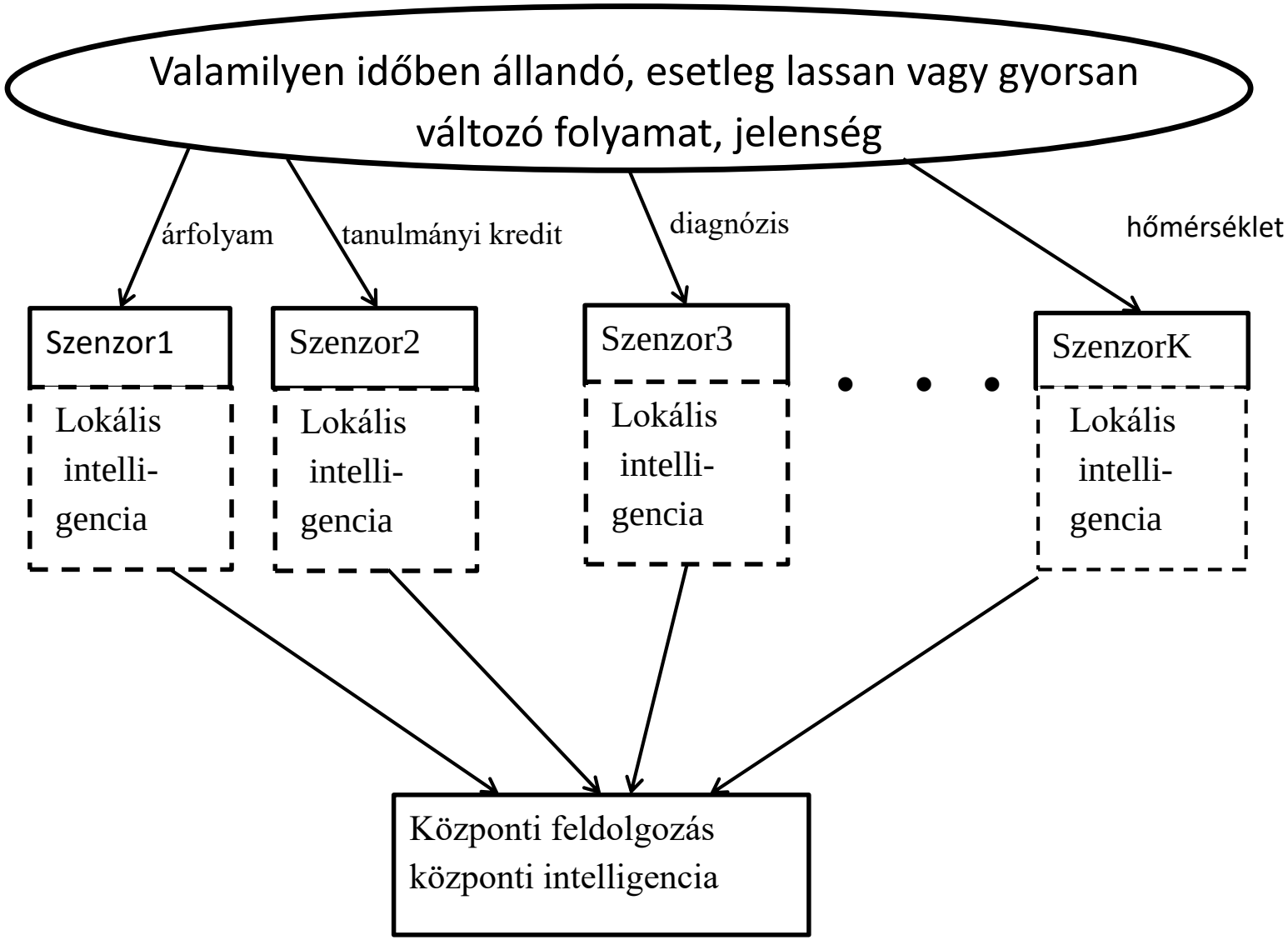
Intelligens elosztott rendszerek
VIMIAC02
Adatelőkészítés: hihetőségvizsgálat, normálás stb.

Pataki Béla

BME I.E. 414, 463-26-79
pataki@mit.bme.hu,
http://www.mit.bme.hu/general/staff/pataki


Valamilyen időben állandó, esetleg lassan vagy gyorsan változó folyamat, jelenség
árfolyam
tanulmányi kredit
diagnózis
hőmérséklet
Szenzor1
Szenzor2
Szenzor3
SzenzorK
Lokális intelli-gencia
Lokális intelli-gencia
Lokális intelli-gencia
Lokális intelli-gencia
Központi feldolgozás központi intelligencia

Sokszor nem kap kellő hangsúlyt az **adatok előfeldolgozása**

Szemét be ⇒ *Szemét ki*

Tipikus eljárások:

1. Adatok egységesítése
2. Hihetőségvizsgálat
3. Kilógó adatok detektálása, törlése
4. Normálás
5. Adatpótlás
6. Dimenziócsökkentés

stb.

1. Adatok egységesítése

- Mértékegységek (láb – méter, mérföld - km, ^{o}C – ^{o}F stb.)

  Volt űrhajó, amelyik azért tört össze, mert a nemzetközi fejlesztő konzorcium egyik része lábban, a másik része méterben számolt

- Elnevezési, illetve tartalmi kérdések
  - Szomorú aktualitás: https://koronavirus.gov.hu/elhunytak

    Megjelenik: „magasvérnyomás-betegség”, „magas vérnyomás” – azonos???
    Van „cukorbetegség”: 1-es típusú? , 2-es típusú?
  - Nevek, címek:
    - Bárány Boldizsár Dr., Bárány Dr., Dr. Bárány, Dr. B. Bárány stb. stb.
    - Kukori és Kotkoda Kft., Kukori & Kotkoda Kft, Kukori és Kotkoda, Kukori és Tsa Kft., Kuk. és Kotk. Kft stb.
    - Konkrét szolgáltató: volt, hogy egy cégnek 2-3 rekordot is nyitottak az adatbázisukban, mert a leolvasó hol így, hol úgy írta a nevet - soha össze nem találkoztak

- Elnevezési, illetve tartalmi kérdések
  - Kézzel bevitt numerikus adatok:
    - A szaki tudja, hogy az adat 0,02 és 0,08 közt van. Időnként elírja, és 0,05 helyett 0,5-öt, 5-öt vagy akár 50-et ír. Nem izgatja magát, úgyis mindenki tudja, hogy az „5” a lényeg. (A szgépes program izgatja magát!)
    - A gyárban az alakult ki, hogy ha „mindegy”, hogy mekkora keresztmetszetű vezetéket használnak, akkor egy nem létező keresztmetszetet írtak oda.
  - Címkék (gépi tanuláshoz sokszor címkézett adatokat használunk):
    - Orvosi adatok: az orvos hol latinul, hol magyarul, hol rövidítve, hol betűhibával írta fel a diagnózist, változatos szórendet és megfogalmazást használva. Persze minden orvos másképpen.

Az adatbevitelnél kell megfogni! Minimális szabadságot kell hagyni az adatok bevitelénél: legördülő listából kell standard elnevezések közül választani – a szabad szövegnél elvesztünk (természetes nyelvű szöveg gépi értelmezése)…

2. Hihetőségvizsgálat

1. Tudnunk kell valamit a jeleinkről (*a priori* információ)
    - pl. az emberek magasságának, tömegének alsó és felső korlátai
    - a magasság és a tömeg szokásos összefüggése, a lehetséges eltérés mértéke
    - mennyit fogyaszt egy lámpaizzó
2. A mért jeleinkből levonunk következtetést a szokásos limitekről, összefüggésekről (mi a tipikus, mi a kilógó adat?)
    - ha sok adatot mérünk, akkor kereshetünk tipikus csoportokat (klasztereket), limiteket, összefüggéseket az adatokban – lásd később

Egyszerű példa:

Szenzor1: A megfigyelt ember magassága **199 cm**

Szenzor2: A megfigyelt ember tömege **89 kg**

Szenzor3: A megfigyelt ember tömege egy nap alatt **18,2 kg**-t nőtt

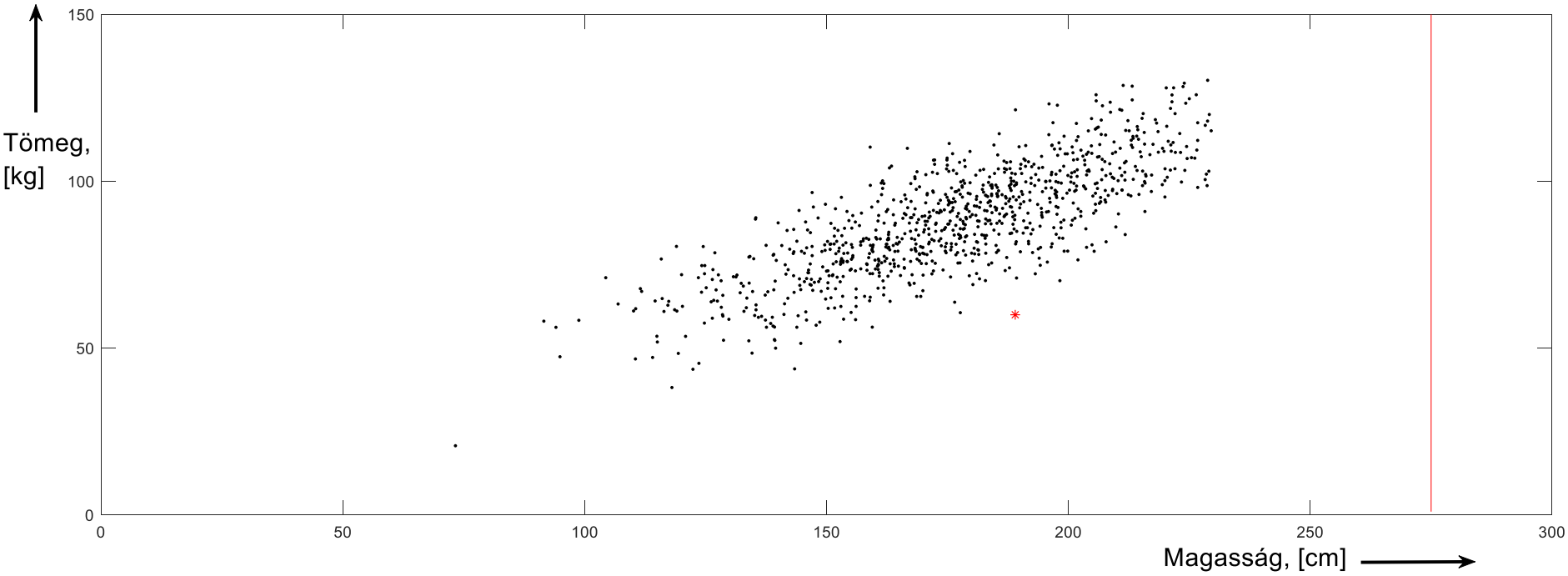
Szenzor4: A megfigyelt ember magassága **195 cm**

