Bevezetés a geostatisztikába

Elektronikus jegyzet

Műszaki Földtudományi Kar
Dr. Szabó Norbert Péter
egyetemi adjunktus

Miskolci Egyetem
2012

Tartalomjegyzék

Bevezetés

A geostatisztika összetett tudományág, melynek alapjait egyrészt a valószínűség-számítás és matematikai statisztika, másrészt az alkalmazott földtudományok (geofizika, geológia, kőzetfizika, geokémia, geográfia stb.) képezik. Statisztikai módszerekkel számos a földtudományi gyakorlatban felmerülő elméleti és gyakorlati kérdés megválaszolható, melyek az adatok feldolgozását és értelmezését szolgálják, például:

Milyen gyakran fordul elő egy bizonyos adat az adatrendszerben? Hogyan modellezhető az adatok gyakorisága? Egy bizonyos érték alatt (vagy a felett) hány adat fordul elő? Mi a legjellemzőbb érték a kutatási területen? Milyen mértékben szórnak az adatok? Hogyan kezeljük a hibás adatokat? Hogyan becsülhetjük meg be nem mért tartományok értékeit a többi mérés ismeretében? Mi az adatok együttes előfordulásának a valószínűsége? Milyen kapcsolatban van egy bizonyos adat a többivel? Milyen erős az adatrendszerek közti kapcsolat és milyen arányosság áll fenn? Hogyan írjuk le a változók közötti függvénykapcsolatot? Nagyszámú változó esetén hogyan csökkenthető hatékonyan a probléma „mérete”? Hogyan osztályozhatjuk az adatokat valamely hasonlósági kritérium alapján? Hogyan következtethetünk az adatokból a földtani modell jellemzőire? Mekkora a következtetés hibája és megbízható-e az eredmény?

A jegyzet tananyaga a fenti kérdéseket igyekszik megválaszolni és bevezető ismereteket kíván nyújtani a leendő földtudományi szakembereknek. Az első nyolc fejezetet a klasszikus statisztikai módszerek megalapozására ajánljuk, mely a Miskolci Egyetem Műszaki Földtudományi Karán oktatott BSc tananyagot fedi le. Elsőként az adatrendszerek hisztogrammal történő ábrázolását tárgyaljuk és átvesszük a legfontosabb sűrűségmodelleket. Az adatrendszer legjellemzőbb értékének becslését a számtani átlagszámítás, a mediánképzés, valamint a robusztus leggyakoribb érték módszerén keresztül mutatjuk be. Az adatrendszerben rejlő bizonytalanság jellemzésével foglalkozó fejezet a hibák világába vezeti be az olvasót. A statisztikai becsléseket a maximum likelihood módszeren keresztül mutatjuk be, ahol rámutatunk a mérési adatszám növelésének kedvező hatására. A statisztikai próbák és illeszkedés-vizsgálatok alapjairól áttekintő képet nyújtunk. A kovariancia és a korreláció fogalmán keresztül az adatrendszerek közötti összefüggéseket vizsgáljuk. Ezután a geostatisztika talán leggyakrabban publikált területével, a krigeléssel foglalkozunk, mely a földtani információt igen hatékonyan hasznosítja az adatok interpolációja során. A krigelés a korszerű térképszerkesztő szoftverek alapeleme, mely gyakran nyer alkalmazást az ásványi nyersanyagok felkutatása során. Végül a lineáris és nemlineáris regresszióval zárjuk a jegyzet első részét.

A következő hat fejezetet MSc szakos hallgatóknak ajánljuk, melyek alapvetően a sokváltozós adat-modell kapcsolatok statisztikai tárgyalásával foglalkoznak. A többdimenziós adateloszlások jellemzése céljából bevezetjük a tulajdonságmátrix (adatmátrix) fogalmát és az adatok elrendezésével, skálázásával foglalkozunk. Ezután a többváltozós adatelemzés dimenzió-csökkentő módszereivel, a faktor- és főkomponens analízissel ismerkedünk meg, melyek az adatokban rejlő információ kiemelésével, valamint a nem mérhető háttérváltozók

származtatásával segítik az adatok értelmezését. Az adatok csoportosítására alkalmas klaszterelemző eljárásokat gyakorlati példákkal illusztrálva mutatjuk be. A földtudományi gyakorlat legkorszerűbb, statisztikai elven alapuló adatfeldolgozó eljárásai közé tartoznak az inverziós módszerek. E problémakör feladataival és megoldási módszereivel ismerkedhet meg az olvasó. Elsőként a lineáris inverz feladat megoldásával foglalkozunk, melynél a Gauss-féle legkisebb négyzetek módszerét, ill. annak az adat- és modelltérben súlyozott változatait adjuk meg. Az inverziós módszerek alkalmazása során igen fontos az eredmények statisztikai minősítése. A „modellbecslés pontosságának és megbízhatóságának jellemzése" c. fejezet ennek lehetőségeit tárgyalja. Végül a tananyagot a nemlineáris inverz feladat robusztus megoldására alkalmas véletlenkereső módszerekkel, az ún. globális optimalizációs eljárások (Simulated Annealing és Genetikus Algoritmus) elméletével és alkalmazásával foglalkozunk.

A geostatisztika tantárgy oktatása során fontosnak tartjuk, hogy a hallgatók a statisztikai módszereket és eljárásokat a gyakorlatban is alkalmazni tudják, ezért az egyes fejezetekben bemutatott tananyag gyakorlati példákat is tartalmaz. A tudományos és ipari feladatok megoldása ma már elképzelhetetlen megfelelő szintű programozási ismeretek nélkül. Ennek elősegítése céljából a gyakorlati példák mellett gyakran megtaláljuk a számítógépes megvalósításnak megfelelő programkódot, melyet MathWorks® MATLAB rendszerben futtathatunk. E számítógépes programok nemcsak a mélyebb megértést segítik elő, hanem a későbbi szakmai munka során is felhasználhatók.

Miskolc, 2012. március 29.

A szerző

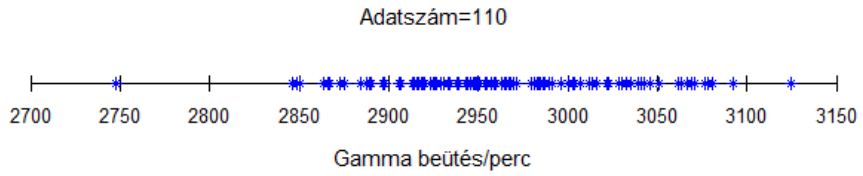
Első rész

BSc tananyag

1. Az adatok eloszlásának modellezése

A megfigyelések kulcsfontosságú szerepet játszanak a földtudományban, ugyanis mérések nélkül aligha lennének igazolhatók a geológiai jelenségekre vonatkozó elméletek és állítások. Megfigyelési eredményeinket **adat**ok formájában rögzítjük, majd a mérés elvégzése után rendszerbe foglaljuk őket. A földtudományi adatrendszerek általában többféle fizikai elven alapuló mérésfajta adataiból épülnek fel, de gyakori az azonos műszerrel történő adatgyűjtés is. Tételezzük fel, hogy lehetőségünk van a földtani objektumon (pl. egy kőzetmintán) egy bizonyos fizikai jellemzőre nézve ismételt méréseket végeznünk. Mivel a műszerzaj, valamint egyéb külső események (pl. környezeti hatások megváltozása) is közrejátszanak, ezért a mérések ismétlése során eltérő mérési eredményeket kapunk. Az így előálló adatrendszert statisztikai nyelven **mintá**nak nevezzük, melyben a véletlennek köszönhetően egyes mintaelemek (adatok) ritkábban, mások pedig gyakrabban fordulnak elő. Az adatok előfordulási gyakoriságát az **adatsűrűség modell**ek jellemzik, melyek ismerete alapvető fontosságú a statisztikai módszerek alkalmazásánál.

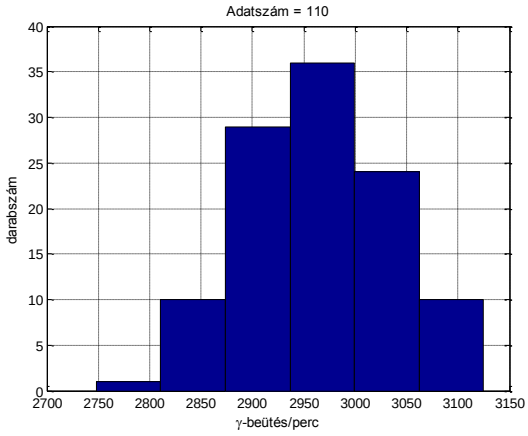
Tekintsünk egy ismételt radioaktív mérésből származó adatsort! Az 1. ábrán az adatokat számegyenesen ábrázolva jelenítettük meg. A terepi mérés ugyanazon földrajzi helyen és műszerrel, viszont különböző időpontokban történt. Megfigyelhetjük, hogy a regisztrálási idő (1 perc) alatt különböző számú γ-részecskét érzékelt a műszer. Ennek oka a fizikában keresendő, mivel az atommagok bomlása során azonos idő alatt kibocsájtott γ-részecskék száma nem állandó. Azt tapasztaljuk, hogy a mért értékek egy jellemző érték körül szórnak. Ez a jellemző érték jelen esetben 2960, ami önmagában nem ad egyértelmű választ arra a kérdésre, hogy „mennyi a γ-intenzitása a kőzetmintának?”. A jellemző érték mellett azt az értéktartományt is meg kell megadnunk, ahol a mérési adatok megtalálhatók. Ráadásul az adatok különböző intervallumokban eltérő számban fordulnak elő, ezért az adatok számegyenesen történő ábrázolása nem igazán praktikus, mivel az adatok gyakoriságára nem kapunk számszerű információt.



1. ábra *γ-sugárzás intenzitás adatok (Mályi, 2002)*

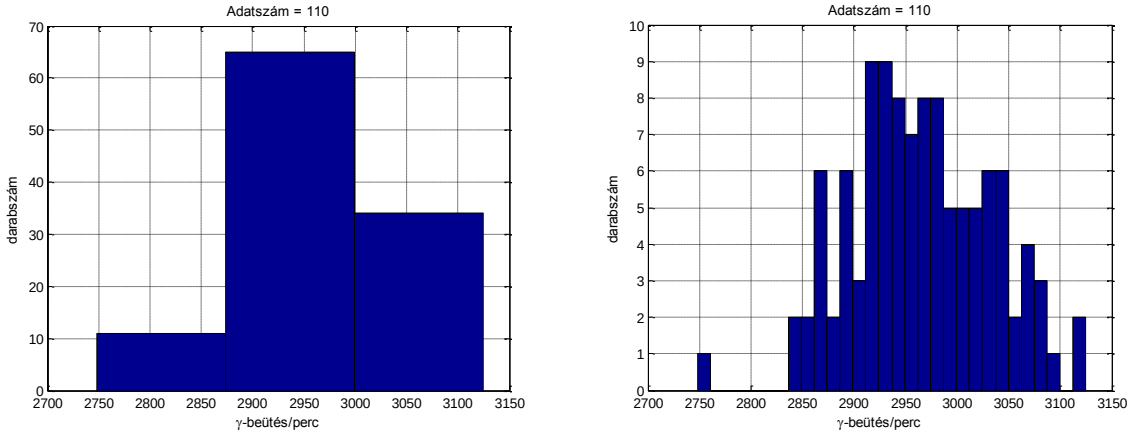
Az adatok táblázatos felsorolása, vagy számegyenesen történő ábrázolása nem elegendő tehát az előfordulási gyakoriságok jellemzésére. Ez különösen igaz a nagyméretű földtudományi adatrendszerek esetén. Az adateloszlás számszerű jellemzésére céljából megfelelőbb az alábbi módszert követnünk. Jelöljük n-nel az összes mérésre vonatkozó adatszámot ill. n_i-vel az i-edik részintervallumba eső adatok számát! Ábrázoljuk az adatok darabszámát h hosszúságú részintervallumonként! Húzzunk az ordináta tengelyen az adott darabszámnak megfelelő magasságban az abszcisszával párhuzamos egyenest az egyes

részintervallumokon! A kapott lépcsős függvényt **hisztogram**nak (tapasztalati sűrűségfüggvénynek) nevezzük. Az 1. ábrán szereplő γ-sugárzás intenzitás adatrendszer hisztogramja a 2. ábrán látható.



2. ábra *γ-sugárzás intenzitás adatok hisztogramja (Mályi, 2002)*

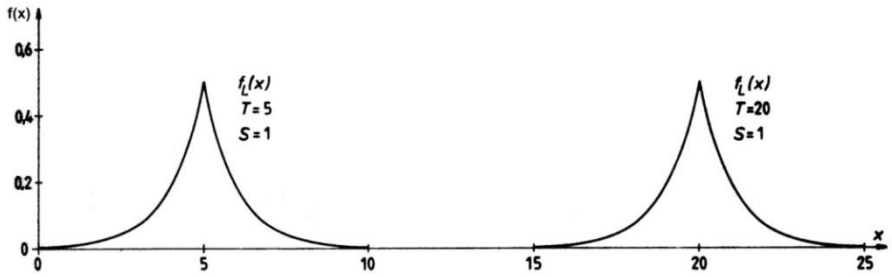
Az optimális intervallumhossz megválasztása lényeges szempont a hisztogram szerkesztésénél. Túl nagy h esetén ugyanis torzul a globális adatsűrűség kép, túl kis h-knál viszont a nagy amplitúdójú zavarok kezelése problematikus. A 3. ábrán a mért értékek tartományát először h=125 (bal oldali ábra) választással három, majd h=12.5-el harminc részintervallumra (jobb oldali ábra) bontottuk. A kapott eredményekkel szemben látható, hogy hat részintervallummal a legoptimálisabb a felosztás (ld. 2. ábra).



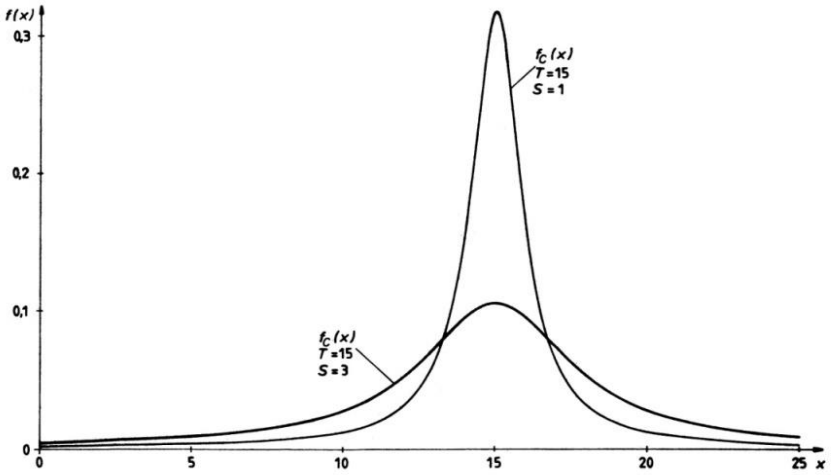
3. ábra *Kedvezőtlen intervallumhossz-választás eredményei*

A hisztogram ordinátáján a darabszám helyett általában az (n_i/n) arányszámot ábrázoljuk, melyet **relatív gyakoriság**nak nevezünk. Ekkor a hisztogram az adatszámtól független lesz és az adatsűrűség jellege sem változik meg. A $100\cdot(n_i/n)$ mennyiség megadja, hogy az összes adat hány százaléka esik az i-edik részintervallumba. További előnyt jelent, hogyha az ordinátán az $n_i/(nh)$ mennyiséget ábrázoljuk. Ekkor a hisztogram oszlopainak összterülete 1 lesz, ahol az i-edik téglalap területe arányos az i-edik részintervallumra eső adatszámmal. Legyen x az abszcissza és y az ordinátatengely jelölése. Illesszünk függvénygörbét a hisztogram (x_i,y_i) adatpárjainak pontjaihoz! A pontokhoz legjobban illeszkedő $f(x,T,S)$

függvényt az adott adateloszlás **sűrűségfüggvény**ének nevezzük. E függvény lesz a hisztogram folytonos megfelelője, melyet növekvő adatszámnál a hisztogram egyre pontosabban közelít. A sűrűségfüggvényt jellemző *T* **helyparaméter** kijelöli a sűrűségfüggvény helyét az *x*-tengelyen, mely egyben a maximális adatsűrűség helye is (ld. 4. ábra). Szimmetrikus eloszlásnál *T* a szimmetriapontot jelöli (aszimmetrikus adateloszlásnál ez nem áll fenn). Az *S* **skálaparaméter** a sűrűségfüggvény szárnyszélességét jellemzi (ld. 5. ábra). Növekvő *S* értékeknél az adatok egyre nagyobb mértékben szórnak, tehát ez a mennyiség az adatok bizonytalanságával áll kapcsolatban.



4. ábra *Laplace eloszlás sűrűségfüggvénye T=5 és T=20 mellett (Steiner nyomán, 1990)*



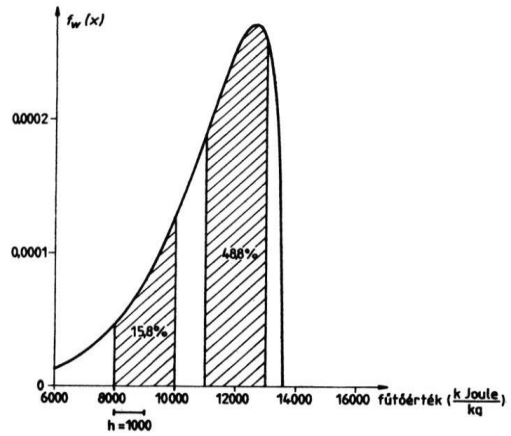
5. ábra *Cauchy eloszlás sűrűségfüggvénye S=1 és S=3 mellett (Steiner nyomán, 1990)*

A hisztogramhoz hasonlóan a sűrűségfüggvény görbe alatti területe is 1, mivel az adat biztos, hogy [-∞,∞] közötti értéket vesz fel. Annak a valószínűsége, hogy az adat a mérés során az [*a*,*b*] intervallumba esik

$$P(a \le x \le b) = \int_a^b f(x)dx.$$

A 6. ábrán pl. leolvasható, hogy 8000-10000 kJ/kg fűtőértékű szén ~16%-ban van jelen az adott mintában, míg 11000-13000 kJ/kg fűtőértékű szén képviseli majdnem a minta mennyiségének a felét.

A sűrűségfüggvényeket kétféle alakban tárgyalja a statisztikai irodalom. A sűrűségfüggvény **standard alak**járól akkor beszélünk, amikor a *T* szimmetriapont zérus és a szélességét szabályzó *S* paraméter egységnyi. A sűrűségfüggvény **általános alak**ját a standard alakból az $x \rightarrow (x-T)/S$ és $f(x) \rightarrow f(x)/S$ transzformációval képezzük. Ekkor a szimmetriapont $x=T$-be kerül, így a sűrűségfüggvény képe *S*-szeresen nyújtott görbe lesz az *x*-tengely mentén.



6. ábra *Szenek fűtőérték szerinti eloszlása (Steiner nyomán, 1990)*

Ezek után nézzük a legnevezetesebb adateloszlások sűrűségfüggvényeinek standard és általános alakját! **Egyenletes eloszlás**ról akkor beszélünk, amikor az adatok az *S*-hosszúságú intervallumban egyenletes valószínűséggel helyezkednek el (pl. lottósorsolás). Az egyenletes eloszlás *f(x)* sűrűségfüggvénye (ld. 7. ábra)

$$f(x)=\begin{cases}\frac{1}{S}, & \text{ha } T-\frac{S}{2}<x<T+\frac{S}{2}\\ 0, & \text{egyébként.}\end{cases}$$

A fenti sűrűségfüggvény teljes számegyenesre vett integrálja 1. A **Gauss (normális) eloszlás** a mérési hibák tipikus (elfogadott) eloszlása, melynek sűrűségfüggvénye egy harang alakú görbét ír le. A Gauss sűrűségfüggvény standard alakja a következő

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

míg általános alakja az alábbi

$$f(x)=\frac{1}{S\sqrt{2\pi}}e^{-\frac{(x-T)^2}{2S^2}}.$$

A Gauss eloszlás sűrűségfüggvényének maximuma a legnagyobb adatsűrűség helyét jelöli ki, ami egyben a függvény *T* szimmetriapontja is (unimodális eloszlás). Egy függvény szimmetrikus, ha a *T* szimmetriahely esetén $f(T-\Delta x)=f(T+\Delta x)$ teljesül. A több maximum-hellyel rendelkező (multimodális) eloszlásoknál ez gyakorlatilag nem áll fenn.

A **Laplace eloszlás**t a Gauss eloszlásnál szélesebb „szárnyú” sűrűségfüggvény jellemzi, mivel az x^2 szerinti gyors csökkenés helyett x szerint (lassabban) csökkennek zérusra a függvényértékek. A Laplace sűrűségfüggvény standard és általános alakja (ld. 7. ábra)

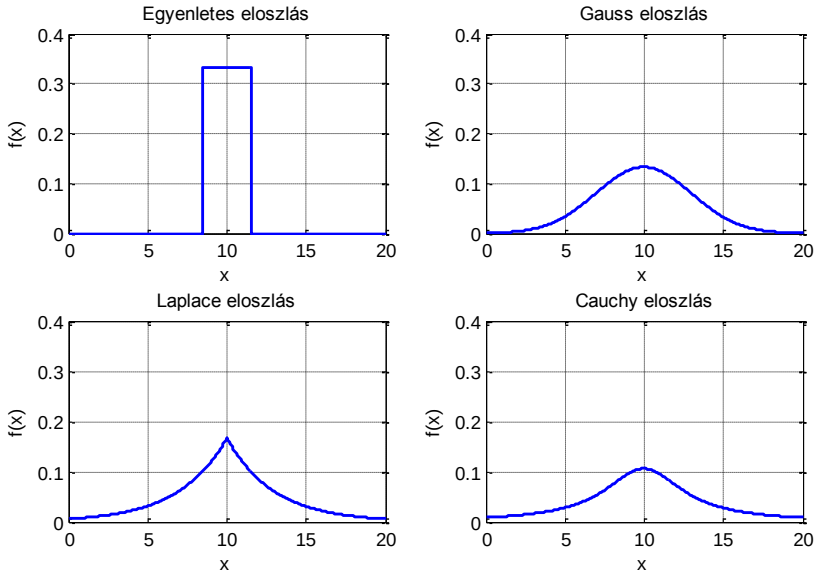
$$f(x)=\frac{1}{2}e^{-|x|}$$

$$f(x)=\frac{1}{2S}e^{-\frac{|x-T|}{S}}.$$

Kedvező tulajdonságai vannak a **Cauchy eloszlás**nak, melynek sűrűségfüggvényét a Laplace eloszláshoz képest kevésbé hegyes csúcs és súlyosabb szárnyak jellemzik. A standard és általános alakú Cauchy sűrűségfüggvények (ld. 7. ábra)

$$f(x)=\frac{1}{\pi}\frac{1}{1+x^2}$$

$$f(x)=\frac{1}{\pi S}\frac{1}{1+\left(\frac{x-T}{S}\right)^2}=\frac{1}{\pi}\frac{S}{S^2+(x-T)^2}.$$



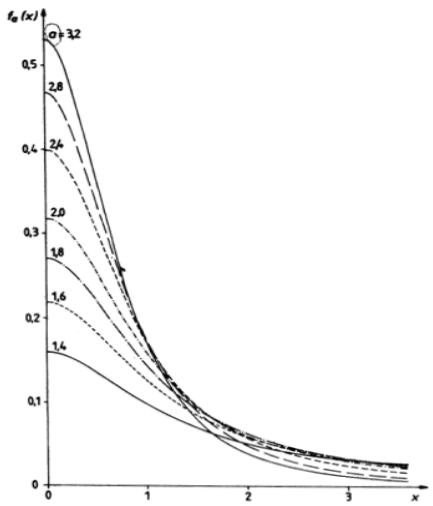
7. ábra *Nevezetes adateloszlások T=10 és S=3 esetén*

A hasonló tulajdonsággal rendelkező sűrűségfüggvényeket modellcsaládokba sorolhatjuk. Az egymástól csak egy (vagy néhány) konstans értékű ún. **típusparaméter**ben különböző sűrűségfüggvények csoportját **szupermodell**nek nevezzük. A szupermodellek lehetnek szimmetrikusak vagy aszimmetrikusak. Például az $f_a(x)$ szimmetrikus szupermodell sűrűségfüggvényének általános alakja $a>1$ típusparaméter mellett

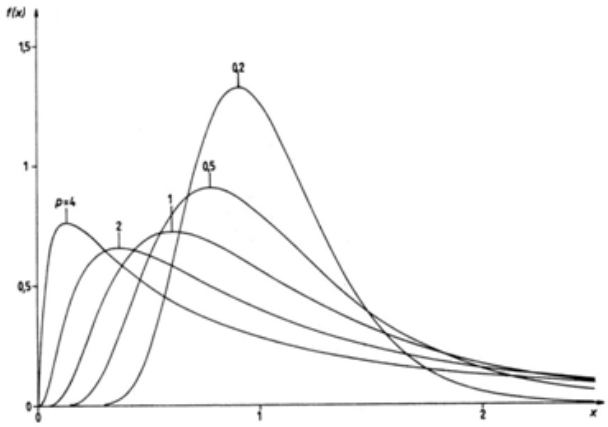
$$f_a(x)=\frac{\Gamma(a/2)}{\sqrt{\pi}\cdot\Gamma((a-1)/2)}\cdot\frac{1}{\left(\sqrt{x^2+1}\right)^a}$$

ahol Γ a gamma függvény. E modellcsalád a=5 értékhez tartozó tagja földtudományi adatrendszerek modellezésénél gyakran alkalmazható (pl. 28. ábra). Az $f_a(x)$ szupermodell néhány tagját a 8. ábrán láthatjuk, ahol a szimmetria miatt csak az x>0-hoz tartozó függvényértékek szerepelnek. Aszimmetrikus szupermodellek közül példaként említhetjük a **lognorm** modellcsaládot, mely jól alkalmazható érctartalommal összefüggő adatok vizsgálata esetén. A sűrűségfüggvény általános alakja p>0 típusparaméter és x>0 mellett

$$f_{ln}(x) = \frac{1}{x\sqrt{\pi p}} e^{-\frac{(\ln x)^2}{p}}.$$



8. ábra *Az $f_a(x)$ modellcsalád néhány eleme (1.4≤a≤3.2) (Steiner nyomán, 1990)*



9. ábra *A lognorm modellcsalád néhány eleme (0.2≤p≤4) (Steiner nyomán, 1990)*

A sűrűségmodell illesztésének követelménye az, hogy a hisztogram összes pontja a lehető legközelebb legyen a sűrűségfüggvény görbéjéhez. Jelöljük x_i-vel az i-edik adatot, $y_i=n_i/(nh)$-vel az i-edik relatív gyakoriságot (azaz a hisztogram pontjait)! Keressük meg a legmegfelelőbb $f(x,T,S)$ kiegyenlítő (analitikus) sűrűségfüggvényt! A feladatot leggyakrabban a **legkisebb négyzetek elve** *(Least SQuares method)* szerint oldjuk meg. Az LSQ

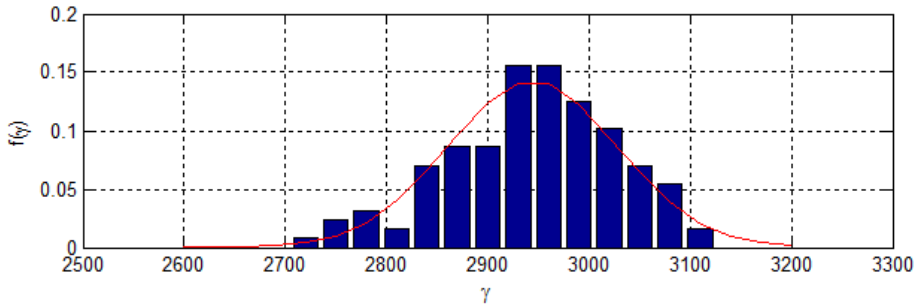
módszer alapján az illeszkedés annál a $[T,S]$ értékpárnál a legjobb, ahol a mérésből meghatározott y_i-k (ahol N a hisztogram pontok száma) és az $f(x_i,T,S)$ modellből számított relatív gyakoriság értékek eltéréseinek négyzetösszege minimális

$$\Phi = \sum_{i=1}^{N} \left(y_i - f\left(x_i, T, S\right)\right)^2 = \min.$$

A Φ célfüggvény minimumát alkalmasan megválasztott szélsőérték-kereső (optimalizációs) eljárással határozzuk meg, melynél megkapjuk az $f(x)$ sűrűségfüggvény optimális T és S paramétereit. (Az LSQ megoldás levezetését az 13. fejezetben találjuk meg).

Példa. Tekintsük az 1. ábrához tartozó adatrendszert! A γ-intenzitás adatok Gauss eloszlását feltételezve meghatároztuk a hely- és skálaparamétert (ld. 10. ábra), mellyel az adatok eloszlását jellemző sűrűségfüggvény a következőnek adódott

$$f(\gamma) = \frac{1}{84\sqrt{2\pi}} e^{-\frac{(\gamma-2944)^2}{14112}}.$$



10. ábra *γ-intenzitás adatok hisztogramja és sűrűségfüggvénye*

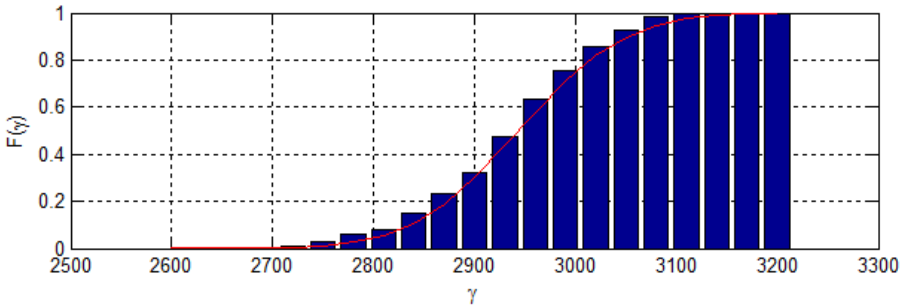
Vizsgáljuk meg, milyen arányban várhatók egy kitüntetett értéknél kisebb adatok a mérés során! A **kumulatív gyakorisági hisztogram** (tapasztalati eloszlásfüggvény) az a lépcsős függvény, mely minden x értéknél megadja hány ennél kisebb adatunk van az adatrendszerben (megj.: a kumulatív szó jelentése „halmozott"). A 11. ábrán ábrázoltuk az 1. ábrához tartozó adatrendszer kumulatív gyakorisági hisztogramját. Megfigyelhető, hogy minden új mérési adat megjelenése esetén a gyakoriság „ugrásszerűen" megnő. Illesszünk függvénygörbét a kumulatív gyakorisági hisztogramhoz! A kapott folytonos görbét **eloszlásfüggvény**nek nevezzük, mely megadja, hogy milyen gyakorisággal vesz fel az x változó kisebb értéket, mint x_0. Az $F(x)$ eloszlásfüggvény helyettesítési értéke az x_0-helyen megegyezik az $f(x)$ sűrűségfüggvény görbe alatti területével az $[-\infty, x_0]$ intervallumon

$$F(x_0) = \int_{-\infty}^{x_0} f(x)dx$$

amiből az adódik, hogy a sűrűségfüggvény az eloszlásfüggvény első deriváltja

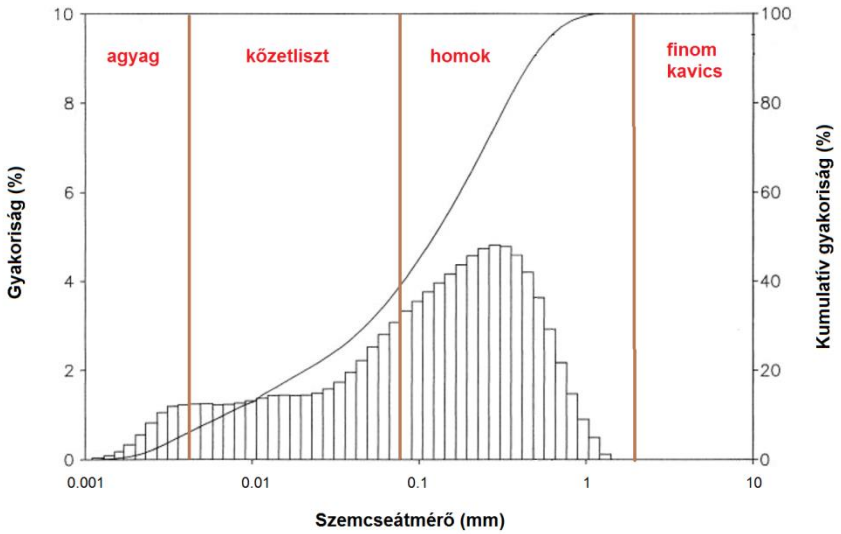
$$\frac{dF(x)}{dx} = f(x).$$

Ábrázoljuk a mályiban mért adatsor eloszlásfüggvényét! A 11. ábrán látható, hogy $F(x)$ monoton növekvő, mivel $F(x_1) \leq F(x_2)$, ha $x_1 < x_2$. Az $f(x)$ függvény egységnyi értékre normált, ezért $F(x)$ értékkészlete $0 \leq F(x) \leq 1$. Továbbá $F(x)$ megadja azt, hogy milyen arányban fordulnak elő x-nél kisebb értékű adatok, ezért $1-F(x)$ azt mutatja meg, hogy milyen arányban vannak annál nagyobbak. A 6. ábrához tartozó kérdés is megválaszolható az eloszlásfüggvény segítségével, ugyanis az $F(b)-F(a)$ különbség megadja, hogy milyen arányban fordulnak elő tetszőleges $[a,b]$ intervallumon adatok. Ha pedig százalékos értelemben szeretnénk tudni azt, hogy adataink hány százaléka kisebb, mint x, akkor az eredmény $100 \cdot F(x)$.



11. ábra *γ-intenzitás adatok tapasztalati és folytonos eloszlásfüggvénye*

Példa. A sűrűség- és eloszlásfüggvények gyakorlati alkalmazására a szemcseeloszlás (röv. szemeloszlás) görbéket említjük meg. A törmelékes üledékes kőzetek szemcséinek mérettartomány szerinti súlyszázalékos megoszlását szemeloszlásnak nevezzük. Az általánosan elfogadott kőzettani kategóriákat szemcseméret tartományokhoz kötjük (kolloid, agyag, kőzetliszt, homok, kavics és tömb). Ennek megállapításán túl az $f(d)$ sűrűségfüggvény kifejezi, hogy a d átmérőjű szemcséből mennyi van a kőzetmintában, míg az $F(d)$ eloszlásfüggvény arról tájékoztat, hogy a d méretnél kisebb szemcse milyen mennyiségben van jelen a mintában. Ezt szitálással lehet pontosan meghatározni. A 12. ábrán egy kőzetminta szemeloszlási görbéire láthatunk példát.



12. ábra *Szemeloszlás görbék*

2. A legjellemzőbb érték meghatározása

Ismeretes, hogy zajmentes mérés a valóságban nem létezik, így az adatrendszerben jelenlévő bizonytalanság rányomja bélyegét a mért mennyiség legjellemzőbb értékének a meghatározására. Többféle becslési eljárás ismeretes a mintát legjobban jellemző érték megadására, melyek közül némelyek igen zajérzékenyek, viszont léteznek olyan módszerek is, melyek kevésbé azok. Ez utóbbiakról azt mondjuk, hogy **robusztus**ak, ami azt jelenti, hogy ezen eljárások tág eloszlástípus-tartományon belül képesek megbízható eredményt szolgáltatni. E tulajdonságot gyakran keverik a **rezisztencia** fogalmával, mely azt fejezi ki, hogy a becslési eljárás az adatrendszerben jelenlévő kiugró értékekre kevésbé érzékenyen reagál, és azok megtartásával is megbízható eredményt ad. E két fogalmat természetesen más statisztikai algoritmusok esetén is használhatjuk, jelentésük nem korlátozik csupán a legjellemzőbb érték becslésének problémájára.

