^T x)$

Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{1}{2}\|Ax - b\|_2^2$.

1) $df = \frac{1}{2} d(\langle Ax-b, Ax-b \rangle) = \frac{1}{2} \cdot 2 \cdot \langle Ax-b, d(Ax-b) \rangle$

$= \langle Ax-b, A\,dx \rangle = \langle A^T(Ax-b), dx \rangle$ $\quad \boxed{\nabla f(x) = A^T(Ax-b)}$

2) $g(x) = \langle A^T(Ax-b), dx_1 \rangle$

$dg = \langle d(A^T(Ax-b)), dx_1 \rangle = \langle A^T d(Ax-b), dx_1 \rangle =$

$= \langle A^T A\,dx, dx_1 \rangle = \langle dx, (A^T A)^T dx_1 \rangle = \langle A^T A\,dx_1, dx \rangle$

$d(df) = \langle H_f\, dx_1, dx_2 \rangle$ $\quad dx = dx_2$

$H_f = A^T A$

Example 3

Find $\nabla f(x), f''(x)$, if $f(x) = -e^{-x^T x}$.

1) $df = d(-e^{-x^T x}) = -\left(e^{-x^T x} \cdot d(\langle x, x \rangle)\right) =$

$= e^{-x^T x} \cdot 2\langle x, dx \rangle \Rightarrow \boxed{\nabla f(x) = e^{-x^T x} \cdot 2x}$

2) $g(x) = \langle 2e^{-x^T x} x, dx_1 \rangle$

$dg = 2\langle d(e^{-x^T x} \cdot x), dx_1 \rangle = 2\langle d(e^{-x^T x}) \cdot x + e^{-x^T x} \cdot dx, dx_1 \rangle =$

$= 2\langle \langle -e^{-x^T x} \cdot 2x, dx \rangle \cdot x + e^{-x^T x} \cdot dx, dx_1 \rangle =$

$= 2\langle -2 \cdot e^{-x^T x} \langle x, dx \rangle \cdot x + e^{-x^T x} dx, dx_1 \rangle = 2e^{-x^T x} \langle -2 \cdot x^T dx \cdot x + dx, dx_1 \rangle =$

$= 2e^{-x^T x} \langle -2\, x x^T dx + dx, dx_1 \rangle =$

$= 2e^{-x^T x} \cdot \langle (-2xx^T + I)\, dx, dx_1 \rangle =$

$= 2e^{-x^T x} \langle (I - 2xx^T)\, dx_1, dx \rangle$

$\boxed{H_f = 2e^{-x^T x} \cdot (I - 2xx^T)}$

Example 4

Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if

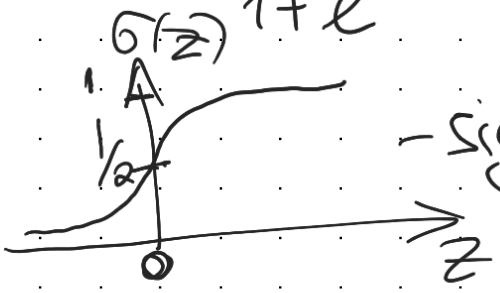
$f(x) = \ln(1 + \exp\langle a, x \rangle)$

$f(x) = \ln(1+\exp\langle a,x\rangle)$

$df = \frac{d(1+\exp\langle a,x\rangle)}{1+\exp\langle a,x\rangle} = \frac{\exp\langle a,x\rangle \cdot d(\langle a,x\rangle)}{1+\exp\langle a,x\rangle} =$

$\nabla f = \frac{\exp\langle a,x\rangle}{1+\exp\langle a,x\rangle} \cdot a$

$\sigma(z) = \frac{1}{1+e^{-z}}$



- Sigmoid

$\sigma'(z) = \sigma(z)\cdot(1-\sigma(z))$

$\sigma(\langle a,x\rangle)$

Example 5

Find $f'(X)$, if $f(X) = \det X$

$X \in S^n_{++}$

Note: here under $f'(X)$ assumes first order approximation of $f(X)$ using Taylor series:

$f(X+\Delta X) \approx f(X) + \mathbf{tr}(f'(X)^\top \Delta X) = \langle \nabla f(X), \Delta X\rangle$

- $f(X+dX) = \det(X+dX) = \det(X(I+X^{-1}dX)) = \det X \cdot \det(I+X^{-1}dX) =$

$= \det X \cdot \prod_{i=1}^n \lambda_i^z = \det X \prod_{i=1}^n (1+\lambda_i)$ $\quad z = I + X^{-1}dX$

assume $\|X^{-1}dX\| \to 0$

- $f(X+dX) - f(X) = \det X \prod_{i=1}^n (1+\lambda_i) - \det X = \det X\left(\prod_{i=1}^n (1+\lambda_i) - 1\right) =$

$\approx \det X \cdot \left(1 + \sum_{i=1}^n \lambda_i - 1\right) =$

$\approx \det X \sum \lambda_i = \det X \cdot \mathrm{tr}(X^{-1}dX)$

$f(x+dx) - f(x)$

$\langle \nabla f(X), dX\rangle \Rightarrow$

$\nabla f(X) = \det X \cdot (X^{-1})^\top$

$\langle \det X \cdot X^{-1}, dX\rangle$

Example 6

Calculate: $\frac{\partial}{\partial X}\sum \mathrm{eig}(X)$, $\frac{\partial}{\partial X}\prod \mathrm{eig}(X)$, $\frac{\partial}{\partial X}\mathrm{tr}(X)$, $\frac{\partial}{\partial X}\det(X)$

$$\frac{\partial\, tr(X)}{\partial X} = \frac{\partial\, tr(I\cdot X)}{\partial X} = \frac{\partial \langle X, I\rangle}{\partial X} =$$

$$= I \qquad tr(X) = \sum_i x_{ii}$$

$$\frac{\partial f}{\partial x_{kp}} = \frac{\partial\left(\sum_i x_{ii}\right)}{\partial x_{kp}} = \begin{cases} 1, & k=p \\ 0, & k\neq p \end{cases}$$

Example 7

Find $\nabla f(X)$, if $f(X) = \langle S, X\rangle - \log\det X$

$\det X \neq 0$

$$f(X) = tr(S^* X) - \log\det X$$

$$df = \langle S, dX\rangle - \frac{d(\det X)}{\det X} = \langle S, dX\rangle - \frac{\langle \det X\, X^{-1}, dX\rangle}{\det X} =$$

$$= \langle S, dX\rangle - \langle X^{-1}, dX\rangle =$$

$$= \langle S - X^{-1}, dX\rangle$$

$$\nabla f(X) = S - X^{-1}$$

Example 8

$$\langle A, B\rangle = tr(A^* B) = tr(B^* A)$$

Calculate the derivatives of the loss function with respect to parameters $\frac{\partial L}{\partial W}$, $\frac{\partial L}{\partial b}$ for the single object x_i (or, $n = 1$)