Szenzor5: A megfigyelt ember tömege 60 kg

Szenzor6: A megfigyelt ember magassága **275 cm**

Szenzor7: A megfigyelt ember magassága **197 cm**

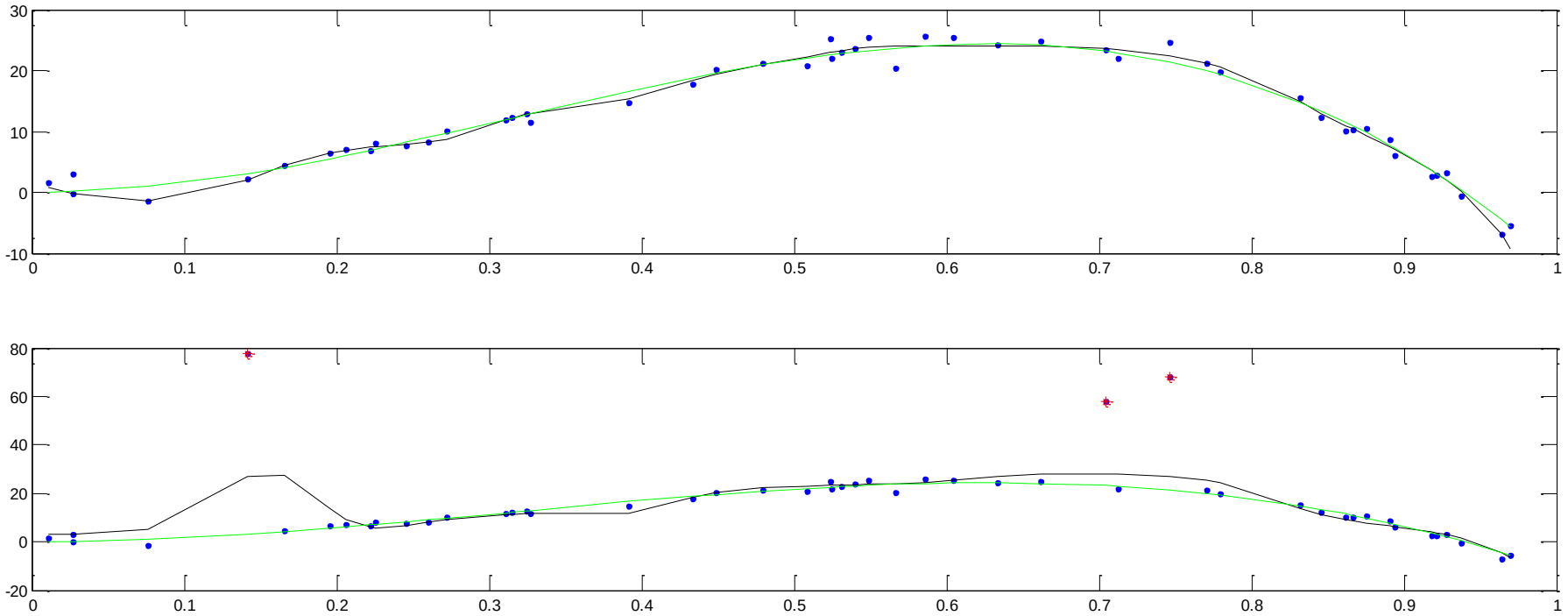
Szenzor8: A megfigyelt ember tömege egy nap alatt 0,7 kg-t csökkent



3. Kilógó (outlier) adatok detektálása

(lehet hihető, de kilóg)

Pl. gépi tanulásnál a kilógó adatok erősen meghamisíthatják a tanítás eredményét

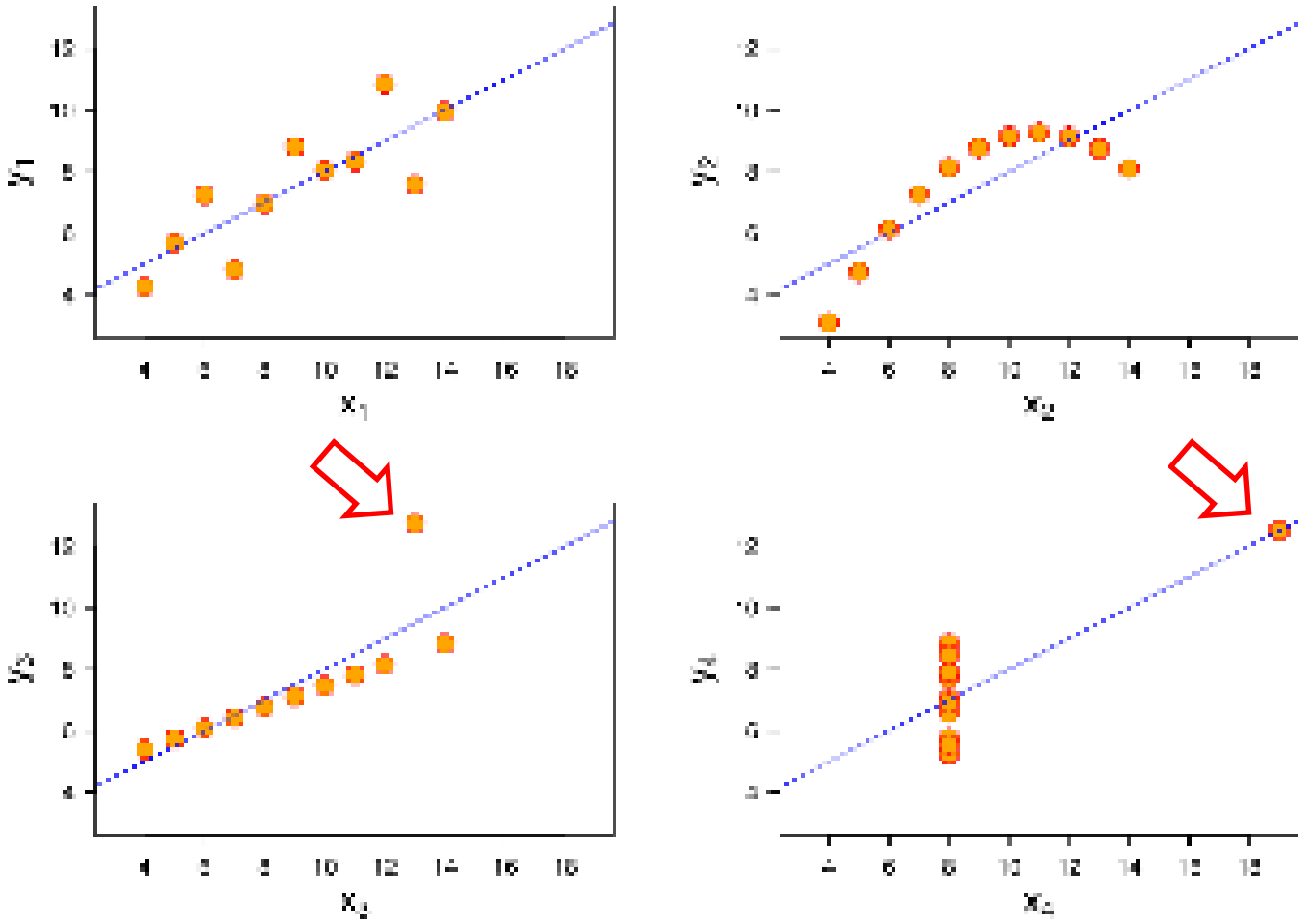


zöld: igazi görbe, fekete: a neuronháló által tanult modell

Az alsó ábrán 3 erősen kilógó (outlier) adatpont miatt torzított a tanítás eredménye (a többi pont, illetve a háló mérete azonos)

Múlt órai előadás: Anscombe kvartett

A két alsónál kilógó adattal hozott létre becsapós számértékeket.



3.1. Az adatok szórását felhasználva

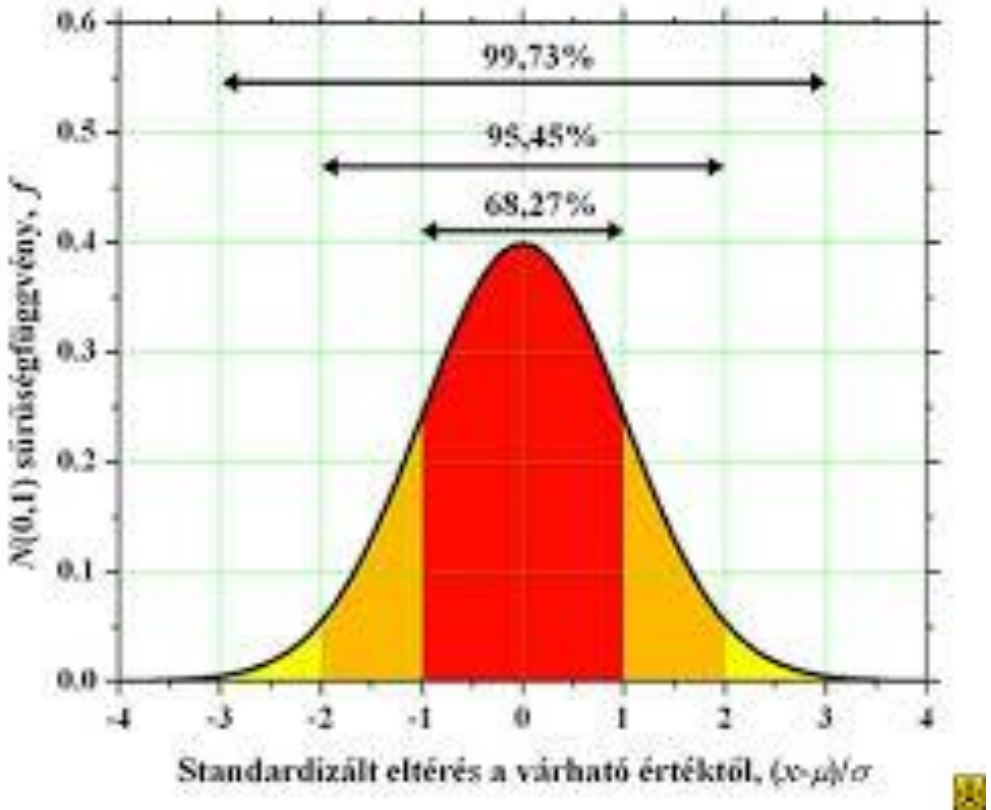
- Adatmodellünk: az adat egy átlagos érték körül szór, és minél távolabb van az átlagtól, annál kevésbé valószínű, hogy reális (nem kilógó) - pl. emberek magassága, bankszámlabetét, vagy pénz kivétel, elektromos fogyasztás stb.
- Normális (0 átlagú, várható értékű, egységnyi szórású Gauss)

$P(\sigma<|x|)\cong 32\%$

$P(2\sigma<|x|) \cong 4{,}6\%$

$P(3\sigma<|x|) \cong 0{,}28\%$

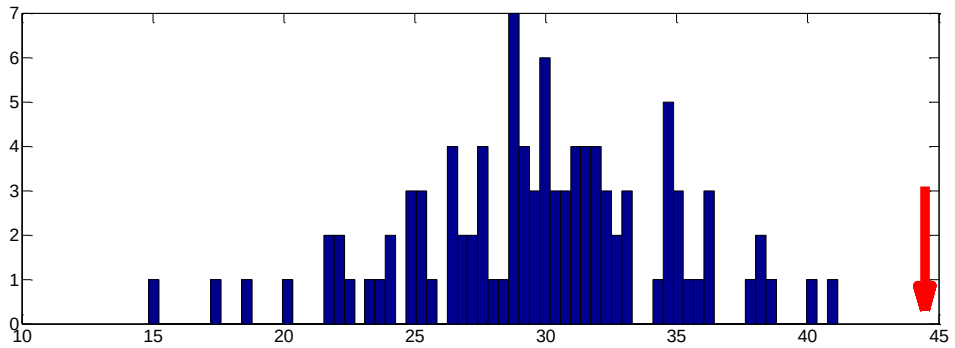
$P(4\sigma<|x|) \cong 0{,}006\%$



Az adatok szórását felhasználva

Példa: adatsorunknál (100 adat) az átlagtól több mint 4σ-val eltérő adatot kilógónak minősítjük.