A jellemző érték valószínűség-elméleti bevezetéséhez tekintsünk néhány alapfogalmat! A relatív gyakorisággal már az 1. fejezetben megismerkedtünk, mely az *A* esemény (azaz adat) bekövetkezésének az összes kísérlethez (összes mérési adat) viszonyított arányát (n_A/n) adja meg. Egyre több kísérlet esetén a relatív gyakoriság egy adott számérték körül ingadozik, mely megadja, hogy az *A* esemény az összes kísérletnek várhatóan hányad részében következik be. Ezt a *P(A)* értéket **valószínűség**nek nevezzük. Annak a valószínűsége, hogy a diszkrét valószínűségi változó x_k értéket vesz fel

$$p_k = P(x = x_k)$$

ahol az összes (*n* számú) lehetséges esemény bekövetkezésének valószínűségére fennáll

$$\sum_{k=1}^{n} p_k = 1.$$

A **valószínűségi változó** olyan mennyiség, melynek számértéke valamilyen véletlen esemény kimenetelétől függ. Azt az értéket, amely körül a valószínűségi változó megfigyelt értékeinek (mérési adatok) átlagértéke ingadozik, **várható érték**nek (E_n) nevezzük és az alábbi módon számítjuk ki

$$E_n = \sum_{k=1}^{n} x_k p_k = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

A várható érték tehát mintajellemző mennyiség, melynek legfontosabb tulajdonságai az *x* és *y* valószínűségi változók esetén

$$E(cx) = cE(x) \qquad \text{ahol } c \text{ konstans}$$
$$E(xy) = E(x)E(y) \qquad \text{ahol } x \text{ és } y \text{ független}$$
$$E(x + y) = E(x) + E(y) \qquad \text{ahol } x \text{ és } y \text{ nem független}$$
$$E(ax + b) = aE(x) + b \qquad \text{ahol } a \text{ és } b \text{ konstans.}$$

Fejezzük ki a várható értéket az *f(x)* sűrűségfüggvény ismeretében! A 13. ábra alapján látható, hogy az adat $[x_0,x_0+h]$ intervallumba esésének valószínűsége megegyezik a kijelölt téglalap területével

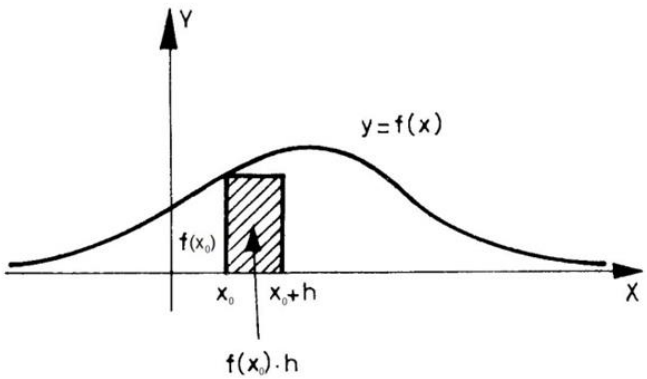
$$P(x_0 \le x < x_0 + h) \cong f(x_0)h$$

$$f(x_0) \cong \frac{P(x_0 \le x < x_0 + h)}{h}.$$

Mivel a fenti *P* valószínűség közelítőleg megegyezik az (n_0/n) relatív gyakorisággal (ahol n_0 és *n* az intervallumba eső és az összes adat száma), ezért a várható érték az *f(x)* sűrűségfüggvénnyel könnyen kifejezhető

$$f(x_0) \cong \frac{n_0}{nh}$$

$$h\sum_{k=1}^{n} x_k f(x_k) = \sum_{k=1}^{n} x_k p_k = E_n.$$



13. ábra *Az $[x_0,x_0+h]$ intervallumba esés valószínűsége*

Vegyük sorra a gyakorlatban leggyakrabban alkalmazott mintajellemzőket! A 4. fejezetben bizonyítani fogjuk, hogy az adatok Gauss eloszlása esetén a várható érték megegyezik a **számtani átlag**gal (mintaátlag)

$$\bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

ahol x_i az *i*-edik adatot jelöli (*n* az adatok száma). Ekkor az adatokat azonos súllyal vesszük figyelembe. Az adatrendszerben jelenlévő zaj véletlen jellege miatt általában az egyes adatokat eltérő hiba terheli. Ennek ismeretében kedvezőbb, ha az adatokat előzetesen megadott (a priori) súlyokkal vesszük figyelembe és **súlyozott átlag**ot számolunk

$$\bar{x}_{n,w} = \frac{\sum_{k=1}^{n} w_k x_k}{\sum_{k=1}^{n} w_k} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

ahol w_i az i-edik adathoz tartozó súly. A **medián** a minta középső elemét adja meg, azaz azt az értéket, melynél nagyobb és kisebb elem ugyanannyi van a mintában

$$\mathrm{med}_n = \begin{cases} x_{(n+1)/2} & \text{ha } n \text{ páratlan} \\ \dfrac{x_{n/2} + x_{(n+2)/2}}{2} & \text{ha } n \text{ páros.} \end{cases}$$

A fenti statisztikai jellemzők a MATLAB (MATrix LABoratory) programrendszerben gyorsan számíthatók. A jegyzetben szereplő példaprogramok ebben a korszerű és hatékony programozási környezetben íródtak. A MATLAB gazdag matematikai eszköztárral rendelkezik, ezek közül a *Statistics Toolbox* nevű csomag számos statisztikai eljárást tartalmaz, melyeket közvetlenül hívhatunk mind a parancsablakon *(Command Window)* keresztül, mind pedig az *Editor*-ban történő programozás során. A MATLAB adatstruktúrájának alaptípusa a komplex számokból álló $N \times M$ méretű mátrix (ahol N a mátrix sorainak és M az oszlopainak a száma). A geostatisztikában ez különösen előnyös, mivel a többdimenziós adatrendszereket általában nagyméretű mátrixokban tároljuk és a matematikai eljárások is ezekkel számolnak. Itt fontosnak tartjuk megemlíteni a **mátrixok közötti szorzás** alapvető szabályát, melyre sokat fogunk még később hivatkozni. Ez kimondja, hogy két mátrix akkor és csak akkor szorozható össze egymással, ha a szorzat bal oldalán álló mátrix oszlopainak a száma megegyezik a szorzat jobb oldalán álló mátrix sorainak a számával. E szorzás eredménye olyan mátrix lesz, mely sorainak száma megegyezik a bal oldali mátrix sorainak számával, oszlopainak száma pedig a jobb oldali mátrix oszlopainak a számával.

Példa. Képezzük az 5×4 méretű $\underline{\underline{A}}$ és 4×3 méretű $\underline{\underline{B}}$ véletlen mátrixokat! Számítsuk ki a $\underline{\underline{C}} = \underline{\underline{A}}\,\underline{\underline{B}}$ és $\underline{\underline{E}} = \underline{\underline{A}}\,\underline{\underline{C}}$ mátrixszorzatokat!

```
>> A=rand(5,4), B=rand(4,3), C=A*B, E=A*C
A =
   0.4389   0.2622   0.2967   0.2625
   0.1111   0.6028   0.3188   0.8010
   0.2581   0.7112   0.4242   0.0292
   0.4087   0.2217   0.5079   0.9289
   0.5949   0.1174   0.0855   0.7303

B =
   0.4886   0.9631   0.4889
   0.5785   0.5468   0.6241
   0.2373   0.5211   0.6791
   0.4588   0.2316   0.3955

C =
   0.5570   0.7814   0.6835
   0.8462   0.7883   0.9638
   0.6516   0.8653   0.8696
   0.8747   0.9947   1.0505
   0.7140   0.8508   0.7111

Error using ==> mtimes
Inner matrix dimensions must agree
```

Látható, hogy az $\underline{\underline{A}}\underline{\underline{B}}$ szorzás megfelelően végrehajtható, viszont az $\underline{\underline{A}}\underline{\underline{C}}$ szorzat esetén a szorzási szabály nem teljesül és hibaüzenetet kapunk. Azonban, ha az $\underline{\underline{A}}$ mátrix sorait és oszlopait felcseréljük (transzponálás), akkor $\underline{\underline{A}}^{\mathrm{T}}\underline{\underline{C}}$ már kiszámítható

```
>> F=A'*C
F =
   1.2889   1.5665   1.4838
   1.3974   1.6160   1.6952
   1.2167   1.4280   1.4732
   2.1770   2.4071   2.4719
```

A MATLAB rendszerben való programozás további elemeinek megismeréséhez ajánljuk *Gisbert (2005)* könyvét és *Szabó (2006)* oktatási segédletét. E rövid kitérő után térjünk vissza a mintajellemzők számítógéppel történő meghatározásához.

Feladat. Számítsuk ki egy 5 elemű adatsorra a fejezetben bemutatott mintajellemző értékeket! Az adatokat tároljuk az *x* oszlopvektorban (5×1 méretű mátrixban)! A mintaátlag és medián számításához a **mean** és a **median** beépített függvényeket is alkalmazhatjuk

```
>> x=[-1 2.2 3.6 4 9.8]'
x =
  -1.0000
   2.2000
   3.6000
   4.0000
   9.8000

>> mean(x)
ans =
   3.7200

>> median(x)
ans =
   3.6000
```

Feladat. A súlyozott átlag meghatározásához definiáljunk egy 5 elemű *w* vektort, melyben az *x* vektor adataihoz rendelt súlyokat (sorrendben) tároljuk! A súlyok összegét a **sum** eljárás adja meg. A vektorok szorzási szabályainak megfelelően az eredmény

```
>> w=[0.5 1 2 1 0.5]'
w =
   0.5000
   1.0000
   2.0000
   1.0000
   0.5000

>> (w'*x)/sum(w)
ans =
   3.5600
```

A fenti adatsor nem tartalmazott kiugró értéket, így a három mintajellemző értéke közel egyformának adódott. Azonban előfordulhatnak **kiugró adat**ok *(outlier)* is a mérés során, melyek forrása lehet műszerhiba, egy elrontott (szakszerűtlen) mérés, adattovábbítás vagy

adatrögzítés stb. Ha a terepen nem ismételhetjük meg a mérést, akkor az adatfeldolgozás során kell ezen adatokat eltávolítani, vagy súlyukat a lehető legkisebbre csökkenteni a becslés során. Utóbbi esetben kapnak vezető szerepet a rezisztens statisztikai eljárások.

Feladat. Példaként tekintsük az előző adatsort (ld. *x* vektor), melynek harmadik elemét cseréljük 3.6-ről 120-ra! Súlyozzuk úgy az adatokat, hogy először emeljük ki (w_3=100), majd ezután nyomjuk el (w_3=0.0001) a kiugró adat hatását! Az alábbi példa azt mutatja, hogy a kiugró adat nagymértékben torzítja a legjellemzőbb érték becslését és csak a második esetben van remény reális mintajellemző megadására

```
>> w=[0.5 1 100 2 1 0.5]'
w =
   0.5000
   1.0000
 100.0000
   2.0000
   1.0000
   0.5000

>> (w'*x)/sum(w)
ans =
  114.4552

>> w=[0.5 1 0.0001 2 1 0.5]'
w =
   0.5000
   1.0000
   0.0001
   2.0000
   1.0000
   0.5000

>> (w'*x)/sum(w)
ans =
   3.5623
```

Feladat. Írjunk saját fejlesztésű programot a számtani átlag, a súlyozott átlag és a medián számítására!

```
clc;
clear all;
x=[-1 2.2 3.6 4 9.8]',
w=[0.5 1 2 1 0.5]',
n=length(x);
atl=0;
for k=1:n
   atl=atl+x(k);
end
atl=atl/n;
sulyatl=0;
seg=0;
for k=1:n
   sulyatl=sulyatl+(w(k)'*x(k));
   seg=seg+w(k);
end
sulyatl=sulyatl/seg;
x=sort(x);
```

```
if mod(n,2)==0
   med=0.5*(x(n/2)+x((n+2)/2));
else
   med=x((n+1)/2);
end
```

Elmentve a fenti programot egy **.m* kiterjesztésű (script) file-ba az alábbi futtatási eredményt kapjuk (mely nagy pontossággal megegyezik a beépített MATLAB eljárások által adott eredményekkel)

```
 x =
  -1.0000
   2.2000
   3.6000
   4.0000
   9.8000

w =
   0.5000
   1.0000
   2.0000
   1.0000
   0.5000

atl =
   3.7200

sulyatl =
   3.5600

med =
   3.6000
```

A következőkben ismerkedjünk meg egy robusztus becslési eljárással, mely ugyancsak mintajellemző értéket szolgáltat. Képezzünk súlyozott átlagot a $\varphi(x)$ szimmetrikus súlyfüggvénnyel! Az adatok zömétől távol eső pontoknak adjunk kis súly értékeket, míg a nagyobb adatsűrűségi helyeken levőkhöz rendeljünk nagyobb súlyokat! A súlyozott átlagérték jelölése legyen M, melynél a $\varphi(x)$ súlyfüggvénynek maximuma van (ld. 14. ábra)

$$M = \frac{\sum_{i=1}^{n} x_i \varphi_i}{\sum_{i=1}^{n} \varphi_i} \quad \text{ahol} \quad \varphi_i = \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M)^2}.$$

A 14. ábrán látható, hogy túl nagy ε értékek esetén a súlyfüggvény minden adathoz közel ugyanakkora súlyt rendel (ld. 1. és 2. eset), ekkor a kieső (kiugró) adatok elrontják az M jellemző érték becslését. Kis ε értékek alkalmazásánál viszont vigyázni kell, nehogy a centrumhoz közeli adatok figyelmen kívül maradjanak (ld. 4. eset). Látható, hogy az ε paraméter megválasztása nagymértékben befolyásolja a súlyfüggvény alakját és a becslés eredményét. Az ε a **dihézió** nevet viseli, mely egy az adatok tömörödésével (kohézióval) fordítottan arányos skálaparaméter jellegű mennyiség.

A fentiek alapján belátható, hogy *M* mintajellemző egy helyparaméter jellegű mennyiség, melyet **leggyakoribb érték**nek vagy MFV-nek *(Most Frequent Value)* nevezünk. Mivel *M* a súlyozott átlagképzés egyenletének mindkét oldalán szerepel, ezért meghatározása iterációs eljárásban lehetséges. Az alábbi iterációs módszer lényege az, hogy az *M*-et és az ε-t együttesen határozzuk meg. Az első (j=1) iterációs lépésben M_1-et a mintaátlaggal vagy a mediánnal helyettesítjük és a dihéziót a mintaterjedelemből az alábbi formula alapján becsüljük (ahol i=1,2,…,n az adatszám indexe)

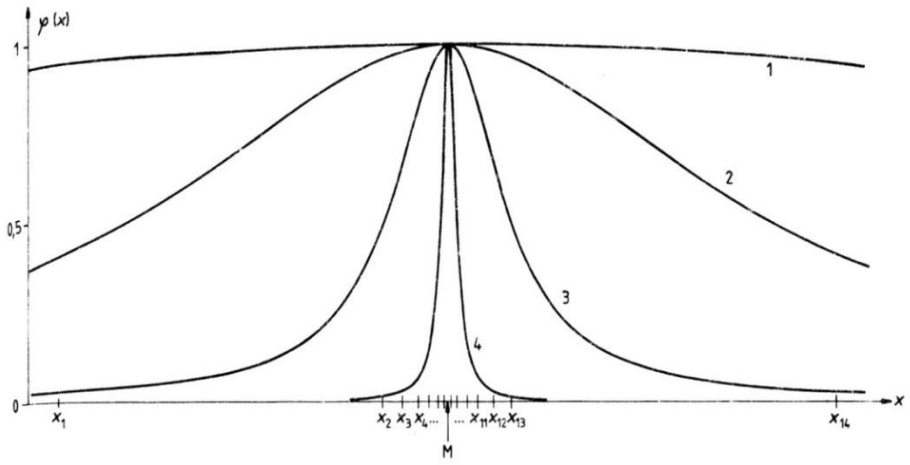
$$\varepsilon_1 \le \frac{\sqrt{3}}{2}\left(\max(x_i) - \min(x_i)\right).$$

Az ezt követő iterációs lépésekben *M*-et és ε-t egymásból származtatjuk

$$\varepsilon_{j+1}^2 = \frac{3\sum_{i=1}^{n} \frac{(x_i - M_j)^2}{\left[\varepsilon_j^2 + (x_i - M_j)^2\right]^2}}{\sum_{i=1}^{n} \frac{1}{\left[\varepsilon_j^2 + (x_i - M_j)^2\right]^2}}$$

$$\updownarrow$$

$$M_{j+1} = \frac{\sum_{i=1}^{n} \frac{\varepsilon_{j+1}^2}{\varepsilon_{j+1}^2 + (x_i - M_j)^2} x_i}{\sum_{i=1}^{n} \frac{\varepsilon_{j+1}^2}{\varepsilon_{j+1}^2 + (x_i - M_j)^2}}.$$

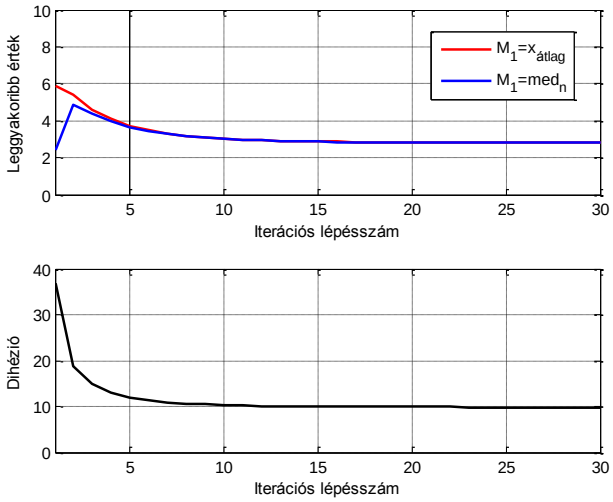


14. ábra *A φ(x) súlyfüggvény különböző ε értékek esetén*
(a görbék melletti számok eseteket, nem pedig dihézió értékeket jelölnek)
(Steiner nyomán, 1990)

Feladat. Tekintsük az -12.5, -6.7, -2, -1.5, 0.1, 2.4, 6.8, 9.8, 15, 23.5, 30 adatsort és számítsuk ki M_n-t és ε-t a fenti iterációs eljárás alapján! (Mivel a leggyakoribb érték mintajellemző mennyiség, ezért a továbbiakban M_n-el fogjuk jelölni).

```
x=[-12.5 -6.7 -2 -1.5 0.1 2.4 6.8 9.8 15 23.5 30];
M1=mean(x);
epsilon1=0.5*sqrt(3)*(max(x)-min(x));
itermax=30;
for j=1:itermax
  szaml=0; szaml2=0;
  nev=0; nev2=0;
  seg=0; seg2=0;
if j==1
    epsilon(j)=epsilon1;
    M(j)=M1;
  else
    for i=1:length(x)
      seg=(x(i)-M(j-1))^2;
      szaml=szaml+3*((seg)/(((epsilon(j-1)^2)+seg)^2));
      nev=nev+(1/((epsilon(j-1)^2)+seg)^2);
    end
    epsilon(j)=sqrt(szaml/nev);
    for i=1:length(x)
      seg2=(epsilon(j-1)^2)/((epsilon(j-1)^2)+((x(i)-M(j-1))^2));
      szaml2=szaml2+(seg2*x(i));
      nev2=nev2+seg2;
    end
    M(j)=szaml2/nev2;
  end
end
```

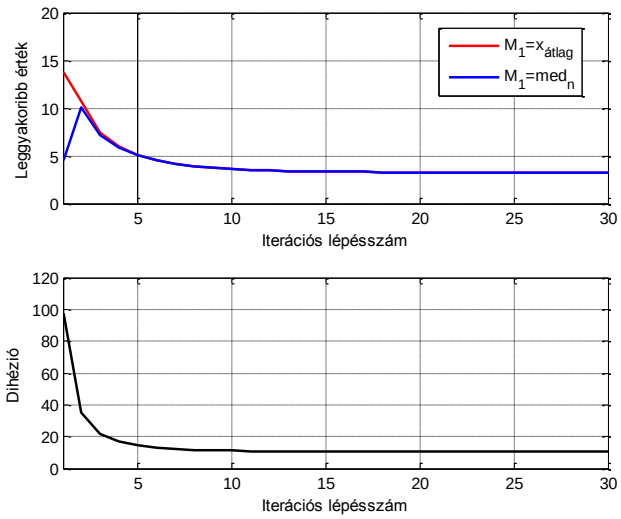
A 15. ábrán látható, hogy M_1 megválasztásától függetlenül M_n a 2.82 értékhez konvergál (a számtani átlag 5.1 és a medián 2.4). Megfigyelhető az is, hogy a dihézió értéke az iterációs lépésszám növekedésével fokozatosan csökken, azaz egyre kisebb súlyt kapnak az M_n-től távolabb elhelyezkedő adatok a becslés során. Vizsgáljuk meg mennyire rezisztens az M_n-t számító eljárás kiugró adatokkal szemben! Adjunk a fenti adatsorhoz még egy elemet, melynek értéke legyen 100! A 16. ábrán megfigyelhető, hogy M_n új értéke (30 iterációs lépés után) 3.25-nek adódott. Tehát míg a számtani átlag 5.1-ről 13.74-re romlott, addig M_n csak ~0.4-et változott. Ez azt jelenti, hogy a leggyakoribb érték számítását csak kismértékben befolyásolja a kiugró adat jelenléte. A medián jobb becslést ad a számtani átlagnál, viszont „kevésbé" rezisztens eljárás, mint a leggyakoribb érték módszer.



15. ábra *Leggyakoribb érték meghatározása (kiugró adat nélkül)*

3. Az adatrendszerben rejlő bizonytalanság jellemzése

Az adatokat terhelő hibákat három fő típusba soroljuk. A rendszeres vagy más néven **szisztematikus hiba** nagysága és előjele azonos körülmények között végzett méréseknél nem változik. Ilyenek a mérőeszköz tökéletlenségéből származó (a működés, ill. hitelesítés pontatlanságai), a mérési módszerek specifikus (pl. graviméterek lineáris járása vagy drift), vagy a környezeti hatásokból eredő (nyomás, hőmérséklet, páratartalom stb.) hibák. Az utóbbi esettől eltekintve a szisztematikus hibák jól korrigálhatók. Más a helyzet a **véletlen hibá**val, mely a mérést befolyásoló véletlen (sztochasztikus) folyamatok együttes következményeként lép fel, és minden egyes mérésnél másképp jelentkezik. Előjele egyaránt lehet negatív és pozitív. Véletlenszerűen fellépő külső hatások, a mérőműszer működési hibája, beállítási- és leolvasási pontatlanságok stb. együttesen képezik a véletlen hibát. Nem küszöbölhető ki teljes mértékben, csak az átlagos hatása becsülhető. A harmadik fő típusra már az 1. fejezetben láttunk példát. Az 1. ábrán látható radioaktív mérésből származó adatokat **statisztikus hiba** terheli, mely nagyszámú egymástól független esemény megfigyelésekor lép fel. A részecske számlálásnál észlelt hiba (statisztikus ingadozás) a mérési adatszám növelésével hatékonyan csökkenthető.



16. ábra *Leggyakoribb érték meghatározása (kiugró adat mellett)*

Az empirikus (tapasztalati) hibajellemzők bevezetése előtt tisztázzunk néhány alapfogalmat! Induljunk ki abból, hogyha ismernénk valamely fizikai mennyiség pontos értékét (x_p), majd egyetlen mérést végeznénk erre a mennyiségre nézve, akkor mérésünk x eredményének a pontos értéktől való eltérése $\delta=|x-x_p|$ lenne, amit valódi hibának nevezhetnénk. Mivel a természetben nincs olyan mennyiség, melynek pontos értéke ismert, ezért x_p-t az előző fejezetben megismert valamely mintajellemző értékkel (számtani átlag, medián vagy leggyakoribb érték) a mérések alapján közelítjük. Ezen mennyiségek az adatsorra nézve bizonyos mértékben eltérnek egymástól (ld. 2. fejezet), ebből következik, hogy a belőlük származtatott hibajellemzők értéke is különbözik.

Az x **adat távolság**át az x_0 jellemző értéktől úgy definiálhatjuk, hogy a δ mennyiséget megadó formulában x_p értékét x_0-ra cseréljük

$$|x - x_0|^p \quad \text{ha} \quad p > 0.$$

A statisztikai minta n számú adatot tartalmaz. Képezzük az x_1, x_2 , ... , x_n **adatrendszer távolság**át az x_0 értéktől! Legegyszerűbb, ha a fenti adattávolságokat összegezzük

$$\sum_{i=1}^{n} |x_i - x_0|^p$$

mely p=2 esetben az eltérések négyzetösszegét, vagy p=1 mellett az abszolút értékek összegét adja meg

$$\sum_{i=1}^{n} |x_i - x_0|^2 \quad \text{vagy} \quad \sum_{i=1}^{n} |x_i - x_0|.$$

Látható, hogy ha a fenti összegben szereplő x_i érték távol van a leggyakrabban előforduló x értékek tartományától (azaz x_0-tól), akkor az $(x_i$-$x_0)$ távolság relatíve nagy lesz, és az összegben a kiugró adat hatása fog dominálni. A nagy eltérések torzító hatását csökkenthetjük alkalmasan választott ε^2-el és szorzással is

$$\prod_{i=1}^{n} \left[\varepsilon^2 + (x_i - x_0)^2\right].$$

Függetlenítsük a jellemző távolságot az n adatszámtól és a kapott távolság mértékegységét harmonizáljuk x mértékegységével! Ekkor az $(x_1 - x_0),(x_2 - x_0),\ldots,(x_n - x_0)$ eltéréseket tartalmazó vektorra felírhatók a gyakorlatban leggyakrabban alkalmazott vektornormák. Általános esetben előáll az **L_p-norma**

$$L_p = \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} |x_i - x_0|^p}$$

mely p=1 esetben az **L_1-normá**nak és p=2 esetben az **L_2-normá**nak felel meg

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_0| \quad \text{és} \quad L_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - x_0)^2}.$$

A gyakorlatban a fenti két normatípust számítjuk a legtöbbször, és a p>2 eseteket már ritkán alkalmazzák. Látni fogjuk majd később azt is, hogy a norma kiválasztása nagymértékben befolyásolja a statisztikai becslés hatékonyságát. Az eltérések szorzatával előálló távolság-jellegű mennyiséget **P_k-normá**nak nevezzük

$$P_k = \varepsilon\left\{\prod_{i=1}^{n}\left[1 + \left(\frac{x_i - x_0}{k\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}.$$

Ábrázoljuk a fenti vektornormák értékeit az x_0 függvényében (ld. 17-18. ábra)! Az így kapott függvények minimumhelyein nevezetes mintajellemző mennyiségek adódnak. Ha ezeket a mintajellemzőket a megfelelő vektornormákban x_0 helyébe írjuk, akkor olyan távolságjellegű mennyiségeket kapunk, melyek az adatoknak a jellemző értéktől való átlagos eltéréseit jellemzik. E távolságok pedig a **bizonytalanság**gal állnak kapcsolatban, ugyanis ha mérési adatainkra nagy távolságot számítunk, az azt jelenti, hogy az adatok a centrumtól távolabb helyezkednek el (jobban szórnak), így a bizonytalanság mértéke x mennyiségre nézve nagyobb szemben azzal az esettel, amikor kis távolságot kapunk (kisebb az adatok szórása). Ha tehát elfogadunk egyetlen adatot jellemző értéknek, akkor a fenti vektor-normákkal számított távolságok mind a **hiba** mértékének tekinthetők. Megjegyezzük, hogy ebben az esetben nem a med_n, E_n vagy M_n mintajellemzők hibájáról, hanem az egyes adatok hibájáról (az adatrendszer bizonytalanságáról) van szó. Az L_1-norma x_0-szerinti minimum-helye a medián (med_n), az L_2-norma x_0-szerinti minimumhelye a számtani átlag (E_n), valamint a P_1-norma (P_k-norma k=1 esetben) x_0-szerinti minimumhelye a leggyakoribb érték (M_n). Az így definiált hibajellemző mennyiségek a **közepes eltérés**

$$d_n = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - med_n\right|$$

a **tapasztalati szórás**

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - E_n\right)^2}$$

és a **tapasztalati határozatlanság**

$$U_n = \varepsilon\left\{\prod_{i=1}^{n}\left[1+\left(\frac{x_i - M_n}{2\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}.$$

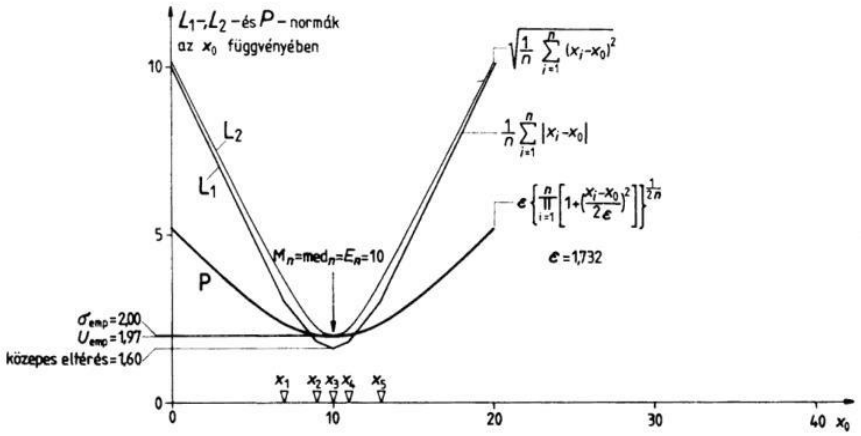
Példa. A fenti empirikus hibajellemzők összehasonlítása céljából számítsuk ki egy hatelemű adatsorra az (x_i-x_0) eltérések L_1-, L_2- és P_2-normáját különböző x_0 értékek esetén! A 17. ábrából kiderül, hogy a normák x_0-szerinti minimumhelyei és a hozzátartozó hiba értékek közel esnek egymáshoz, ha az adatrendszer nem tartalmaz kiugró adatot. A 18. ábrán viszont látható, hogy egyetlen kiugró adat (x_6) jelenléte esetén is a jellemző értékek és hibák jelentősen eltérnek egymástól. Megállapítható, hogy az L_2-norma reagál legérzékenyebben a kiugró adatra, ugyanis a számtani átlagot 10 helyett 15-nél kaptuk és 2-ről 11.3-ra növekedett az empirikus szórás értéke. Az ábrából az is kiderül, hogy az L_1-norma kevésbé érzékeny a kiugró adat megjelenésére. A P_2-norma rezisztens becslést ad, ahol a leggyakoribb érték 10-ről 10.15-re növekedett 1.97 és 2.8 empirikus határozatlanság mellett. Észrevehető, hogy az ε mennyiség értéke is majdnem változatlan (1.73-ről 1.75-re nőtt). E mennyiség megegyezik a 2. fejezetben megismert dihézióval.

Az empirikus hibajellemzőket kis adatszám (n) mellett számítjuk. Abban az esetben, amikor $n\to\infty$, akkor a fenti hibajellemzők elméleti értékéről beszélünk. Például σ_n esetén σ-val fogjuk az elméleti szórást jelölni. Mivel az adateloszlásokat általában Gauss eloszlással

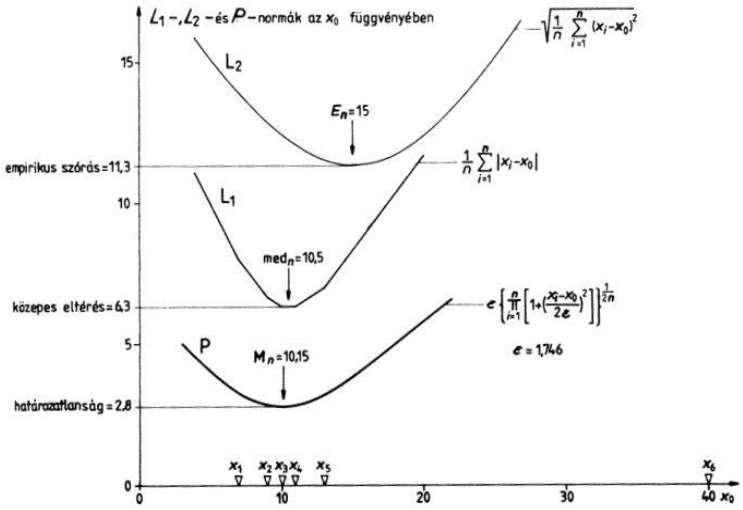
közelítjük, ezért a három bevezetett hibajellemző közül az empirikus szórásnak kiemelt jelentősége van. Felvetődik a kérdés, hogy miért éppen a szórás jellemzi a Gauss eloszlású adatok hibáját a legjobban? Maximum likelihood becsléssel bebizonyítható (ld. 4. fejezet), hogy optimális esetben a Gauss eloszlás sűrűségfüggvényének skálaparamétere (S) megegyezik a szórással (σ) és a helyparamétere (T) pedig a várható értékkel (E)

$$f(x) = \frac{1}{S\sqrt{2\pi}} e^{-\frac{(x-T)^2}{2S^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-E)^2}{2\sigma^2}}.$$

A 19. ábra alapján is belátható, hogy Gauss eloszlás esetén nagy σ értéknél nagy az adatrendszerben rejlő bizonytalanság (az adatok jobban szórnak). Ha pl. x mennyiség mérése során ilyen „széles szárnyú" sűrűségfüggvény áll elő, akkor a mérést nagyobb hiba (több kieső érték) terheli, mint kis σ-nál, ahol az egyedi mérések a várható érték szűkebb környezetébe esnek (azaz pontosabb a mérés).



17. ábra *L_1-, L_2-, P_2-normák az x_0 függvényében (kiugró adat nélkül) (Steiner nyomán, 1990)*



18. ábra *L_1-, L_2-, P_2-normák az x_0 függvényében (kiugró adat mellett) (Steiner nyomán, 1990)*

A **szórásnégyzet (variancia)** a valószínűség-elmélet alapján a valószínűségi változó várható értéktől való (átlagos négyzetes) eltérés mértékét jellemzi. Diszkrét valószínűségi változó esetén a szórásnégyzet

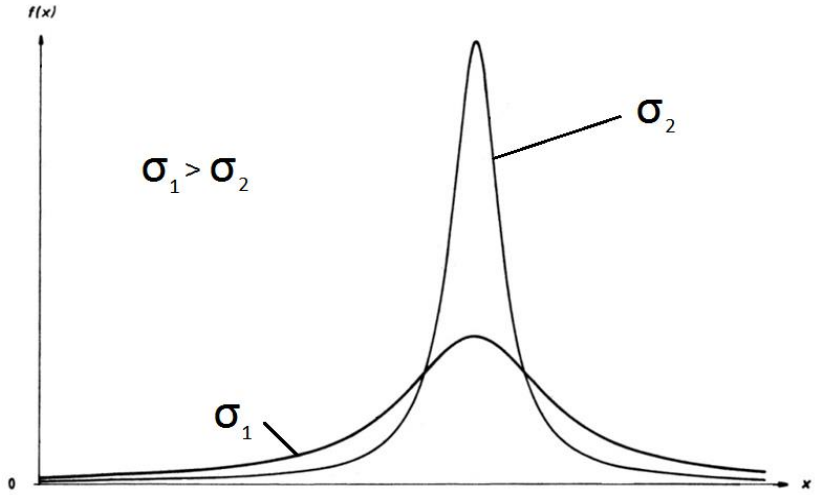
$$\sigma_n^2 = \sum_{k=1}^{n} (x_k - E_n)^2 p_k$$

ahol p_k a $P(x=x_k)$ valószínűséget jelöli. A varianciára vonatkozó gyakori tételek x és y valószínűségi változók esetén

$$\sigma^2(x) = E\left[(x - E(x))^2\right] = E(x^2) - E^2(x)$$
$$\sigma^2(ax + b) = a^2\sigma^2(x) \qquad \text{ahol } a \text{ és } b \text{ állandó}$$
$$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y) \qquad \text{ahol } x \text{ és } y \text{ független.}$$

A **Csebisev-egyenlőtlenség** a valószínűségi változó várható érték körüli szóródására ad felvilágosítást (ahol λ tetszőleges küszöbérték)

$$P(|x - E(x)| \geq \lambda) \leq \frac{\sigma^2(x)}{\lambda^2}.$$



19. ábra *Az adatsűrűség és a hiba viszonya*

Az empirikus szórás (σ_n) **torzított becslés**e az elméleti szórásnak (σ), mivel a (korrigálatlan) σ_n várható értéke nem egyezik meg az elméleti értékkel

$$E(\sigma_n^2) = \frac{n-1}{n}\sigma^2.$$

Képezzük a **korrigált empirikus szórás**t az alábbi módon

$$\sigma_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - E_n)^2}$$

melynek négyzete a korrigálatlan szórásnégyzettel az alábbi kapcsolatban van

$$\sigma_{n-1}^2 = \frac{n}{n-1}\sigma_n^2.$$

A korrigált empirikus szórás formulájának nevezőjében (*n-1*) szerepel. Belátható, hogy a szórás meghatározása (*n-1*) független adatból történik (a számtani közép függ a minta-elemektől és egy adatot kiszámíthatóvá tesz), ezért a normálást ennek megfelelően kell végezni. A korrigált empirikus szórás ebben a formában már **torzítatlan becslés**e az elméleti szórásnak, mivel várható értéke megegyezik a szórás elméleti értékével. Ennek rövid bizonyítása

$$E\left(\sigma_{n-1}^2\right) = E\left(\frac{n}{n-1}\sigma_n^2\right) = \frac{n}{n-1}E\left(\sigma_n^2\right) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Feladat. Írjunk MATLAB programot, amely az *x* adatsorra kiszámítja a korrigált és korrigálatlan empirikus szórást, ill. szórásnégyzetet!

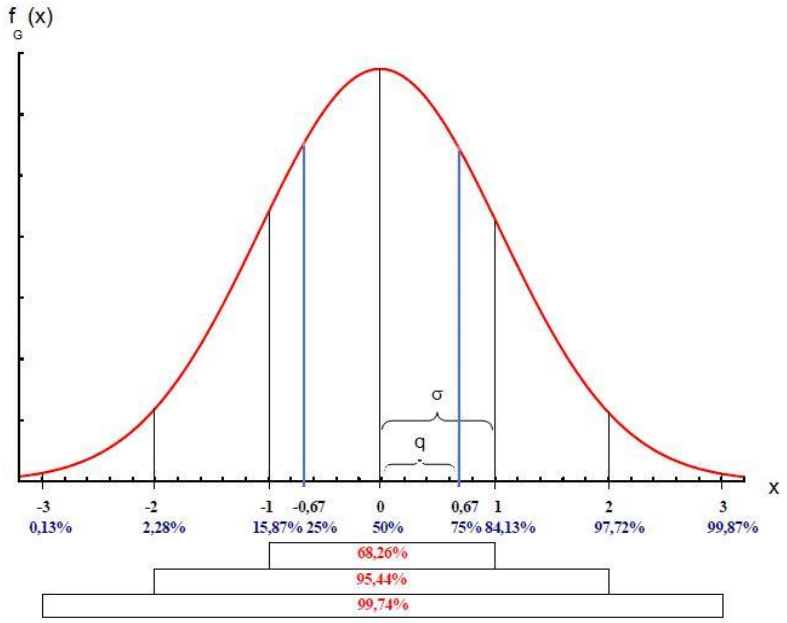
```
clc;
clear all;
x=[-1 2.2 3.6 4 9.8 11.6 15 16.7 17 18.1]',
n=length(x);
atl=0;
for i=1:n
   atl=atl+x(i);
end
atl=atl/n;
szoras=0;
for i=1:n
   szoras=szoras+(x(i)-atl)*(x(i)-atl);
end
Szoras=sqrt(szoras/n),
KorrSzoras=sqrt(szoras/(n-1)),
Variancia=szoras/n,
KorrVariancia=szoras/(n-1),
```

A programfuttatás eredménye:

```
x =
   -1.0000
    2.2000
    3.6000
    4.0000
    9.8000
   11.6000
   15.0000
   16.7000
   17.0000
   18.1000

Szoras =
    6.6708
```

```
KorrSzoras =
   7.0317

Variancia =
   44.5000

KorrVariancia =
   49.4444
```

Gauss eloszlású minta esetén a szórás arról is tájékoztat, hogy az adatok hány százaléka várható a szórás valamilyen többszörösét kitevő hosszúságú intervallumon. Ezt a százalékos előfordulási gyakoriságot **konfidenciaszint**nek, valamint a hozzá tartozó intervallumot **konfidencia-intervallum**nak nevezzük. A 20. ábrán látható, hogy standard normális eloszlás esetén a $0\pm\sigma$ intervallumba az adatok 68.3%-a esik, a $0\pm2\sigma$ intervallumban azok 95.4%-a várható, míg $0\pm3\sigma$ intervallumban majdnem az összes adat (99.7%) beletartozik.



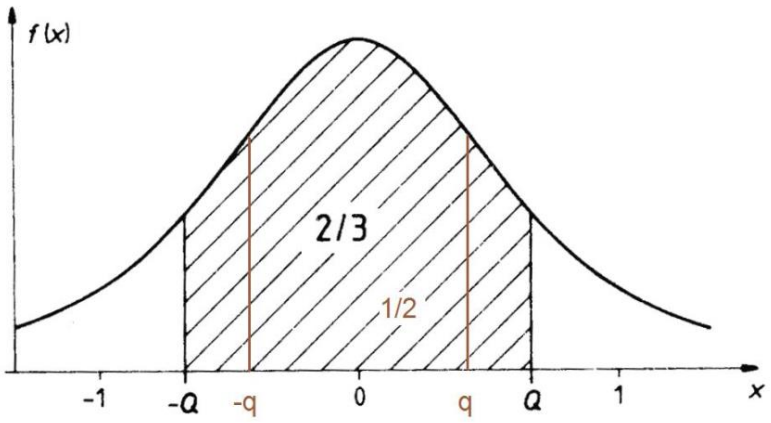
20. ábra *Standard Gauss eloszlás nevezetes konfidencia intervallumai*

A 21. ábrán egy tetszőleges adateloszlást jellemző konfidencia-intervallumok láthatók. Az $[Q,Q]$ interszextilis intervallumban az adatok 2/3 része (66% konfidenciaszint), a $[-q,q]$ interkvartilis intervallumban azok fele (50% konfidenciaszint) várható. Így akárcsak a szórás, hibajellemző mennyiségként viselkedik az **interszextilis félterjedelem** (Q) és az **interkvartilis félterjedelem** (q). Az interkvartilis félterjedelem a 20. ábrán is látható, melynek értéke standard Gauss eloszlás esetén kisebb a szórásnál. A $-q$ értéket alsó kvartlisnek (az adatok 25%-a ennél kisebb), a q értéket felső kvartilisnek (az adatok 25%-a ennél nagyobb) nevezzük. A $-Q$ az alsó szextilis (az adatok 1/6-a ennél kisebb), és Q a felső szextilis (az adatok 1/6-a ennél nagyobb). Általánosan az n-edik percentilis (latinul „per centum" százalék) fejezi ki azt az értéket, melynél az adatok n%-a kisebb. Például az alsó kvartilis a 25%, a medián az 50%, valamint a felső kvartilis a 75% percentilisnek felel meg.

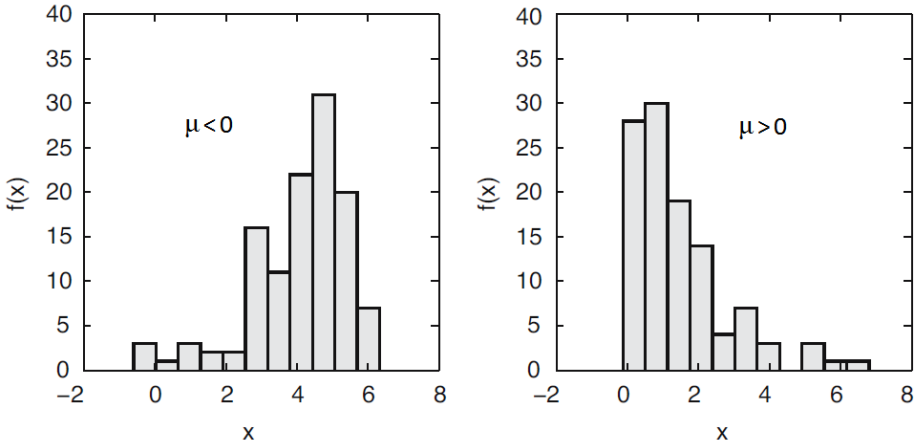
A statisztikai momentumok fontos szerepet játszanak az adateloszlások jellemzésében. Definiáljuk a **k-adik centrális momentum**ot $E((x - E(x))^k)$ formulával, ahol k pozitív egész szám. A szórásnégyzet értelmezhető úgy is, mint a második centrális momentum (k=2). Léteznek magasabb rendű statisztikai momentumok is. A **ferdeség** (*skewness*) a harmadik centrális momentum és a szórás köbének hányadosa, mely a szimmetriától való eltérés mérőszáma

$$\mu = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}.$$

A ferdeség az adateloszlás sűrűségfüggvényének szimmetriájáról tájékoztat, mely az adatrendszerből közvetlenül számítható. Ha μ=0, akkor a sűrűségfüggvény szimmetrikus, ellenkező esetben aszimmetrikus. Ez utóbbi esetben, ha μ<0, akkor a sűrűségfüggvény alakja a szimmetrikushoz képest balra, μ>0 esetén pedig jobbra „nyúlik” el (ld. 22. ábra).



21. ábra *Nevezetes konfidencia-intervallumok*

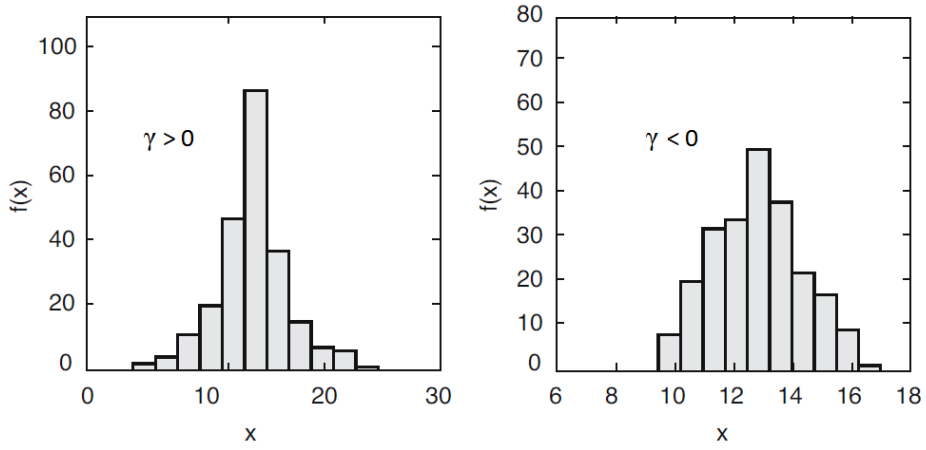


22. ábra *Empirikus sűrűségfüggvények ferdesége*

A **lapultság** (*kurtosis*) a negyedik centrális momentum és a variancia négyzetének hányadosa, mely azt mutatja meg, hogy a vizsgált sűrűségfüggvény alakja „csúcsosság” vonatkozásában hogyan viszonyul a Gauss sűrűségfüggvényhez

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3.$$

A γ=0-hoz tartozó esetekben a sűrűségfüggvény Gauss eloszlású. Ellenkező esetben, a γ>0 értékeknél az a normál eloszlástól csúcsosabb, ill. γ<0 esetén a normál eloszlástól lapultabb lesz (ld. 23. ábra). A 7. ábrán látható, hogy pl. a Laplace eloszlást a Gaussnál csúcsosabb, míg a Cauchy eloszlást annál laposabb sűrűségfüggvény jellemzi.



23. ábra *Empirikus sűrűségfüggvények lapultsága*

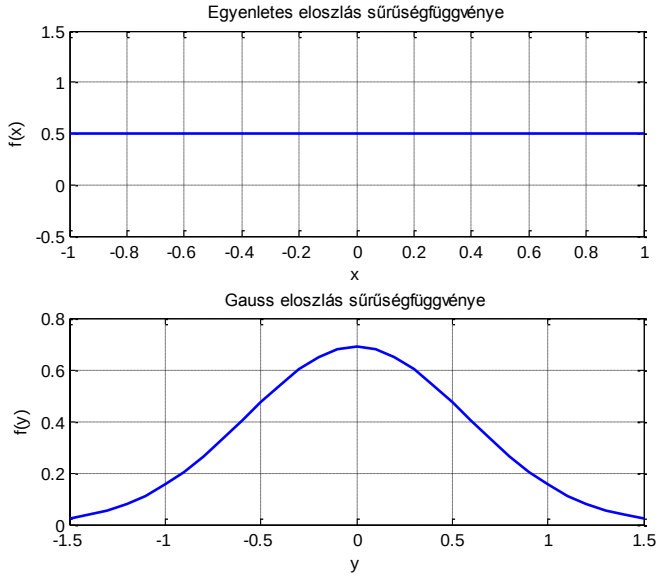
Feladat. Összegezzük eddigi tudásunkat! Vizsgáljuk meg két véletlen adatrendszer empirikus jellemzőit! Az egyenletes eloszlás véletlen adatait az **unifrnd**, míg a Gauss eloszlásúakat **normrnd** paranccsal generáljuk! A szórást és a varianciát az **std** és **var** beépített MATLAB függvények hívásával adjuk meg!

```
x=unifrnd(-1,1,200,1);
y=normrnd(0,1/sqrt(3),200,1);
subplot(2,1,2);
t=normpdf([-1.5:.1:1.5],0,1/sqrt(3));
plot([-1.5:.1:1.5],t);
subplot(2,1,1);
k=unifpdf([-1:.1:1],-1,1);
plot([-1:.1:1],k);
z=[x,y];
szkozep=mean(z),
med=median(z),
empvar=var(z),
szoras=std(z),
terjed=range(z),
lapult=kurtosis(z)-3,
```

A 24. ábrán a két adateloszlás sűrűségfüggvénye látható, melyet a **plot** utasítás segítségével ábrázoltunk. A numerikus eredmények a következők

```
szkozep =
   0.0710   -0.0213

med =
   0.0957   -0.0201

empvar=
   0.3091    0.3457

szoras=
   0.5560    0.5880

terjed=
   1.9703    3.4610

lapult=
  -1.0855    0.6581
```

Látható, hogy a két adatsor átlagértéke és szórása jó közelítéssel megegyezik, viszont két teljesen különböző sűrűségfüggvényről van szó. Az egyenletes eloszlás a Gauss eloszlásnál lapultabb és ebben az esetben kisebb **mintaterjedelem** (a maximális és minimális érték különbsége) jellemzi.



24. ábra *Az egyenletes és Gauss eloszlású minta sűrűségfüggvénye*

A fejezetet a mérési hiba terjedésének törvényével zárjuk. Tételezzük fel, hogy egy *q* mennyiség függ más mennyiségektől pl. *x*-től és *y*-tól, azaz *q=q(x,y)*! A kérdés az, hogy az *x* és *y* mérésével, valamint a Δx ill. Δy mérési (véletlen) hibák ismeretében meg tudjuk-e adni *q*-t (amely nem mérhető mennyiség) és annak hibáját. Alkalmazzunk lineáris közelítést! Az átlagérték a *q=q(x,y)* függvénykapcsolat ismeretében helyettesítéssel könnyen megadható,

mert $\bar{q} = q(\bar{x},\bar{y})$. A Δq **maximális abszolút hiba** (azaz az x és y mennyiségből származtatott q mennyiség hibája) pedig sorfejtésből (a lineáris tagok megtartásával) adódik

$$q = \bar{q} \pm \Delta q \quad \text{ahol} \quad \Delta q = \left|\frac{\partial q}{\partial x}\right| \Delta x + \left|\frac{\partial q}{\partial y}\right| \Delta y.$$

Független valószínűségi változók (ξ_i) esetén érvényes az alábbi összefüggés (ahol c_i konstans)

$$\sigma^2\left(c_1\xi_1 + c_2\xi_2 + \cdots + c_N\xi_N\right) = \sum_{i=1}^{N} c_i^2 \sigma^2\left(\xi_i\right)$$

amelyből adódik, hogy a σ_x^2 és σ_y^2 varianciák ismeretében σ_q^2 számítható

$$\sigma_q^2 = \left|\frac{\partial q}{\partial x}\right|^2 \sigma_x^2 + \left|\frac{\partial q}{\partial y}\right|^2 \sigma_y^2.$$

A fenti összefüggés a Gauss-féle hibaterjedési törvény néven ismert, amelyből a Δq **kvadratikus abszolút hiba** származtatható

$$\Delta q = \sqrt{\left|\frac{\partial q}{\partial x}\right|_{\bar{x},\bar{y}}^2 \Delta x^2 + \left|\frac{\partial q}{\partial y}\right|_{\bar{x},\bar{y}}^2 \Delta y^2}.$$

4. Statisztikai becslések

Tegyük fel, hogy ismerjük az adateloszlást jellemző $f(x)$ sűrűségfüggvény típusát és S skálaparaméterét! Adjuk meg ezen a priori információ ismeretében a sűrűségfüggvény T helyparaméterét! Azt a statisztikai eljárást, mely a minta ismeretében valamely minta-jellemzőt állít elő, **statisztikai becslés**nek nevezzük. Belátható, hogy a fenti probléma esetén azt a T értéket kell becsléssel előállítanunk, melynél az n számú adat bekövetkezése a legnagyobb valószínűséggel megy végbe. A maximális valószínűség feltételezésén alapuló klasszikus becslési eljárást **maximum likelihood módszer**nek nevezzük. A becslés eredménye nagymértékben függ az adatszámtól (nagyobb minta pontosabb becslést tesz lehetővé).

Tekintsünk egy Cauchy-eloszlásból ($S=1$) származó adatsort ($x_1,x_2,\ldots,x_{10}$)! Válasszunk alkalmas Δx-et és képezzük az x_i adathelyeken az $f(x_i)\Delta x$ valószínűségeket! A maximum likelihood elv értelmében a teljes adatsorra képzett valószínűségek szorzat maximumánál adódik az optimális T érték (ld. a 25. ábrán a T=18.6-hoz tartozó esetet). Az optimalizációs feladat célfüggvénye

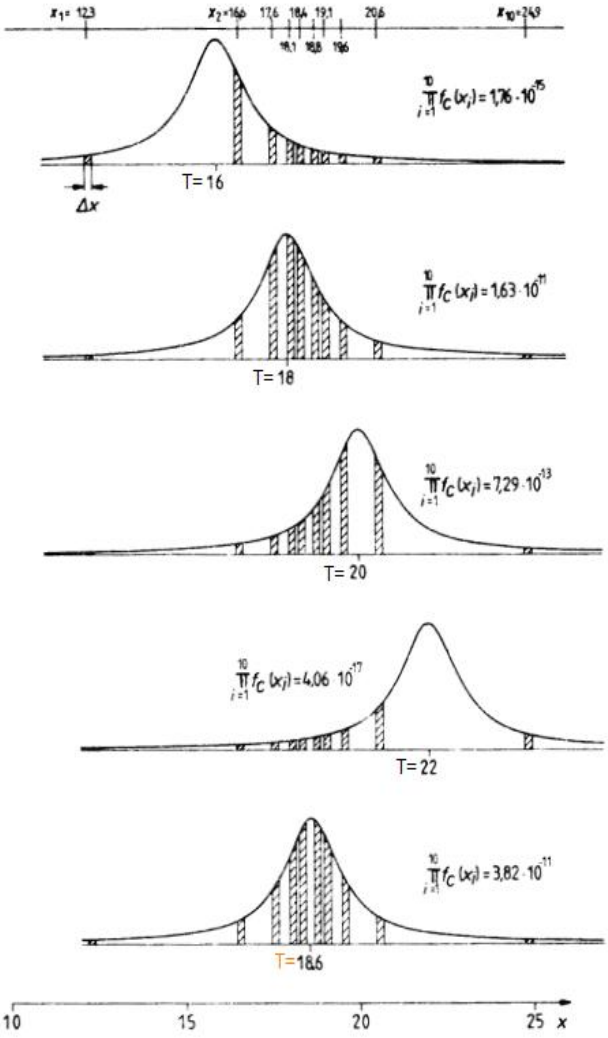
$$L = \prod_{i=1}^{n} f\left(x_i, T\right) = \max$$

melyet **likelihood függvény**nek nevezünk (az $f(x_i)\Delta x$ valószínűségek n-szeres szorzatában megjelenő Δx^n szorzótényezőt elhagyhatjuk, mivel az T-től független konstans). Vegyük az L

függvény természetes alapú logaritmusát és jelöljük L^*-al! Az így kapott célfüggvényt **log-likelihood függvény**nek nevezzük

$$L^* = \sum_{i=1}^{n} \ln\left[f(x_i, T)\right] = \max.$$

Az L^* (ill. L) függvénynek ott van maximuma, ahol annak T paraméter szerinti deriváltja zérus ($dL^*/dt=0$), azaz ahol az $L^*(T)$ függvény érintőjének meredeksége zérus. A fenti feltételből származó egyenlet megoldásával kapjuk a keresett (optimális) T értéket.



25. ábra *Likelihood függvények különböző T értékek esetén (Steiner nyomán,1990)*

Példa. Tekintsük a Gauss eloszlás skála- és helyparaméter értékének maximum likelihood módszerrel történő becslését. Az L likelihood függvény Gauss eloszlás esetén

$$L = \prod_{i=1}^{n} f(x_i, S, T) = \prod_{i=1}^{n} \frac{1}{S\sqrt{2\pi}} e^{-\frac{1}{2S^2}\cdot(x_i - T)^2} = \frac{1}{\left(S\sqrt{2\pi}\right)^n} e^{-\frac{1}{2S^2}\sum_{i=1}^{n}(x_i - T)^2}.$$

Vegyük az L célfüggvény természetes alapú logaritmusát és képezzük az L^* log-likelihood függvényt

$$L^* = \left(-n \ln S - \frac{n}{2} \ln 2\pi \right) - \left(\frac{1}{2S^2} \sum_{i=1}^{n} (x_i - T)^2 \right) = \max.$$

Képezzük az L^* függvény T és S ismeretlenek szerinti parciális deriváltjait és tegyük zérussal egyenlővé őket! A T változó szerinti deriválásból az adódik, hogy optimális esetben a Gauss eloszlás helyparamétere megegyezik a számtani átlaggal

$$\frac{\partial L^*}{\partial T} = \frac{1}{S^2} \sum_{i=1}^{n} (x_i - T) = 0$$
$$(x_1 - T) + (x_2 - T) + \cdots + (x_n - T) = 0$$
$$T = \frac{1}{n} \sum_{i=1}^{n} x_i = E_n.$$

Az L^* függvény S változó szerinti deriválása pedig arra vezet, hogy a Gauss eloszlás skálaparamétere pedig az empirikus szórás (mely egyben a Gauss eloszlás hibajellemző mennyisége, ld. 19. ábra)

$$\frac{\partial L^*}{\partial S} = -\frac{n}{S} + \frac{1}{S^3} \sum_{i=1}^{n} (x_i - T)^2 = 0$$
$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - T)^2} = \sigma_n.$$

A statisztikai becslés eredményének eloszlása növekvő minta elemszám ($n \to \infty$) esetén egy ún. **határeloszlás**hoz tart. Határozzuk meg tetszőleges eloszlású mintából származó számtani átlagok (mintaátlagok) határeloszlásának átlagértékét

$$E\left(\frac{x_1 + \ldots + x_n}{n} \right) = \frac{1}{n} (E(x_1) + \ldots + E(x_n)) = \frac{1}{n} n E(x) = E(x)$$

mely azt jelenti, hogy a minta és a **mintaátlagok átlagértéke** megegyezik

$$E(\bar{x}) = E(x).$$

Most vizsgáljuk meg, hogy mi lesz a mintaátlagok szórásnégyzete

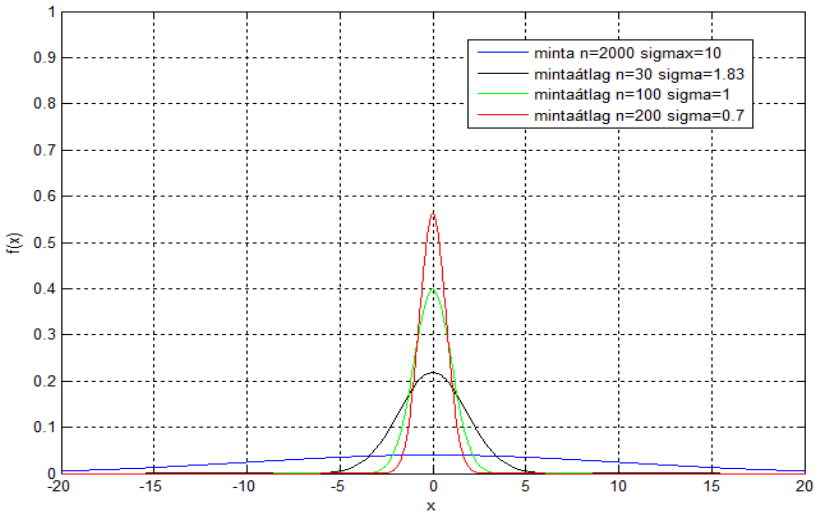
$$\sigma^2\left(\frac{x_1 + \ldots + x_n}{n} \right) = \frac{1}{n^2} (\sigma^2(x_1) + \ldots + \sigma^2(x_n)) = \frac{1}{n^2} n \sigma^2(x) = \frac{\sigma^2(x)}{n}.$$

Ez viszont azt jelenti, hogy a **mintaátlagok szórása** nem egyezik meg a minta szórásával, sőt az adatszám függvényében változik

$$\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{n}}.$$

A fenti összefüggés kimondja, hogy az átlagképzés nagy adatszám (ill. véges szórás) esetén $\sqrt{n}$-el arányos pontosságnövekedést mutat. Ez a **nagy számok törvénye**. A mérések során tehát érdemes nagy adatszámra törekedni, melynek már csak a rendelkezésre álló idő és az anyagi lehetőségek szabhatnak határt. A fenti eredmények összhangban vannak a **centrális határeloszlás tétel**lel is, mert látjuk, hogy az átlagok (mint becslések) eloszlása határesetben (véges szórás esetén) a fenti paraméterekkel jellemzett Gauss-eloszlást közelíti. Megemlítjük, hogy ha egy becslés eloszlása $\sigma_A / \sqrt{n}$ szórású Gauss-eloszlás, akkor σ_A-t **aszimptotikus szórás**nak nevezzük.

Példa. A 26. ábrán egy 2000 elemű Gauss eloszlásból származó minta sűrűség-függvényét (ld. kék görbe) és különböző adatszám mellett előállított mintaátlagok sűrűségfüggvényeit ábrázoltuk. Látható, hogy a minta és a mintaátlagok átlagértéke az adatszámtól függetlenül egybeesik ($E_n=0$). Viszont a mintaátlagok szórása az adatszám növelésével nagymértékben csökken. Ennek megfelelően egyre csúcsosabb és „keskenyebb szárnyú" sűrűségfüggvényeket kapunk.



26. ábra *Mintaátlagok átlagértéke és szórása különböző adatszám esetén*

5. Statisztikai próbák

A statisztikai próba olyan teszt eljárás, mely valamilyen statisztikai feltevésnek az ellenőrzését teszi lehetővé a mintából származó információ alapján. Ezeket az eljárásokat általában két nagy csoportba soroljuk. **Paraméteres próbá**ról akkor beszélünk, ha ismert eloszlástípus esetén döntünk az eloszlás ismeretlen paramétereire tett feltevés elfogadásáról. Ennek több fajtája is van: egymintás (egy adatsor esetén), kétmintás (két adatsor esetén) és többmintás próbák (varianciaanalízis). A másik csoportot a **nemparaméteres próbák** képezik, melyeket ismeretlen eloszlástípus esetén alkalmazhatjuk. Például vizsgálhatjuk azt, hogy a mérési adatokból előállított empirikus sűrűségfüggvény (hisztogram) egy megadott elméleti sűrűségfüggvénnyel leírható-e. Ezt **illeszkedés-vizsgálat**nak nevezzük. Megállapíthatjuk azt is,

hogy két különböző mérési eljárásból származó adatsor független-e egymástól. Ezt **függetlenség-vizsgálat**nak hívjuk. Végül a **homogenitás-vizsgálat** keretében eldönthetjük, hogy két különböző adatsor azonos eloszlást követ-e. A geostatisztikában elvileg mindegyik eset előfordulhat, alkalmazást azonban leggyakrabban az illeszkedés-vizsgálatra látunk, amikor el kell dönteni, hogy egy adatrendszer leírható-e az általunk feltételezett elméleti eloszlással.

A paraméteres próbák széles körben elterjedtek az alkalmazott statisztikában. Sokféle módszer ismeretes, melyek részletezésére nem törekszünk ebben a tananyagban. E helyett inkább egy áttekintő képet szeretnénk nyújtani és tisztázni az ebben a tárgykörben előforduló legfontosabb alapfogalmakat. Ilyen alapvető fogalom a **statisztikai hipotézis**, mely a megfigyelt mennyiség eloszlásának a típusára, vagy az eloszlás paramétereire tett feltevést jelenti. **Nullhipotézis**ről (H_0) akkor beszélünk, amikor az előzetes feltevést igaznak tételezzük fel (ekkor a vizsgált eltérés 0). A nullhipotézissel szembenálló (bármilyen!) más feltételezést **ellenhipotézis**nek (H_1) nevezzük. Vegyünk egy egyszerű példát! Legyen az x mennyiség eloszlása σ szórású Gauss eloszlás (ahol $\bar{x}$ a mintaátlagot jelöli). Tegyük fel, hogy a teljes sokaság várható értéke T_0 és vizsgáljuk meg azt, hogy igaz-e ez az állítás! A feltevésnek megfelelő nullhipotézis és ellenhipotézis

$$H_0: E(x) = T_0$$
$$H_1: E(x) \neq T_0.$$

Mivel nincs a teljes sokaság a birtokunkban, ezért relatíve kisszámú mérési adatra tudjuk csak a nullhipotézis fennállását ellenőrizni. Kérdésünk tehát az, hogy a mintabeli tapasztalat alátámasztja-e a nullhipotézist. A **statisztikai függvény** (röviden statisztika) olyan számítási utasítás, mely egyetlen értéket számít n számú adat alapján. A statisztikai próba feladata megtalálni azt az alkalmas statisztikai függvényt, amelynek eloszlását H_0 fennállása esetén ismerjük. Válasszuk az alábbi statisztikát

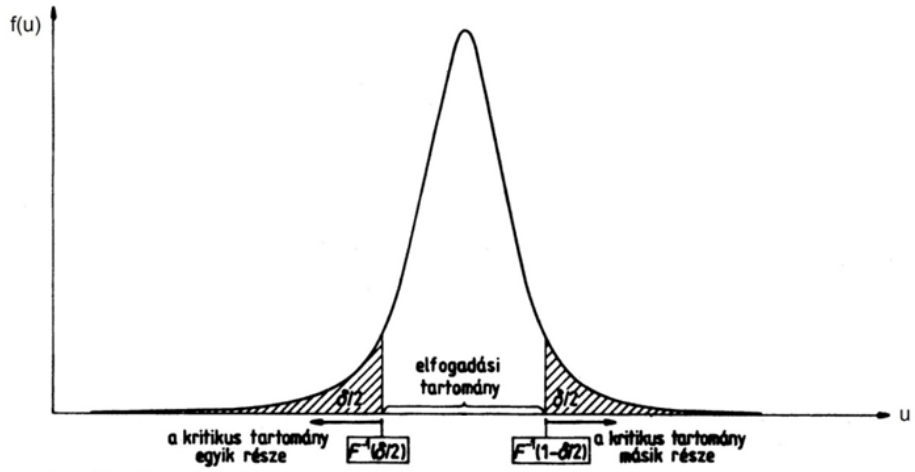
$$u = \frac{\bar{x} - T_0}{\sigma / \sqrt{n}}$$

mely előállítja az u véletlen változót (ahol $\bar{x}$-ot az adatokból számítjuk). Látható, hogy az u mennyiség is Gauss-eloszlást követ, mivel u az $\bar{x}$-nak a standardizáltja (ld. skálázás a 10. fejezetben). Az u érték nagy valószínűséggel a **megbízhatósági (konfidencia) intervallum**ba esik. Ennek a valószínűségét **szignifikancia szint**nek ($1-\delta$) nevezzük

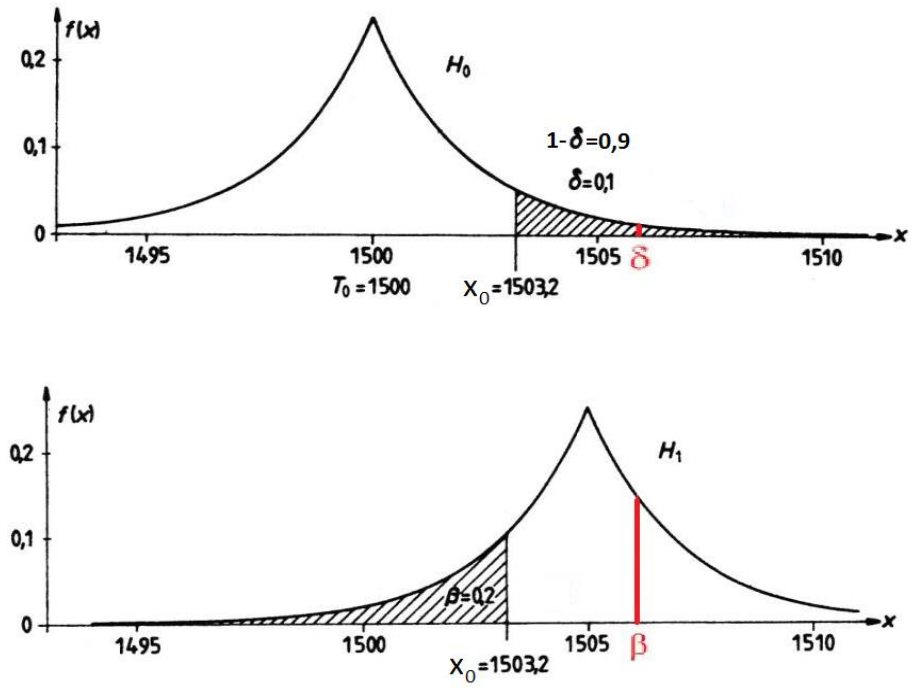
$$P(-u_\delta \leq u \leq u_\delta) = 1 - \delta$$

ahol δ a **kritikus tartomány**ra esés valószínűsége (ld. 27. ábra). **Egymintás u-próba** esetén, ha a H_0 nullhipotézis igaz, akkor az u nagy ($1-\delta$) valószínűséggel esik a megbízhatósági tartományba, azaz kis (δ) valószínűséggel tartozik a kritikus tartományba. Ezért, ha u a megbízhatósági tartományon belül van, akkor H_0 nullhipotézist elfogadjuk. Ellenkező esetben, amikor u a kritikus tartományban van, akkor H_0-t elvetjük. A szignifikancia szintet mi szabjuk meg a statisztikai próba során. Nézzük meg, hogy mekkora hibát követünk el adott szignifikancia szint alkalmazása mellett, ha rosszul döntünk az elfogadásról. Tegyük fel, hogy

u a kritikus tartományba esik és H_0-t elvetjük, annak ellenére, hogy H_0 mégis igaz! Ekkor δ valószínűséggel követünk el hibát, melyet **elsőfajú hibá**nak nevezünk (ld. 28. ábra felső sűrűségfüggény). Olyan eset is lehetséges, amikor H_0 nem igaz, de azt mégis elfogadjuk δ valószínűséggel, ekkor **másodfajú hiba** (β) keletkezik (ld. 28. ábra alsó sűrűségfüggény). A H_0 elfogadása annál nagyobb kockázattal jár, minél nagyobb az (1-δ) valószínűség értéke. Például a 28. ábra alsó részén látható, hogy jelentősen lecsökkentve δ értékét (piros jelzés), a β nagymértékben megnő. Ebből látható, hogy nem célszerű a biztonsági szintet túl magasra állítani.



27. ábra *Egymintás u-próba statisztikájának elméleti sűrűségfüggvénye (Steiner nyomán, 1990)*

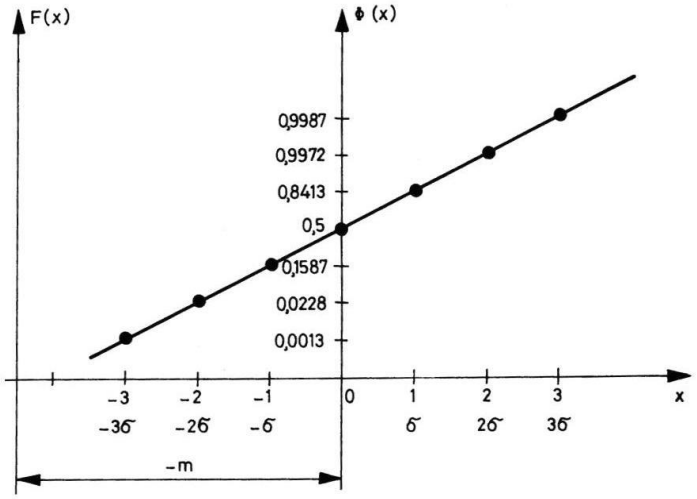


28. ábra *Első-, és másodfajú hibák különböző szignifikancia szinteken (Steiner nyomán, 1990)*

Az illeszkedés-vizsgálatok gyors elvégzésére általában grafikus módszereket alkalmazunk. Segítségükkel el tudjuk dönteni, hogy az adataink megfelelnek-e egy általunk feltételezett eloszlásnak, továbbá ezen eloszlás paramétereire vonatkozó információt is megkaphatjuk. Azonban, ha az adatok nem a kívánt eloszlást követik, akkor a teszt sajnos nem képes megmondani, hogy milyen más eloszlásból származhatnak. A **grafikus normalitás vizsgálat** eldönti, hogy a minta Gauss eloszlásból származik vagy sem. Ennek eszköze a Gauss-papír, melynek abszcisszáján az *x* független változó értékei, az ordinátán pedig a *Φ(x)* standard Gauss eloszlásfüggvény átskálázott értékei szerepelnek (ld. 1. táblázat). Ábrázoljuk az ordinátán a pontokat úgy, hogy *Φ(0)*-tól *Φ(1)* egy távolságegységgel feljebb, *Φ(-1)* egy távolságegységgel lejjebb, *Φ(2)* kettővel feljebb, *Φ(-2)* kettővel lejjebb legyen (ld. 29. ábra)!

1. táblázat

x	Φ(x)
-3	0.0013
-2	0.0228
-1	0.1587
0	0.5000
1	0.8413
2	0.9972
3	0.9987



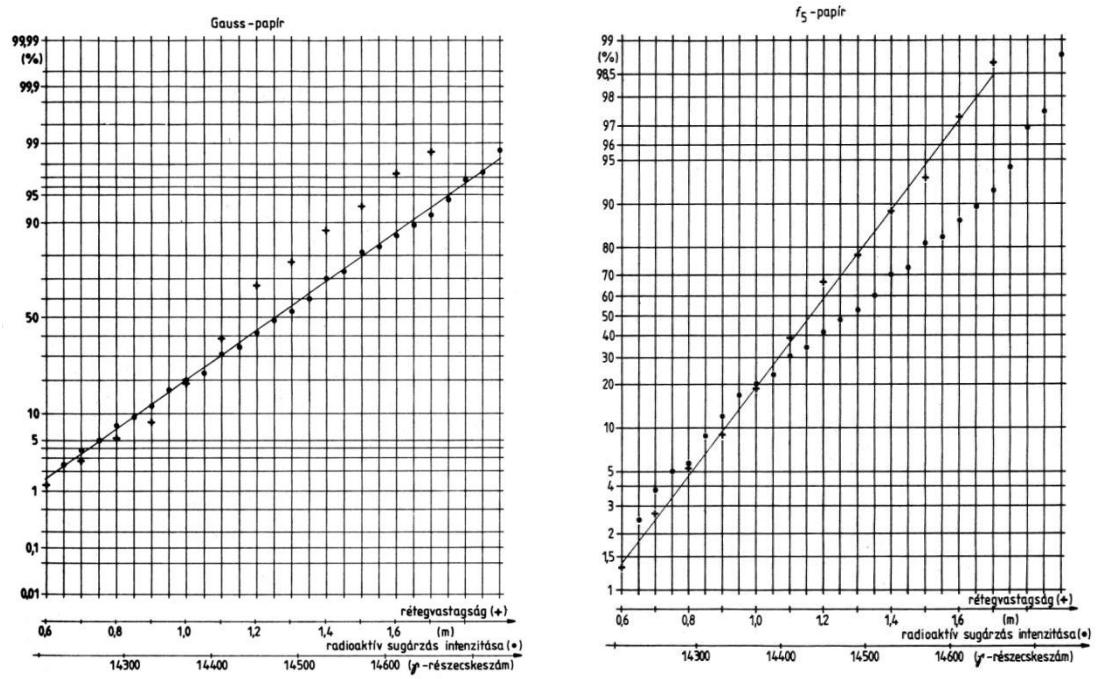
29. ábra *A Gauss-papír koordináta-rendszere*

Képezzük a *Φ(x)* függvényből az *F(x)* függvényt az *x*-tengely menti egységek *σ*-val való szorzásával és az ordinátatengely *-m* értékkel való eltolásával! Ezzel előáll a Gauss-papír koordináta-rendszere

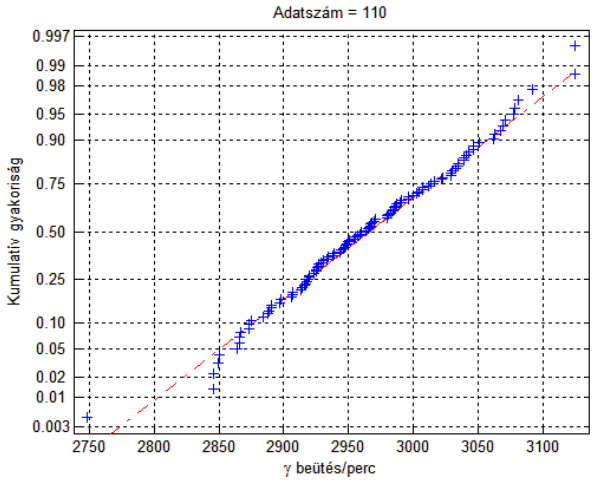
$$F(x)=\Phi\left(\frac{x-m}{\sigma}\right).$$

A fenti transzformációt könnyen ellenőrizhetjük, ha kiszámítjuk az *F(x)* függvényt néhány nevezetes helyen, pl. $F(m)=\Phi(0), F(m\pm\sigma)=\Phi(\pm 1)$. Az így előálló Gauss-papíron az *m* várható értékű és *σ* szórású Gauss eloszlású adatsor képe egyenes, ezért ha x_i (i=1,2,…,n) adataink egy egyenesre esnek, akkor megállapítható, hogy az *x* változó Gauss eloszlású.

Példa. Tekintsünk két adatrendszert *Steiner (1990)* alapján. Az egyik egy kőzetmintán laboratóriumi körülmények között mért ismételt radioaktív mérés eredménye (I. adatsor), a másik egy borsodi térségben található széntelep több fúrásban mért rétegvastagságait tartalmazza (II. adatsor). A szerző először Gauss-papíron ábrázolja a két adatrendszert (ld. 30. ábra). Ebből megállapítható, hogy az I. adatsor jó közelítéssel Gauss eloszlású, azonban a II. adatrendszer attól lényegesen eltér. A II. adatrendszerről további illeszkedés-vizsgálatok után kiderül, hogy az adatok a=5 érték melletti $F_a(x)$ eloszlást követnek (ld. 1. fejezet). Ezt az eredményt mutatja a jobb oldali ábrán az f_5-papír, melynek ordinátáját az $F_5(x)$ eloszlás-függvény értékei szerint skálázzuk. A MATLAB rendszerben a **normplot** függvény alkalmazásával végezhetünk normalitás-vizsgálatot. A 31. ábrán a Mályiban mért adatok (ld. 1. ábra) tesztje látható. A radioaktív adatsor jól közelíthető Gauss eloszlással, melyet megerősítenek a -0.02-es ferdeség, ill. -0.04-es lapultság értékek is.