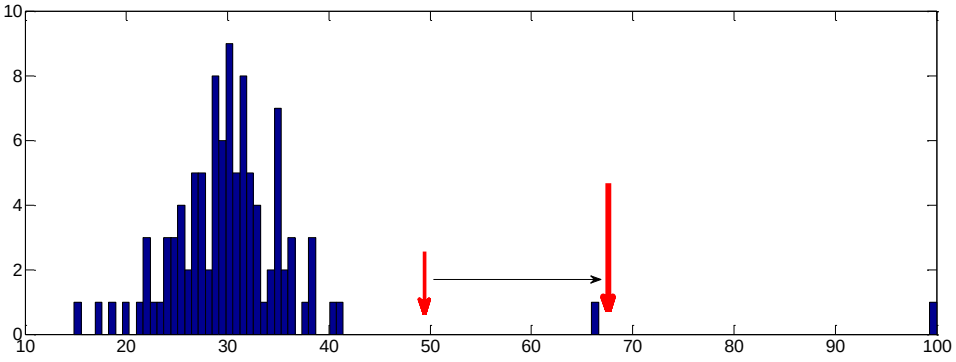
N adatszám	átlag	szórás	kilógó adat limit
100	29,7	4,85	49,1 (10,3)



Jön még két adat (2 kilógó):

X(101)=66 X(102)=100

N adatszám	átlag	szórás	kilógó adat limit
100+2	30,7	9,17 !!!	67.4 (-5,98)



Mi a baj sokszor a szórásalapú kilógóadat-detektálással?

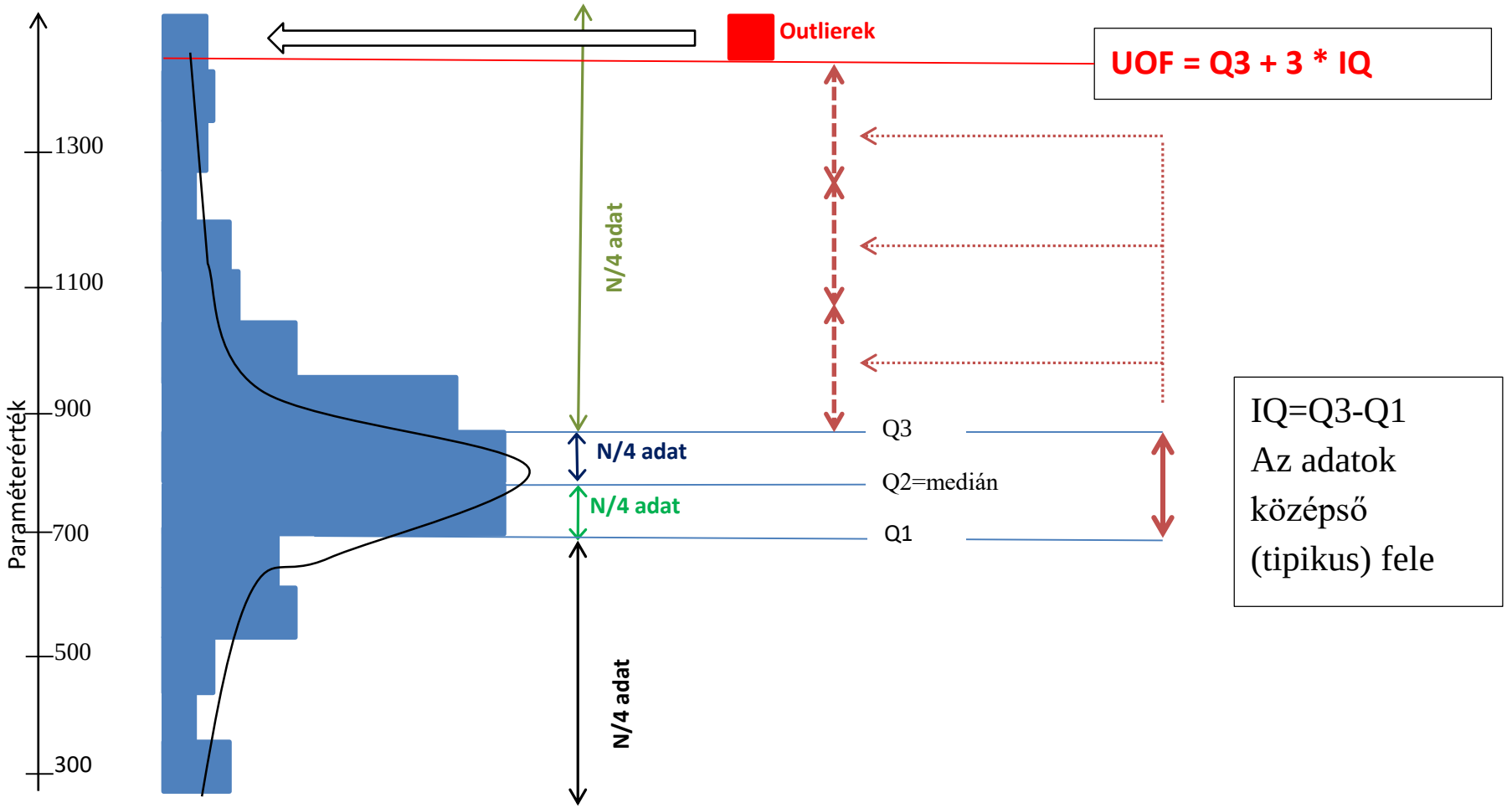
Átlag P adat esetén

$$\overline{x} = \frac{1}{P}\sum_{p=1}^{P} x^{(p)}$$

Szórás P adat esetén

$$\hat{\sigma}^2 = \frac{1}{P-1}\sum_{p=1}^{P} (x^{(p)} - \overline{x})^2$$

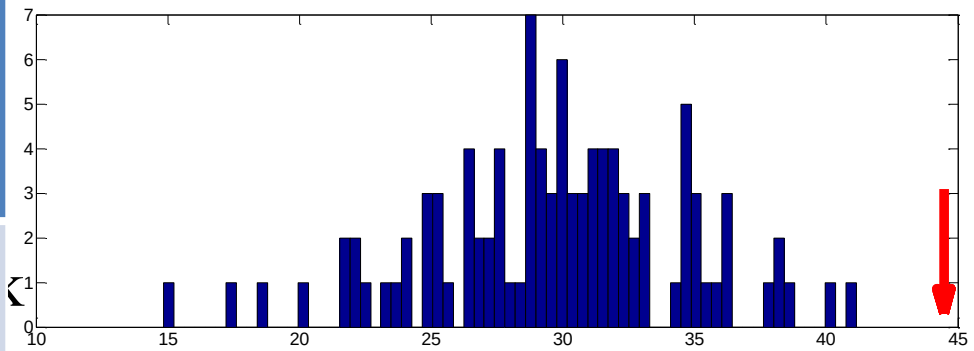
3.2 Az adatok sorrendezését felhasználva (Order statistic filtering)



Az adatok sorrendezését felhasználva

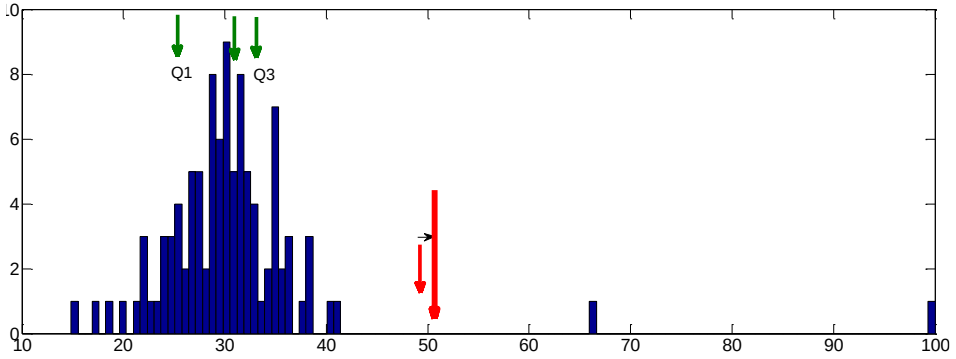
Ugyanaz a példa, mint az előbb: 100 adat, majd plusz kettő; 66 és 100):