30. ábra *Az I. (●) és II. (+) adatrendszer illeszkedés-vizsgálata*



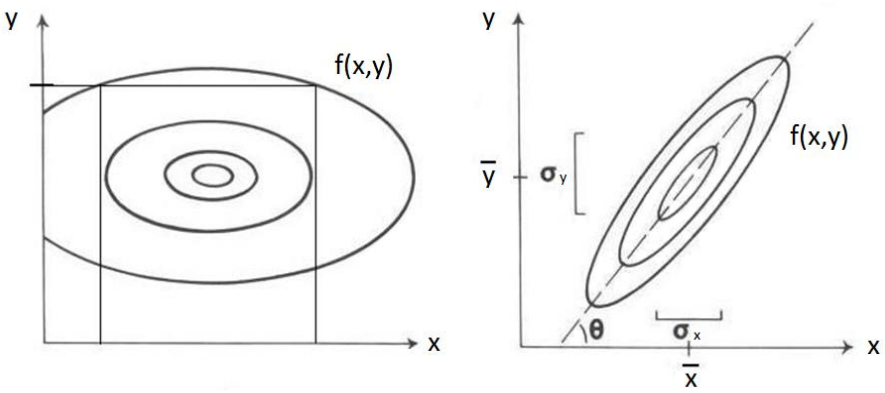
31. ábra *γ-intenzitás adatok normalitás-vizsgálata*

6. Az együttváltozás mérőszámai

Legyen x és y két különböző fizikai mennyiség. Végezzünk méréseket a két (valószínűségi) változóra egymástól független eljárásban! A két változó együttes előfordulásának jellemzésére alkalmas az $f(x_0,y_0)$ **együttes sűrűségfüggvény**, mely megadja, hogy az első mérés milyen valószínűséggel esik x_0, a második pedig y_0 környezetébe. Abban az esetben, amikor az x változó y-tól függetlenül változik, akkor azt mondjuk, hogy az adatok **korrelálatlan**ok és az együttes sűrűségfüggvény az egyedi sűrűségfüggvényekkel kifejezve egyszerűen számítható

$$f(x,y)=f(x)f(y).$$

A 32. ábra bal oldalán látható, hogy függetlenség esetén adott y érték előfordulási valószínűsége kis és nagy x-eknél is nagyjából ugyanaz. Ekkor az x és y mennyiség megváltozása nem követi egymást (nincs trend jellegű változás). **Korrelált** adatok esetén viszont adott nagyságú y értékekhez ugyanolyan valószínűséggel csak bizonyos x-ek tartoznak. A jobb oldali ábrán látható, hogy az x és y változók egyenes arányban állnak egymással és az „együttváltozás” mértéke a Θ szöggel arányos.



32. ábra *Együttes sűrűségfüggvény független (baloldalon) és függő (jobboldalon) változók esetén*

Definiáljuk az együttváltozás mértékét jól kifejező mérőszámot! Osszuk fel az $\bar{x}\bar{y}$ síkot négy negyedre! Ezután képezzük az adatokból az $(x-\bar{x})(y-\bar{y})$ függvényt (mely különböző előjelű a szomszédos síknegyedekben)! Szorozzuk össze ezt a függvényt az $f(x,y)$ sűrűség-függvény értékekkel, majd a részeredményeket adjuk (előjelesen) össze! Az így kapott **kovariancia** (*cov*) nevű mennyiség a két valószínűségi változó együttes változásának legegyszerűbben képzett mérőszáma. A 33. ábrán látható, hogy korrelálatlan változók esetén a kovariancia zérus, mivel a négy síknegyedre azonos nagyságú értékek esnek. Korrelált változók esetén a kovariancia zérustól különböző. Ekkor két eset lehetséges. Ha a kovariancia pozitív, akkor azonos irányú, negatív esetben pedig ellentétes irányú a változás. A kovariancia tapasztalati úton (relatíve kis n adatszám esetén) előálló formulája

$$\mathrm{cov}_n=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}).$$

Valószínűség-elmélet szerinti tárgyalásban a kovariancia definíciója

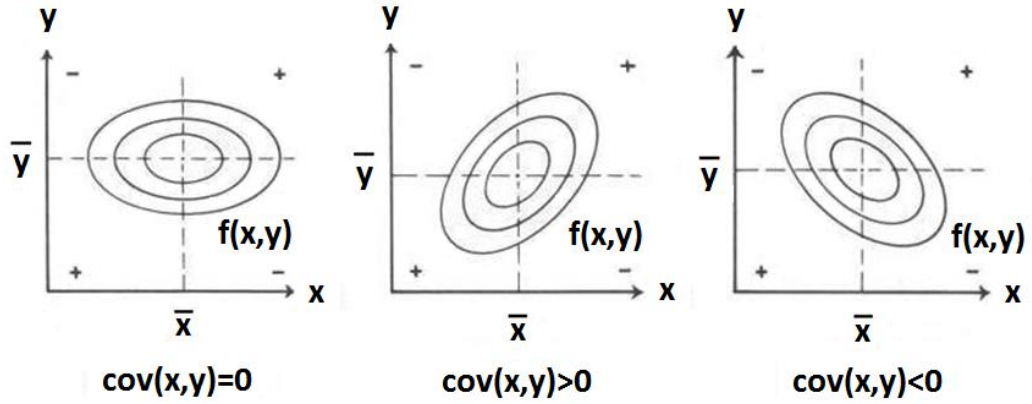
$$\operatorname{cov}(x,y) = E\big((x - E(x))(y - E(y))\big) = E(xy) - E(x)E(y)$$

melynek legfontosabb tulajdonságai

$$\operatorname{cov}(x,x) = \sigma^2(x)$$
$$|\operatorname{cov}(x,y)| \le \sigma(x)\sigma(y)$$
$$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y) + 2\operatorname{cov}(x,y).$$

Látható, hogy $x=y$ esetén a kovariancia megegyezik a varianciával, amely empirikusan is könnyen igazolható

$$\operatorname{cov}(x,x) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(x_k - \bar{x}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \sigma_n^2.$$



33. ábra *A kovariancia előjele és az „együttváltozás” iránya*

Normáljuk a kovarianciát az x és y változók szórásának szorzatával! Az ily módon előálló mennyiséget **korrelációs együttható**nak nevezzük

$$r(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma(x)\sigma(y)}$$

mely a két változó közötti lineáris kapcsolat szorosságát méri. A korrelációs együttható empirikusan kifejezve

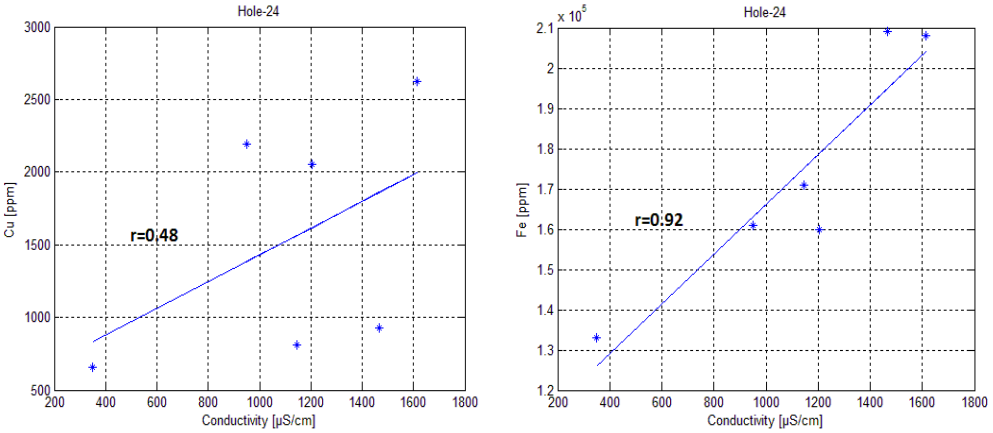
$$r_n = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

ahol r_n egy [-1,1] intervallumba eső szám. Ha $|r_n|=1$, akkor teljes korrelációról, $r_n=0$ esetén pedig lineáris függetlenségről beszélünk. A változók közötti kapcsolat szorosságát (a korreláció erősségét) a 2. táblázatban szereplő gyakorlati intervallumokkal jellemezhetjük. A korrelációs együttható nagysága mellett annak előjele is lényeges, mely (akárcsak a kovariancia előjele) a két mennyiség együttváltozásának az irányáról tájékoztat.

Példa. A 34. ábrán egy finnországi fúrásban mért *Fe* és *Cu* érctartalomnak a fajlagos vezetőképességgel mutatott kapcsolatát és korrelációs együtthatóját láthatjuk. A kisszámú mérési adat ellenére jól látszik, hogy ebben a geológiai környezetben a fajlagos vezetőképesség szorosabb kapcsolatban van a vastartalommal (r_n=0.92), mint a réz-tartalommal (r_n=0.48). Utóbbi esetben az adatok szórása is nagyobb, mely azt igazolja, hogy a korrelációs együttható és a szórás fordított arányban áll egymással. Mivel a korreláció lineáris kapcsolatot feltételez, ezért nagyobb r_n esetén pontosabban illeszthetünk egyenest az adatokra. (Az egyenes illesztés problémáját a 8. fejezetben tárgyaljuk).

2. táblázat

korreláció	**min|r_n|**	**max|r_n|**
gyenge	0	0.4
közepes	0.4	0.7
erős	0.7	1



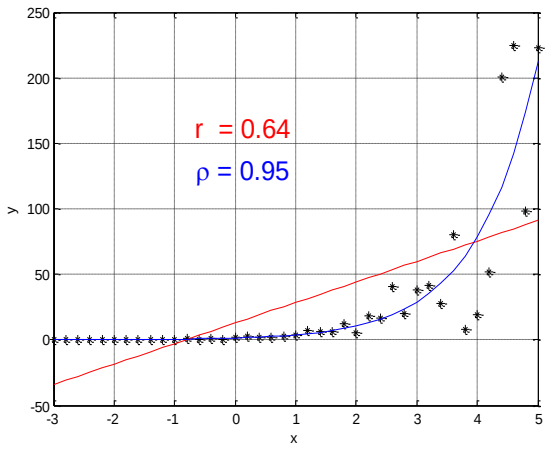
34. ábra *Réz- (baloldalon) ill. vasérctartalom (jobboldalon) kapcsolata a fajlagos vezetőképességgel*

A korrelációs együttható jól jellemzi a lineáris kapcsolat erősségét. Előfordulhat azonban olyan eset, amikor r_n~0 ellenére a két változó szorosan összefügg egymással. Ekkor a változók nemlineáris kapcsolatáról beszélhetünk. Ráadásul ugyanazon r_n érték mellett számos különféle nemlineáris függvénykapcsolat állhat fenn (ekvivalencia probléma). A fenti probléma feloldására egy alkalmasabb korrelációs mennyiséget célszerű alkalmazni. Rendezzük az x_i (i=1,2,…,n) adatokat növekvő sorrendbe! A legkisebb érték kapjon 1-es rangot, a legnagyobb érték pedig n-et. Végezzük el hasonló módon az y_i (i=1,2,…,n) adatok rangsorolását, majd számítsuk ki a kapott rangértékek átlagértékét és szórását! Ezzel előállíthatjuk az x és y változók **rang korrelációs együttható**ját

$$\rho_n = \frac{\sum_{k=1}^{n}\left(\text{rang}(x_k)-\overline{\text{rang}(x)}\right)\left(\text{rang}(y_k)-\overline{\text{rang}(y)}\right)}{\sigma_{\text{rang}(x)}\sigma_{\text{rang}(y)}}$$

mely nemlineáris függvénykapcsolat esetén alkalmasabb r_n-nél a korreláció jellemzésére, mert kevésbé befolyásolják a kiugró értékű adatpárok.

Példa. A 35. ábrán látható exponenciális kapcsolatban álló (zajjal terhelt) adatok hagyományos korrelációs együttható értéke 0.64-nek adódott, mely közepes erősségű kapcsolatot jelöl. Látható, hogy különösen nagy *x* értékek esetén az adatokra illesztett egyenes nem ad optimális közelítést, így a kapott korrelációs együttható érték sem túl nagy. Az exponenciális függvénnyel történő (nemlineáris) közelítés azonban jobban modellezi a két változó kapcsolatát. A rang korrelációs együttható 0.95-re nőtt, mely erős kapcsolatot mutat a két változó között.



35. ábra *Exponenciális függvénykapcsolat korrelációja*

Többváltozós lineáris összefüggések esetén a kovariancia és a korrelációs együttható segítségével a változók kapcsolatát páronként kell megvizsgálnunk. Ez azt jelenti, hogy az összes változó minden paraméter-kombinációjára ki kell számítanunk a fenti mennyiségeket. Tekintsük az $x(x_1,x_2,\ldots,x_n)$ **n-dimenziós valószínűségi (vektor)változó**t (ahol x_i ebben az esetben az *i*-edik fizikai változót jelöli, nem pedig az *i*-edik adatot), és tételezzük fel, hogy ismerjük az egyes változók várható értékeit és szórásait! A **kovariancia mátrix** a változók páronkénti együttváltozásának mértékét adja meg

$$\underline{\underline{\mathrm{COV}}} = \begin{pmatrix} \mathrm{cov}(x_1,x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_n) \\ \mathrm{cov}(x_2,x_1) & \mathrm{cov}(x_2,x_2) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots \\ \mathrm{cov}(x_n,x_1) & \cdots & \cdots & \mathrm{cov}(x_n,x_n) \end{pmatrix}.$$

Mivel azonos változók esetén a kovariancia megegyezik a varianciával, ezért a kovariancia mátrix főátlójában az egyes változók szórásnégyzetei szerepelnek

$$\underline{\underline{\mathrm{COV}}} = \begin{pmatrix} \sigma^2(x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_n) \\ \mathrm{cov}(x_2,x_1) & \sigma^2(x_2) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots \\ \mathrm{cov}(x_n,x_1) & \cdots & \cdots & \sigma^2(x_n) \end{pmatrix}.$$

A kovariancia mátrix négyzetes (a sorok és oszlopok száma megegyezik) és szimmetrikus, mivel $cov(x_i,x_j)=cov(x_j,x_i)$ (ahol i és j a mátrix sor- és oszlopindexe). A kovariancia mátrix elemeit az aktuális változópár szórásainak szorzatával osztva kapjuk a **korrelációs mátrix**ot, mely a változók kapcsolatának az erősségét adja meg

$$\underline{\underline{R}} = \begin{pmatrix} 1 & r(x_1,x_2) & \cdots & r(x_1,x_n) \\ r(x_2,x_1) & 1 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots \\ r(x_n,x_1) & \cdots & \cdots & 1 \end{pmatrix}.$$

Az R mátrix szimmetrikus, azaz $r(x_i,x_j)=r(x_j,x_i)$. Az R i-edik főátlóbeli eleme az x_i változó önmagával képzett korrelációs együtthatója, azaz $r(x_i,x_i)=\sigma^2(x_i)/\sigma(x_i)\sigma(x_i)=1$. A MATLAB rendszerben a **cov** és **corrcoef** beépített függvénnyel számíthatjuk ki a fenti mennyiségeket.

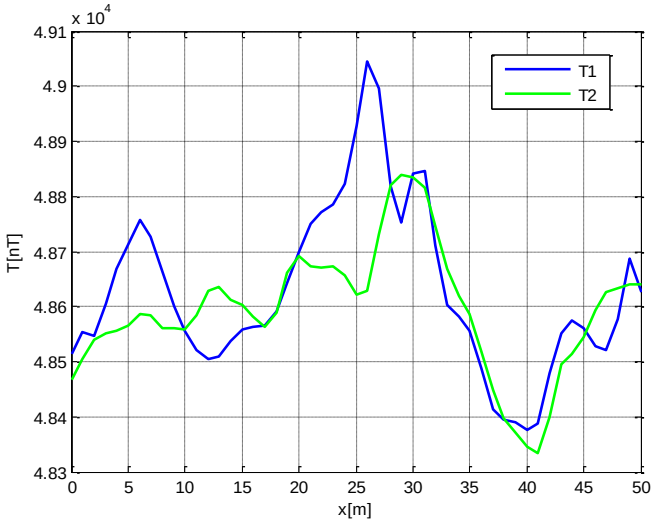
Feladat. Írjunk saját fejlesztésű programot tetszőleges x és y adatsor kovariancia és korrelációs mátrixának számítására!

```
x=[1 2 3 4 5]',
y=[-1 3 5 6 9.4]',
N=length(x);
xatls=0;
for i=1:N
   xatls=xatls+x(i);
end
xatl=xatls/N;
yatls=0;
for i=1:N
   yatls=yatls+y(i);
end
yatl=yatls/N;
s1=0;
for k=1:N
   s1=s1+((x(k)-xatl)^2);
end
kov11=s1/(N-1);
szorasx=sqrt(kov11),
s2=0;
for k=1:N
   s2=s2+((x(k)-xatl)*(y(k)-yatl));
end
kov12=s2/(N-1);
kov21=s2/(N-1);
s3=0;
for k=1:N
   s3=s3+((y(k)-yatl)^2);
end
kov22=s3/(N-1);
szorasy=sqrt(kov22),
kovariancia=[kov11 kov12;kov21 kov22],
korr11=kov11/(szorasx*szorasx);
korr12=kov12/(szorasx*szorasy);
korr21=korr12;
korr22=kov22/(szorasy*szorasy);
korrelacio=[korr11 korr12;korr21 korr22],
```

Mivel két változót vizsgálunk, ezért a kovariancia és a korrelációs mátrix 2×2-es méretű. A futtatási eredmények azt mutatják, hogy az *x* és *y* adatsor igen szoros kapcsolatban áll egymással, és az egyedi szórások is kicsik

```
x =
     1
     2
     3
     4
     5

y =
   -1.0000
    3.0000
    5.0000
    6.0000
    9.4000

szorasx =
    1.5811

szorasy =
    3.8408

kovariancia =
    2.5000    5.9500
    5.9500   14.7520

korrelacio =
    1.0000    0.9798
    0.9798    1.0000
```

Példa. A Borsodban található nyékládházi hulladéklerakó felett 2004-ben végzett földmágneses mérések a felszín alatt elhelyezkedő (mágnesezhető) fémtárgyak helyének megkeresését szolgálták. A proton-precessziós magnetométerrel mért *T[nanoTesla]* totális mágneses térerősség adatokat párhuzamos vonalak mentén mértük, melynek anomáliái jól jelzik a felszín alatti mágnesezhető hatók jelenlétét. A 36. ábra két szomszédos (Δy=3m), egymással párhuzamos szelvény mentén mért adatrendszert (*T1* és *T2*) mutat. Számítsuk ki a korrelációs mátrixot és jellemezzük a két adatsor közötti kapcsolatot!



36. ábra *Mágneses mérési adatok szelvényei (Nyékládháza, 2004)*

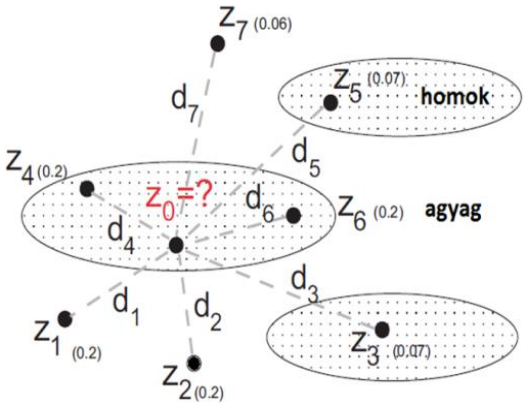
A futási eredmény a következő

```
szorasT1 =
  151.9839

szorasT2 =
  114.1975

kovariancia =
   23099   12635
   12635   13041

korrelacio =
   1.0000   0.7280
   0.7280   1.0000
```

A fenti eredmények és a 36. ábra alapján megállapítható, hogy a két adatsor nem független, mivel a maximumok és minimumok azonos előjellel követik egymást. A korrelációs együttható értéke 0.73, ami a 2. táblázat szerinti besorolás alapján (éppen csak) erős kapcsolatot jelöl. Ennek oka a fizikában keresendő. A mágneses térerősség a hatótól való távolsággal rohamosan csökken, emiatt a *T1* szelvény alatt esetlegesen elhelyezkedő mágnesezhető tárgyak hatása a 3m-el odébb található *T2* szelvény mentén már csak gyengébben jelentkezik. Ez fordítva is igaz, a *T2* szelvény alatt vagy közelében elhelyezkedő felszín alatti hatók már lehet, hogy túl messze vannak a *T1* szelvényhez képest és hatásuk csak kis amplitúdóval jelentkezik (vagy egyáltalán nem jelenik meg) a *T1* adatsorban. Emellett a mérések bizonyos mértékű zajt mindig tartalmaznak, így ezek együttesen azt eredményezik, hogy a térbeli korreláció a két szelvény között (még ilyen rövid állomástávolság mellett) nem teljes.

7. A krigelés

Az interpolációs eljárások fontos szerepet játszanak a földtudományi gyakorlatban, legyen szó valamely fizikai (pl. mágneses térerősség, nehézségi gyorsulás, fajlagos ellenállás), kőzetfizikai (pl. porozitás, víztelítettség, agyagtartalom, érctartalom, hidraulikus permeabilitás) vagy geometriai (topográfiai adatok, rétegvastagság) mennyiség becsléséről egy adott területen. Tekintsük a 37. ábrát, ahol Z mennyiség ismert a Z_i (i=1,2,…,7) mérési pontokban! Az interpoláció feladata az ismeretlen Z_0 érték megadása, azaz a be nem mért térrész adott tulajdonságának meghatározása.



37. ábra *Az interpoláció alapproblémája*

A matematikából ismert hagyományos interpolációs eljárásokkal könnyen meghatározható Z_0 értéke úgy, hogy a Z_i értékeket a Z_0-tól való d_i távolság szerint súlyozzuk, majd a kapott w_i súlyokkal átlagértéket számolunk

$$Z_0 = \sum_{i=1}^{n} w_i Z_i$$

ahol az *i*-edik mért adathoz tartozó súly

$$w_i = \frac{\frac{1}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}.$$

A súlyozás eredményeként azok a Z_i mérések, melyek a Z_0-tól távolabb helyezkedik el, kisebb súlyt kapnak, mint a közelebb levők. A 37. ábrán látható, hogy a hagyományos súlyozás esetén minden olyan pontnak, mely Z_0-tól egyenlő távolságra van, azonos súlyt adunk (pl. Z_1, Z_2, Z_4 és Z_6 esetén w=0.2). Azonban látjuk azt is, hogy pl. a Z_4 és Z_6 pontoknak nagyobb súlyt kellene adnunk, mint Z_1 és Z_2-nek, mivel azok Z_0-al azonos földtani egységbe (homok) tartoznak. A földtani problémák jellemzője, hogy a vizsgált fizikai mennyiségek térbeli korrelációt mutatnak (ld. pl. 36. ábra adatrendszere). Ez olyan fontos a priori információ, melyet előnyös lenne figyelembe vennünk az interpoláció során. A kérdés az, hogy létezik-e olyan eljárás, ami érvényesíteni tudja ezt a földtani (többlet) információt és sokkal valóságosabb megoldást képes előállítani, mint a hagyományos (csak a pontok közötti távolságra alapozott) interpolációs eljárások. A krigelés olyan módszer, mely képes a térbeli összefüggések figyelembe vételére, ezért igen népszerű a geostatisztika gyakorlatában.

Definiáljuk azt a mennyiséget, mely a *h* eltolás függvényében megadja a *Z* mennyiség értékkülönbség négyzetösszegének a felét

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z(r_i) - Z(r_i + h)]^2$$

ahol *h* a két vizsgált pont távolsága, *n(h)* az egymástól *h* távolságban lévő pontpárok száma, $Z(r_i)$ a vizsgált mennyiség értéke az r_i helyzetű pontban, $Z(r_i+h)$ a vizsgált mennyiség értéke az r_i ponttól *h* távolságra elhelyezkedő pontban (r_i az *i*-edik pont helyzete). Az így kapott $\gamma(h)$ görbét **félvariogram**-nak (röv. variogram) nevezzük. Belátható, hogy amikor két pont helyet cserél a térben, akkor a $Z(r_i)$-$Z(r_i+h)$ különbség -1-szeres értékre vált. Ezért a különbségek átlagértéke zérus. Mivel az egyedi különbségek az átlagértéktől való eltérésként értelmezhetők, így a variogram jelentése nem más, mint az eltérések empirikus szórásnégyzetének a fele

$$\gamma(h) = \frac{1}{2} \mathrm{VAR}[Z(r) - Z(r + h)].$$

A kovariancia és a variogram változása a *h* távolság függvényében a 38. ábrán látható. Látható, hogy a két mennyiség fordítottan arányos egymással. Mivel a távolság

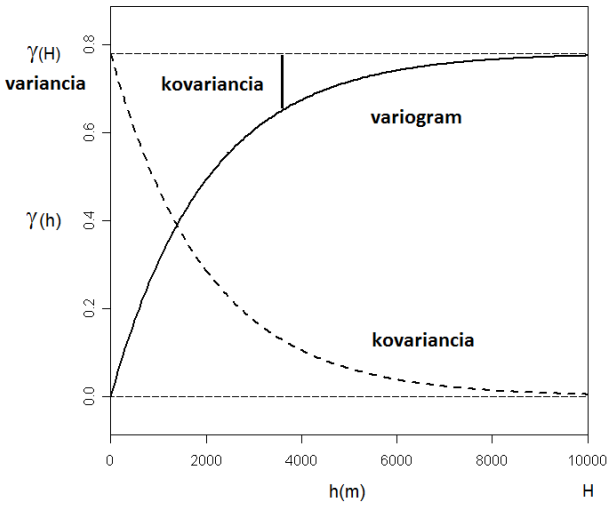
növekedésével a térbeli korreláció csökken, ezért a korrelációval arányos kovariancia is csökken. Ugyanakkor nagyobb távolságok esetén a *Z* értékkülönbségek nőnek és a szórás is növekszik. Megfigyelhető, hogy a variogram görbéje aszimptotikusan tart egy bizonyos *γ(H)* értékhez. Ez a *h*=0 helyen kijelöli a *Z(r)* szórásnégyzetét

$$\mathrm{VAR}[Z(r)] = \mathrm{COV}[Z(r), Z(r)] = \gamma(H)$$

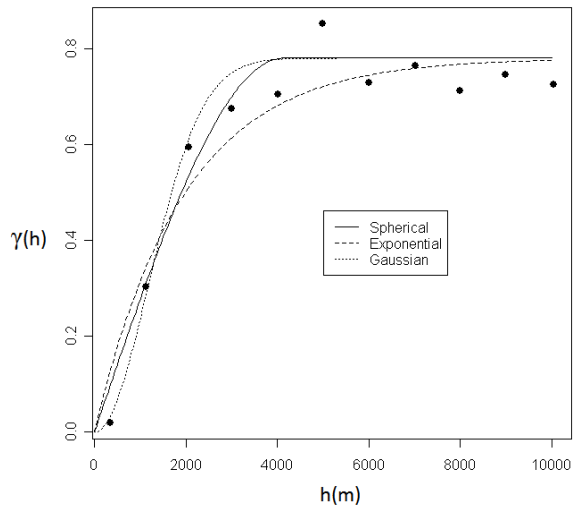
ahol *H* az ún. **hatástávolság**. A korreláció két pont között csak ezen a távolságon belül áll fenn, azaz csakis e távolságon belül lehet a pontokat megválasztani az interpolációhoz. A kovariancia értéke a hatástávolsággal kifejezve

$$\mathrm{COV}[Z(r), Z(r+h)] = \gamma(H) - \gamma(h).$$

A mérési eredményekből számított **tapasztalati variogram** pontjaira elméleti függvények ún. **variogram modell**ek illeszthetők. A 39. ábrán egy porozitás adatsor empirikus variogram pontjai és háromféle azokra illeszkedő variogram modell látható.



38. ábra *Az elméleti variogram*



39. ábra *Tapasztalati variogram értékek és variogram modellek*

A 39. ábrán látható **szférikus, exponenciális** és **Gauss variogram modell**ek formulái a következők

$$\gamma_{sz}(h)=\begin{cases} C\left[1.5\frac{h}{H}-0.5\left(\frac{h}{H}\right)^3\right] & \text{ha} \quad h \le H \\ C & \text{ha} \quad h > H \end{cases}$$

$$\gamma_e(h)=C\left[1-e^{-\frac{h}{H}}\right]$$

$$\gamma_G(h)=C\left[1-e^{-\left(\frac{h}{H}\right)^2}\right].$$

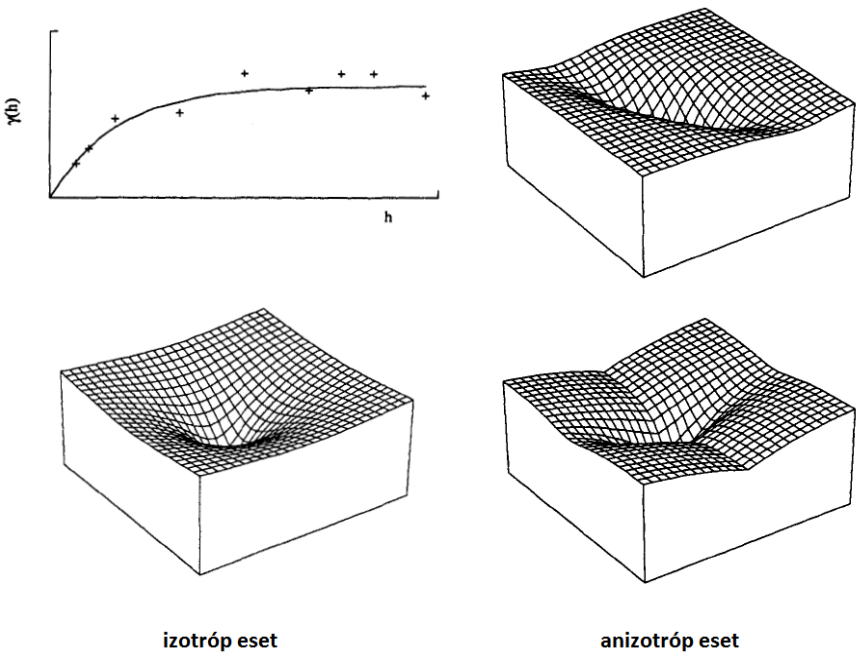
A fenti elméleti $\gamma(h)$ görbék egy bizonyos C értékhez tartanak

$$C=\gamma(H)=VAR[Z(r)]$$

melyből H értéke a tapasztalati variogram pontjai ismeretében kiegyenlítéssel számítható. A krigeléshez szükséges kovarianciákat a variogramból számíthatjuk

$$COV[Z(r),Z(r+h)]=C-\gamma(h).$$

A variogram általános esetben **anizotróp** (irányfüggő) mennyiség, tehát a térbeli korreláció vizsgálatánál nem mindegy, hogy milyen irányban rajzoljuk fel a $\gamma(h)$ görbét. A 40. ábrán látható, hogy **izotróp** (irányfüggetlen) esetben a variogram alakja minden irányban egyforma, míg anizotrópia esetén az egyes irányokhoz eltérő karakterisztikájú variogramok adódnak.



40. ábra *Irányfüggetlen (baloldalon) és irányfüggő (jobboldalon) variogramok (Isaaks és Srivastava nyomán,1989)*

A **krigelés** robusztus becslési eljárás, mely nem érzékeny a variogram modellre, valamint tekintetbe veszi annak irányfüggését is. Közelítsük P_0 pontban az ismeretlen $Z(P_0)$ mennyiség értékét n számú (közeli) P_i pont ismert $Z(P_i)$ értékének súlyozott átlagával

$$Z(P_0) = \sum_{i=1}^{n} w_i Z(P_i).$$

A becslés akkor torzítatlan, hogyha a súlyokra előírjuk

$$\sum_{i=1}^{n} w_i = 1.$$

Ez azért szükséges, mert ha pl. minden környező érték egyforma lenne, akkor csak ebben az esetben kapnánk a kérdéses pontban is ugyanazt az értéket. A krigelés feladata tehát a w_i súlyok és azon keresztül a $Z(P_0)$ érték (és más kérdéses pontok értékének) meghatározása. Kössük a súlyok meghatározását a becslés szórásnégyzetének (azaz a valódi és a becsült érték eltérésének varianciája) minimumához

$$\mathrm{VAR}\left[Z(P_0) - \sum_{i=1}^{n} w_i Z(P_i) \right] = \min.$$

A fenti optimalizációs feladat megoldását a Lagrange-féle multiplikátorok módszerével állíthatjuk elő (a részletes levezetést *Steiner (1990)* könyvében találjuk meg), mely a $KW=D$ lineáris egyenletrendszerre vezet (ahol $\underline{\underline{K}}$ az ún. Krige-mátrix, és μ a Lagrange-féle multiplikátor)

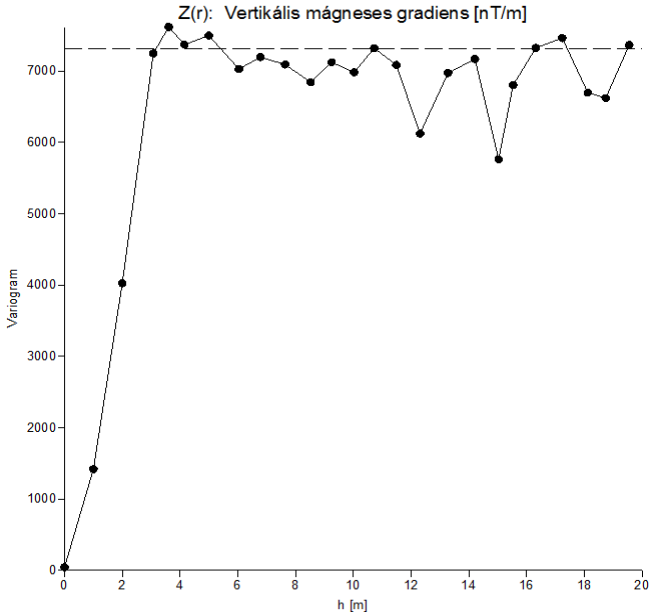
$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} & 1 \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3n} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} c_{01} \\ c_{02} \\ c_{03} \\ \vdots \\ c_{0n} \\ 1 \end{pmatrix}.$$

A μ paraméter a súlyokra tett kikötést érvényesíti, tehát a $\sum w=1$ feltételt írja elő a szélsőérték-keresés során. Az egyenletrendszerben található kovarianciákat a variogramból számítjuk, így az alábbi mátrix elemek ismert mennyiségek

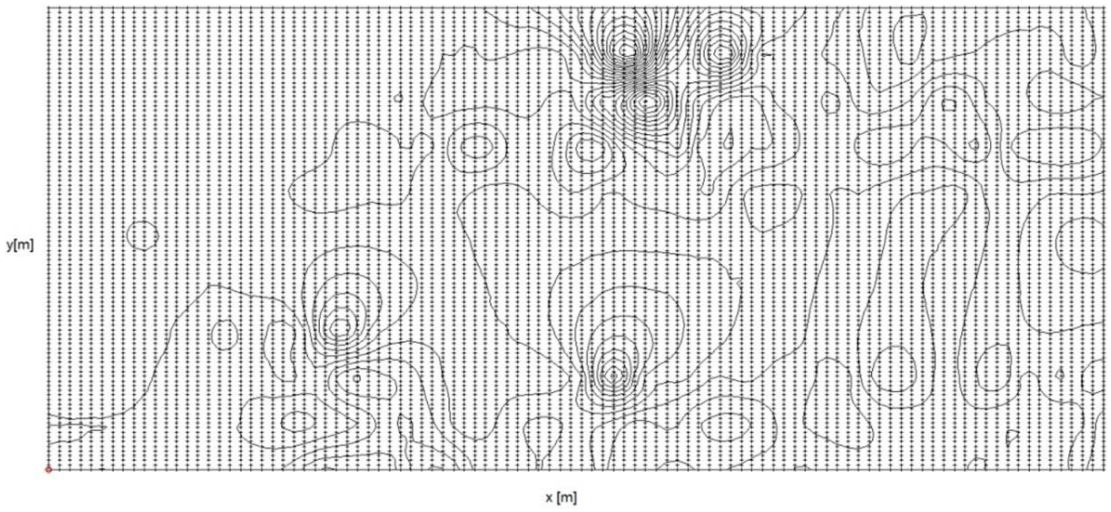
$$c_{ij} = \mathrm{COV}\left[Z(P_i), Z(P_j)\right] = C - \gamma\left(h(P_i, P_j)\right)$$
$$c_{ii} = \mathrm{VAR}\left[Z(P_i)\right] = C$$
$$c_{0i} = \mathrm{COV}\left[Z(P_0), Z(P_i)\right] = C - \gamma\left(h(P_0, P_i)\right).$$

Az ismeretlen súlyokat (és a Lagrange multiplikátort) a $W=K^{-1}D$ egyenletrendszerből határozzuk meg (ahol $\underline{\underline{K}}^{-1} = adj(\underline{\underline{K}})/det(\underline{\underline{K}})$ a Krige-mátrix inverze). A becslési hibát (becslés szórásnégyzetét) $\sigma=W^TD$ segítségével kapjuk meg (ahol T a transzponált jelölése).

Példa. Tekintsük ismét a 36. ábrán bemutatott nyékládházi mérési szelvényeket! Az interpoláció célja az, hogy mágneses térképet szerkesszünk a területről és kijelöljük a felszín alatt eltemetett fémhulladékok helyét. A teljes mérési terület 50m×35m volt. Az adatokat *x* és *y* irányban egyaránt 1m-es mintavételi távolság mellett gyűjtöttük. A mágneses térerősség adatokból a feldolgozás során kiszámítottuk a *dT/dz [nanoTesla/m]* vertikális mágneses gradiens értékeket, melyek különösen érzékenyek a felszínközelben elhelyezkedő mágneses hatókra (a mágnesezhető fémhulladékok a felszín alatti pár m-es tartományban helyezkedtek el). Először kiszámítottuk a vertikális mágneses gradiens adatok variogramját (ld. 41. ábra). Majd a lineáris egyenletrendszer megoldásával kapott súlyokkal minden kérdéses pontban meghatároztuk az adatok interpolált értékét. Az eredményül adódó izovonalas térképet és az interpolációs pontokat a 42. ábra mutatja, melyen több mágneses anomália felfedezhető.



41. ábra *Vertikális mágneses gradiens adatok variogramja*



42. ábra *Vertikális mágneses gradiens adatok interpolált térképe*

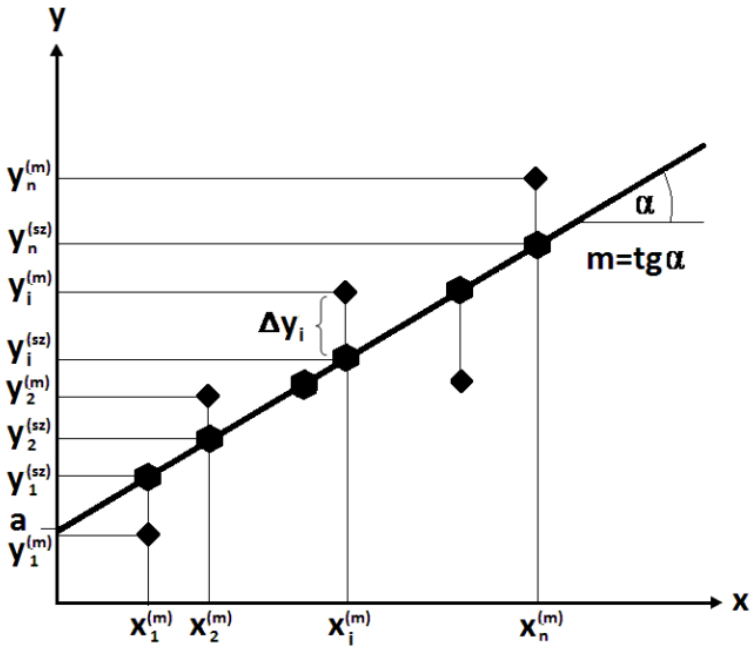
8. Lineáris és nemlineáris regresszió

A földtani kutatásban gyakori feladat két (vagy több) fizikai mennyiség között fennálló függvénykapcsolat meghatározása. Az egyváltozós regressziós vizsgálatok célja, hogy az *x* és *y* tapasztalati úton megfigyelt (mért) mennyiségek kapcsolatát leíró *y=f(x)* **regressziós függvény**t megtaláljuk. Az *f(x)* függvény sok esetben lineáris, ekkor a mérési adatokra legjobban illeszkedő egyenest keressük. E feladatot **lineáris regresszió**nak nevezzük. Abban az esetben viszont, amikor a regressziós függvény nemlineáris, akkor **nemlineáris regresszió**ról beszélünk.