N	Q1 Q2 Q3 Q4	IQ	Q3+3*IQ
100	26,7 29,9 32,4 41,1	5,6	49,2

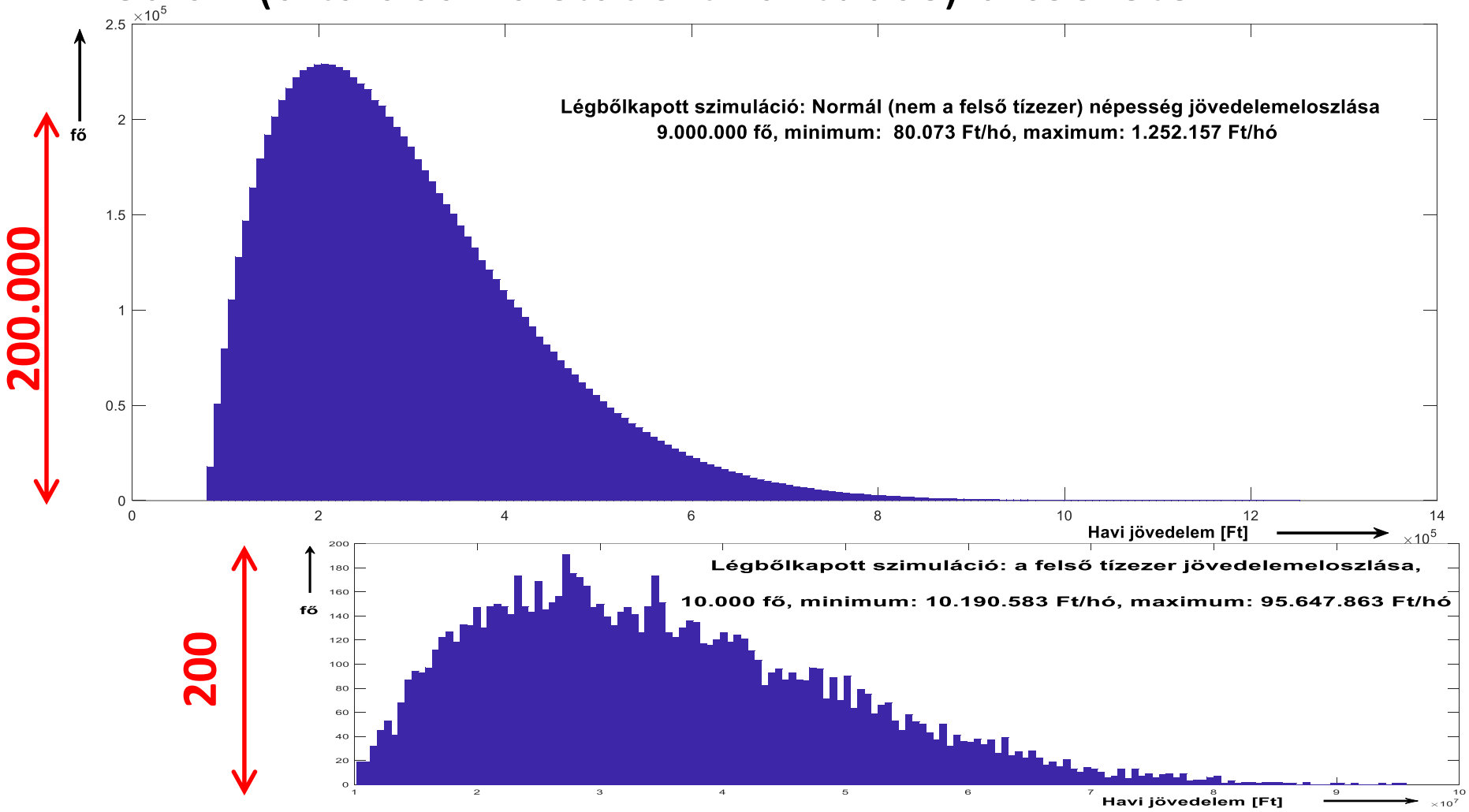


adatok Q3 körül: 32.22 , 32.36, 32.58

N	Q1 Q2 Q3 Q4	IQ	Q3+3*IQ
102	26,8 29,9 32,6 100	5,8	50,1



A medián (általában a statisztikai tudás) dicsérete



N0=9000000 P0Atlag: 294.238 P0Median: 268.673 P0Arany: 1.10

NOssz=9010000 POsszAtlag: 332.742 POsszMedian: 268.856 POsszArany: 1.24

POsszAtlag/P0Atlag: 1.13 POsszMedian/P0Median: 1.00

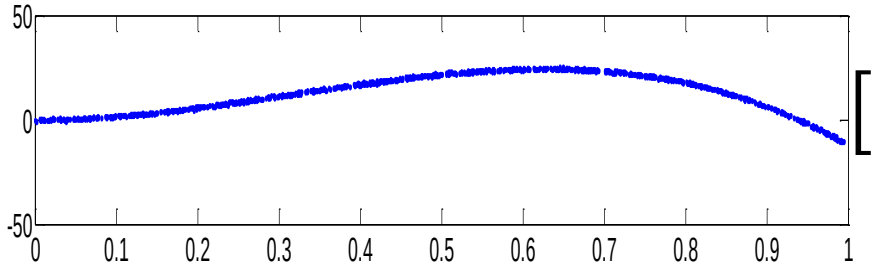
Magyarországon a hivatalos, 2014-es adatsor alapján a mediánjövedelem az átlagbér 76 százaléka volt.

4. Normálás

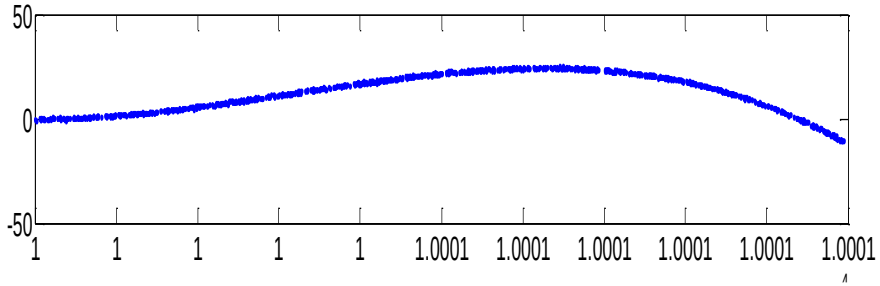
A normálás sokszor nagymértékben gyorsítja, pontosabbá, esetenként lehetővé teszi a feldolgozást

Demópélda: polinomiális regresszió

$$y = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + noise$$



$$[p_0, p1, p2, p3] = polyfit(x, y, 3)$$



$$[p_0, p1, p2, p3] = polyfit(10000 + x, y, 3)$$

$$y = p_0 + p_1(x - x0) + p_2(x - x0)^2 + p_3(x - x0)^3 + noise$$

először $\quad x \in (0,1) \quad x0 = 0$

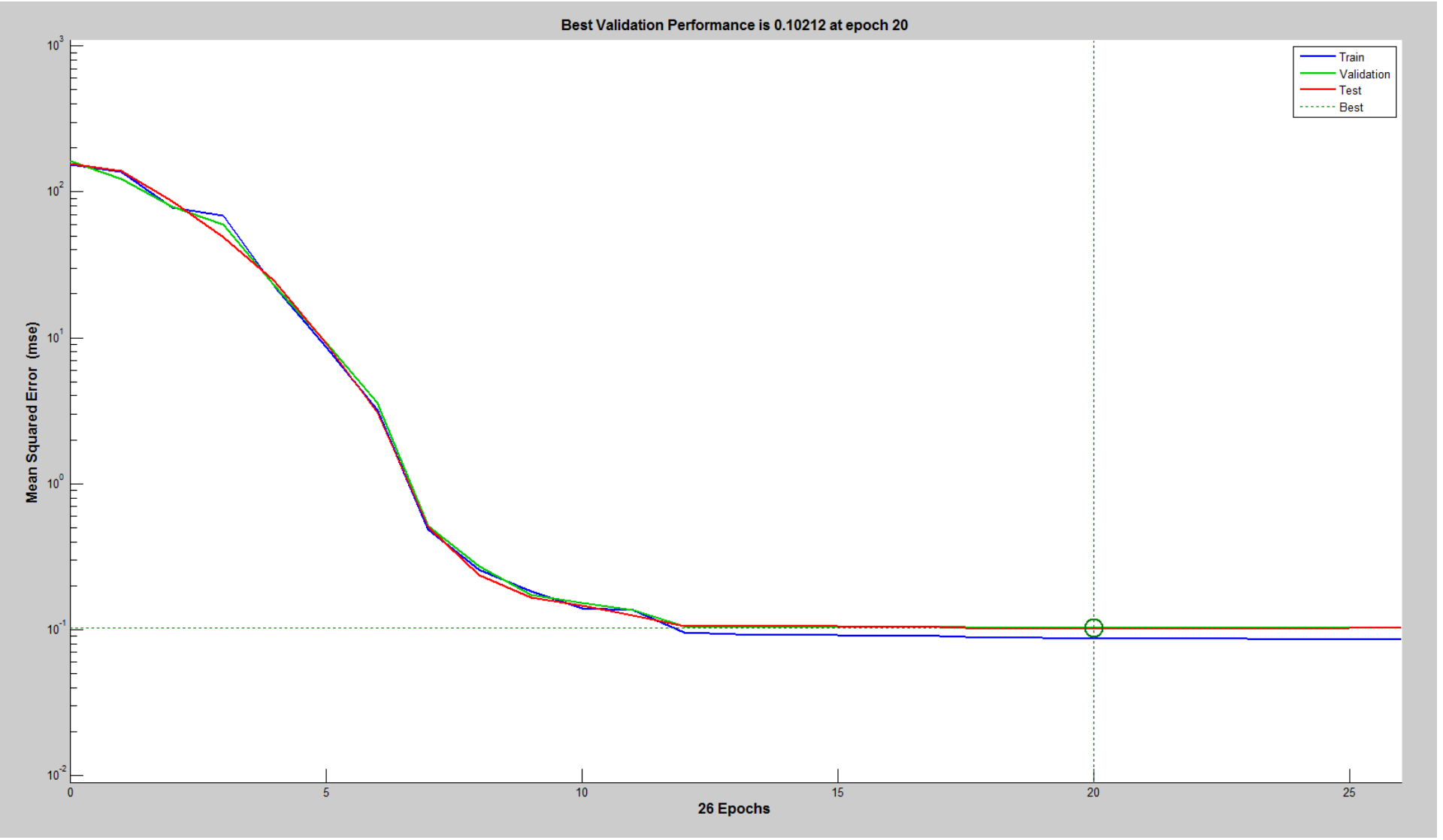
másodszor $\quad x \in (10.000,10.001) \quad x0 = 10.000$

Valódi értékek	p_0=0	p_1=-1	p_2=189	p_3=-200
Becsült - x ϵ (0,1)-ben	-0,0068	-1,209	189,7	-200,56
Becsült - x ϵ (10000,10001)	19.5	-58.613.481.990	5.861.166	-195.38

Warning: Polynomial is badly conditioned. Add points with distinct X values, reduce the degree of the polynomial, or *try centering and scaling* as described in HELP POLYFIT.

Ugyanez a demópélda neuronhálós tanítással eltolás nélkül x0=0

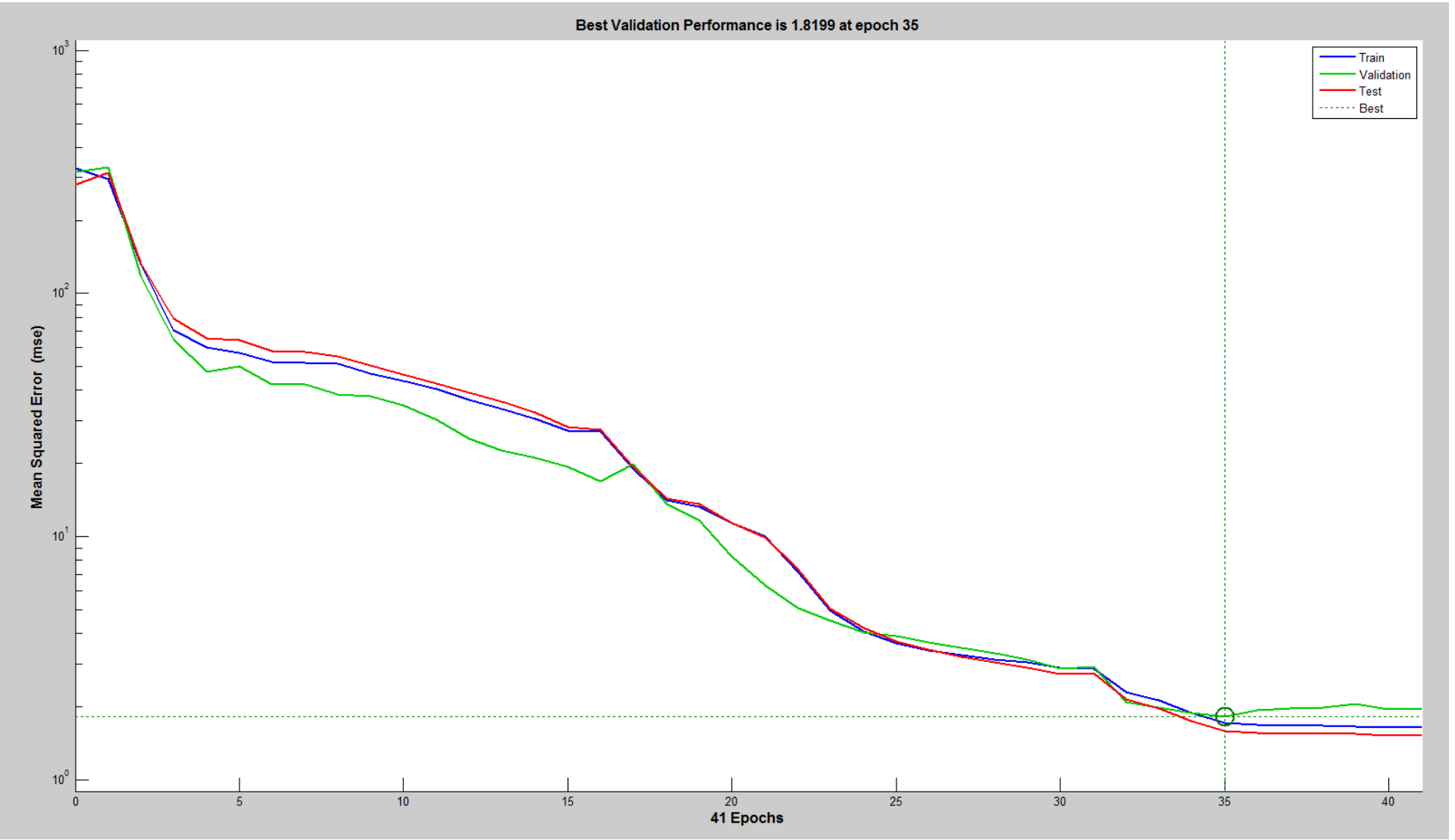
a 20. epochnál éri el a legjobb közelítést, 0,10212 MSE



Eltolással x0=10.000

A 35. epochnál éri el a legjobb közelítést, 1,8199 MSE

Több mint egy nagyságrenddel rosszabb közelítés! (Egyes esetekben ennél lényegesen lassabb futást is tapasztalhatunk, sokkal rosszabb végeredménnyel!)



Egyszerű normálási eljárások

[0,1] tartományra normálás (skalár komponensenként, i=1,2,…,N)

$$\tilde{x}_i^{(p)} = \frac{x_i^{(p)} - \min\{x_i^{(p)}\}}{\max\{x_i^{(p)}\} - \min\{x_i^{(p)}\}}$$

Nulla várhatóértékűre és egységnyi szórásúra normálás (skalár komponensenként eset)

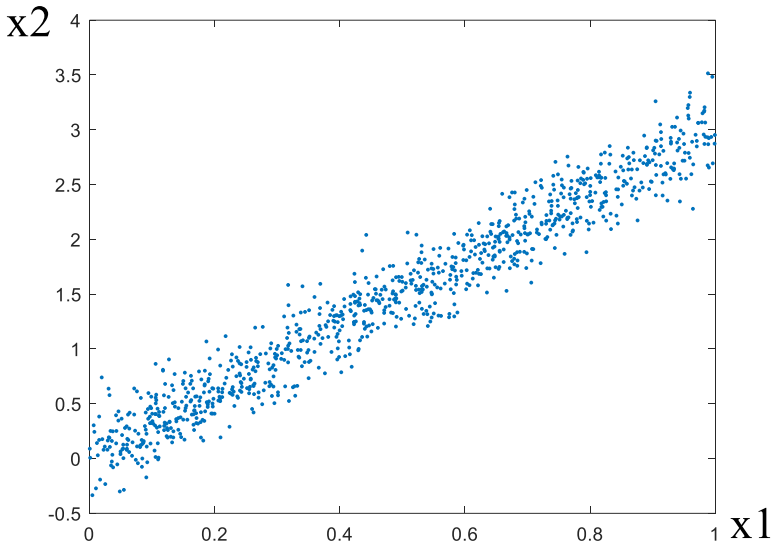
$$\bar{x}_i = \frac{1}{P}\sum_{p=1}^{P} x_i^{(p)}$$

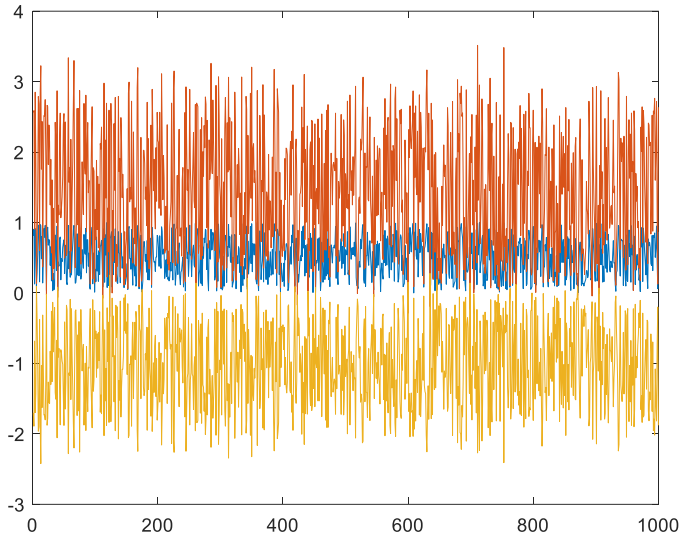
$$\hat{\sigma}_i^2 = \frac{1}{P-1}\sum_{p=1}^{P}(x_i^{(p)} - \bar{x}_i)^2$$

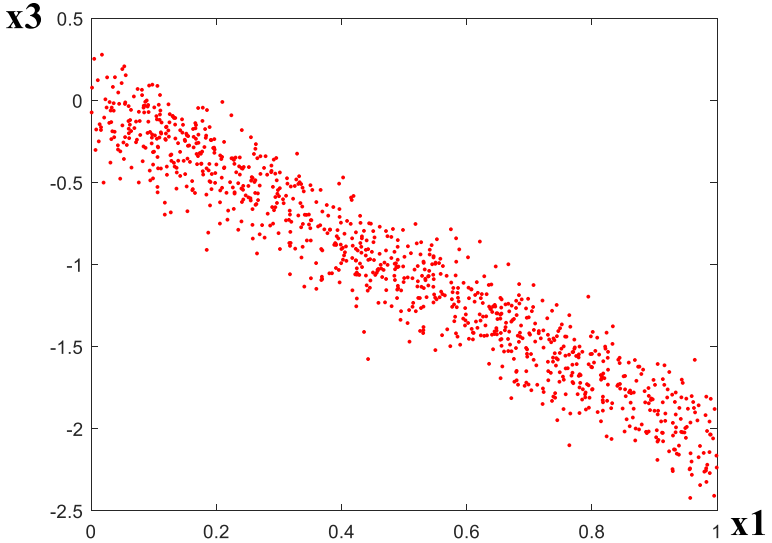
$$\tilde{x}_i^{(p)} = \frac{x_i^{(p)} - \bar{x}_i}{\hat{\sigma}_i}$$

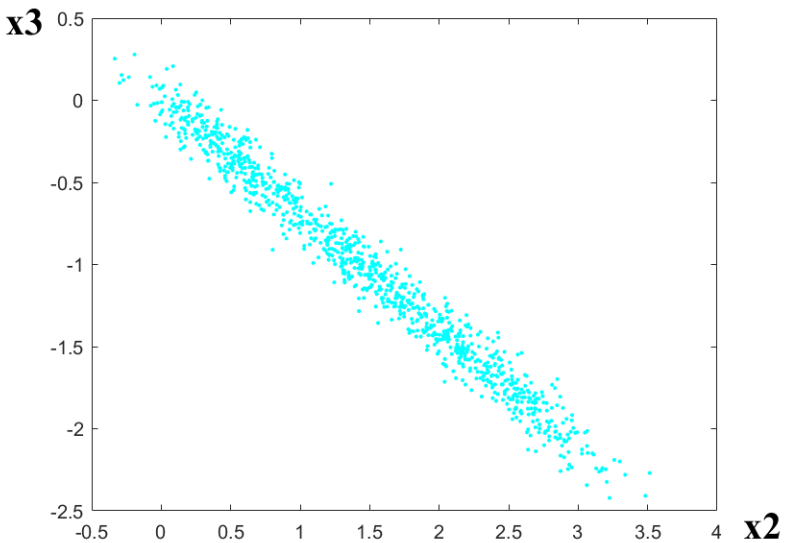
A paramétervektor skalár komponenseinek együttes normálása

$$\mathbf{x} = \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix}$$









A paramétervektor skalár komponenseinek együttes normálása

$$\bar{\mathbf{x}} = \frac{1}{P}\sum_{p=1}^{P}\mathbf{x}^{(p)}$$

Kovarianciamátrix

0.080	0.2394	-0.1682
0.239	0.7540	-0.529
-0.1682	-0.529	0.381

$$\mathbf{C} = \frac{1}{P-1}\sum_{p=1}^{P}(\mathbf{x}^{(p)} - \bar{\mathbf{x}})(\mathbf{x}^{(p)} - \bar{\mathbf{x}})^{\mathrm{T}}$$

$$\mathbf{C}\boldsymbol{\varphi}_j = \lambda_j \boldsymbol{\varphi}_j$$

0.924	0.289	-0.252
-0.369	0.490	-0.790
-0.105	0.823	0.559

$$\Phi = \begin{bmatrix} \boldsymbol{\varphi}_1 & \boldsymbol{\varphi}_2 & \ldots\ldots & \boldsymbol{\varphi}_N \end{bmatrix}$$

0.0038	0	0
0	0.0071	0
0	0	1.2043

$$\Lambda = diag\begin{pmatrix} \lambda_1 & \lambda_2 & & \lambda_N \end{pmatrix}$$

$$\tilde{\mathbf{x}}^{(p)} = \Lambda^{-1/2}\Phi^{\mathrm{T}}(\mathbf{x}^{(p)} - \bar{\mathbf{x}})$$

$\tilde{\mathbf{x}}^{(p)}$ nulla várhatóértékű és kovarianciamátrixa egységmátrix

5. Hiányos, hibás adatok, adatpótlás

Mivel mindig kevés az adat: ha hiányzik egyik-másik komponense, rendszerint akkor sem éri meg eldobni, jobb pótolni a hiányt.