Az $(x_i^{(m)},y_i^{(m)})$ mérési adatok az $\overline{xy}$ síkon *n* számú pontot jelölnek ki (ld. 43. ábra). Ha a két mennyiség korrelál egymással és lineáris kapcsolatot feltételezünk, akkor a pontokra legjobban illeszkedő egyenes egyenletét az ismert alakban keressük

$$y = mx + a$$

ahol *m* az egyenes meredeksége (iránytangense), és *a* az egyenes ordináta-metszete. Például fúrólyukbeli hőmérséklet mérések esetén, ha azt feltételezzük, hogy a hőmérséklet a mélységgel lineárisan növekszik, akkor *m* a geotermikus gradienst (ami földi átlagban 3^0C/100m, de Magyarországon 5^0C/100m), *a* pedig a kezdeti hőmérsékletet (ami a felszínen átlagosan 10^0C) adja meg. A fenti egyenlet (regressziós modell) segítségével egy $y_i^{(sz)}$ számított adatsort állíthatjuk elő, melynek az $y_i^{(m)}$ mért adatoktól való eltérése az (*m, a*) paraméterpár megválasztásától függ. A 43. ábrán a mérési és számított adatok, valamint a regressziós egyenes és paraméterei láthatók. Tapasztalati tény, hogy az adatokat terhelő zaj (véletlen hiba) miatt a legjobb egyenes sem haladhat át mindegyik mérési ponton egyszerre.



43. ábra *A regressziós egyenes és annak paraméterei*

A mérési adatokra legjobban illeszkedő egyenest a mért és számított adatok egyedi eltéréseinek ($\Delta y_i = y_i^{(m)} - y_i^{(sz)}$ ahol $i=1,2,\ldots,n$) minimális értékeinél kapjuk. Határozzuk meg m és a paraméterek optimális értékét a legkisebb négyzetek (LSQ) módszerével

$$E = \sum_{i=1}^{n} \left(y_i^{(m)} - y_i^{(sz)} \right)^2 = \sum_{i=1}^{n} \left(y_i - m x_i - a \right)^2 = \min.$$

A minimumhely ott található, ahol az E célfüggvény m és a paraméterek szerinti parciális deriváltjai (egyidejűleg) zérussal egyenlők

$$\left.\begin{aligned} \frac{\partial E}{\partial a} &= -2\sum_{i=1}^{n} \left(y_i - m x_i - a \right) = 0 \\ \frac{\partial E}{\partial m} &= -2\sum_{i=1}^{n} \left(y_i - m x_i - a \right) x_i = 0 \end{aligned}\right\}.$$

A szorzás elvégzése után kapjuk a következő egyenletrendszert

$$\left.\begin{aligned} \sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - an &= 0 \\ \sum_{i=1}^{n} x_i y_i - m\sum_{i=1}^{n} x_i^2 - a\sum_{i=1}^{n} x_i &= 0 \end{aligned}\right\}.$$

Az első egyenletet x átlagértékével beszorozva kapjuk

$$\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \frac{m}{n}\left(\sum_{i=1}^{n} x_i \right)^2 - a\sum_{i=1}^{n} x_i = 0$$

melyet az egyenletrendszer második tagjából kivonva előáll

$$\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - m\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right] = 0.$$

Ebből m regressziós koefficienst kifejezhetjük

$$m = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

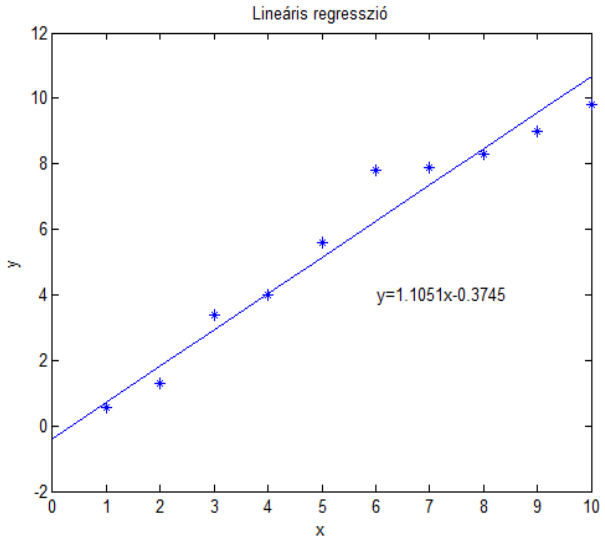
ahol r a korrelációs együtthatót, σ pedig a szórást jelöli. Az a regressziós koefficienst az egyenletrendszer első egyenletéből származtatjuk, és m függvényében számíthatjuk

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{m}{n}\sum_{i=1}^{n} x_i = \bar{y} - m\bar{x}.$$

A MATLAB rendszerben regressziós számításokat a **polyfit** és **polyval** függvényekkel végezhetünk. Az előbbi az adatok ismeretében megadja a regressziós függvény együtthatóit, az utóbbi pedig a független változó értékek helyén a számított adatokkal tér vissza.

Feladat. Végezzünk lineáris regressziót 11 adat esetén, majd ábrázoljuk a mérési adatokat és a kapott regressziós egyenest! Az eredmény a 44. ábrán látható.

```
clc;
clear all;
x=[0 1 2 3 4 5 6 7 8 9 10];
y_mert=[-1 0.56 1.3 3.4 4 5.6 7.8 7.9 8.3 9 9.8];
eh=polyfit(x,y_mert,1);
y_szam=polyval(eh,x);
plot(x,y_mert,'*');
hold on;
plot(x,y_szam);
xlabel('x');
ylabel('y');
title('Lineáris regresszió');
m=eh(1),
a=eh(2),
```



44. ábra *A lineáris regresszió eredménye*

A legkisebb négyzetek módszerén alapuló regressziónak jelentős hátránya, hogy igen érzékenyen reagál a kiugró adatokra és az adatok Gausstól eltérő eloszlása esetén nem ad optimális eredményt. Tekintsük a mért és a regresszióval számított adatok L_p-normáját

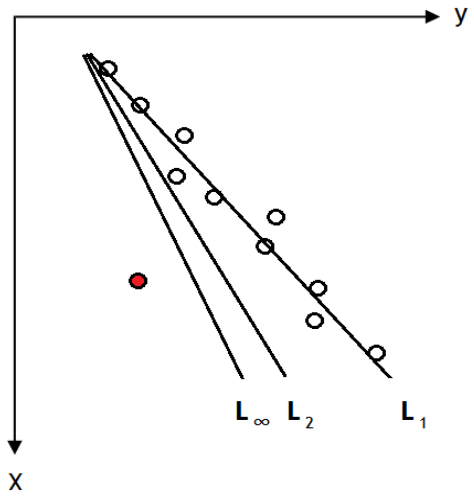
$$L_p = \left[\frac{1}{n}\sum_{i=1}^{n}\left|y_i^{(m)} - y_i^{(sz)}\right|^p\right]^{1/p} = \left[\frac{1}{n}\sum_{i=1}^{n}\left|y_i^{(m)} - f\left(x_i\right)\right|^p\right]^{1/p}$$

melynek négyzete p=2 esetben (L_2-norma) egy konstanstól (n adatszám) eltekintve megegyezik a legkisebb négyzetek módszerén alapuló regressziós feladat célfüggvényével

$$nL_2^2 = \sum_{i=1}^{n}\left(y_i^{(m)} - f\left(x_i\right)\right)^2 = E.$$

Az *E* célfüggvény minimalizálásán alapuló regressziós eljárás nem rezisztens. A 45. ábrán látható, hogy a becslés kiugró adatok jelenlétében igen pontatlan. Ráadásul a *p* paraméter növelésével egyre rosszabb eredményt kapunk. A legszélső esetet az L_∞-norma alkalmazása képviseli, mely számszerűen a mért és a számított adatok eltérésének maximumával tér vissza. Ahhoz, hogy rezisztens eljárást kapjunk, érdemes $p<2$ esethez tartozó célfüggvényt választani. Az ábrán látható, hogy $p=1$ esetre vonatkozó (L_1-norma) kiegyenlítési eljárás nem érzékeny a kiugró adatokra és jobb becslést ad a regressziós paraméterekre nézve. Durva mérési hibák és a regressziós paraméterekre vonatkozó *R* számú $A(\vec{m})=0$ mellékfeltétel esetén az L_1-normán alapuló célfüggvény (ahol $\vec{m}$ a regressziós paraméterek vektora és λ_r az *r*-edik Lagrange-féle multiplikátor) a következő

$$L_1^* = \sum_{i=1}^{n} \left| y_i^{(m)} - f(x_i, \vec{m}) \right| + \sum_{r=1}^{R} \lambda_r A(\vec{m}) = \min.$$



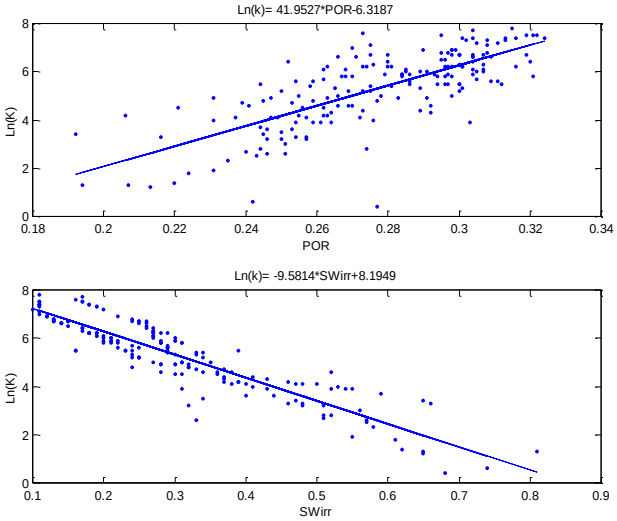
45. ábra *Lineáris regresszió különböző célfüggvények alkalmazása esetén*

Tételezzük fel, hogy a változók közötti kapcsolat nemlineáris! Az alkalmazott függvénytípus sokféle lehet, mely mindig az aktuális földtani szituációtól függ. Gyakran végzünk hatványfüggvények szerinti kiegyenlítést, ahol a regressziós függvény egy *M*-edfokú polinom (ahol m_0 a nulladfokú taghoz tartozó konstans és m_i az *x* változó *i*-edik hatványához tartozó sorfejtési együttható)

$$f(x, \vec{m}) = f(x, m_0, m_1, ..., m_M) = m_0 + \sum_{i=1}^{M} m_i x^i.$$

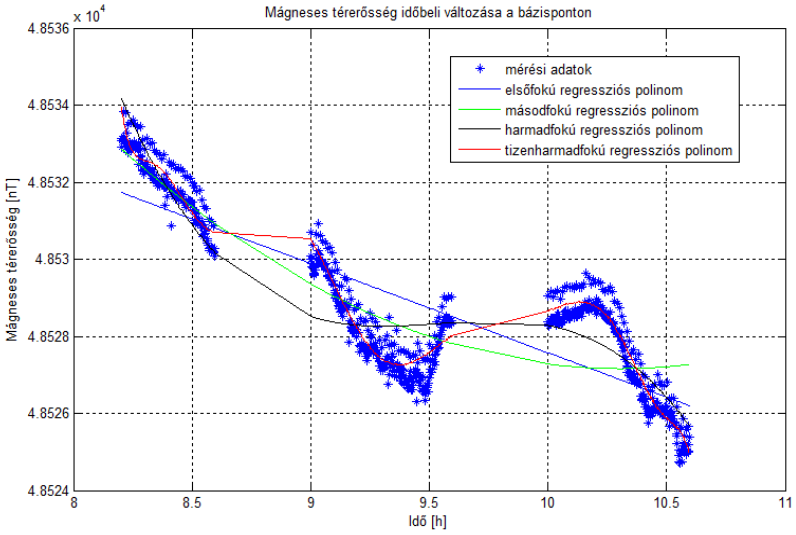
A nemlineáris regressziós feladat **linearizálás**sal is megoldható, melynél az eredeti változók helyett velük összefüggő, de egymással lineáris kapcsolatban lévő változókat vezetünk be. Például legyen a nemlineáris regressziós függvény $y = ae^{bx}$ alakú! Vegyük mindkét oldal természetes alapú logaritmusát, majd a kapott $lny = lna + bx$ függvény változóit helyettesítsük új paraméterekkel $Y = lny, X = x$! Ekkor a regressziós kapcsolat lineáris: $Y = A + BX$. Elvégezve a lineáris regressziós feladatot, a nemlineáris regressziós függvény paraméterei könnyen előállíthatók $a = e^A, b = B.$

Példa. Tekintsünk egy hazai fúrásból származó kőzetmintákon mért porozitás (*POR*), kötött víztelítettség (SW_{irr}) és áteresztőképesség (*K*[*mDarcy*]) adatrendszert! A 46. ábrán a fenti kőzetfizikai változók regressziós kapcsolatát figyelhetjük meg. Látható, hogy a területen jelenlévő üledékes kőzetek áteresztőképessége a porozitással egyenes, viszont a pórustérben elhelyezkedő agyagszemcsék felületén megkötött víz térfogatával fordítottan arányos. Ez az eredmény a szénhidrogén-tárolók minősítését és kitermelését szolgálja.



46. ábra *Permeabilitás - porozitás (felül) és permeabilitás - kötött víztelítettség (alul) adatok linearizált regressziója*

Példa. A 47. ábra a nyékládházi bázisállomáson mért mágneses adatsor különböző fokszámú hatványfüggvények szerinti regresszióját mutatja. A bázismérések alapvetőek a mágneses adatok feldolgozása szempontjából, ugyanis ezek segítségével korrigálhatjuk a nyers adatokat a mágneses tér 24h-ás periódusú változásának hatásával. Ez azért szükséges, mert így állíthatunk elő a napi hatástól mentes (csak a felszín alatti mágnesezhető hatóktól származó) mágneses anomáliákat térkép formájában.



47. ábra *Mágneses térerősség adatok nemlineáris regressziója*

Második rész

MSc tananyag

9. Az adatok jellemzése és skálázása

A földtani kutatás során gyűjtött adatok gyakran többféle (különböző fizikai elven alapuló) mérésből származnak. Például a felszíni geofizika módszerei magukban foglalják a gravitációs, földmágneses, egyenáramú elektromos, elektromágneses, szeizmikus, radioaktív és termikus méréseket. A mérési területen gyűjtött kőzetminták laboratóriumi vizsgálatával az ásványok mennyiségének területi eloszlását adjuk meg (pl. egyes ércek feltérképezése céljából). A fúrásos kutatás módszereivel nemcsak nagy mélységből származó kőzetmagot vagy fluidum-mintát hozhatunk a felszínre, melyeket laboratóriumban ugyancsak mérések alá vethetünk, hanem számos kőzetfizikai paramétert tudunk „in-situ" megmérni, melyek a szénhidrogének és egyéb ásványi nyersanyagok mennyiségével állnak kapcsolatban. A fenti néhány példából látható, hogy a terepen igen sokféle adatot gyűjthetünk. Ezeket az adatokat az összehasonlíthatóság miatt rendszereznünk, valamint hasonló „alakra" kell hoznunk. E műveletnek az a célja, hogy az adatfeldolgozás során az adatokból a lehető legtöbb földtani információt tudjunk kinyerni.

A többváltozós statisztika terminológiája szerint az adatoknak két fő jellemzőjük van. **Objektum**nak a földtani képződmények sokaságának egy elemét, **tulajdonság**nak pedig a sokaság elemeihez tartozó fizikai jellemzőt (változót) nevezzük. A változókat tapasztalati úton (méréssel) vizsgáljuk, melyek jellemzése a statisztika módszereivel történik. Rendezzünk I számú objektum J különböző tulajdonságát egy $I \times J$ méretű mátrixba

$$\underline{\underline{D}} = \begin{pmatrix} d_{11} & \cdots & d_{1j} & \cdots & d_{1J} \\ \vdots & & \vdots & & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & d_{iJ} \\ \vdots & & \vdots & & \vdots \\ d_{I1} & \cdots & d_{Ij} & \cdots & d_{IJ} \end{pmatrix}$$

ahol $i=1,2,\ldots,I$ a sor-, és $j=1,2,\ldots,J$ az oszlopindexet jelöli. A fenti mennyiséget **adatmátrix**nak nevezzük, melynek i-edik sora (vagy objektumvektora) az i-edik objektum tulajdonságait tartalmazza

$$\vec{d}_i^{(o)} = \left[d_{i1}, d_{i2}, \ldots, d_{iJ}\right]$$

valamint j-edik oszlopa (vagy tulajdonságvektora) a j-edik tulajdonság különböző objektumoknál megvalósult értékeit rögzíti

$$\vec{d}_j^{(t)} = \left[d_{1j}, d_{2j}, \ldots, d_{Ij}\right]^T.$$

Például fúrási geofizikai mérések során az i-edik objektum lehet pl. egy kőzetréteg vagy akár egy mélységpont, és a j-edik tulajdonság pedig egy bizonyos fizikai mennyiség (pl. természetes potenciál, természetes gamma, sűrűség, akusztikus terjedési idő vagy fajlagos ellenállás), melyet minden mélységpontban egymás után egy speciális mérőberendezéssel

(szonda) regisztrálunk. A regisztrátumot szelvénynek nevezzük, mely a mélység függvényében ábrázolja a kőzetsorozatra vonatkozó mérési eredményeket.

A tulajdonságvektorok különböző nagyságrendű és mértékegységgel rendelkező jellemzőket tartalmazhatnak. A statisztikai számítások számára olyan egységesített adatrendszerre van szükségünk, melynek adatai azonos nagyságrendűek, ill. dimenzió nélküliek. Azt a transzformációt, mely a nyers mérési adatokat a fenti céloknak megfelelően alakítja át, **skálázás**nak (léptékváltás) nevezzük. A leggyakrabban alkalmazott skálázási eljárás a **centrálás**, mely zérus középértékre vonatkozó eltolást (az adatmátrix elemek konstans értékkel való eltolását) jelent. Ekkor a *j*-edik tulajdonságvektor (az adatmátrix *j*-edik oszlopa) elemeinek számtani közepe zérus lesz, viszont az adatok szórása nem változik

$$d'_{ij} = d_{ij} - \bar{d}_j \quad \text{ahol} \quad \bar{d}_j = \frac{1}{I}\sum_{i=1}^{I} d_{ij}.$$

Feladat. Végezzünk centrálást a MATLAB rendszer segítségével! Legyen *d* egy tetszőleges ötelemű (tulajdonság-) vektor és *duj* a megegyező méretű skálázott oszlopvektor! Ellenőrizzük a két vektor számtani közepét és szórását! A feladat megoldása MATLAB parancsablakban

```
>>  d=[1 3.2 -5.6 5 8 11]'
d =
    1.0000
    3.2000
   -5.6000
    5.0000
    8.0000
   11.0000

>> datl=mean(d)
datl =
    3.7667

>> dszor=std(d)
dszor =
    5.7875

>> duj=d-datl
duj =
   -2.7667
   -0.5667
   -9.3667
    1.2333
    4.2333
    7.2333

>> mean(duj)
ans =
 -5.9212e-016

>> dujszor=std(duj)
dujszor =
    5.7875
```

A **standardizálás** olyan skálázási eljárás, mely zérus középértékre és egységnyi szórásra történő léptékváltást valósít meg. (Emlékezzünk, hogy az 1. fejezetben a sűrűségfüggvények standard alakja T=0 és S=1 mellett áll elő, mely az általános alak standardizáltjának tekinthető. A 10. fejezetben a faktor- és főkomponens elemzésről lesz szó, melyben az adatokat jelen eljárással standardizáljuk). A művelet a konstans eltolás mellett egyben nyújtást is jelent. A standardizált változó a korrigált empirikus szórás (ld. 3. fejezet) alkalmazása mellett dimenzió nélküli (és torzítatlan) mennyiség

$$d'_{ij} = \frac{\left(d_{ij} - \bar{d}_j\right)}{\sigma_j} \quad \text{ahol} \quad \sigma_j = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\left(d_{ij} - \bar{d}_j\right)^2}.$$

Feladat. Az előző MATLAB példa *d* vektorának standardizáltja a *duj2* vektor.

```
>> duj2=(d-mean(d))/szigma
duj2 =
   -0.4780
   -0.0979
   -1.6184
    0.2131
    0.7315
    1.2498

>> mean(duj2)
ans =
 -1.1102e-016

>> std(duj2)
ans =
     1
```

A **maximum-skálázás** végrehajtása után a *j*-edik tulajdonságvektor elemeinek maximális értéke 1 lesz. A művelet konstans zsugorítást jelent, ahol a skálázott változó dimenziótlan

$$d'_{ij} = \frac{d_{ij}}{\max(d_{ij})}.$$

Feladat. A MATLAB rendszerbeli *d* tulajdonságvektor elemeinek és a *duj3* skálázott vektor elemeinek maximális értéke

```
>> max(d)
ans =
   11

>> duj3=d/max(d)
duj3 =
    0.0909
    0.2909
   -0.5091
    0.4545
    0.7273
    1.0000

>> max(duj3)
ans = 1
```

A **terjedelem-skálázás** alapesete az, amikor [0,1] intervallumba transzformáljuk az adatokat. A művelet konstans eltolást és zsugorítást jelent, mellyel a *j*-edik tulajdonságvektor elemei [0,1] határok közé kerülnek. A skálázás dimenziótlan változót ad eredményül

$$d'_{ij} = \frac{d_{ij} - \min(d_{ij})}{\max(d_{ij}) - \min(d_{ij})}.$$

Feladat. A MATLAB rendszerbeli *d* tulajdonságvektorból képzett *duj4* skálázott vektor minimuma 0 és maximuma 1

```
>> duj4=(d-min(d))/(max(d)-min(d))
duj4 =
   0.3976
   0.5301
        0
   0.6386
   0.8193
   1.0000

>> min(duj4)
ans =
     0

>> max(duj4)
ans =
     1
```

A terjedelem-skálázás általánosított műveletével egy tetszőleges [*A*,*B*] intervallumba transzformálhatjuk a tulajdonságvektor elemeit. A *j*-edik tulajdonságvektor skálázása az A_j (alsó) és B_j (felső) határok előírása mellett

$$d'_{ij} = A_j + (B_j - A_j)\frac{d_{ij} - \min(d_{ij})}{\max(d_{ij}) - \min(d_{ij})}.$$

Feladat. Legyen *A*=-10 és *B*=20! Ekkor a MATLAB rendszerbeli *d* tulajdonság-vektorból képzett *duj5* skálázott vektor

```
>> duj5=A+((B-A)*(d-min(d))/(max(d)-min(d)))
duj5 =
    1.9277
    5.9036
  -10.0000
    9.1566
   14.5783
   20.0000

>> min(duj5)
ans =
   -10

>> max(duj5)
ans =
    20
```

10. Faktor- és főkomponens elemzés

A nagyméretű adatrendszerek sokszor áttekinthetetlenek, ezért a többváltozós adatelemzésnél gyakran célravezető lehet a probléma méretének (dimenziójának) a csökkentése. Ennek keretében a statisztikai mintában felhalmozott információ nagy részének megtartásával ugyanazt a jelenséget kevesebb változóval írjuk le. Az új változók magukban foglalják az objektumok lényeges tulajdonságait, valamint új (az adatokkal rejtett kapcsolatban lévő) jellemzőket is szolgáltatnak. A dimenziócsökkentést faktor- és főkomponens elemzéssel végezzük el. A **faktoranalízis** során nagyszámú, egymással összefüggő vagy független valószínűségi változót kevesebb számú korrelálatlan változóval helyettesítünk, ahol a keletkezett új változókat közvetlen módon nem tudjuk megfigyelni. A **főkomponens analízis**sel a valószínűségi változókat szintén kisebb számú korrelálatlan változóba transzformáljuk át. Az új változók koordináta-rendszerében az lesz az első főirány (főkomponens), ahol az összes irány közül legnagyobb a minta varianciája, a második főkomponens az elsőre merőleges irányok közül a maximális varianciájú irányt képviseli. Az első néhány főkomponens jól tükrözi a mintában rejlő információt (az adatok varianciájának legnagyobb részét írja le), ezáltal a többi változó elhanyagolhatóvá válik.

Elsőként tekintsük át a faktoranalízis elméletét! Jelöljük a 9. fejezetben bevezetett $\underline{\underline{D}}$ mátrixot $\underline{\underline{X}}$-el! Az adatokat tartalmazó $\underline{\underline{X}}$ mennyiséget **tulajdonság-mátrix**nak nevezzük, mely N számú sorból és M számú oszlopból áll

$$\underline{\underline{X}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}.$$

A mátrix i-edik sora az i-edik objektumot, j-edik oszlopa pedig a j-edik kiindulási (mért) változót jelöli. A faktoranalízis végrehajtásának a feltétele az, hogy a tulajdonság-mátrixban ne legyen 5%-nál nagyobb adathiány, ill. ne hiányozzon egy sorból vagy oszlopból a mátrix elemeknek több mint a fele. Ha mégis ez az eset állna fenn, akkor a hiányzó helyekre a sorok vagy az oszlopok átlagát szoktuk beírni, azonban ez további hibával terheli a zaj miatt egyébként is közelítő megoldást. A faktoranalízis modellje alapján a (skálázott) kiindulási változókat tartalmazó $\underline{\underline{X}}$ mátrixot felbonthatjuk az $\underline{\underline{A}}$ **közös komponens** mátrix (nem mérhető változók) és az $\underline{\underline{E}}$ **hibakomponens** mátrix összegére

$$\underline{\underline{X}} = \underline{\underline{A}} + \underline{\underline{E}}$$

ahol $\underline{\underline{A}}$ és $\underline{\underline{E}}$ mérete egyaránt $N \times M$. A faktoranalízis célja, hogy az $\underline{\underline{A}}$ mátrixot $a<M$ számú tényezőre, ún. faktorra bontsuk fel (ahol a előre rögzített szám). Alapvetően négyféle faktortípust különböztetünk meg egymástól. **Általános faktor**nak nevezzük az összes kiindulási változóhoz kapcsolódó faktort. A **közös faktor** legalább két mérhető változót

befolyásol, míg az **egyedi faktor** csak egyet. Végül **maradékfaktor**nak nevezzük a mérési vagy a becslési hibából származó egyedi faktort. Bontsuk fel a közös komponens mátrixot *a* számú faktor lineáris kombinációjával

$$\underline{\underline{A}} = \underline{\underline{F}}\underline{\underline{L}}^{T}$$

ahol $\underline{\underline{F}}$ az $N\times a$ méretű **faktorok** mátrixa és $\underline{\underline{L}}$ az $M\times a$ méretű **faktoregyütthatók** mátrixa. Alkalmazzuk a 2. fejezetben megismert szorzási szabályt! Mivel az $\underline{\underline{F}}$ mátrix oszlopainak a száma megegyezik $\underline{\underline{L}}^{T}$ transzponált ($a\times M$ méretű) mátrix sorainak a számával, ezért az eredményül adódó $\underline{\underline{A}}$ mátrix mérete $N\times M$ (melyet az $\underline{\underline{F}}$ sorainak és $\underline{\underline{L}}^{T}$ oszlopainak a száma adja ki). Legyen pl. 10 objektumunk, 5 mért változónk és 2 faktorunk! Ekkor a fenti egyenlet általános mátrixelemekkel kifejezve

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \\ A_{41} & A_{42} & A_{43} & A_{44} & A_{45} \\ A_{51} & A_{52} & A_{53} & A_{54} & A_{55} \\ A_{61} & A_{62} & A_{63} & A_{64} & A_{65} \\ A_{71} & A_{72} & A_{73} & A_{74} & A_{75} \\ A_{81} & A_{82} & A_{83} & A_{84} & A_{85} \\ A_{91} & A_{92} & A_{93} & A_{94} & A_{95} \\ A_{101} & A_{102} & A_{103} & A_{104} & A_{105} \end{pmatrix} = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \\ F_{31} & F_{32} \\ F_{41} & F_{42} \\ F_{51} & F_{52} \\ F_{61} & F_{62} \\ F_{71} & F_{72} \\ F_{81} & F_{82} \\ F_{91} & F_{92} \\ F_{101} & F_{102} \end{pmatrix} \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} & L_{15} \\ L_{21} & L_{22} & L_{23} & L_{24} & L_{25} \end{pmatrix}.$$

Az $\underline{\underline{F}}$ mátrixot a faktorok különböző objektumoknál megvalósult értékei (*factor scores*), míg az $\underline{\underline{L}}$ mátrixban a kiinduló változókkal kapcsolatban lévő faktorok súlyai (*factor loadings*) alkotják. Tételezzük fel, hogy az $\underline{\underline{A}}$ és $\underline{\underline{E}}$ mátrix korrelálatlan ($\underline{\underline{A}}^{T}\underline{\underline{E}} = \underline{\underline{E}}^{T}\underline{\underline{A}} = 0$) és $\underline{\underline{E}}^{T}\underline{\underline{E}}/N = \underline{\underline{\Psi}}$ ismert mennyiség! Ekkor a tulajdonságmátrix standardizált (átlagértékekkel eltolt és egységnyi szórásokkal osztott) adatainak az $M\times M$ méretű korrelációs mátrixa

$$\underline{\underline{R}} = \frac{1}{N}\underline{\underline{X}}^{T}\underline{\underline{X}} = \frac{1}{N}\underline{\underline{A}}^{T}\underline{\underline{A}} + \underline{\underline{\Psi}}.$$

Legyenek a faktorok lineárisan függetlenek ($\underline{\underline{F}}^{T}\underline{\underline{F}}/N = \underline{\underline{I}}$ az egységmátrix), ekkor a korrelációs mátrix elemei kifejezhetők a faktoregyütthatókkal

$$\underline{\underline{R}} = \frac{1}{N}\left(\underline{\underline{F}}\underline{\underline{L}}^{T}\right)^{T}\left(\underline{\underline{F}}\underline{\underline{L}}^{T}\right) + \underline{\underline{\Psi}} = \underline{\underline{L}}\underline{\underline{L}}^{T} + \underline{\underline{\Psi}}$$

ahol az $M\times M$ méretű diagonális $\underline{\underline{\Psi}}$ mátrix a kiinduló változók szórásnégyzeteinek a közös faktorokkal nem értelmezhető részét képviseli. Az $\underline{\underline{R}}$ korrelációs mátrix főátlóbeli elemei 1-

gyel egyenlők, melyet a mért változók standardizált szórásnégyzetei adnak ki. Képezzünk a faktoregyütthatókkal korrelációs mátrixot! A **redukált korrelációs mátrix** elemei a mért változók és a faktorok közötti korrelációs együtthatók

$$\underline{\underline{L}}\underline{\underline{L}}^{T} = \underline{\underline{R}} - \underline{\underline{\Psi}} = \begin{pmatrix} h_1^2 & r_{12} & \cdots & r_{1M} \\ r_{12} & h_2^2 & \cdots & r_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1M} & r_{2M} & \cdots & h_M^2 \end{pmatrix} \quad \text{ahol} \quad h_i^2 = \sum_{j=1}^{M} L_{ij}^2 \le 1.$$

Az $M \times M$ méretű redukált korrelációs mátrix főátlójában szereplő elemeket **kommunalitás**-oknak (h^2) nevezzük, melyek a mért változók standardizált szórásnégyzeteinek a közös faktorokkal leírható részeit képviselik. A maradékfaktorok által képviselt részt a kommunalitások ismeretében számíthatjuk ki

$$\underline{\underline{\Psi}} = \underline{\underline{I}} - \underline{\underline{H}}^2.$$

A kommunalitások $M \times M$ méretű $\underline{\underline{H}}^2$ mátrixának elemei, ha sokkal kisebbek 1-nél, akkor a mért változóknak kevés közük van a faktorokhoz. A faktoregyütthatók $\underline{\underline{L}}$ mátrixának meghatározása a szimmetrikus mátrixok spektrál-felbontásának módszerével történhet

$$\underline{\underline{L}}\underline{\underline{L}}^{T} = \underline{\underline{Z}}\underline{\underline{\Lambda}}\underline{\underline{Z}}^{T}$$

ahol $\underline{\underline{Z}}$ az $\underline{\underline{R}} - \underline{\underline{\Psi}}$ mátrix sajátvektorainak $M \times a$ méretű mátrixa (a sajátvektorokat a mátrix oszlopai tartalmazzák), $\underline{\underline{\Lambda}}$ az $\underline{\underline{R}} - \underline{\underline{\Psi}}$ sajátértékeinek az $a \times a$ méretű mátrixa (melynek főátlója tartalmazza az a számú λ sajátértéket). A faktoregyütthatók mátrixa

$$\underline{\underline{L}} = \underline{\underline{Z}}\underline{\underline{\Lambda}}^{1/2}$$

ahol az i-edik sajátértékkel képezzük a fenti $\Lambda_{ii}^{1/2} = \sqrt{\lambda_i}$ mátrixelemet. A faktoregyütthatók és a mért változók ismeretében kiszámíthatjuk a faktorértékeket. Ennek legegyszerűbb módja a főkomponens elemzés, mely az $\underline{\underline{E}}$ hibakomponens-mátrixot elhanyagolja, és a redukált korrelációs mátrix helyett a mért változók korrelációs mátrixát hozza kapcsolatba a faktoregyütthatókkal. Mivel a változók szórásnégyzeteinek a közös faktorokkal nem értelmezhető részét ebben az esetben elhanyagoljuk, így a kapott főkomponensek nem a teljes varianciát teszik ki. Más módszerek az $\underline{\underline{E}}$ hibakomponens-mátrix figyelembe vételével, optimalizációs feladat keretében jutnak megoldásra, ilyen pl. a maximum likelihood becslés (Móri, 1999). Ha a $\underline{\underline{\Psi}}$ mátrixot ismeretlennek tételeztük fel, akkor becslést kell végezni rá. Ez úgy történik, hogy az $\underline{\underline{R}}$ korrelációs mátrix elemekkel közelítő számítást végzünk a kommunalitásokra, majd a $\underline{\underline{H}}^2$ mátrixból kifejezzük a $\underline{\underline{\Psi}}$ mennyiséget. Ennek módszereit Horvai (2001) könyve részletezi.

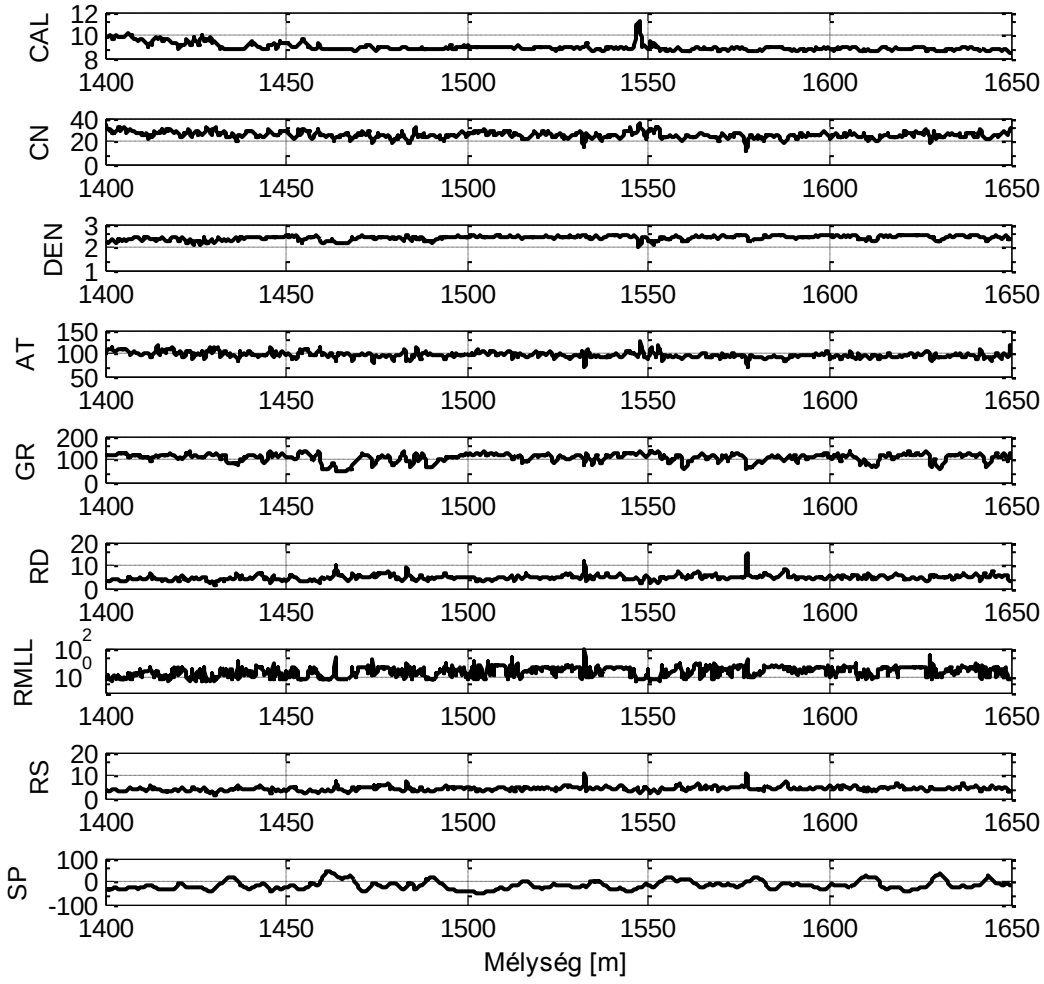
A *k*-adik faktor értelmezése az L_{ik} faktoregyüttható alapján történik. Minél nagyobb a fenti együttható értéke, annál szorosabban kapcsolódik a *k*-adik faktor az *i*-edik mért változóhoz. Egyszerű struktúra (1-hez és 0-hoz közeli faktoregyütthatók) esetén a faktorok könnyen értelmezhetők. Abban az esetben, amikor azok fizikai tartalommal nem hozhatók kapcsolatba, akkor lehetőségünk van ún. **forgatási módszerek**et alkalmazni, mellyel szemléletesebb jelentésű faktorokká alakíthatjuk át őket. Az ortogonális rotációs módszerek korrelálatlan faktorokat eredményeznek. Például a *varimax* módszer célja, hogy minél több zérushoz közeli faktorsúlyt állítson elő. Ekkor kialakul az egyszerű struktúra és azon változók száma kevés lesz, melyhez sok faktor nagy súllyal kapcsolódik. Az eredeti változó ekkor egy vagy csak kisszámú faktorhoz kapcsolódik és mindegyik faktor kevés számú változót reprezentál. A forgatási módszereket Horvai (2001) könyvében szintén megtaláljuk.

Példa. Tekintsünk egy fúrási geofizikai alkalmazást! A vizsgált földtani szerkezet egy magyarországi szénhidrogén-kutatófúrás agyagos homokkő rétegsora, melyben víztároló és impermeábilis rétegek váltakoznak. A mérést az alábbi szondákkal végezték: *CAL*(inch) lyukátmérő, *CN*(%) kompenzált neutron, *DEN*(g/cm^3) sűrűség, *AT*(μs/ft) akusztikus terjedési idő, *GR*(API) természetes gamma, *RD*(ohmm) mély-, *RMLL*(ohmm) kis-(mikrolaterolog) és *RS*(ohmm) sekélybehatolású fajlagos ellenállás, *SP*(mV) természetes potenciál. A szelvények hossza 250m, a mintavételi távolság 0.1m volt. Az $\underline{\underline{X}}$ mátrix *i*-edik sora az *i*-edik mélységpontban (objektum) mért 9-féle szelvényadatot (mért változókat vagy tulajdonságokat) tartalmazza. A mérési szelvényeket a 48. ábrán láthatjuk, melyek átlagos korrelációs együtthatója 0.10. A faktorok számát 2-nek választottuk, mivel e két tényező magyarázta a mért változók varianciájának több mint 90%-át. A faktoranalízist MATLAB rendszerben végeztük el a **factoran** függvény (maximum likelihood módszer) alkalmazásával. A kapott faktorsúlyokat a 3. táblázatban, valamint a korrelálatlan faktorok szelvényeit (a faktoroknak az egyes mélységpontokban előálló értékeit) a 49. ábrán láthatjuk.

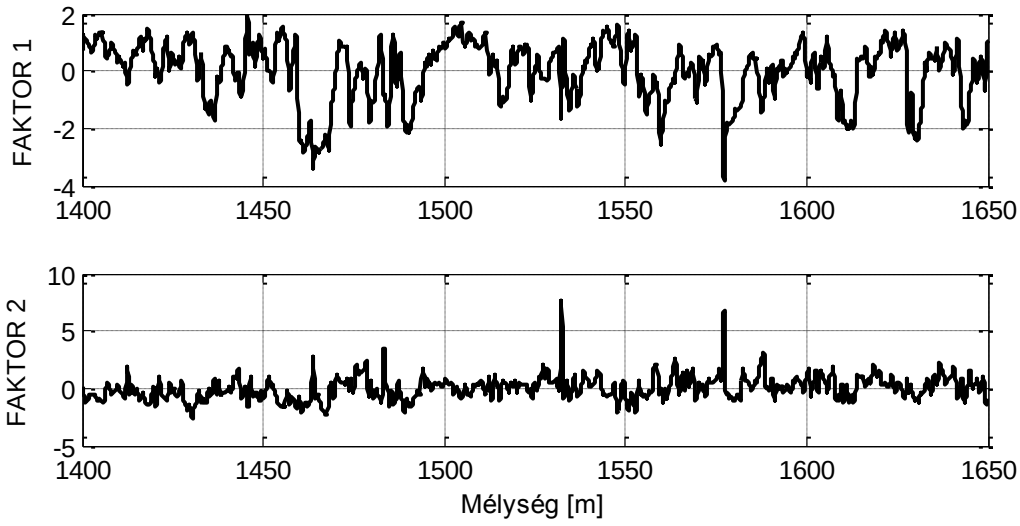
3. táblázat

Szelvények	Faktor 1	Faktor 2
CAL	0.47	-0.14
CN	0.68	-0.37
DEN	0.36	0.59
AT	0.55	-0.58
GR	0.93	0.10
RD	-0.46	0.85
RMLL	0.04	0.57
RS	-0.18	0.98
SP	-0.85	-0.15

A faktorok és a mért adatok szelvényeit összehasonlítva azt tapasztaltuk, hogy az első faktor (FAKTOR 1) a kőzettípussal áll szoros kapcsolatban. A fenti táblázat megerősíti, hogy az 1-es faktor súlyai a *GR* és *SP* (litológiai) szelvények esetén a legnagyobbak (ld. *GR* 0.93 ill. *SP* -0.85 faktoregyütthatókat). A fúrási adatokon végzett agyagtartalom számítási eredményeket bevonva az 50. ábrán látható, hogy az első faktor erős (lineáris) kapcsolatban áll a kőzetrétegek agyagtartalmával (*VSH*).

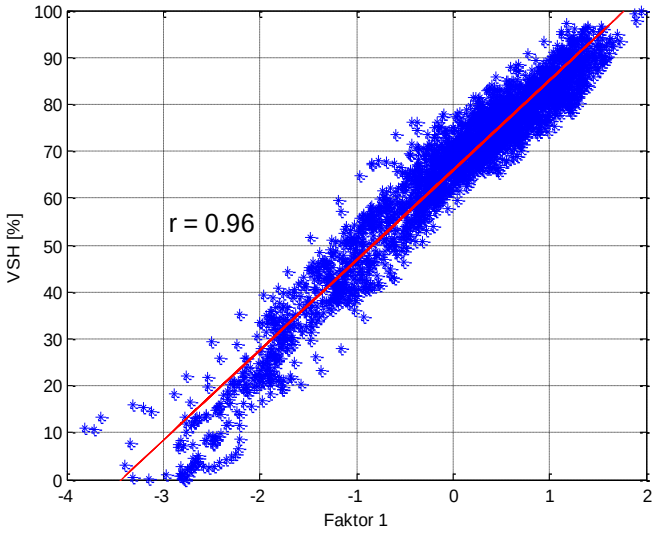


48. ábra *Fúrási geofizikai szelvények (mért változók)*



49. ábra *Faktor szelvények (új változók)*

A főkomponens elemzés nemcsak a faktoranalízis gyakorlati megvalósítása, hanem önállóan alkalmazható adatstruktúra elemző módszer, mely az $\underline{\underline{X}}$ tulajdonságmátrix változóit (tulajdonság-vektorait) kevesebb számú változóvá transzformálja. A fenti transzformáció ortogonális, ezért az új változók korrelálatlanok. Az új változókat **főkomponens**eknek nevezzük, melyeket úgy rendezzük sorba, hogy közülük az első néhány az eredeti változók varianciájának ($\underline{\underline{X}}$ összes elemére számított szórásnégyzet) legnagyobb részét magyarázza.



50. ábra *A mérési adatokból számított agyagtartalom és az első faktor kapcsolata*

Fejezzük ki az $N\times M$ méretű $\underline{\underline{X}}$ tulajdonság-mátrixot az $N\times r$ méretű $\underline{\underline{T}}$ **főkomponens** mátrix és az $M\times r$ méretű $\underline{\underline{P}}$ **főkomponens együttható** mátrix transzponáltjának a szorzatával, ahol $r<M$ a főkomponensek számát jelöli

$$\underline{\underline{X}} = \underline{\underline{T}}\underline{\underline{P}}^{T}.$$

A fenti lineáris egyenletrendszernek mindig létezik egyértelmű megoldása, ahol a j-edik főkomponenst a T_{ij} mátrixelemek ($\underline{\underline{T}}$ mátrix j-edik oszlopa) adják meg (i=1,2,…,N). Legyen pl. 8 objektumunk, 4 mért változónk és 2 főkomponensünk, ekkor a fenti egyenlet általános mátrixelemekkel felírva

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \\ X_{51} & X_{52} & X_{53} & X_{54} \\ X_{61} & X_{62} & X_{63} & X_{64} \\ X_{71} & X_{72} & X_{73} & X_{74} \\ X_{81} & X_{82} & X_{83} & X_{84} \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \\ T_{31} & T_{32} \\ T_{41} & T_{42} \\ T_{51} & T_{52} \\ T_{61} & T_{62} \\ T_{71} & T_{72} \\ T_{81} & T_{82} \end{pmatrix} \begin{pmatrix} P'_{11} & P'_{12} & P'_{13} & P'_{14} \\ P'_{21} & P'_{22} & P'_{23} & P'_{24} \end{pmatrix}.$$

Látható, hogy az $\underline{\underline{X}} = \underline{\underline{T}}\underline{\underline{P}}^T$ egyenlet szerkezete hasonlít a faktoranalízis modellegyenletére, azzal a különbséggel, hogy ebben az esetben a hibakomponens mátrix zérus, ezért baloldalon a közös komponensek mátrixa helyett a tulajdonságmátrix áll. Szorozzuk meg jobboldalról az egyenletet $\underline{\underline{P}}$ mátrix-szal! Mivel a $\underline{\underline{P}}$ mátrix ortonormált ($\underline{\underline{P}}^T\underline{\underline{P}} = \underline{\underline{I}}$), ezért a főkomponens-mátrix könnyen képezhető

$$\underline{\underline{T}} = \underline{\underline{X}}\underline{\underline{P}}.$$

Az előző példa esetén az eredményül adódó lineáris egyenletrendszer

$$\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \\ T_{31} & T_{32} \\ T_{41} & T_{42} \\ T_{51} & T_{52} \\ T_{61} & T_{62} \\ T_{71} & T_{72} \\ T_{81} & T_{82} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \\ X_{51} & X_{52} & X_{53} & X_{54} \\ X_{61} & X_{62} & X_{63} & X_{64} \\ X_{71} & X_{72} & X_{73} & X_{74} \\ X_{81} & X_{82} & X_{83} & X_{84} \end{pmatrix} \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \\ P_{31} & P_{32} \\ P_{41} & P_{42} \end{pmatrix}.$$

A főkomponensek tehát az eredeti (mért) változók lineáris kombinációi. A $\underline{\underline{T}}$ mátrix első oszlopa megadja az első főkomponenst alkotó elemeket

$$\left.\begin{aligned} T_{11} &= X_{11}P_{11} + X_{12}P_{21} + X_{13}P_{31} + X_{14}P_{41} \\ T_{21} &= X_{21}P_{11} + X_{22}P_{21} + X_{23}P_{31} + X_{24}P_{41} \\ &\vdots \\ T_{81} &= X_{81}P_{11} + X_{82}P_{21} + X_{83}P_{31} + X_{84}P_{41} \end{aligned}\right\}$$

míg a második oszlop a második főkomponens elemeit tartalmazza

$$\left.\begin{aligned} T_{12} &= X_{11}P_{12} + X_{12}P_{22} + X_{13}P_{32} + X_{14}P_{42} \\ T_{22} &= X_{21}P_{12} + X_{22}P_{22} + X_{23}P_{32} + X_{24}P_{42} \\ &\vdots \\ T_{82} &= X_{81}P_{12} + X_{82}P_{22} + X_{83}P_{32} + X_{84}P_{42} \end{aligned}\right\}.$$

A fenti egyenletrendszer megoldásához szükség van a P_{ij} főkomponens-együtthatók értékeire. Centráljuk az $\underline{\underline{X}}$ tulajdonságmátrix elemeit (ld. 9. fejezet)! Ekkor $\underline{\underline{X}}$ mátrix j-edik tulajdonságvektora (j-edik oszlopa) elemeinek számtani közepe zérus lesz (viszont az adatok szórása nem változik). A kiindulási (mért) változók $M \times M$ méretű $\underline{\underline{COV}}$ kovariancia mátrixa

$$\underline{\underline{COV}} = \underline{\underline{X}}^T\underline{\underline{X}}.$$

Határozzuk meg a $\underline{\underline{COV}}$ mátrix sajátvektorait és sajátértékeit a szimmetrikus mátrixok spektrál-felbontásának módszerével

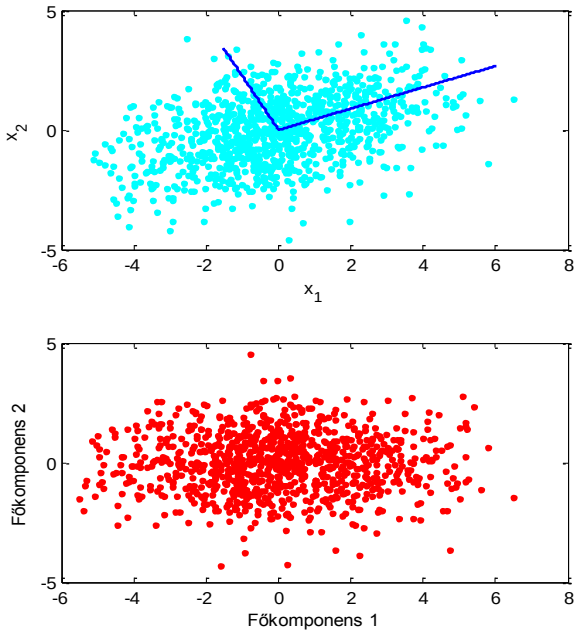
$$\underline{\underline{COV}} = \underline{\underline{Z}}\,\underline{\underline{\Lambda}}\,\underline{\underline{Z}}^T$$

ahol $\underline{\underline{Z}}$ a sajátvektorokat (oszlopaiban) tartalmazó $M \times r$ méretű mátrix, $\underline{\underline{\Lambda}}$ a sajétértékek $r \times r$ méretű diagonális mátrixa. Mivel $\underline{\underline{Z}}$ mátrix ortonormált ($\underline{\underline{Z}}^T\underline{\underline{Z}} = \underline{\underline{I}}$), ezért a kovariancia mátrix $\underline{\underline{COV}} = \underline{\underline{\Lambda}}\underline{\underline{I}}$, ahol $\underline{\underline{\Lambda}}$ nagyságrendben tartalmazza a $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_i \geq \ldots \geq \lambda_r \geq 0$ sajátértékeket. A kovariancia mátrix főátlóbeli elemei a varianciák, innen jön az alapvető összefüggés

$$\lambda_i = \sigma_i^2.$$

A fentiek alapján látható, hogy a főkomponenseket a kovariancia mátrix sajátértékeinek nagysága alapján állítjuk sorrendbe. **Első főkomponens**nek azt az irányt nevezzük, amely mentén legnagyobb az eredeti változók szórása. Ezt a legnagyobb sajátértékhez tartozó sajátvektor iránya jelöli ki. A **második főkomponens**t a második legnagyobb sajátértékhez tartozó sajátvektor adja meg, ahol az első főirányra merőleges irányok közül legnagyobb a szórás, és így tovább. A főkomponens elemzés az eredeti objektumok koordinátáit a főkomponensek által kifeszített új koordináta-rendszerben adja meg, azaz a főtengelyek irányába forgatja az eredeti változókat.

Példa. Határozzuk meg tetszőleges x_1 és x_2 változók esetén a sajátértékeket, és ábrázoltuk a főkomponenseket! A feladatot MATLAB rendszerben a **princomp** és **pcacov** függvények alkalmazásával végezhetjük el. A 51. ábra egy 1000 elemű véletlen minta főkomponenseit mutatja, ahol az első főkomponens az eredeti változók 74%-át, a második pedig azok 26%-át magyarázzák ($\sigma_1=\lambda_1=4.8$, $\sigma_2=\lambda_2=1.5$).

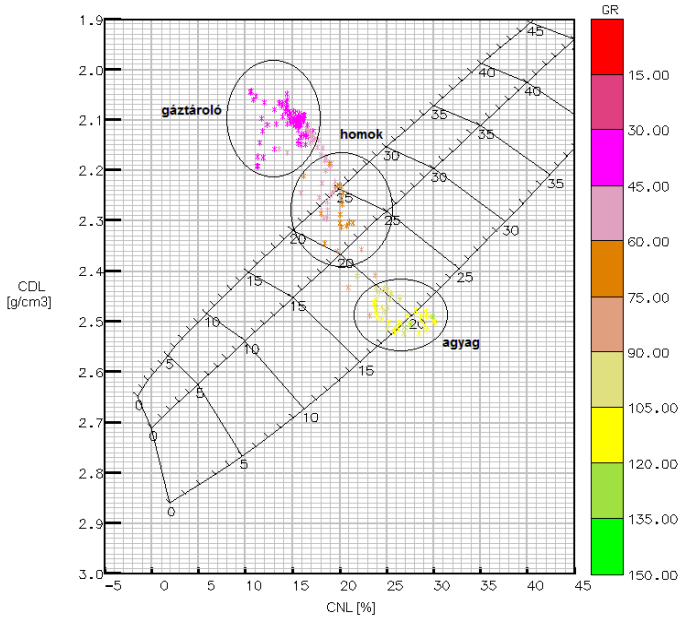


51. ábra *Az x_1 és x_2 változók főkomponensei (ld. sötétkék egyenesek)*

11. Klaszterelemzés

Ebben a fejezetben a többváltozós adatelemzés csoportosítási módszereiről lesz szó. A mérési adatok alkotta tér azon pontjait (objektumok), melyek egy jól definiált szempont szerint nézve hasonlóak ugyanazon **klaszter**be (csoport) sorolhatjuk. A csoportosítás alapját egy adott metrika szerinti közelség, azaz valamilyen távolságdefiníció képezi. Ez azt jelenti, hogy ha két objektum távolsága kicsi, akkor hasonlónak tekintjük, és azonos csoportba soroljuk őket. Nagy távolság esetén az objektumok eltérőek, így nem tartoznak bele ugyanabba a csoportba. Megjegyezzük, hogy túl nagy távolság esetén figyelembe kell azt is vennünk, hogy a klaszterelemző eljárások nem rezisztensek (kiugró adatokra érzékenyek), így pl. az eltérő nagyságrendű adatok torzíthatják a becslést. Emellett az $\underline{\underline{X}}$ tulajdonságmátrix részhalmazokra való bontása során teljesülnie kell azoknak a feltételeknek, hogy minden elem tartozzon bele egy csoportba, ill. egy elem csak egy csoportba tartozzon, valamint ne legyen olyan csoport, amely nem tartalmaz egyetlen elemet sem.

Példa. Egy csoportosítási példát említhetünk a mélyfúrási geofizika területéről. Az 52. ábrán az azonos mélységponthoz tartozó kompenzált neutron- (*CNL*) és sűrűség- (*CDL*) szondával mért adatokat egy koordináta-rendszerben ábrázoltuk. A neutron-porozitás - sűrűség crossplot egy nagyobb mélységintervallum (szénhidrogén-tároló zóna) adatait ábrázolja, ahol a kialakított csoportok alapvető kőzettani információt szolgáltatnak. Az adathalmaz elemeinek elhelyezkedése alapján meg tudjuk mondani, hogy milyen kőzetek fordulnak elő az adott mélységtartományban, azok milyen rétegtartalmúak (van-e szénhidrogén) és mekkora a porozitásuk (ld. a litológiai vonalakat a hézagtérfogat szerint skálázva). Az ábrán a harmadik változó (*GR*) a természetes gamma intenzitás, mely üledékes rétegsorban az agyagtartalomra érzékeny mennyiség.



52. ábra *Neutron-sűrűség crossplot*
(MOL Nyrt. jóvoltából)

Tekintsük a csoportok jellemző tulajdonságait! A klaszterjellemzők közül az **átmérő** a csoport két legtávolabbi elemének a távolságát adja meg. A **súlypontvektor** a klaszter középpontjához húzott helyvektor, mely a csoport helyét adja meg a térben. A **sugár** a csoport súlypontja és legtávolabbi elemének a távolsága. A **centroid**ot, azaz a *K*-adik csoport c_K súlypontját a csoport elemeinek számtani átlagaként értelmezzük

$$c_K = \frac{1}{m_K}\sum_{i=1}^{m_K} x_i^{(K)},$$

ahol m_K a csoport elemszáma. A csoportképzést az objektumoknak a változók száma által meghatározott dimenziójú térben való elhelyezkedése alapján hajtjuk végre. Többváltozós probléma esetén a különböző fizikai elv felhasználásával mért adatokat egyetlen vektorba soroljuk. Legyen két objektumunk, melyet az *n*-dimenziós adattérben $\vec{x} = [x_1, x_2, \ldots, x_n]^T$ és $\vec{y} = [y_1, y_2, \ldots, y_n]^T$ vektorok képviselnek (*n* a mérőberendezések száma). A két objektum közötti távolságot többféleképpen definiálhatjuk. Például a **Minkowski-távolság** alatt a fenti két vektor különbségének az L_p-normáját értjük (ld. 3. fejezet)

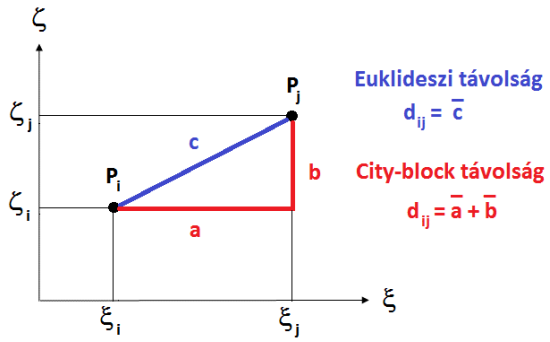
$$d(\vec{x}, \vec{y}) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}.$$

Az **Euklideszi-távolság** ebben az esetben is a különbségvektor L_2-normájával (ld. 3. fejezet) egyezik meg

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Az 53. ábrán látható az Euklideszi-távolság geometriai jelentése, mely koordináta-geometriából mindenki számára ismert. Az ún. **City-block (Manhattan) távolság** értelmezését ugyancsak tartalmazza az ábra, mely két változó esetén a különbségvektor L_1-normájaként számítható (ld. 3. fejezet)

$$d(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$



53. ábra *Az Euklideszi és a City-block távolság értelmezése*

A fenti távolság definícióknál azt feltételeztük, hogy a mért változók függetlenek egymástól. Ha azok korreláltak, akkor célszerű a **Mahalanobis-távolság**ot alkalmazni

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \underline{\underline{COV}}^{-1}(\vec{x}-\vec{y})}.$$

A fenti kifejezésben a kovariancia mátrix (főátlójában a szórásnégyzetekkel) inverze, normálási tényezőként jelenik meg. A normálást súlyozásként foghatjuk fel, mellyel az adatok korreláltságának hatását kívánjuk csökkenteni. Abban az esetben, amikor a kovariancia-mátrix egységmátrix (a változók függetlenek), akkor a Mahalanobis-távolság az Euklideszi távolságot adja vissza. A fenti távolság definíció alkalmazása akkor előnyös, amikor a változók nagyságrendje és dimenziója különböző. Ekkor ui. a távolságok nem összemérhetők. A klaszterelemek páronkénti távolság értékeit egy $m \times m$-es mátrixba rendezhetjük (ahol m az objektumok teljes száma). Az így kialakított mennyiséget **távolságmátrix**nak nevezzük

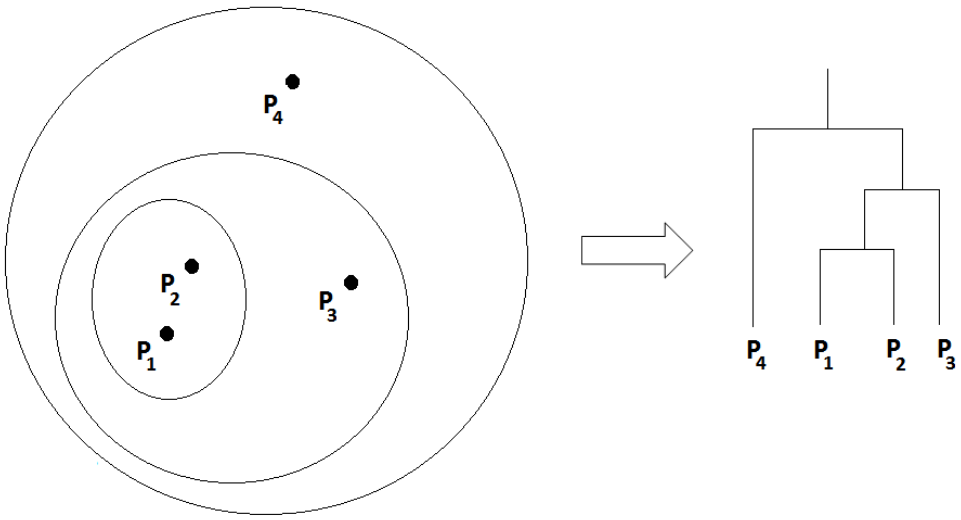
$$\underline{\underline{D}} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1m} \\ d_{21} & 0 & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & 0 \end{pmatrix},$$

melyben a d_{ij} elem megadja az i-edik és j-edik adatpont közötti távolságot. A klaszterelemzés során arra törekszünk, hogy a csoporton belüli elemek között a távolság minimális, de a csoportok közötti távolság maximális legyen.

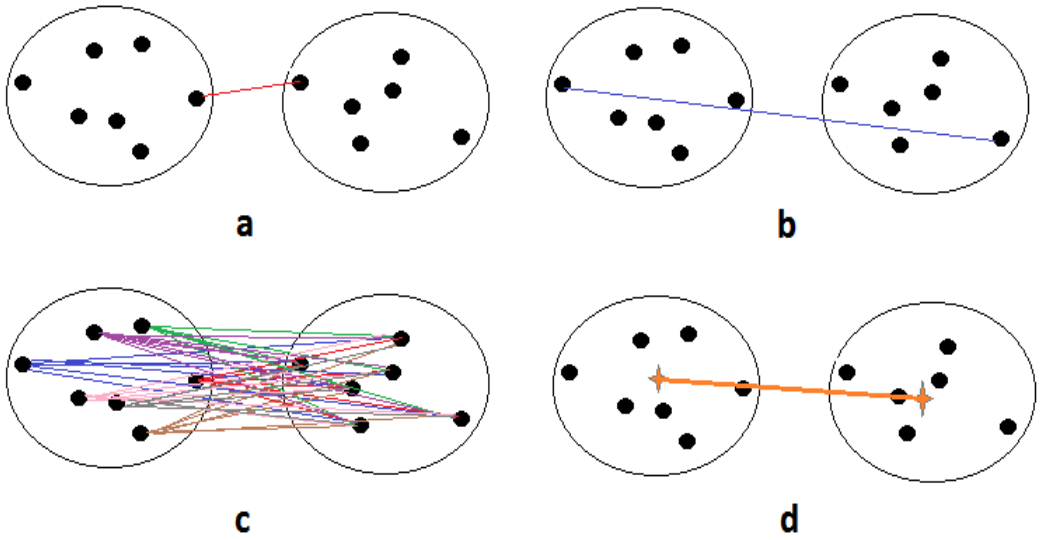
A klaszterelemző eljárásokat két csoportra osztjuk: hierarchikus és nem hierarchikus eljárásokra. Az egymásba ágyazott, ún. **hierarchikus klaszterezés** előnye az, hogy nem kell előre ismernünk a létrehozandó klaszterek számát. Hátránya az időigényesség, ezért csak kis mintaelem szám esetén használjuk őket (ui. tárolni kell a távolságmátrix elemeit). A hierarchikus **agglomeratív eljárás** esetén kezdetben az elemszámmal megegyező számú (m db egyelemű) klaszterünk van. Első lépésben kiszámítjuk a $\underline{\underline{D}}$ távolságmátrixot, majd a két egymáshoz legközelebb álló klasztert egyesítjük. Ez által eggyel csökken a klaszterek száma. Mivel a korábban egyesített klaszterek együtt maradnak, ily módon minden lépésben eggyel csökken a klaszterek száma. A távolságmátrixot is minden lépésben újra kell számítani. Az eljárás végén egy klaszter marad, amely az összes elemet tartalmazza (ld. 54. ábra).

A csoportok egyesítése során különböző módon értelmezhetjük a klaszterek közötti hasonlóságot. Nézzük az 55. ábrát! **Egyszerű lánc módszer** *(Simple Linkage)* alkalmazása esetén a csoportok legközelebbi elemeinek a távolságát vizsgáljuk. Ezzel ellentétes a **teljes lánc módszer** *(Complete Linkage)*, mellyel a legtávolabbi elemek távolságát számítjuk. A **csoportátlag módszer**nél *(Average Linkage)* a két csoport összes eleme közötti távolságok átlagát tekintjük alapul. A **súlypont módszer**nél *(Centroid Linkage)* a csoportok súlypontjainak (centroidok) távolságát vizsgáljuk. Végül a **Ward-módszer**rel *(Ward Linkage)* a csoporton belüli (x_i-c_g) eltérések négyzetösszegét (szórás) minimalizáljuk (ahol c_g a g-edik csoport súlypontja és i=1,2,…,m_g).

A hierarchikus klaszterező eljárás az adatelemeket egy jellegzetes fastruktúrába rendezi, melyet **dendrogram**nak nevezünk. Az 54. ábrán látható, hogy a fa minden belső ága megfelel egy-egy klaszternek, melynek végein találhatók az összetartozó csoportelemek. A fa az elemek összetartozását és egymáshoz való viszonyát (hierarchiáját) szemlélteti, viszont nem alkalmas a csoportok térbeli elhelyezkedésének a szemléltetésére. A fastruktúrán jól követhetők a klaszterezés egyes lépései. A számítási példáknál a dendrogram vízszintes tengelyén az adatok sorszáma szerepel az összekapcsolódás sorrendjében, a függőleges tengelyen pedig a centroidok (csoportközpontok) közötti távolság értékek vannak feltüntetve.



54. ábra *A hierarchikus klaszterezés sémája és a dendrogram*

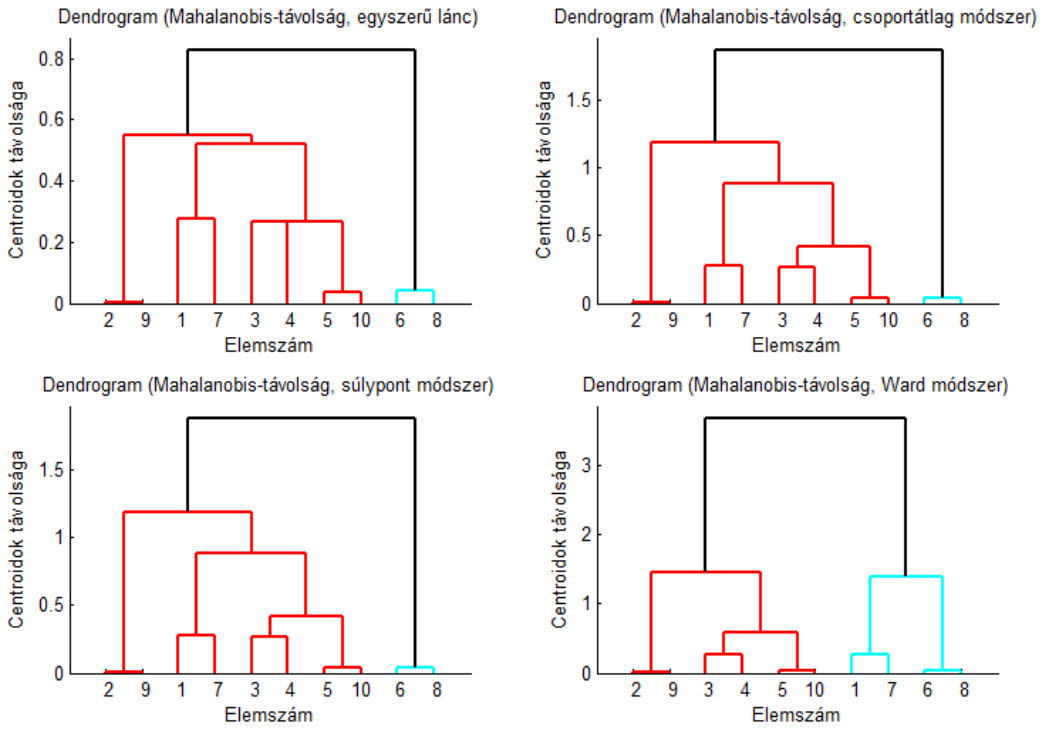


55. ábra *Klaszterek hasonlósági definíciói*
(a: egyszerű lánc, b: teljes lánc, c: csoportátlag, d: súlypont)

Példa. Hierarchikus klaszterezésre alkalmas MATLAB rendszerben a **dendrogram** függvény, mely megkívánja az adatok távolságát számító **pdist** és a fastruktúrát a megadott hasonlósági definíció alapján létrehozó **linkage** függvény előzetes használatát. (Az ábrát maga a dendrogram függvény hozza létre). Nézzünk egy egyszerű példát! Egy 10 elemű véletlen adatsort generáltunk (ld. 4. táblázat), melynek a csoportosítását négy különböző hasonlósági mérték alkalmazásával is elvégeztük! A távolság számításánál a Mahalanobis-formulát alkalmaztuk. Az 56. ábrán látható, hogy először a legközelebbi elemek összevonása történt meg (ld. 2-es és 9-es sorszámú elemek), majd később a távoli elempárokat is összevontuk. A négy különböző módszerből három ugyanazt a lépéssorozatot hajtotta végre, míg a Ward-módszer egy lépésben eltért (ld. 1-7 és 6-8 elempár összevonása).

4. táblázat

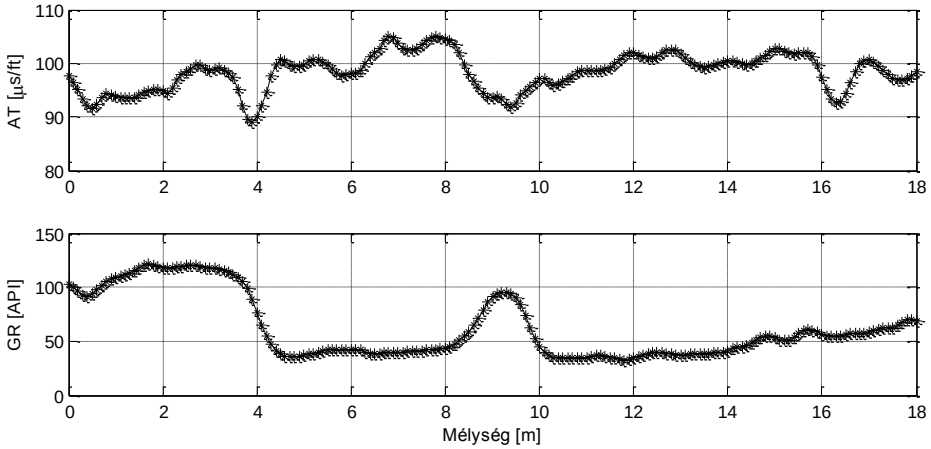
Adat sorszáma	1	2	3	4	5	6	7	8	9	10
Mért érték	6.81	2.34	4.56	3.85	5.39	9.92	7.55	9.80	2.35	5.29



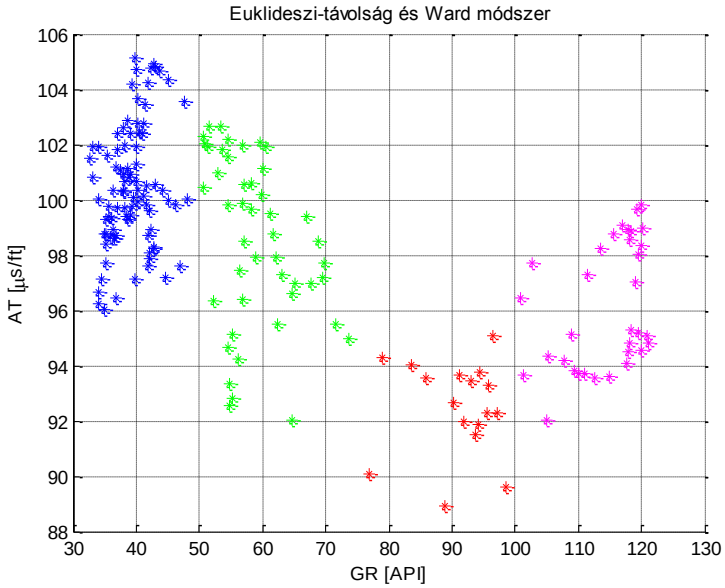
56. ábra *Hierarchikus klaszterezés eredménye 10 elemű véletlen minta esetén*

Példa. A klaszterelemzés igen jól alkalmazható fúrási geofizikai szelvényadatok csoportosítására. A MATLAB rendszerben található **cluster** függvény számára meg kell adnunk a maximális (kialakítandó) klaszterszámot. Az eljárás úgy végez hierarchikus csoportosítást, hogy a fát elvágja ennél az előírt klaszterszámnál, és az így kialakított csoportosítást tekinti végeredménynek. Az 57. ábrán egy négyréteges (konszolidálatlan) agyag-homok szénhidrogén-tároló szerkezetben mért akusztikus terjedési idő és természetes gamma szelvények láthatók. A klaszteranalízis bemenő (diszkrét) adatsorának elemeit a szelvényeken csillaggal jelöltük. Mivel a két szelvény csak gyengén korrelált és nem

tartalmaztak kiugró adatot, ezért az Euklideszi távolságot vettük alapul. A maximális klaszterszámot négynek választottuk. A Ward-módszeren alapuló csoportosítás eredménye az 58. ábrán látható. A crossploton látható, hogy a kialakított csoportoknak kőzettani jelentése van. Például a kék színnel jelölt klaszter elemei a homokhoz, a rózsaszínnel jelöltek az agyaghoz tartoznak. A zöld és piros csoportok adatait eltérő agyag-, ill. kőzetliszt-tartalmú rétegekben mérték. Ezt az eredményt fel lehet használni a fúrási adatok kőzetfizikai értelmezése során, mivel az egyes kőzettípusokhoz jellemző mérési értékeket tudunk rendelni (pl. homokrétegek esetén *GR*≈40API, *AT*≈100µs/ft, valamint „tiszta" agyagoknál *GR*≈110API, *AT*≈96µs/ft), mely fontos a priori információt jelent az inverz modellezés során (ld. 12. fejezet).



57. ábra *A természetes gamma és akusztikus terjedési idő szelvény*



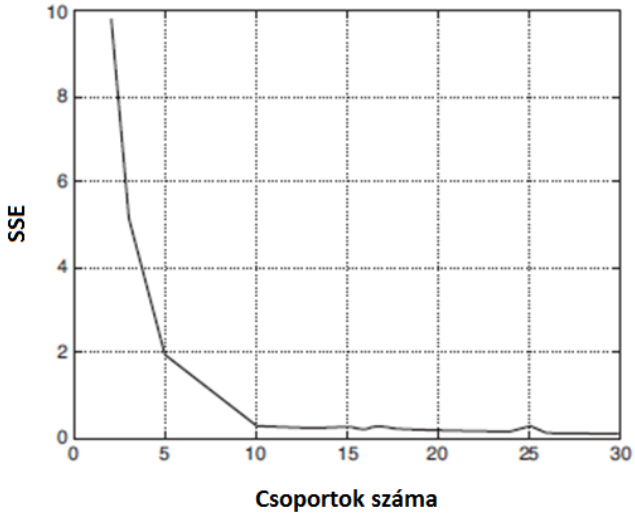
58. ábra *Természetes gamma és akusztikus terjedési idő adatok csoportosítása*

A klaszterező eljárások másik csoportját alkotó partícionáló vagy más néven **nemhierarchikus klaszterezési** módszerek fő jellemzője, hogy előre meg kell adnunk a

kialakítandó klaszterszámot. Az optimális csoportszám megadása céljából számítsuk ki az összes elem hozzá legközelebb eső centroidtól mért távolságának négyzetösszegét. Az így kapott mennyiséget *SSE*-vel *(Sum of Squared Error)* jelöljük

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} d^2(c_i, x_j).$$

Az *SSE* a szóródás mérőszáma, mely a klaszterek számának növekedésével aszimptotikusan csökken (ld. 59. ábra). Ez alapján azt mondhatjuk, hogy az optimális klaszterszámot úgy kell megválasztanunk, hogy az relatíve kis szám legyen és hozzá kis *SSE* érték tartozzon.