Ennek alkalmas eszköze lehet a *klaszterezés*, csoportok keresése a mért adatok közt.

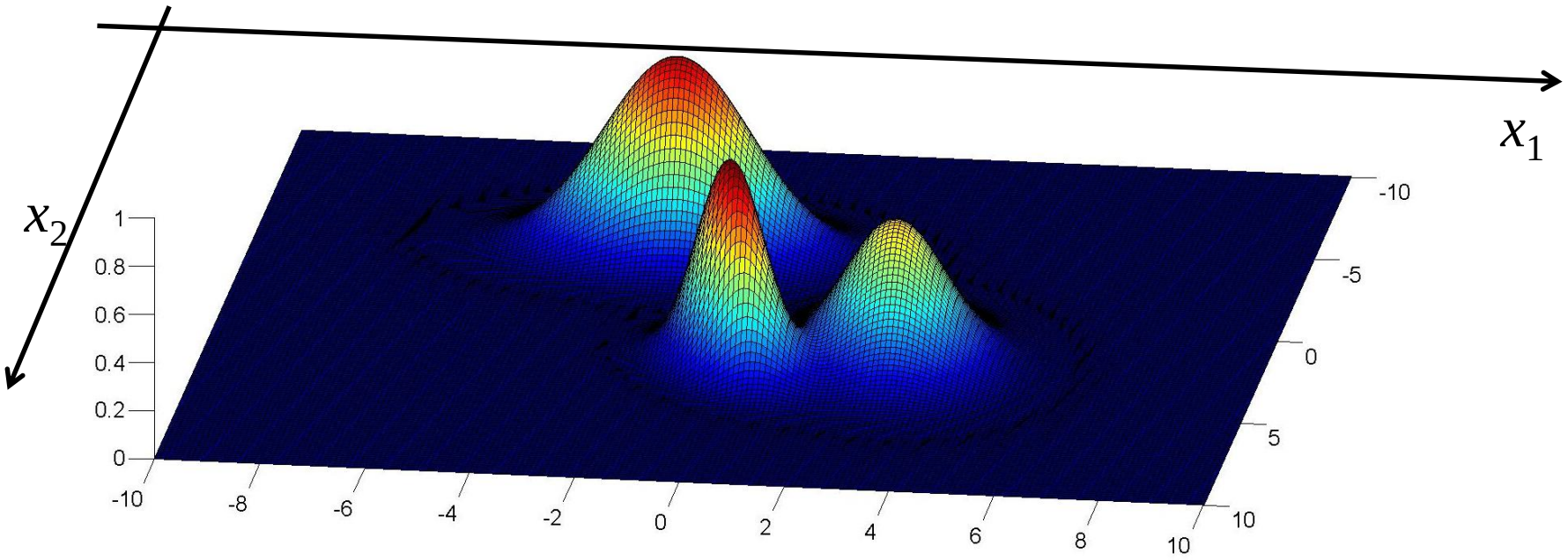
Demópélda:

Két paraméterrel jellemezzük a mintáinkat: $p_k = \begin{bmatrix} x_{k1} \\ x_{k2} \end{bmatrix}$

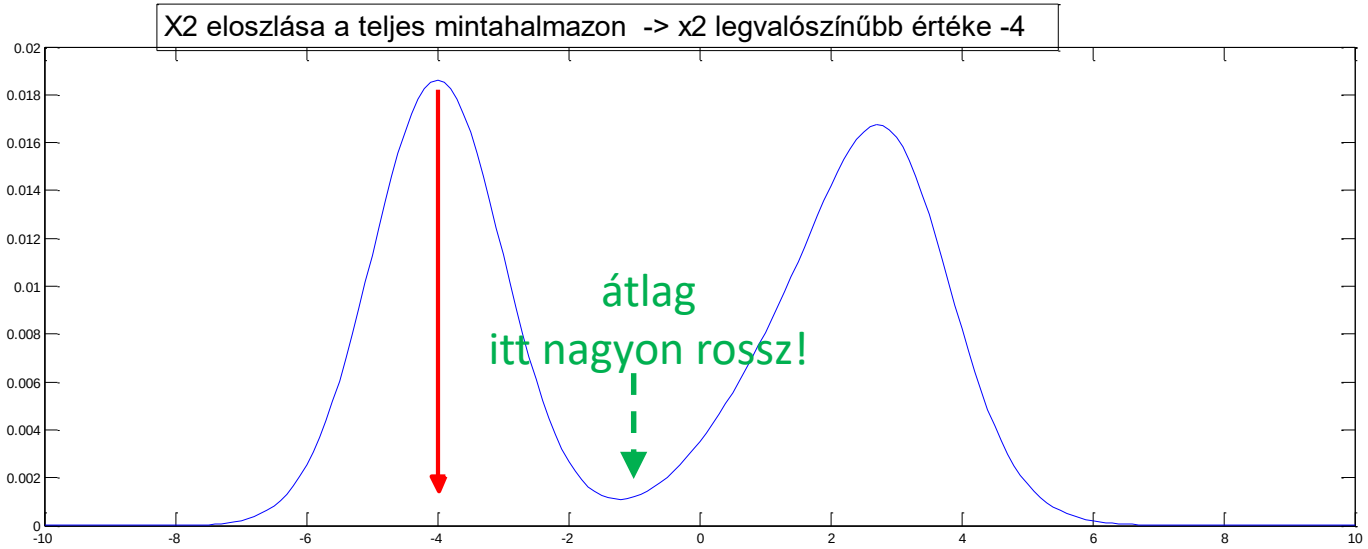
Egyes mintáknak hiányzik az egyik komponense (vagy annyira torz, hogy nem vesszük figyelembe).

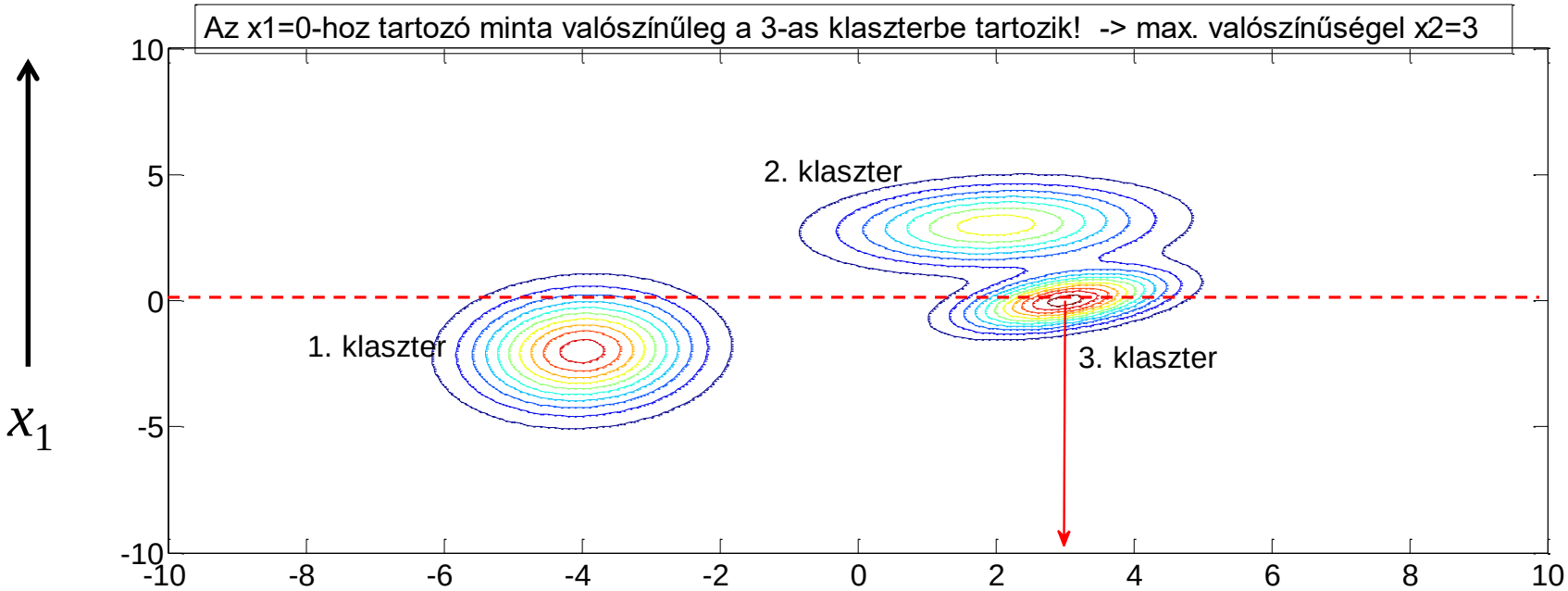
Például az *n*-dik minta második paramétere hiányzik: $p_n = \begin{bmatrix} x_{n1} \\ ? \end{bmatrix}$

A kétdimenziós adathalmaz eloszlása



n-dik minta: x_{n1}=0, de hiányzik az x_{n2} paraméter. Nézzük meg, melyik x_{n2} legvalószínűbb értéke az x_2 eloszlás alapján:

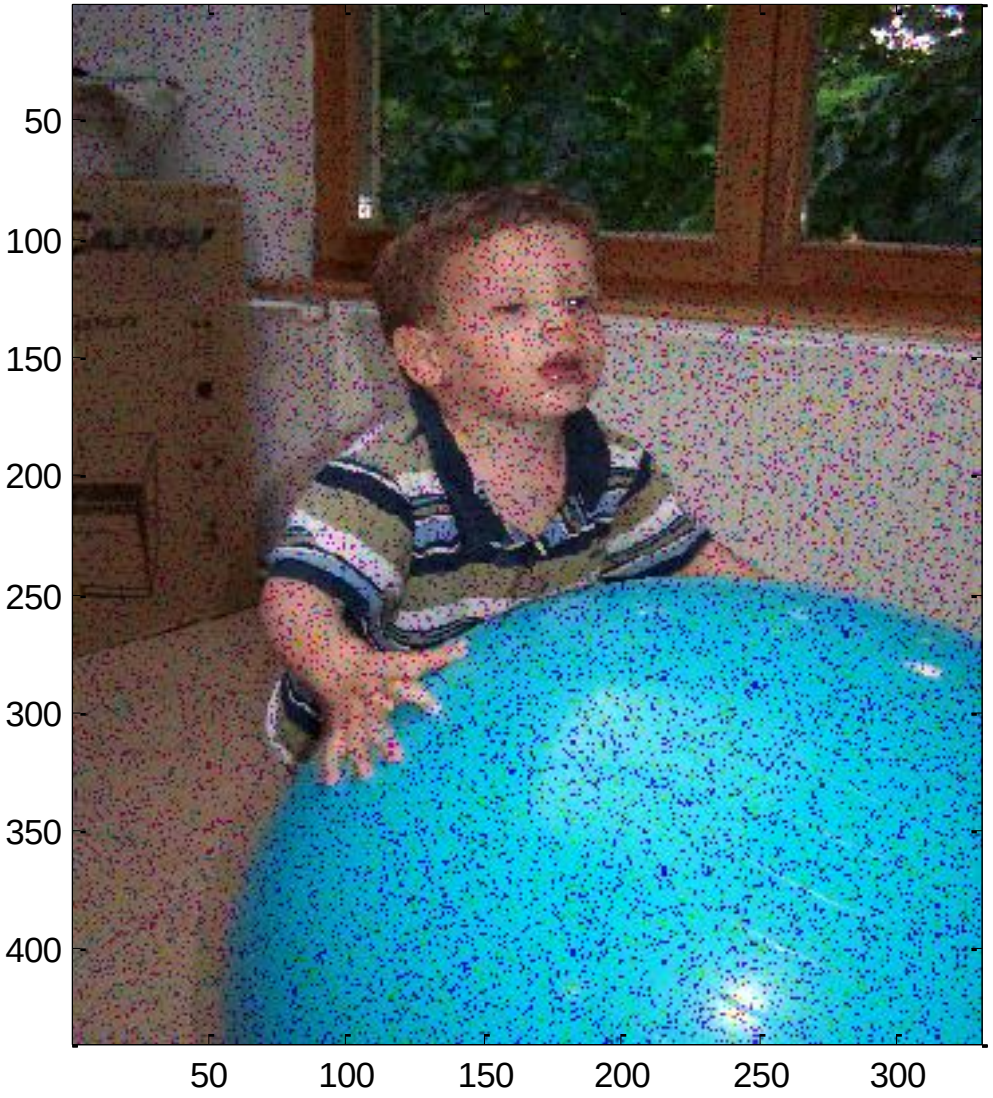




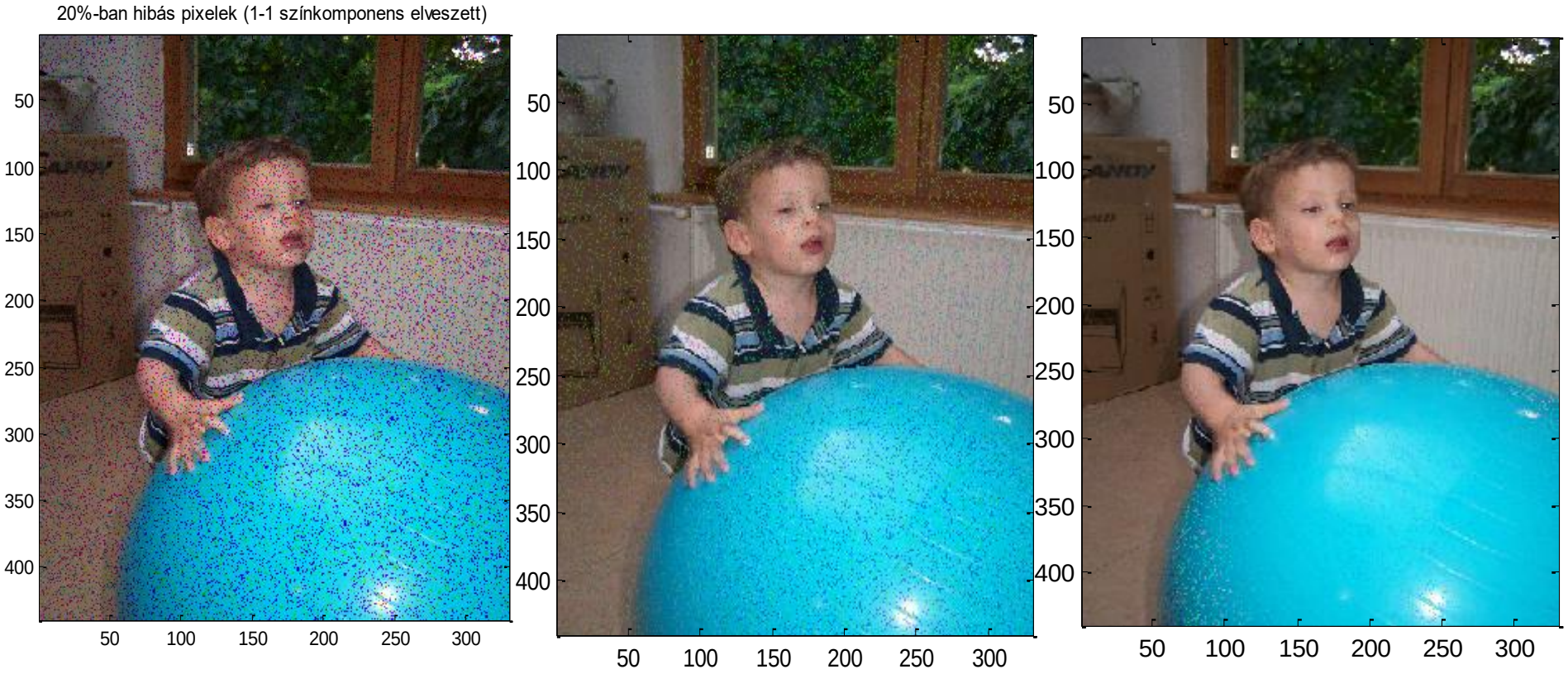
Második demópélda

Eredeti kép

20%-ban hibás pixelek (1-1 színkomponens elveszett)



Balról-jobbra: a hibás kép, a globális paraméterekkel javított és a klaszterezett, majd klaszterenkénti paraméterekkel javított



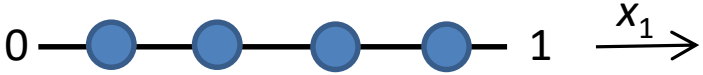
6. Dimenziócsökkentés (lényegkiemelés)

A dimenzió átka (*curse of dimensionality*)

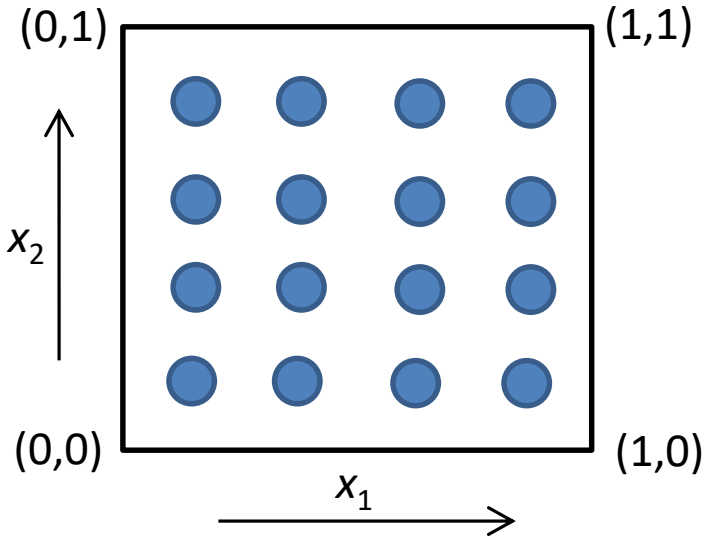
Ha az adatok sokdimenziósak

$\Rightarrow$ Hány mintapont kell, hogy kb. egyenletesen lefedjük a mintateret? (Tfh. normáltuk [0,1]-re az összes komponenest)

D=1, N=M mintapont



D=2, N=M^2 mintapont



Általánosságban *:* N=M^D mintapont

A felrajzolt esetben M=4, ha D=20, akkor $N \cong 10^{12}$ pont kell,

ha M=15 és D=20, akkor $N \cong 3 \cdot 10^{23}$ pont kell

A dimenzió átka (*curse of dimensionality*)

Ugyanaz a probléma másképp: az ismeretlen jövőbeli mintára valószínűleg akkor adunk jó eredményt, ha vannak a közelében ismert mintáink.

D- dimenziós egyenletes eloszlású minták, egységnyi oldalhosszúságú *D*-dimenziós hiperkockába normálva (minden x_i a [0,1]-re), *N* mintapontunk van, és azt szeretnénk, hogy az ismeretlenhez *k* db. ismert mintapont legyen „közel”.

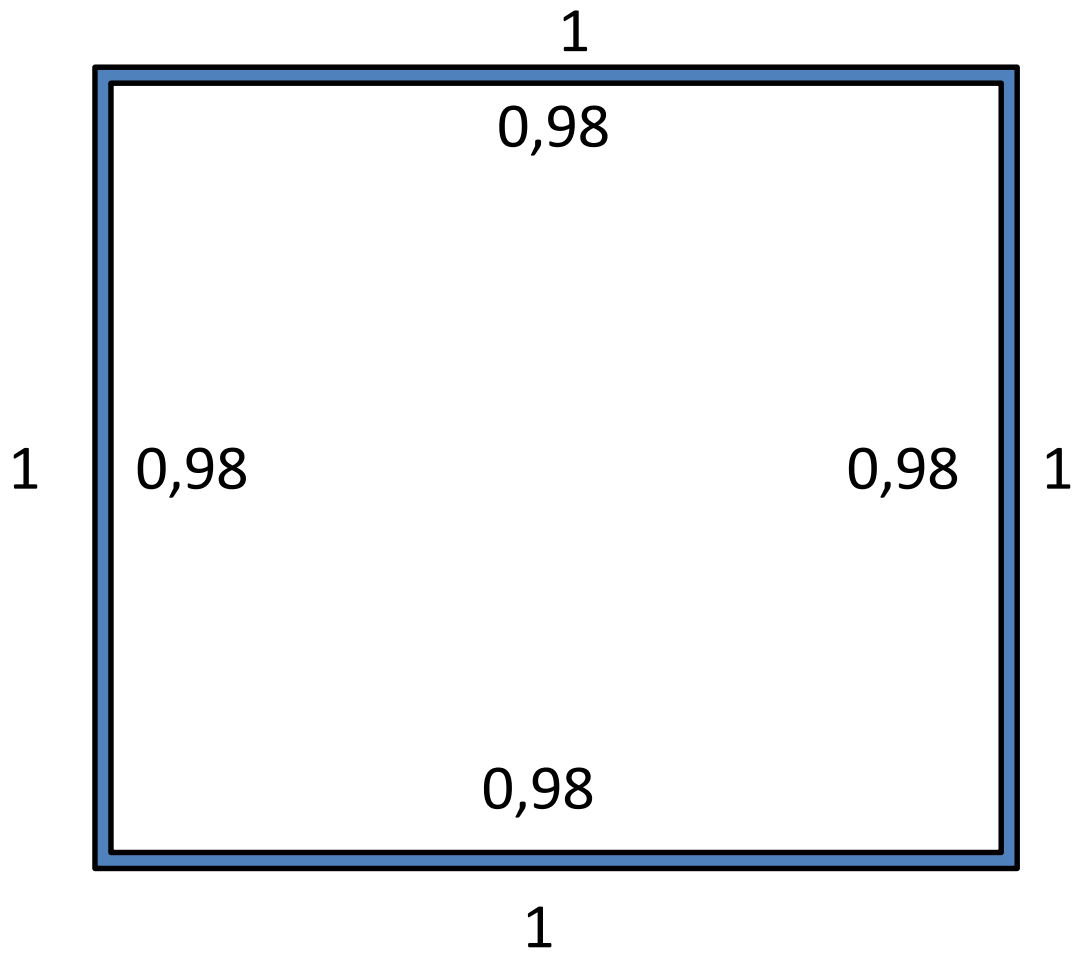
A k pontot tartalmazó, ℓ oldalhosszúságú "kis" hiperkocka térfogata: $V_k = \ell^D$

Az egységnyi oldalhosszúságú hiperkockában N pont van: $\ell^D = \frac{k}{N}$

Tehát: $\ell = \left(\frac{k}{N}\right)^{\frac{1}{D}} = \sqrt[D]{\frac{k}{N}}$ (de $\frac{k}{N} \ll 1$!)