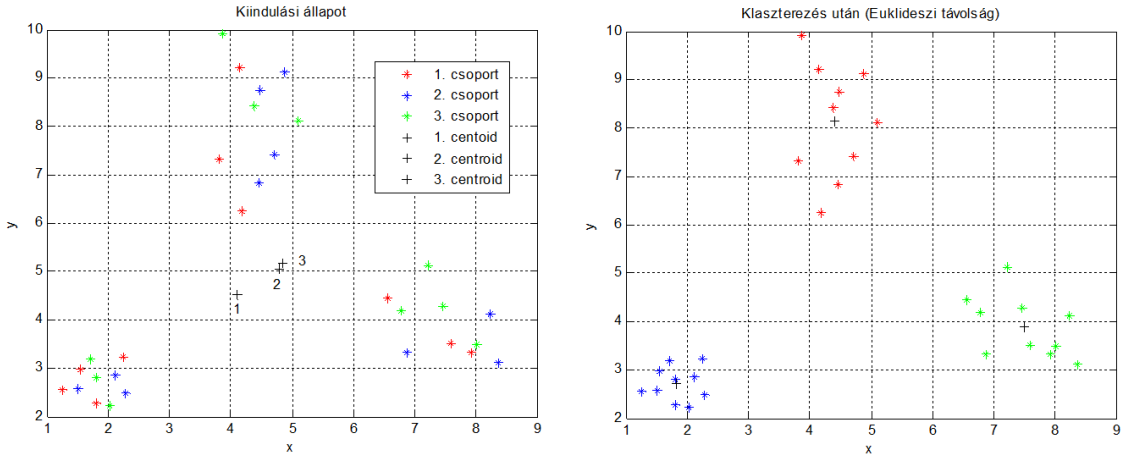
59. ábra *A pontok és centroidok „távolságának" változása a klaszterek számával*

Az iterációs elven alapuló nemhierarchikus klaszterezési eljárások gyorsak, viszont meglehetősen zajérzékenyek, és az eredményt nagymértékben befolyásolja a centroidok kezdeti megadása. Közülük a legelterjedtebb módszer a **K-középpontú klaszterezés**. Válasszuk ki a klaszterek számát és *K* számú kezdő centroidot! Alakítsunk ki *K* számú csoportot úgy, hogy minden egyes elemet soroljunk a hozzá legközelebb eső centroidhoz tartozó klaszterbe! Számoljuk ki az új klaszter középpontokat! A fenti lépéseket stopkritérium teljesüléséig ismételjük!

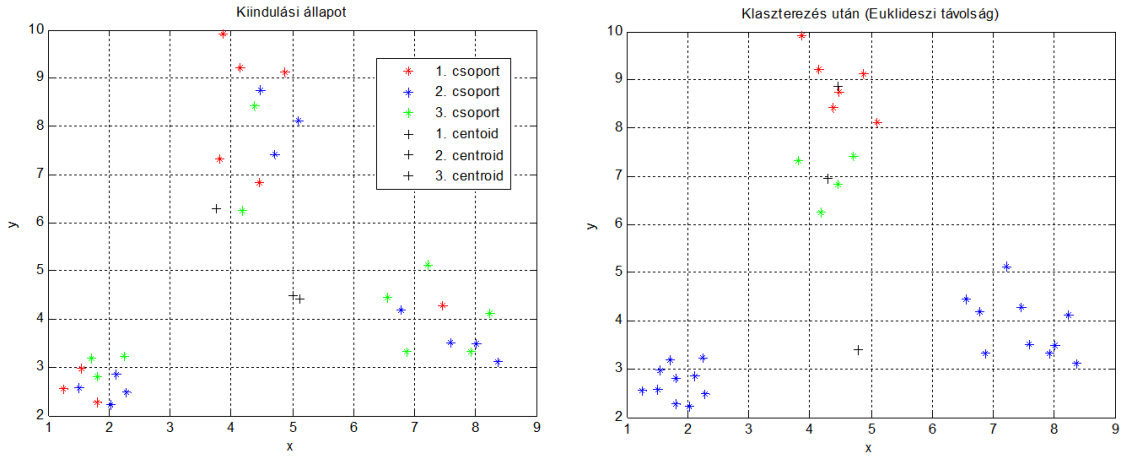
Példa. *K*-középpontú klaszterezést végeztünk szintetikusan generált adatok csoportosítása céljából. A 30 objektum alapján az előírt (maximális) csoportszám 3 volt. Az eljárás ennek megfelelően alakította ki a kezdeti klasztereket és kiszámította a csoport-középpontokat! A 60. és 61. ábrát összehasonlítva azt látjuk, hogy ugyanazt az adatrendszert lényegesen eltérő módon csoportosította a *K*-középpontú klaszterező eljárás. Első esetben, az elemeket kezdetben egyenletesen osztottuk el, így a centroidok középre estek (ld. 60. ábra). Második esetben viszont, az első csoport (7.5,4.3) piros ponttal jelölt eleme mintegy kiugró adatot képezve eltolta a kezdő centroidot (ld. 61. ábra). Első esetben optimális megoldást, a másodikban rossz megoldást kaptunk (az 1. és 3. csoport elemeit összekeverte az eljárás). Tudjuk, hogy az L_2-normán alapuló eljárások a kiugró adatok jelenlétére igen érzékenyek, így az eredmény nem volt váratlan.

12. A lineáris inverz feladat megoldása

A terepen gyakran előfordul, hogy közvetlenül nem mérhető változókról szeretnénk információt szerezni (pl. víztelítettség, porozitás, szeizmikus hullámterjedési sebesség). Ha ezek a mennyiségek kapcsolatba hozhatók egyéb mérhető változókkal (pl. fajlagos ellenállás, neutron beütésszám, szeizmikus futási idő), akkor azokból módunkban áll leszármaztatni őket. E feladat megoldásának első lépése a modellalkotás. A **modell** a vizsgált objektumok tulajdonságait kvantitatív módon írja le, úgy, hogy bizonyos tulajdonságokat elhanyagol és a lényeges vonások megtartásával a valóságot egyszerűbb formában kezeli. A modellt kőzetfizikai és geometriai paraméterek alkotják, melyek egy-, két- vagy háromdimenziósak (1D, 2D, 3D) lehetnek. A dimenziószámot a független geometriai változók száma határozza meg. Manapság már négydimenziós modellekről is beszélünk, ahol a negyedik változó az idő. A 4D problémák adatait az időben ismételt (monitoring) mérések szolgáltatják, melyek gyakoriak pl. szénhidrogén-tárolók telítettségi viszonyainak felmérése során (ennek az a célja, hogy a kitermelés üteméről vagy az esetleges kúthibákról információhoz jussunk).



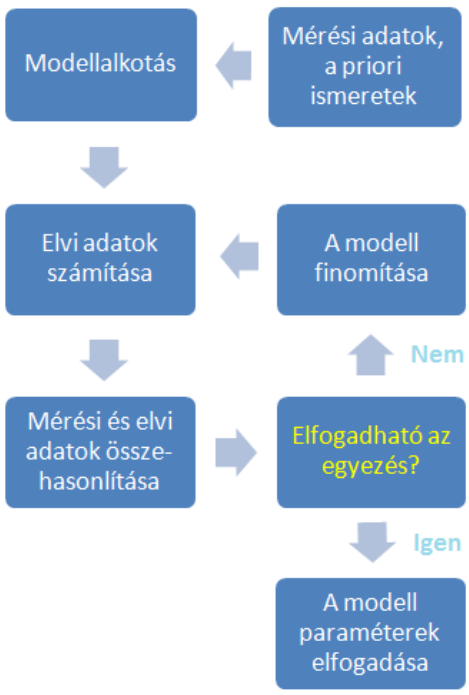
60. ábra *Megfelelően megválasztott kezdeti csoportközpontok*



61. ábra *Nem megfelelően megválasztott kezdeti csoportközpontok*

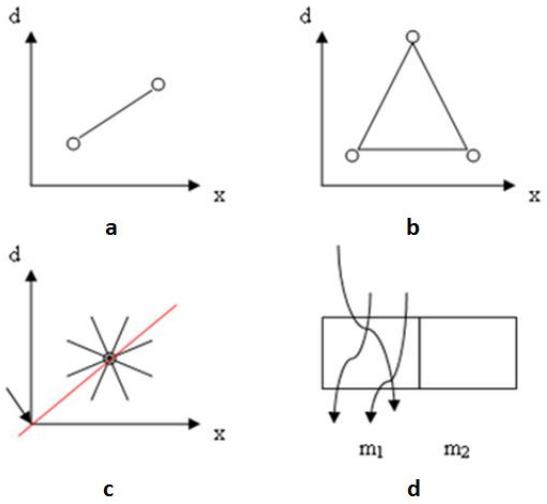
A modell paramétereinek meghatározása érdekében méréseket végzünk a vizsgált földtani objektumokon. Az adatok és a modellparaméterek kapcsolatát leíró összefüggéseket **válaszfüggvény**eknek nevezzük. Ezek ismeretében, a **direkt feladat** (előremodellezés) keretében a modellen elvi adatokat számíthatunk. Más esetben, amikor a mérési adatok ismeretében becsüljük meg az (ismeretlen) modellparaméterek értékeit, akkor az **inverz feladat** megoldásáról beszélünk. Az inverziós eljárás tehát olyan adatfeldolgozási (értelmezési) procedúra, mely a mérések ismeretében határozza meg a földtani objektumok lényeges tulajdonságait. Matematikai szempontból nézve az inverzió olyan optimalizációs eljárás, mely a mért és számított adatok illesztésével állítja elő a földtani valóságnak leginkább megfelelő (optimális) modellt.

Az inverziós eljárás folyamatát a 62. ábra mutatja. Vegyünk a geofizikából egy példát és kövessük az egyes lépéseket! Egy földtani objektum (pl. érctest vagy akár egy üreg) felkutatása érdekében gravitációs méréseket végzünk. A gravitációs anomália és a terület geológiájának ismeretében megalkotjuk a legvalószínűbb sűrűség modellt. A potenciál-elméletből levezetett összefüggések (válaszfüggvények) alapján a hatóra elvi gravitációs adatsort számítunk. Az elvi és a mérési adatsor egyezését vizsgálva kiderül, hogy az általunk feltételezett modell mennyire felel meg a valóságnak. Amíg rossz az elméleti és a mérési adatok egyezése, addig a modellparamétereket (sűrűség értékeket) javítanunk kell. Ez minden iterációs lépésben a direkt feladat (újra) számítását igényli, mivel azzal állíthatjuk elő az aktuális modellre vonatkozó elméleti adatsort. A fenti lépéseket addig ismételjük, míg egy előre megadott kilépési feltétel (előírt pontosság) nem teljesül. Az inverz feladat megoldásának az utolsó lépésben elfogadott modellt tekintjük, melynek paraméterei (jelen esetben a ható helyzete, mélysége, kiterjedése, valamint a ható és környezetének sűrűségkülönbsége) alapján feltérképezhetjük a földtani szerkezetet.



62. ábra *Az inverziós eljárás folyamatábrája*

Az N számú független adat és M számú ismeretlen (modellparaméter) aránya alapján az inverz problémák négy típusba sorolhatók. Nézzük a 63. ábrát! Lineáris regressziós feladat esetén az ismeretlenek száma 2 (regressziós egyenes meredeksége és ordináta-metszete). Abban az esetben, amikor éppen 2 adat áll rendelkezésünkre (ld. 63a ábra), akkor a feladat **egyértelműen meghatározott**. Ekkor az adatok és ismeretlenek száma megegyezik ($N=M$), így az inverz probléma algebrai úton (egyértelműen) megoldható. Amikor az adatszám nagyobb, mint az ismeretlenek száma ($N>M$), akkor **túlhatározott** feladatról beszélünk, melynek nincs egyértelmű (csak közelítő) megoldása (ld. 63b ábra). Általában ez a probléma jól kezelhető, mivel a mérési adatszám növelésével a becslés pontossága (az adatokat terhelő zaj mértékétől függően) növelhető. **Alulhatározott** probléma esetén kevesebb adatunk van, mint ismeretlenünk ($N<M$), és végtelen számú egyenértékű (ekvivalens) megoldás lehetséges. A megfelelő megoldás kiválasztása a priori információ (ismeret a földtani objektumról egyenletben megfogalmazva) bevonásával lehetséges. Például a 63c ábrán előzetesen előírjuk, hogy a megoldásnak egy origón átmenő egyenesnek kell lennie. Ettől az inverz probléma egyértelműen meghatározottá válik. Végül a legnehezebben kezelhető esetet a **kevert határozottság**ú feladat képezi, mely részben túlhatározott, részben pedig alulhatározott. A 63d ábrán látható, hogy egy tomográfiai probléma objektumának egyik cellájában több sugár (adat) is áthalad, egy másikban pedig egyetlen egy sem. Ekkor m_1 ismeretlenre (modell paraméter) nézve az inverz feladat túlhatározott, viszont m_2 nézve alulhatározott.



63. ábra *Az inverz feladat típusai*

Rendezzük az inverz feladat M számú modell-paraméterét az $\vec{m} = [m_1, m_2, \ldots, m_M]^T$ **modellvektor**ba, az N számú mérési adatot pedig a $\vec{d} = [d_1, d_2, \ldots, d_N]^T$ **adatvektor**ba! Az adatok és a modellparaméterek közötti kapcsolat felírható

$$\vec{\mathrm{d}} = \vec{\mathrm{g}}(\vec{\mathrm{m}})$$

mely általános esetben egy nemlineáris vektor-vektor függvény. A **lineáris (linearizált) inverziós eljárás**ok a nemlineáris inverz feladatot lineáris problémák sorozatára vezetik vissza. Az iterációs eljárás keretében a modelltér egy megoldáshoz közeli pontjából indítjuk az

eljárást (melyet a priori információk alapján határozunk meg), majd a modellt az alábbi formula alapján lépésenként javítjuk

$$\vec{m} = \vec{m}_0 + \delta\vec{m}$$

ahol $\vec{m}_0$ az ún. kezdeti- (start, később az előző lépésbeli) modell és $\delta\vec{m}$ a modellkorrekció-vektor. Alkalmazzunk Taylor-sorfejtést a startmodell környezetében és hanyagoljuk el a magasabb rendű deriváltakat (linearizáljunk!). Ekkor a *k*-adik számított adat

$$d_k(\vec{m}) = g_k(\vec{m}^{(0)}) + \sum_{i=1}^{M} \left.\frac{\partial g_k}{\partial m_i}\right|_{\vec{m}_0} \delta m_i \quad \text{ahol} \quad k = 1,2,\ldots,N.$$

Az egyenlet jobb oldalának első tagját jelöljük $d_k^{(0)}$-al, és vezessük be az $\delta d_k = d_k - d_k^{(0)}$ ill. $G_{ki} = (\partial g_k / \partial m_i)_{\vec{m}_0}$ jelöléseket! Az $N \times M$ méretű $\underline{\underline{G}}$ mátrixot **érzékenységi- (Jacobi) mátrix**nak nevezzük. Az érzékenységi mátrix a (számított) adatok modellparaméterek szerinti parciális (gyakorlatban numerikus) deriváltjait tartalmazza

$$\underline{\underline{G}} = \begin{pmatrix} \frac{\partial d_1}{\partial m_1} & \frac{\partial d_1}{\partial m_2} & \cdots & \frac{\partial d_1}{\partial m_M} \\ \frac{\partial d_2}{\partial m_1} & \frac{\partial d_2}{\partial m_2} & \cdots & \frac{\partial d_2}{\partial m_M} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial d_N}{\partial m_1} & \frac{\partial d_N}{\partial m_2} & \cdots & \frac{\partial d_N}{\partial m_M} \end{pmatrix}.$$

Mivel a $\underline{\underline{G}}$ mátrix független a modellkorrekció-vektortól, ezért az adatok és az ismeretlenek kapcsolata lineáris

$$\delta\vec{d} = \underline{\underline{G}}\delta\vec{m} \quad \text{vagy} \quad \vec{d} = \underline{\underline{G}}\vec{m}.$$

A fenti lineáris egyenletrendszer megoldásával a modell finomítható. Alapfeltevésünk, hogy a mérési adatok mindig tartalmaznak valamilyen mértékű zajt, másrészt a modellezés következtében elhanyagolva a földtani objektumok bizonyos (kevésbé lényeges) tulajdonságait, modellhiba is jelen van. Ez azt jelenti, hogy a mért és a számított adatok eltérését jellemző ún. **eltérés-(hiba) vektor**

$$\vec{e} = \vec{d} - \underline{\underline{G}}\vec{m} \quad \text{azaz} \quad \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} d_1^{(m)} - d_1^{(sz)} \\ d_2^{(m)} - d_2^{(sz)} \\ \vdots \\ d_N^{(m)} - d_N^{(sz)} \end{bmatrix} \neq \vec{0}$$

nem lehet zérus. Az inverz feladatot az $\vec{e}$ eltérésvektor valamely normájának (ld. 3. fejezet) minimalizálásával oldjuk meg. A vektornorma, mint az optimalizációs feladat célfüggvénye egyetlen számot (skalárt) rendel az adatok különbség-vektorához. Az alkalmazott norma

típusa alapján különféle inverziós módszereket hozhatunk létre, melyek megválasztása függ az adatok eloszlásától, a probléma határozottságától, a modell fizikai sajátságaitól stb. Ebben a tananyagban a túlhatározott inverz probléma megoldásával foglalkozunk, mivel a gyakorlatban ez fordul elő a legtöbbször. A többi esetre vonatkozó megoldási módszereket példákkal illusztrálva *Dobróka (2001)* egyetemi jegyzetében találjuk.

A Gauss-féle **legkisebb négyzetek módszere** *(Least SQuares method)* az adatok Gauss-eloszlása esetén ad optimális megoldást. Az optimalizációs feladat E-vel jelölt célfüggvényét, az eltérésvektor L_2-norma négyzete (mért és számított adatok eltéréseinek négyzetösszege) képezi

$$E = \vec{e}^T\vec{e} = [e_1, e_2, \ldots, e_N]\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \sum_{i=1}^{N} e_i^2 = \min .$$

Az inverz feladat megoldása a $\partial E/\partial m_q = 0$ (ahol q=1,2,…,M) szélsőérték-feltételek teljesülése mellett

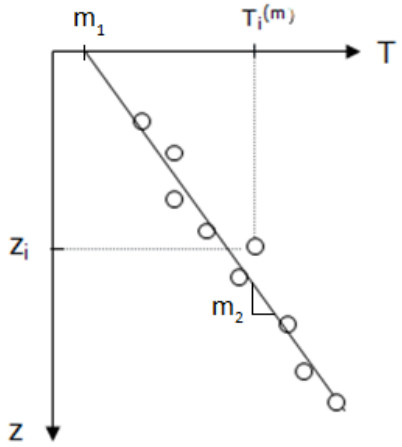
$$\vec{m} = \left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\vec{d}$$

ahol $\vec{m}$ az inverziós eljárással **becsült modell(vektor)**t jelöli. A fenti eredmény részletes levezetése megtalálható *Menke (1984)* könyvében. A továbbiakban részletesen bemutatunk két elméleti és egy gyakorlati példát az LSQ módszer alkalmazására.

Példa. Tételezzük fel, hogy egy adott területen a hőmérséklet-mélység kapcsolat lineáris! Végezzünk hőmérséklet-méréseket egy fúrásban, majd hajtsunk végre lineáris regressziót! A regressziós modell (válaszegyenlet) ennek megfelelően

$$T(z) = m_1 + m_2 z$$

ahol $T(z)$ a z mélységben mért hőmérséklet adat, m_1 és m_2 a regressziós egyenes ordináta-metszete és meredeksége (ld. 64. ábra)



64. ábra *Hőmérséklet adatok lineáris regressziója*

A feladatot oldjuk meg LSQ inverziós eljárással! Az ismeretlen modellparaméterek vektorát két elem alkotja

$$\vec{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

míg az adatvektor *N* számú mérési adatot tartalmaz

$$\vec{d} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_N \end{bmatrix}.$$

Az inverz feladat *N*>>2 esetben nagymértékben túlhatározott, ahol alkalmazhatjuk a legkisebb négyzetek módszerét. A lineáris adat-modell kapcsolat könnyen felírható

$$\vec{d} = \underline{\underline{G}}\vec{m} \quad \text{azaz} \quad \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_N \end{pmatrix} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

ahol a jobb oldali *N*×*2* méretű mátrix megfelel $\underline{\underline{G}}$ érzékenységi mátrixnak. Képezzük az alábbi mátrixszorzatokat

$$\underline{\underline{G}}^T\underline{\underline{G}} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_N \end{pmatrix} \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ & \vdots \\ 1 & z_N \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{pmatrix}$$

$$\underline{\underline{G}}^T\vec{d} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_N \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} T_i \\ \sum_{i=1}^{N} z_i T_i \end{pmatrix}$$

mellyel az inverz (ill. regressziós) feladat megoldása

$$\vec{m} = \left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\vec{d} \quad \text{azaz} \quad \vec{m} = \begin{pmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} T_i \\ \sum_{i=1}^{N} z_i T_i \end{pmatrix}.$$

Példa. A fenti feladatot terjesszük ki kétdimenziós esetre is! Jelöljék az *x* és *y* független változók a mérések helyének koordinátáit és *d* valamilyen fizikai változót! Tételezzük fel, hogy a *d* mennyiség mind az *x*, mind pedig az *y* irányban lineárisan változik. A lineáris regresszió eredménye ebben az esetben egy sík (regressziós sík) lesz, melynek keressük az

egyenletét (ld. 65. ábra). Oldjuk meg a lineáris feladatot a *d* adatrendszer inverziójával! A megoldást a legkisebb négyzetek módszerével keressük. A regressziós sík egyenlete a következő

$$d(x,y) = m_1 + m_2 x + m_3 y$$

ahol az ismeretlen modellparaméterek vektora

$$\vec{m} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}$$

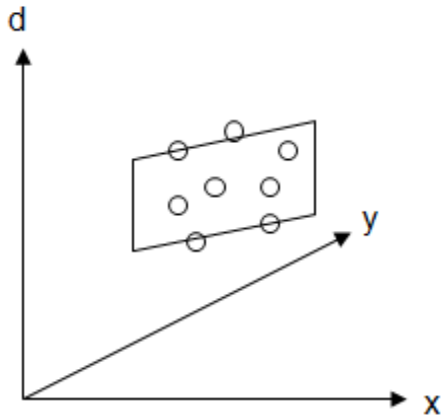
és az *N* számú mérési adat vektora

$$\vec{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix}.$$

A lineáris adat-modell kapcsolat ebben az esetben

$$\vec{d} = \underline{\underline{G}}\vec{m} \quad \text{azaz} \quad \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$$

ahol $\underline{\underline{G}}$ érzékenységi mátrix mérete $N \times 3$. Látható, hogy a $\underline{\underline{G}}$ mátrixban nem szerepel egyetlen modellparaméter sem, így az csak a probléma geometriájától függ. Ez nagyobb méretű (nagyszámú adat és ismeretlen) inverz problémák esetén is így van. A mátrixelemeket általában a $\Delta d/\Delta m$ numerikus deriváltak számításával származtatjuk, ebben az egyszerű esetben viszont nem szükséges a deriválás, mivel $\underline{\underline{G}}$ a koordinátákkal közvetlenül felírható.



65. ábra *Kétdimenziós lineáris regressziós feladat*

Felírva a megfelelő mátrixszorzatokat

$$\underline{\underline{G}}^T\underline{\underline{G}} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{pmatrix} \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^N x_i & \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N y_i^2 \end{pmatrix}$$

$$\underline{\underline{G}}^T\vec{d} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N d_i \\ \sum_{i=1}^N x_i d_i \\ \sum_{i=1}^N y_i d_i \end{pmatrix}$$

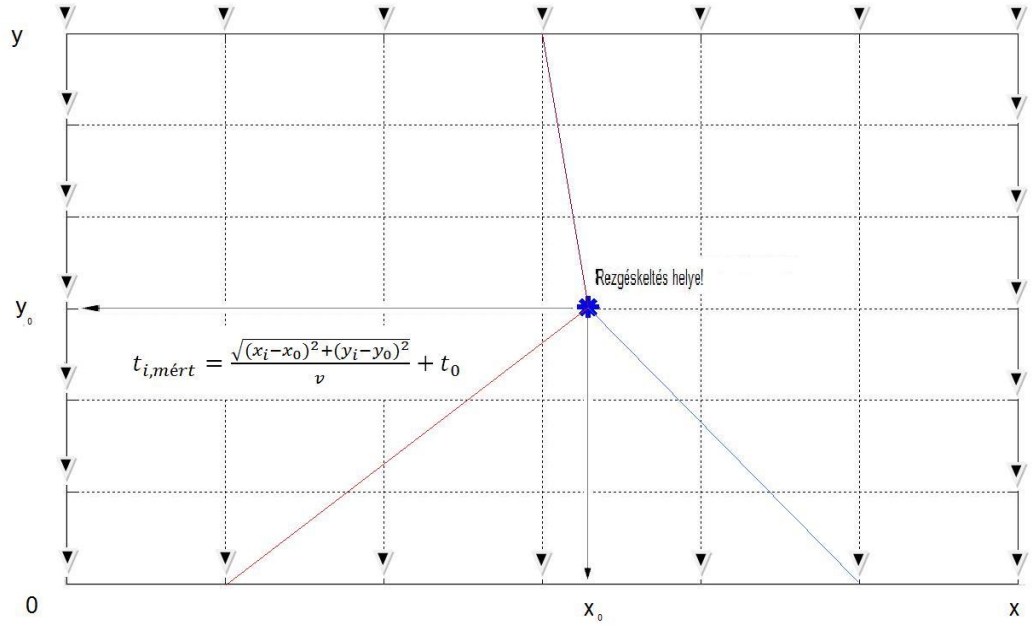
az inverz feladat megoldása

$$\vec{m} = \left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\vec{d} \quad \text{azaz} \quad \vec{m} = \begin{pmatrix} N & \sum_{i=1}^N x_i & \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N y_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^N d_i \\ \sum_{i=1}^N x_i d_i \\ \sum_{i=1}^N y_i d_i \end{pmatrix}.$$

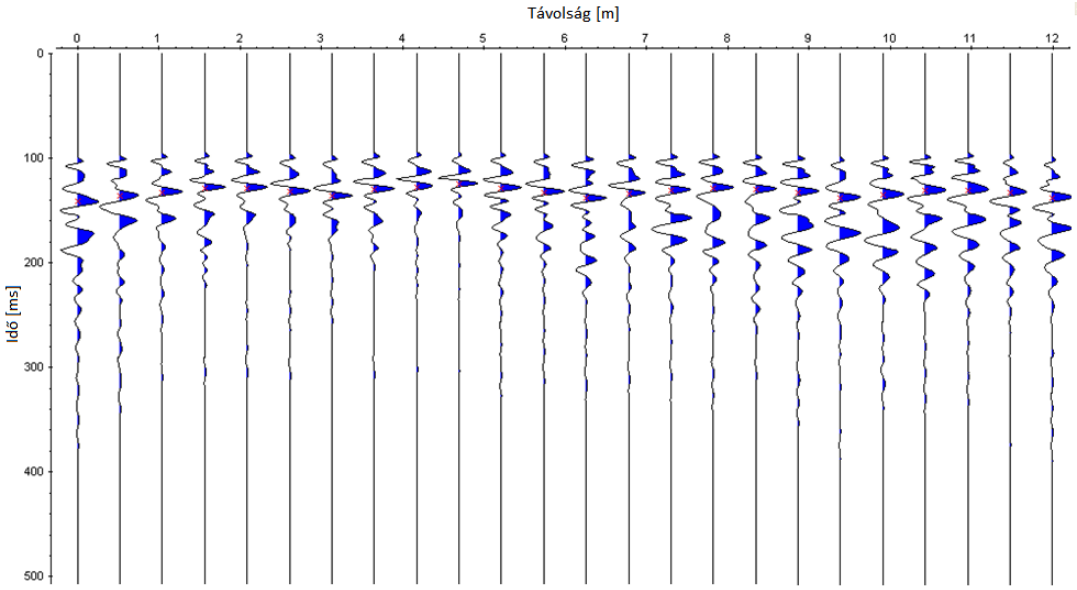
Példa. A szeizmikus mérések nagymennyiségű információt hordoznak a felszín alatti régiók (teljes litoszférát beleértve) földtani objektumairól. Az ásványi nyersanyag- és szénhidrogén-kutatáson kívül a módszer alkalmas a földrengések kipattanási helyének meghatározására. Egy kísérleti terepi mérés eredményét mutatjuk be, ahol *Ormos T. és Szabó N.P. (2010)* kisléptékben modellezte a földrengéseket detektáló obszervatóriumok működését. A szimuláció során az obszervatóriumokat szeizmikus érzékelőkkel (geofon) helyettesítettük, melyek a „rengés” által keltett hullámokat detektálták. A hullámok beérkezési ideje a rezgéskeltés és a geofonok helyének függvényében változik. A direkt feladat megoldásához elegendő az „út-idő-sebesség” összefüggés alkalmazása az adott mérési elrendezés esetén. Az időadatok inverziós feldolgozása lehetővé teszi, hogy a robbantás helyét kijelöljük. A kis területen végzett kísérletnél ellenőrizni lehetett a rezgéskeltés számított koordinátáit, viszont a földgolyót átszelő földrengések esetén ez nem lehetséges (ezért végzünk inverziót).

A 66. ábrán látható, hogy a vizsgált területet (amit tekinthetünk a Föld egy nagykiterjedésű régiójának) 2m-es lépésközzel 24db szeizmométerrel (jelképesen földrengésvizsgáló obszervatóriummal) vettük körül. Az egyszerűség kedvéért a direkt hullám beérkezési időket vettük alapul a feldolgozásnál. A terület érdekessége az volt, hogy nagyon laza talajon mértünk, melynek hullámterjedési sebessége $\approx 250 ms^{-1}$ volt, ami kisebb, mint a hangsebesség ($\approx 330 ms^{-1}$). A mérési adatokat ábrázoló szeizmogramon (ami a távolság függvényében ábrázolja a beérkezési időket) ezért nem az első, hanem a második beérkezéshez tartozó

időket jelöltük ki. A 67. ábrán piros „x” jelek mutatják a kijelölt beérkezési időket. Az x_0=6.6m, y_0=6m helyen keltett rezgéshez tartozó adatrendszert az 5. táblázat tartalmazza. A futási idők és geofon pozíciók kapcsolatát a 66. ábrán feltüntetett képlet adja meg, mellyel az i-edik geofon (x_i, y_i) koordinátái, a sebesség (v) és a regisztrálás kezdetétől függő időeltolódás (t_0) ismeretében ki tudjuk számítani a robbantás (x_0, y_0) koordinátáit. Ez jelenti a direkt feladat megoldását. Ha minden geofonra elvégezzük ezt a számítást, akkor 24db (elvi) időadatot kapunk, amit az inverziós eljárásban összehasonlítunk a 24db mért időadattal.



66. ábra *A szeizmikus mérés geometriája*



67. ábra *Az x_0=6.6m, y_0=6m helyen keltett rezgéshez tartozó szeizmogram*

Fogalmazzuk meg a fenti inverz feladatot! Az adatvektorban tárolt 24db mért beérkezési időadat ismeretében

$$\vec{d} = [t_1, t_2, \ldots, t_{24}]^T$$

becslést végzünk az alábbi modellvektor elemeire

$$\vec{m} = [x_0, y_0, v, t_0]^T.$$

5. táblázat

Geofon	x(m)	y(m)	t(ms)
1	0	0	141.5
2	2	0	136.4
3	4	0	132.5
4	6	0	128.5
5	8	0	128.5
6	10	0	131.9
7	12	0	136.4
8	12	2	130.2
9	12	4	127.4
10	12	6	125.1
11	12	8	129.1
12	12	10	133.0
13	12	12	138.1
14	10	12	133.6
15	8	12	130.8
16	6	12	129.1
17	4	12	130.2
18	2	12	132.5
19	0	12	137.6
20	0	10	135.3
21	0	8	131.3
22	0	6	130.8
23	0	4	133.0
24	0	2	138.1

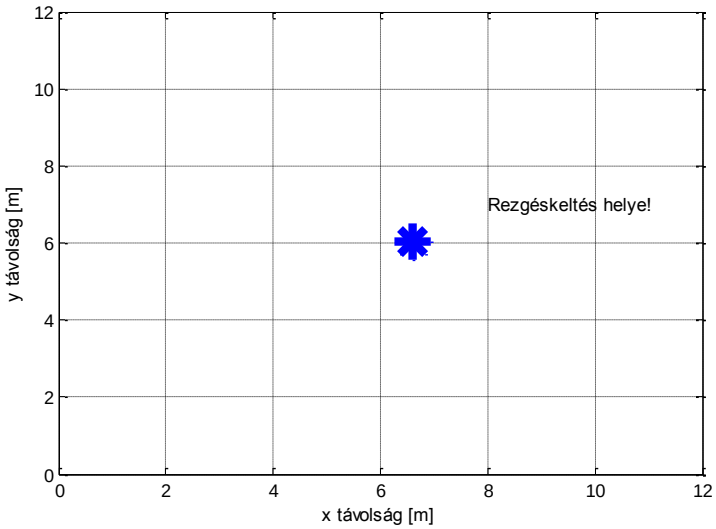
Az inverz probléma nagymértékben túlhatározott volt, mivel 24 adatból négy modell-paramétert kellett meghatározni. Az LSQ eljárás kezdeti modelljét x_0=1m, y_0=5m, v=200ms^{-1}, t_0=300ms paraméterekkel adtuk meg. Az inverziós eljárás a számítógépes futtatás során numerikusan stabilnak bizonyult, és az adatokat terhelő zaj ellenére is pontos (ld. 7. táblázat) megoldást adott: x_0=6.59m, y_0=6.03m, v=253ms^{-1}, t_0=105ms (ld. 68. ábra). A módszer megbízhatóságát az is alátámasztja, hogy azokban az esetekben, amikor a geofonokkal körbevett területen kívül „rengettük meg a Földet”, akkor is visszakaptuk a helyes (negatív előjelű) koordinátákat. Ez a szimuláció nyilván sok közelítést tartalmaz, hiszen a valóságban a hullámok sugárútjai görbültek, a Föld sebességeloszlása sem homogén, sem pedig izotróp, és a Descartes-féle koordináta-rendszert is ritkán alkalmazzák az efféle feladatok megoldásánál. Az inverz probléma is sokkal összetettebb, mivel több adat és ismeretlen képezi, valamint az adatok Gausstól eltérő eloszlása és a kiugró adatok jelenléte miatt robusztus és numerikusan stabilabb (regularizált) inverziós módszerek szükségesek.

A robusztusság felé vezető út első lépését a súlyozott inverziós módszerek bevezetése jelentette. Ismeretes, hogy az inverzióba bevont adatok pontossága (megbízhatósága) eltérő. Ha előzetes információval rendelkezünk az (egyedi) adatok megbízhatóságáról, akkor azt érdemes figyelembe venni az inverz feladat megoldása során. A jobb megoldás érdekében a megbízhatóbb adatoknak nagyobb, míg a kevésbé megbízható adatoknak kisebb súlyt adunk. Konstruáljunk egy olyan $N \times N$-es mátrixot, melynek főátlójában az egyes adatok hibájának megfelelő súlyok szerepeljenek! Az így kialakított $\underline{\underline{W}}^{(d)}$ mennyiséget **adattérbeli súlymátrix**nak nevezzük, mely korrelálatlan adatok esetén diagonális (főátlón kívüli elemei zérusok) mátrix. A legegyszerűbben képzett súlymátrix az, amikor azt mondjuk, hogy pl. a második adat kétszer megbízhatóbb a többinél

$$\underline{\underline{W}}^{(d)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Ha ismerjük az egyes adatok eloszlását, akkor azok hibajellemzőit is bevonhatjuk a súlyozásba (emlékezzünk a 19. ábrára, ahol a sűrűségfüggvény skálaparamétere és a megbízhatóság közötti fordított arányosságot szemléltettük). Például, ha az egyes adatok Gauss eloszlást követnek, akkor az adatok szórásának ismeretében a súlymátrix

$$\underline{\underline{W}}^{(d)} = \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_N^{-2} \end{pmatrix}.$$



68. ábra *Az inverziós eljárással meghatározott koordináták*

Az inverziós eljárás során súlyozhatunk a számított és mért adatok eltérésétől függő mértékben is. Ezt a technikát **iteratívan újrasúlyozás**nak nevezzük, mely a legkisebb négyzetek módszere esetén az *IRLS (Iteratively Reweighted Least Squares)* nevet viseli. Az $\vec{e} = \vec{d}^{(m)} - \vec{d}^{(sz)}$ eltérésvektor L_1-normájával képzett súlyok alkalmazásával rezisztens becslést valósíthatunk meg (ld. 45. ábra). Ezt a **legkisebb abszolút eltérések** módszerének *(Least Absolute Deviations method)* nevezzük, melynek súlymátrixa

$$\underline{\underline{W}}^{(d)} = \begin{pmatrix} |e_1|^{-1} & 0 & \cdots & 0 \\ 0 & |e_2|^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & |e_N|^{-1} \end{pmatrix}.$$

A súlyozott legkisebb négyzetek módszerén *(Weighted Least SQuares method)* alapuló inverziós eljárás az alábbi célfüggvényt minimalizálja

$$\Phi = \vec{e}^{\,T} \underline{\underline{W}}^{(d)} \vec{e} \;= \min$$

mely a következő megoldásra vezet (Menke, 1984)

$$\vec{m} = \left(\underline{\underline{G}}^{T} \underline{\underline{W}}^{(d)} \underline{\underline{G}}\right)^{-1} \underline{\underline{G}}^{T} \underline{\underline{W}}^{(d)} \vec{d}.$$

Vegyük észre, hogy $\underline{\underline{W}}^{(d)} = \underline{\underline{I}}$ esetén (ahol $\underline{\underline{I}}$ az egységmátrix, és azt fejezi ki, hogy az adatok korrelálatlanok és egyforma megbízhatóságúak) a módszer a Gauss-féle legkisebb négyzetek módszerének megoldását adja vissza.