***N*=1.000.000 ; *k*=10**	
D	ℓ
2	0,003
20	0,56
200	0.94

„Hámozzuk le” minden dimenzióban a *D*-dimenziós egységnyi oldalhosszúságú hiperkocka külső 1-1%-át. Mekkora térfogat marad a belső „meghámozott” részen?

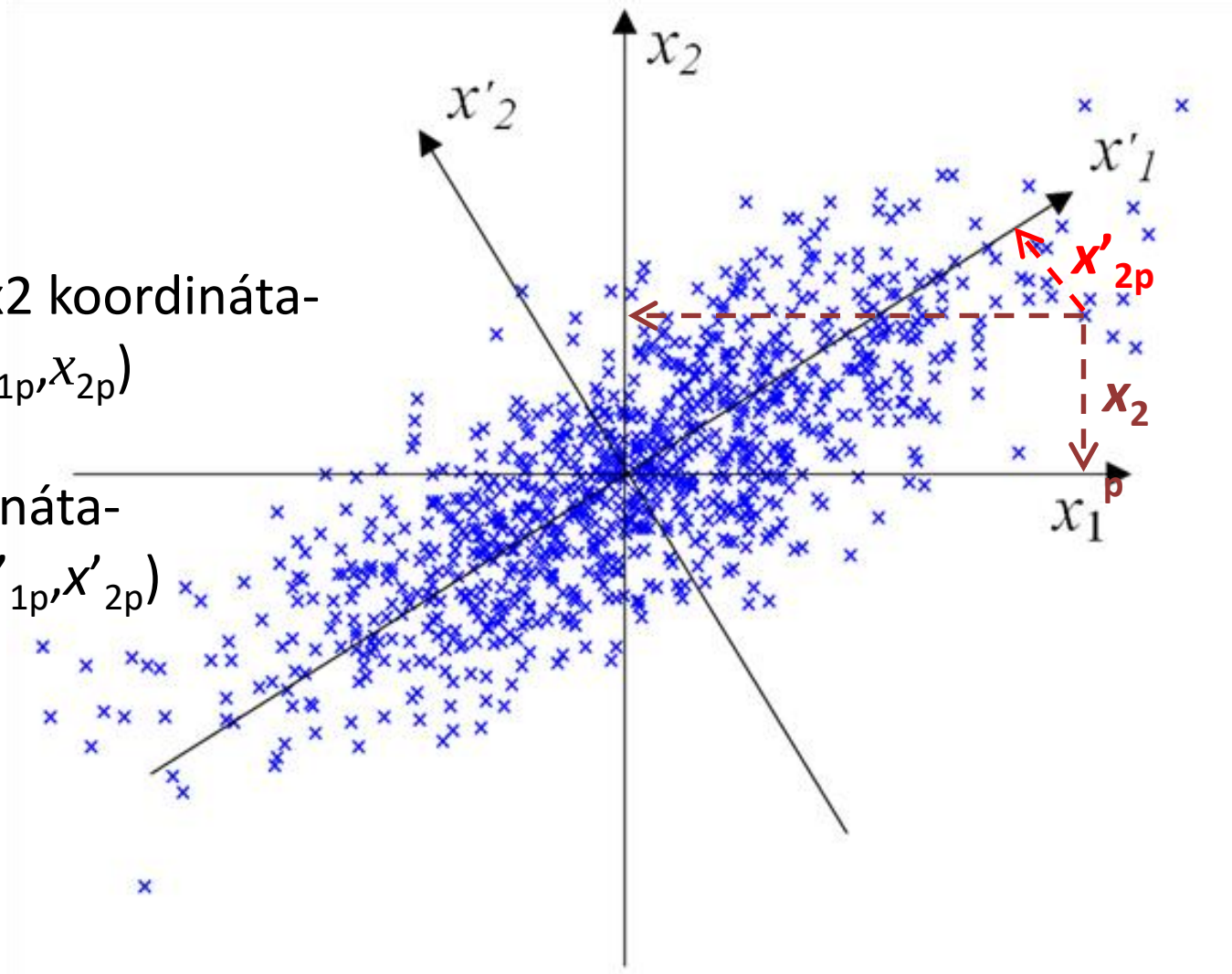


D	*V*
2	0,96
20	0,67
200	0.018 (1,8%!)

Az egyik legelterjedtebb dimenziócsökkentő eljárás: a KL (Karhunen–Loève) transzformáció

A P pont az x1-x2 koordináta-rendszerben: (x_{1p},x_{2p})

Az x_1'-x_2' koordináta-rendszerben: (x'_{1p},x'_{2p})



KL transzf.

$$\mathbf{y} = \mathbf{T} \cdot \mathbf{x} \qquad \mathbf{T} = \begin{bmatrix} \boldsymbol{\varphi}_1^T \\ \boldsymbol{\varphi}_2^T \\ \dots \\ \boldsymbol{\varphi}_N^T \end{bmatrix}$$

ortonormált vektorok: $\boldsymbol{\varphi}_k^T \cdot \boldsymbol{\varphi}_j = \delta_{kj}$

$$\mathbf{x} = \mathbf{T}^{-1} \cdot \mathbf{y} = \mathbf{T}^T \cdot \mathbf{y} = \sum_{i=1}^{N} y_i \cdot \boldsymbol{\varphi}_i$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varphi}_1^T \\ \boldsymbol{\varphi}_2^T \\ \dots \\ \boldsymbol{\varphi}_N^T \end{bmatrix} \cdot \mathbf{x} \quad \Rightarrow \quad y_k = \boldsymbol{\varphi}_k^T \cdot \mathbf{x} = \mathbf{x}^T \cdot \boldsymbol{\varphi}_k$$

Dimenziócsökkentés – elhagyunk komponenseket

$$\mathbf{x} = \sum_{i=1}^{N} y_i \cdot \boldsymbol{\varphi}_i \quad \Rightarrow \quad \hat{\mathbf{x}} = \sum_{i=1}^{M} y_i \cdot \boldsymbol{\varphi}_i$$

Elkövetett hiba:

$$\varepsilon^2 = E\left\{\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|\right\} = E\left\{\left\|\sum_{i=1}^{N} y_i \cdot \boldsymbol{\varphi}_i - \sum_{i=1}^{M} y_i \cdot \boldsymbol{\varphi}_i\right\|\right\} = \sum_{i=M+1}^{N} E\left\{y_i^2\right\}$$

$$\varepsilon^2 = \sum_{i=M+1}^{N} E\left\{y_i^2\right\} = \sum_{i=M+1}^{N} E\left\{\left(\boldsymbol{\varphi}_i^T \cdot \mathbf{x}\right) \cdot \left(\mathbf{x}^T \cdot \boldsymbol{\varphi}_i\right)\right\} = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\} \cdot \boldsymbol{\varphi}_i$$

$\mathbf{C}_{\mathbf{xx}} = E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\}$ a hiba akkor lesz minimális, ha $\mathbf{C}_{\mathbf{xx}} \cdot \boldsymbol{\varphi}_i = \lambda_i \cdot \boldsymbol{\varphi}_i$

$$\varepsilon^2 = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\} \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot \mathbf{C}_{\mathbf{xx}} \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot \lambda_i \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \lambda_i$$

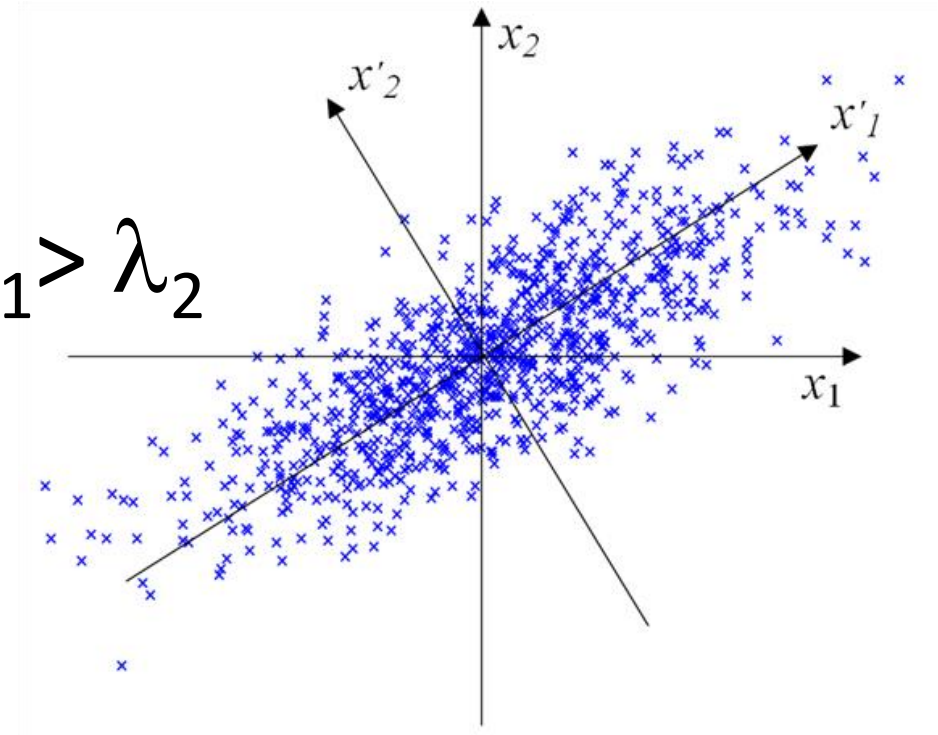
Dimenziócsökkentés – elhagyunk komponenseket

$$\mathbf{x} = \sum_{i=1}^{N} y_i \cdot \boldsymbol{\varphi}_i \quad \Rightarrow \quad \hat{\mathbf{x}} = \sum_{i=1}^{M} y_i \cdot \boldsymbol{\varphi}_i$$

Elkövetett hiba:

$$\varepsilon^2 = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\} \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot \mathbf{C}_{\mathbf{xx}} \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \boldsymbol{\varphi}_i^T \cdot \lambda_i \cdot \boldsymbol{\varphi}_i = \sum_{i=M+1}^{N} \lambda_i$$

$\lambda_1 > \lambda_2$



A mai előadás anyaga legnagyobb részben megtalálható

Szerk. Horváth Gábor

Neurális hálózatok (10.3 és 13. fejezet)

Panem

A tanszék honlapjáról elérhető