Feladat. Végezzünk gyors ellenőrzést arra, hogy a fenti egyenlet jobb oldalán a mátrixok és vektorok sorrendje helyes!

$$(M \times 1) = ((M \times N)(N \times N)(N \times M))(M \times N)(N \times N)(N \times 1)$$
$$(M \times 1) = (M \times M)(M \times 1)$$
$$(M \times 1) = (M \times 1).$$

A fenti módszer esetén kizárólag az adatok súlyozásával javítottuk az inverz feladat megoldását. Léteznek azonban olyan inverziós eljárások is, ahol magukat a modell-paramétereket súlyozzuk (sőt a két módszer a kevert határozottságú inverz feladatok megoldása esetén össze is vonható). E technikával kiemelhetünk vagy elnyomhatunk bizonyos paramétereket, vagy akár bizonyos paraméter-tartományokat teljesen kizárhatunk a megoldásból. Ezt a módszert **„kényszerített" (constrained) inverzió**nak nevezzük, mely elsősorban alulhatározott feladatok (pl. háromdimenziós gravitációs vagy mágneses adatok inverziójánál, ahol az adatokénál nagyságrendekkel nagyobb számú téglatestre osztjuk fel a félteret, ahol az ismeretlen kőzetfizikai paraméterek értékét egyenként kell meghatározni) megoldásánál alkalmazzuk. Ilyen esetet képez, amikor az ismert paraméterekkel megadott (referencia) modell felé „tereljük", vagy egy előírt tartományba kényszerítjük a megoldást. Gyakran pedig a szomszédos modellparaméterek „egyenetlenségeinek" eltűntetése érdekében

simítjuk a megoldást. Az $M \times M$ méretű **modelltérbeli súlymátrix**ot pl. úgy konstruálhatjuk meg, hogy az a modellparaméterek első deriváltját előíró operátor legyen

$$\underline{\underline{W}}^{(m)} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & & 0 & -1 & 1 \end{pmatrix}.$$

A $\underline{\underline{W}}^{(m)}$ súlymátrix-szal képzett $\vec{m}^T \underline{\underline{W}}^{(m)} \vec{m}$ szorzat (a modellparaméterek numerikus deriváltjainak) minimalizálása eléri, hogy a szomszédos geometriájú helyeken (pl. rétegekben vagy cellákban) értelmezett modellparaméter értékek nem térnek el egymástól irreális mértékben, így az inverzióval becsült modellparaméterek térbeli eloszlása (itt a hely-koordináták szerinti eloszlást értjük) megfelelően sima lesz. Az alulhatározott problémát a Lagrange-féle multiplikátorok módszerével oldhatjuk meg

$$\Omega = \vec{m}^T \underline{\underline{W}}^{(m)} \vec{m} + \vec{\lambda}^T \vec{e} = \min$$

ahol $\vec{\lambda}$ az eltérésvektor elemeihez tartozó (azzal megegyező méretű) multiplikátorok vektora. A minimalizálás eredményeként előálló becsült modell *Dobróka (2001)* alapján

$$\vec{m} = \underline{\underline{W}}^{(m)-1} \underline{\underline{G}}^T \left(\underline{\underline{G}} \underline{\underline{W}}^{(m)-1} \underline{\underline{G}}^T \right)^{-1} \vec{d}.$$

Feladat. Végezzünk gyors ellenőrzést arra, hogy a fenti egyenletben a mátrixok és vektorok sorrendje megfelelő!

$$(M \times 1) = (M \times M)(M \times N)((N \times M)(M \times M)(M \times N))(N \times 1)$$
$$(M \times 1) = (M \times N)(N \times N)(N \times 1)$$
$$(M \times 1) = (M \times 1).$$

Példa. A fúrás során harántolt rétegek kőzetfizikai paramétereinek meghatározása hagyományosan mélyfúrási geofizikai szelvényadatok mélységpontonkénti inverziójával valósítható meg. Több ásványt tartalmazó kőzetmodell esetén az ismeretlen kőzetfizikai mennyiségek modellvektora

$$\vec{m} = [POR, SX0, SW, VSH, VMA_i]^T$$

melyben a porozitás (*POR*), a kisepert- (*SX0*) és az érintetlen zóna (*SW*) víztelítettsége, az agyagtartalom (*VSH*), valamint az *n* számú ásványból (pl. kvarc, kalcit, dolomit stb.) felépülő kőzetmátrix részarányok (VMA_i) térfogatjellemző mennyiségek. A fenti kőzetfizikai paraméterek közvetlenül nem mérhetők, ezért meghatározásuk más fizikai mennyiségeket mérő szondák adatainak együttes inverziójával történik. Az inverz feladat $\vec{d}$ adatvektorának tipikus elemeit a 6. táblázat mutatja, ahol a litológiai megjelölésű szelvények elsősorban a kőzettípusra, a porozitás-követők a porozitásra, és a szaturációs szelvények a (víz-, gáz- és olaj-) telítettségi viszonyokra érzékenyek. A mélyfúrási geofizikai mérések rendszerint

bonyolult mérési körülmények (szabálytalan lyukgeometria, iszappal történő elárasztás, szondák eltérő behatolási mélysége és vertikális felbontóképessége, nagy nyomás és magas hőmérséklet, vertikálisan és horizontálisan inhomogén közeg, anizotrópia stb.) között történnek. A mérési hibákat laza rétegekben általában a kavernásodás vagy az iszaplepény kialakulása, továbbá az elektronika (statisztikus ingadozás, ciklusugrás stb.) okozza. Ezért először a mért szelvényeket a szonda környezetének különböző hatásaira korrigálni kell. Azonban az adatkorrekciók is egyfajta hibaforrásnak foghatók fel. A mérések szondahossztól függően különböző behatolásúak, így az együttes kiértékelés céljából néha a nagy felbontóképességű szelvényeket simítjuk. Az olajiparban alkalmazott mélyfúrási geofizikai szelvényezési módszerekről *Kiss és Ferenczy (1993)* jegyzetében bővebben olvashatunk.

6. táblázat

Szelvények	Elnevezés	Típus	Mértékegység
GR	természetes gamma	litológiai	API
K	kálium (spektrális) gamma	litológiai	%
U	urán (spektrális) gamma	litológiai	ppm
TH	tórium (spektrális) gamma	litológiai	ppm
SP	természetes potenciál	litológiai	mV
CN	kompenzált neutron	porozitás-követő	%
DEN	sűrűség (gamma-gamma)	porozitás-követő	g/cm^3
AT	akusztikus terjedési idő	porozitás-követő	µs/m
RMLL	mikrolaterolog	szaturációs	ohmm
RS	sekélybehatolású fajlagos ellenállás	szaturációs	ohmm
RD	mélybehatolású fajlagos ellenállás	szaturációs	ohmm

A mélyfúrási geofizikai inverz feladat kismértékben túlhatározott, mivel az alkalmazott szondák száma alig több az ismeretlenek számánál. Az inverziós kiértékelés kezdetén a mérési adatrendszer és előzetes ismereteink alapján megbecsüljük a modellparaméterek kezdeti értékét. A feladat jellegzetessége, hogy általában elegendő a priori ismeret (pl. fúrómagok laboratóriumi vizsgálati eredményei, crossplot-ok, közeli fúrások mélyfúrási geofizikai szelvényei, felszíni geofizikai mérések és egyéb geológiai ismeretek) áll rendelkezésünkre a megfelelő modellalkotáshoz, így a lineáris inverziós technika jól alkalmazható. Az adatok és modellparaméterek közötti kapcsolat általános esetben nemlineáris, továbbá a rendelkezésre álló válaszfüggvények empirikusak. A megfelelő egyenlet kiválasztása az aktuális földtani felépítéstől, a telep mélységétől, korától, kőzettípustól, rétegtartalmától, kompakciójától függ. Az elméleti válaszfüggvények független változóit a térfogatjellemző kőzetfizikai és egyéb zonális (a túlhatározottság fenntartása érdekében nagyobb mélységintervallumban konstansként kezelt) paraméterek képezik. A *k*-adik elvi adat általános válaszfüggvénye

$$d_k = g_k(\vec{m}, C_{k1}, \ldots, C_{kH})$$

ahol *C* mennyiségek a kőzet texturális és egyéb tulajdonságaitól függő konstansok, melyek értékét labor- és szelvény információk, ill. a szakirodalomban található javaslatok alapján választhatjuk meg. Az adatok kiértékelése során előfordulnak olyan mennyiségek is, melyek a válaszegyenletekben közvetlenül nem jelennek meg, azonban meghatározásuk alapvetően

szükséges (pl. a szénhidrogén-készlet becslése szempontjából). A kitermelhető- és maradék szénhidrogén-telítettséget, valamint az áteresztőképességet az inverziós eredményekből determinisztikus módon határozzuk meg. A direkt feladat megoldása során bizonyos alapvető fizikai feltételeknek is teljesülniük kell, pl.: $0 \le POR \le 1$, $0 \le VSH \le 1$, $0 \le SW \le 1$, $0 \le SX0 \le 1$, $0 \le VMA_i \le 1$. Ezen kívül a kőzetfizikai paraméterekre további tapasztalati kikötések is tehetők, pl. törmelékes üledékes kőzetek esetén a $0 \le POR \le 0.47$, $0.15 \le SW \le 1.0$, $0.50 \le SX0 \le 1.0$ relációkat figyelték meg. Ezen kívül az ismeretlenek közötti regressziós vizsgálatok eredményei alapján egyéb járulékos feltételeket is szabhatunk (pl. a kisepert zóna és az érintetlen zóna víztelítettségének nemlineáris kapcsolatát leíró helyi összefüggés). Végül a kőzetkomponensek fajlagos térfogatösszegére vonatkozó alapvető törvényt (anyagmérleg egyenlet) is figyelembe kell vennünk, mely abban az esetben, amikor a kőzetet három alkotórészre (pórustér, kőzetmátrix és agyag) bontjuk a következő

$$\mathrm{POR} + \mathrm{VSH} + \sum_{i=1}^{n} \mathrm{VMA}_i = 1 .$$

A mélyfúrási geofizikai inverz feladatot mélységpontonként végrehajtott, egymástól független inverziós eljárások sorozatával oldjuk meg. Tehát, az egyes mélységpontokban meghatározzuk a kőzetfizikai paraméterek értékét, majd ezután a kapott eredményeket interpoláljuk, és szelvények formájában jelenítjük meg. Mivel az egyes szelvénytípusok eltérő dimenziójúak és más-más nagyságrendbe esnek, ezért a súlyozott legkisebb négyzetek módszerét (WLSQ) alkalmazzuk. Ennek megfelelően az inverz feladat célfüggvénye

$$\Phi = \sum_{k=1}^{N} \left(\frac{d_k^{(m)} - d_k^{(sz)}}{\sigma_k} \right)^2 = \min$$

ahol $d_k^{(m)}$ és $d_k^{(sz)}$ a pontbeli k-adik mérési és számított adat, σ_k a k-adik szelvénytípus szórása, mellyel a megoldás

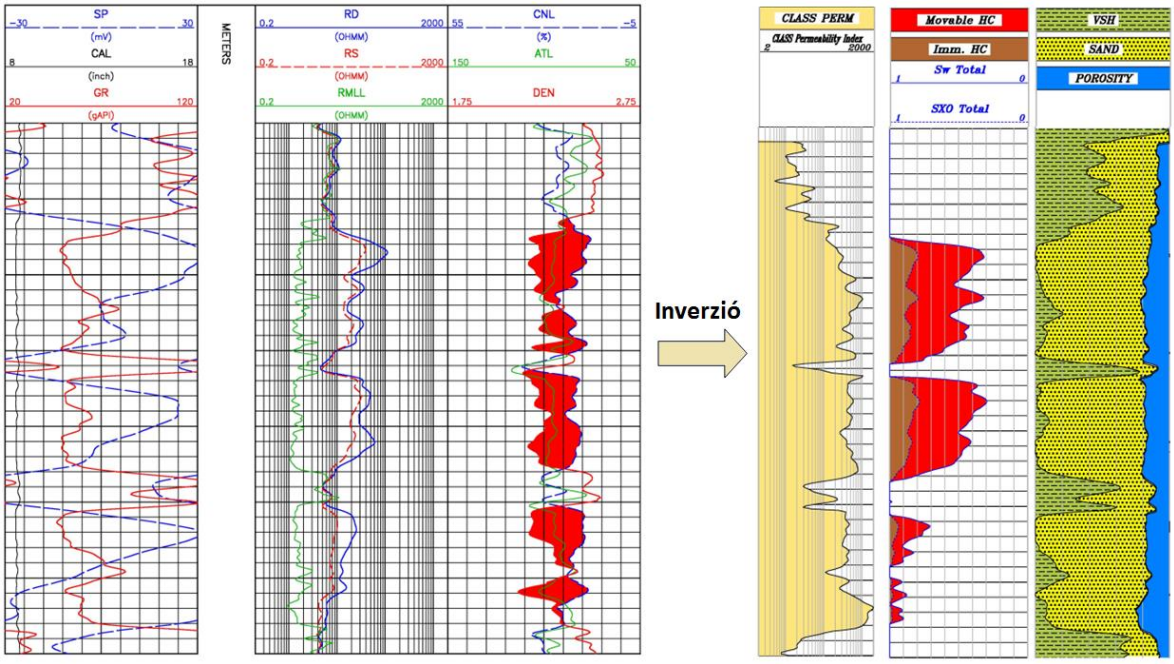
$$\vec{m} = \left(\underline{\underline{G}}^T \underline{\underline{W}}^{(d)} \underline{\underline{G}} \right)^{-1} \underline{\underline{G}}^T \underline{\underline{W}}^{(d)} \vec{d} \quad \text{ahol} \quad \underline{\underline{W}}^{(d)} = \begin{pmatrix} \sigma_{GR}^{-2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_K^{-2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_U^{-2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{RS}^{-2} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_{RD}^{-2} \end{pmatrix}.$$

A 69. ábra egy olajipari példát mutat be a mélységpontonkénti inverzió alkalmazására. A mélyfúrási geofizikai adatrendszerek bőséges „in-situ” és „nagyfelbontású” információt hordoznak a felszín alatti objektumokról. A műszerfejlesztés és számítógépes kapacitás, ill. az elméleti tudás rohamos fejlődésével újabb és újabb inverziós módszerek születnek. A téma egyéb területei iránt érdeklődőknek ajánljuk *Dobróka (2001)* jegyzetét és *Szabó (2004)* doktori értekezését, valamint az *SPWLA (Society of Petrophysicist and Well Log Analyst)* kiadásában kéthavonta megjelenő *Petrophysics* nevű folyóirat tanulmányozását.

13. A becslés pontosságának és megbízhatóságának jellemzése

A becslési eredmények minősítésének (minőség-ellenőrzésének) elmélete kiemelt jelentőséggel bír a gyakorlatban. A most bemutatandó tématerület szervesen kapcsolódik az előző fejezetben megismert lineáris inverziós módszerek elméletéhez. Ebben a fejezetben az inverziós eredmények pontosságával (becslési hiba számítása), az adatok bizonytalanságának az eredmények pontosságára gyakorolt hatásával (adattérbeli hiba megadása), valamint az inverzióval becsült modell megbízhatóságának jellemzésével (korrelációs együtthatók számítása) foglalkozunk.

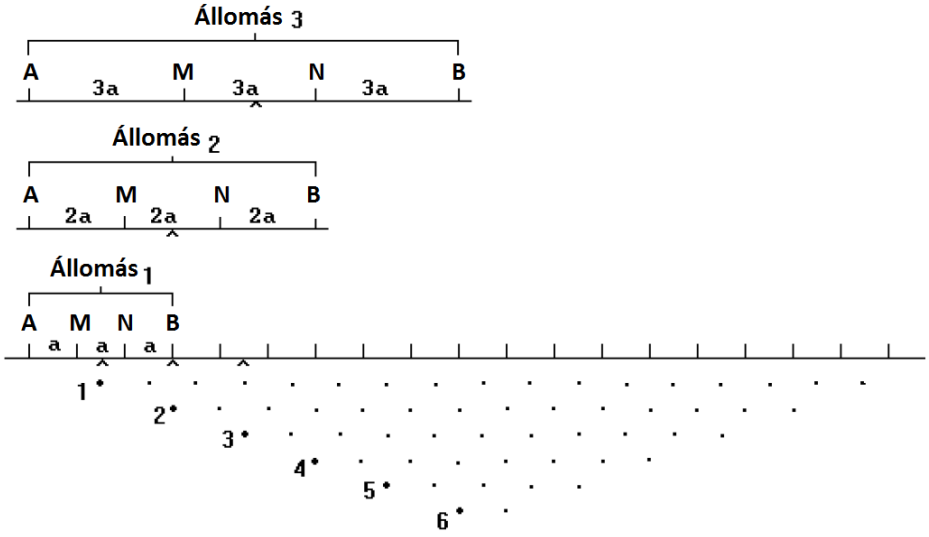
A mérési adatokat terhelő zaj az inverzió során (az adat-modell kapcsolat lineáris egyenletrendszerén keresztül) áttranszformálódik a modelltérbe. Ennek eredményeképpen az inverzióval becsült modellparaméterek a bemenő hiba nagyságával arányos mértékben lesznek pontosak és megbízhatóak. Emellett a modellezés, mely a valóságot nem írja le teljes részletességében (bizonyos tulajdonságokat elhanyagolunk), ugyancsak hibaforrásként fogható fel, mely közelítő válaszfüggvényeken keresztül a számított adatokat terheli az inverziós eljárás során. E két egymástól független hibamennyiség összege alkotja a σ_d-vel jelölt **adathibá**t, mely az inverziós eljárás bemenő hibajellemző mennyisége.



69. ábra *Mélyfúrási geofizikai adatok inverziója*
(MOL NYrt. jóvoltából)

Példa. Egyenáramú elektromos módszerekkel a felszín alatti objektumok ρ(ohmm) fajlagos ellenállását lehet meghatározni. Mivel az egyes kőzetek fajlagos ellenállása igen eltérő lehet (nagy értéktartományt fog át), ezért a vizes rétegek, olajszennyeződések, ércek és egyéb inhomogenitások általában jól elkülöníthetők a környezetüktől. A mérési eszközt két

áram- (*A* és *B*), két potenciál-elektróda (*M* és *N*), a kábelek és az elektronika képezi. Az *A*-*B* elektródapáron egyenáramot táplálunk a talajba, majd az *M*-*N* elektródák között fellépő potenciálkülönbséget regisztrálva a felszín alatti szerkezetek fajlagos ellenállásával arányos mennyiséget kapunk. A mérést többféle elektróda-elrendezésben is elvégezhetjük, melyek eltérő érzékenységgel reagálnak a földtani szerkezet paramétereire. A jelen példában szereplő Wenner-elrendezés jellemzője, hogy a négy elektróda közötti távolság azonos, melyek együttes megváltoztatásával szabályozhatjuk a behatolási mélységet (az elektróda-távolság növelésével nő a behatolás). A 70. ábrán egy gyakran alkalmazott kétdimenziós leképezést megvalósító mérési eljárást, a rétegszelvényezést látjuk. Az ábrán az egyes állomásokhoz tartozó referencia mélységeket az 1,2,3… jelű pontok mutatják. E pontokban lévő adatokat interpolálva állíthatjuk elő a fajlagos ellenállás szelvényt.

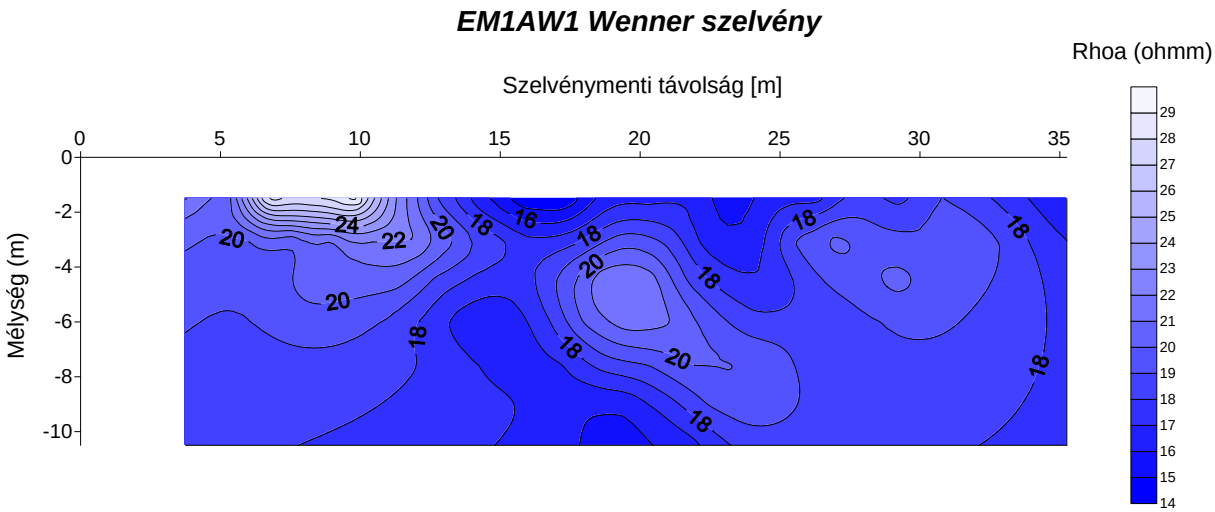


70. ábra *Rétegszelvényezés Wenner-elrendezés esetén*

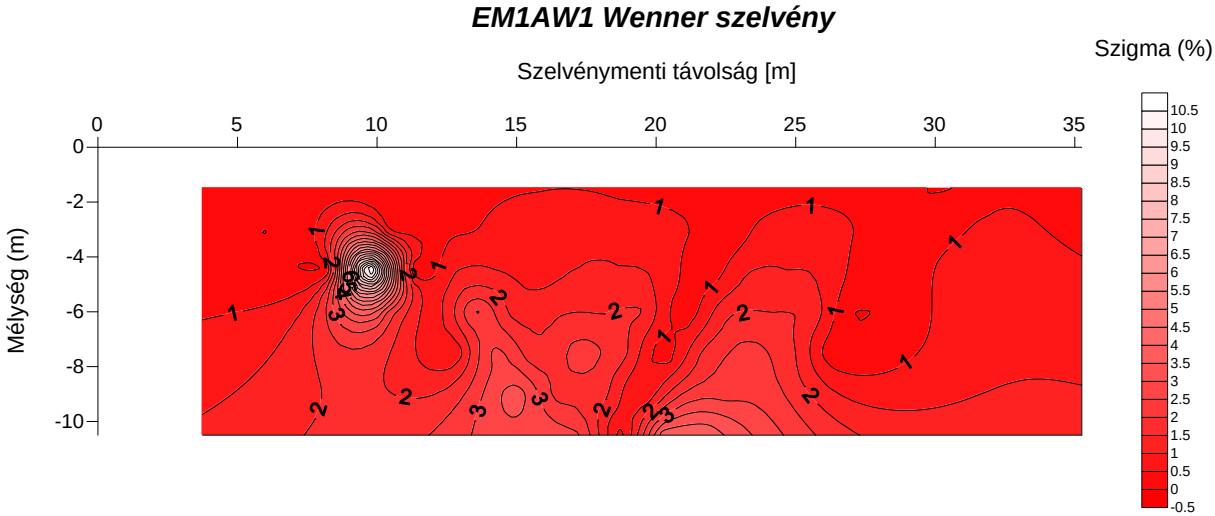
Az elektromos adatokat terhelő hiba megfigyelése céljából *Ormos Tamás (Miskolci Egyetem Geofizikai Tanszék)* ismételt méréseket végzett Emőd községben. Wenner-elrendezést alkalmazva mindegyik állomáson öt alkalommal mért, majd kiszámította a kapott fajlagos ellenállás adatok számtani átlagát és empirikus szórását. A fajlagos ellenállás átlagértékek szelvényét és a hibaszelvényt a 71-72. ábrák mutatják. Az utóbbin látható, hogy az adathiba nagysága átlagosan a mért értékek 2%-a. Az *x*=9m és *z*=5m koordinátákkal jellemzett pontok környezetében megnövekedett hiba annak az eredménye, hogy a mérés során ezen a helyen gyengébb volt a jel/zaj viszony.

A lineáris (linearizált) inverzió elméletében lehetőség van a becsült modellparaméterek hibájának és megbízhatóságának mennyiségi jellemzésére. A modellparaméterek és adatok kapcsolatát az $M \times N$ méretű $\underline{\underline{\mathrm{M}}}$ **általánosított inverz mátrix** adja meg

$$\vec{\mathrm{m}} = \underline{\underline{\mathrm{M}}}\vec{\mathrm{d}} + \vec{\upsilon}.$$



71. ábra *Látszólagos fajlagos ellenállás átlagérték szelvény (Emőd, 1995)*
(Dr. Ormos Tamás jóvoltából)



72. ábra *Látszólagos fajlagos ellenállás hiba szelvény (Emőd, 1995)*
(Dr. Ormos Tamás jóvoltából)

Legyen $\vec{v} = \vec{0}$, ekkor az adat-modell kapcsolat lineáris. Például a Gauss-féle legkisebb négyzetek módszere esetén az általánosított inverz a következő

$$\underline{\underline{M}} = \left(\underline{\underline{G}}^T \underline{\underline{G}}\right)^{-1} \underline{\underline{G}}^T \quad \text{mivel} \quad \vec{m} = \left(\underline{\underline{G}}^T \underline{\underline{G}}\right)^{-1} \underline{\underline{G}}^T \vec{d} = \underline{\underline{M}} \vec{d}.$$

Lineáris kapcsolatot feltételezve a modellparaméter-vektor elemeinek átlagértékével (felülvonással jelölve) együtt is fennáll a fenti összefüggés

$$\vec{m} - \bar{\vec{m}} = \underline{\underline{M}}(\vec{d} - \bar{\vec{d}})$$

amely az *i*-edik és *j*-edik modell-paraméterrel indexhelyes alakban (ahol *i*=1,2,…,*M* és *j*=1,2,…*M*) a következő

$$m_i - \overline{m}_i = \sum_{k=1}^{N} M_{ik}\left(d_k - \overline{d}_k\right) \quad \text{és} \quad m_j - \overline{m}_j = \sum_{l=1}^{N} M_{jl}\left(d_l - \overline{d}_l\right).$$

Képezzük az *i*-edik és *j*-edik modellparaméter kovarianciáját

$$\mathrm{COV}\left(m_i, m_j\right) = \overline{\left(m_i - \overline{m}_i\right)\left(m_j - \overline{m}_j\right)} = \sum_{k=1}^{N}\sum_{l=1}^{N} M_{ik} M_{jl} \overline{\left(d_k - \overline{d}_k\right)\left(d_l - \overline{d}_l\right)}$$

Mivel a *k*-adik és *l*-edik adat kovarianciája

$$\mathrm{COV}\left(d_k, d_l\right) = \overline{\left(d_k - \overline{d}_k\right)\left(d_l - \overline{d}_l\right)}$$

ezért az alapegyenlet

$$\mathrm{COV}\left(m_i, m_j\right) = \sum_{k=1}^{N}\sum_{l=1}^{N} M_{ik}\,\mathrm{COV}\left(d_k, d_l\right) M_{jl}$$

melynek mátrixokkal felírt alakja

$$\underline{\underline{\mathrm{COV}}}\left(\vec{m}\right) = \underline{\underline{M}}\,\underline{\underline{\mathrm{COV}}}\left(\vec{d}\right)\underline{\underline{M}}^{\mathrm{T}}$$

ahol az $M \times M$ méretű $\underline{\underline{\mathrm{COV}}}\left(\vec{m}\right)$ mátrixot a **modellparaméterek kovariancia mátrixá**nak (röviden modell-kovariancia), és az $N \times N$ méretű $\underline{\underline{\mathrm{COV}}}\left(\vec{d}\right)$-t az **adatok kovariancia mátrixá**nak (röviden adat-kovariancia) nevezzük. A fenti összefüggés kimondja, hogy az inverziós eljárásba bemenő adatok

$$\sigma_{d_k} = \sqrt{\underline{\underline{\mathrm{COV}}}\left(d_k, d_k\right)}$$

hibájának (ebben az esetben nem a teljes adatrendszerre számított hibáról beszélünk, hanem az adatok egyedi hibájáról van szó) ismeretében meghatározhatók az inverzióval előállított modellparaméterek **becslési hibá**ja

$$\sigma_{m_i} = \sqrt{\underline{\underline{\mathrm{COV}}}\left(m_i, m_i\right)}.$$

A fenti összefüggésnek fontos gyakorlati jelentősége van, mivel a becslési hibák megadásán keresztül tudjuk az inverziós eljárás **pontosság**át jellemezni. A fenti módszer azt feltételezi, hogy az adatok és a modellparaméterek egyaránt Gauss-eloszlást követnek. Tudjuk, hogy ebben az esetben a legkisebb négyzetek módszere szolgáltatja az optimális becslési eredményeket. Ha az adatok korrelálatlanok és azonos szórásúak (σ_d), akkor az LSQ módszer alkalmazásakor a modell-kovariancia számítása egyszerűbbé válik

$$\underline{\underline{\mathrm{COV}}}\left(\vec{m}\right) = \underline{\underline{M}}\left(\sigma_d^2 \underline{\underline{I}}\right)\underline{\underline{M}}^{\mathrm{T}} = \sigma_d^2 \left(\underline{\underline{G}}^{\mathrm{T}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^{\mathrm{T}}\left[\left(\underline{\underline{G}}^{\mathrm{T}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^{\mathrm{T}}\right]^{\mathrm{T}}.$$

Alkalmazzuk az $\left(\underline{\underline{A}}\underline{\underline{B}}^T\right)^T = \underline{\underline{B}}\underline{\underline{A}}$ algebrai azonosságot, mely az $\underline{\underline{A}} = \left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}$ és $\underline{\underline{B}} = \underline{\underline{G}}$ helyettesítéssel előállítja a modell-kovariancia mátrixot

$$\underline{\underline{COV}}(\vec{m}) = \sigma_d^2\left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1} = \sigma_d^2\left(\underline{\underline{G}}^T\underline{\underline{G}}\right)^{-1}.$$

Az inverzióval becsült modellparaméterek **megbízhatóság**át a modellparaméterek $\underline{\underline{R}}(\vec{m})$ korrelációs mátrixának (ld. 6. fejezet) együtthatói segítségével jellemezzük

$$r(m_i, m_j) = \frac{cov(m_i, m_j)}{\sigma_{m_i}\sigma_{m_j}}.$$

A korrelációs mátrixot egyetlen $0 \le S \le 1$ értéktartományba eső skalárral, az ún. **korrelációs átlag**gal is megadhatjuk

$$S = \sqrt{\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\underline{\underline{R}}(m_i, m_j) - \delta_{ij}\right)^2}$$

ahol δ a Kronecker-delta szimbólumot jelöli (mely $i=j$ esetén 1, egyébként 0). Az alacsony korrelációs együtthatók ($r<0.4$) megbízható megoldást jelentenek, mivel ekkor a modell-paraméterek függetlenek vagy csak kismértékben korrelálnak, így azok egyedileg (egymástól függetlenül) meghatározhatók. Az erős korrelációs kapcsolat nem kedvező, ui. a csatolt paraméterek az inverziós eljárás közben nem tudnak egymástól függetlenül változni, így nem az optimumban stabilizálódik az inverziós eljárás. Erős kapcsolat esetén a modell-paramétereket nem tudjuk önállóan meghatározni, esetleg azok valamely kombinációját (pl. elektromos adatok inverziójánál a rétegek vastagságai és valódi fajlagos ellenállásai csatoltak). Ennek az a következménye, hogy ugyanahhoz a mérési adatsorhoz végtelen számú modell tartozhat. E **többértelműségi (ekvivalencia) probléma** következtében a becsült modell megbízhatatlan lesz. Az ekvivalenciát más (lineáris vagy globális, ld. 14. fejezet) inverziós módszer alkalmazásával sem tudjuk feloldani, az a probléma fizikai sajátossága. Egy megoldási alternatíva az, ha kikötéseket (újabb független egyenleteket vonunk be az eljárásba) teszünk a meghatározandó modellre. Ezáltal egyértelművé tehetjük az inverz problémát. Ennek a sikeressége nagymértékben függ az előzetes (a priori) ismeretek megbízhatóságától. A másik lehetőség az ún. **együttes inverzió** alkalmazása, melynek keretében ugyanazon földtani szerkezeten különböző fizikai elven mért adatrendszereket egyetlen inverziós eljárásban dolgozunk fel. Ennek az egyedi inverziós eljárásokkal (ahol egyfajta adatrendszert invertálunk) szembeni előnye az, hogy pontosabb és megbízhatóbb eredményt kapunk. Az új adatrendszerek bevonása révén ugyanis új információt viszünk be az inverziós eljárásba, melynek ekvivalencia feloldó hatása van.

Feladat. Egy túlhatározott inverz probléma esetén öt adat birtokában három ismeretlent határoztunk meg. Írjuk fel az adatok kovariancia mátrixát korrelált és korrelálatlan adatok esetén! Adjuk meg a modellparaméterek kovariancia-mátrixát korrelált és független paraméterek esetén! Az inverzióval becsült modellparaméterek korrelációs mátrixát is írjuk fel korrelálatlan és korrelált esetben!

Független adatok és modellparaméterek esetén az adat- és modell-kovariancia mátrix diagonális, melyek főátlójában a hibák (varianciák) szerepelnek, míg korrelált esetben a kovariancia mátrix főátlón kívüli elemei (az „együttváltozás” mértékétől függően) zérustól különbözőek. A modellparaméterek korrelációs mátrixa négyzetes ($M \times M$ méretű), mely korrelálatlan esetben megegyezik az egységmátrix-szal. A feladatban mindegyik mátrix szimmetrikus.

1. *korrelálatlan esetben*

$$\underline{\underline{COV}}(\vec{d}) = \begin{pmatrix} \sigma_{d_1}^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{d_2}^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{d_3}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{d_4}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{d_5}^2 \end{pmatrix}$$

$$\underline{\underline{COV}}(\vec{m}) = \begin{pmatrix} \sigma_{m_1}^2 & 0 & 0 \\ 0 & \sigma_{m_2}^2 & 0 \\ 0 & 0 & \sigma_{m_3}^2 \end{pmatrix}$$

$$\underline{\underline{R}}(\vec{m}) = \underline{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2. *korrelált esetben*

$$\underline{\underline{COV}}(\vec{d}) = \begin{pmatrix} \sigma_{d_1}^2 & cov(d_1,d_2) & cov(d_1,d_3) & cov(d_1,d_4) & cov(d_1,d_5) \\ cov(d_2,d_1) & \sigma_{d_2}^2 & cov(d_2,d_3) & cov(d_2,d_4) & cov(d_2,d_5) \\ cov(d_3,d_1) & cov(d_3,d_2) & \sigma_{d_3}^2 & cov(d_3,d_4) & cov(d_3,d_5) \\ cov(d_4,d_1) & cov(d_4,d_2) & cov(d_4,d_3) & \sigma_{d_4}^2 & cov(d_4,d_5) \\ cov(d_5,d_1) & cov(d_5,d_2) & cov(d_5,d_3) & cov(d_5,d_4) & \sigma_{d_5}^2 \end{pmatrix}$$

$$\underline{\underline{COV}}(\vec{m}) = \begin{pmatrix} \sigma_{m_1}^2 & cov(m_1,m_2) & cov(m_1,m_3) \\ cov(m_2,m_1) & \sigma_{m_2}^2 & cov(m_2,m_3) \\ cov(m_3,m_1) & cov(m_3,m_2) & \sigma_{m_3}^2 \end{pmatrix}$$

$$\underline{\underline{R}}(\vec{m}) = \begin{pmatrix} 1 & r(m_1,m_2) & r(m_1,m_3) \\ r(m_2,m_1) & 1 & r(m_2,m_3) \\ r(m_3,m_1) & r(m_3,m_2) & 1 \end{pmatrix}.$$

A fenti hibajellemzőkön kívül illeszkedést jellemző mérőszámokat is bevezethetünk az inverziós eredmények ellenőrzése céljából. E mennyiségek nemcsak a végeredmény „jóságáról” tájékoztatnak, hanem az egyes iterációs lépésekben elért adat-, ill. modelltérbeli egyezést is megadják. Abban az esetben, amikor az iterációs lépésszám növekedésével fokozatosan az optimum felé tartunk (miközben a mérési és számított adatok eltérése fokozatosan csökken) **konvergens** inverziós eljárásról beszélünk. Viszont, ha egyre inkább eltávolodunk az optimumtól (nő a mért és számított adatok távolsága), mely gyakran numerikus instabilitással is párosul, akkor **divergens** inverziós eljárásról van szó. A becsült modell paramétereivel számított és a mért adatok eltérését az ún. **adattérbeli távolság**gal (röviden adattávolság) jellemezzük

$$D_a = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(d_k^{(m)} - d_k^{(sz)}\right)^2}$$

ahol $d_k^{(m)}$ és $d_k^{(sz)}$ a k-adik mért és számított adatot jelöli. Ha a fenti mennyiséget 100-al megszorozzuk, akkor az adattérbeli egyezést százalékos értelemben kapjuk meg. Érdemes megemlíteni, hogy ha az adatok eltérő nagyságrendbe esnek (és különböző mértékegységgel rendelkeznek), akkor előnyös a k-adik mért vagy számított adattal normálni az eltérésnégyzeteket. Ha az adatok eloszlása Gausstól különbözik, akkor pl. az L_p-normával (ld. 3. fejezet) a fentihez hasonló mennyiség definiálható. Például a Laplace eloszlás (ld. 1. fejezet) esetén alkalmazott L_1-normán alapuló adattávolság definíciója a következő

$$D_a = \frac{1}{N}\sum_{k=1}^{N}\left|d_k^{(m)} - d_k^{(sz)}\right|.$$

Az inverziós eljárások alkalmazhatóságát (pontosságát, megbízhatóságát és teljesítményét) gyakran szintetikus adatokon teszteljük. Ennek keretében zajjal terhelt elvi (szintetikus) adatokat invertálunk egy ismert modell paramétereinek meghatározása céljából. A vizsgálat során kiderül, hogy milyen jól tudja az adott inverziós módszer rekonstruálni az ismert modellt, konvergens-e az eljárás, mekkora a zajérzékenysége, és mennyire ad megbízható eredményt. Ekkor új illeszkedési mérőszámot vezethetünk be, melyet **modelltérbeli távolság**nak (röviden modelltávolságnak) nevezünk

$$D_m = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(m_i^{(b)} - m_i^{(e)}\right)^2}$$

ahol $m_i^{(b)}$ és $m_i^{(e)}$ az i-edik modellparaméter becsült és egzakt értéke. Megjegyezzük, hogy a modelltávolság terepi adatok inverziója esetén nem számítható, mivel ott nem ismerjük az inverziós modellt (annak meghatározása az inverzió feladata).

Példa. Az 5. táblázat szeizmikus adatrendszeréből származtatott inverziós eredmények (ld. 68. ábra) pontosságának ellenőrzését mutatjuk be. A szeizmikus adatok hibája normális körülmények között az 5%-ot nem haladja meg, mely általában csak a háttérzaj mértékétől és az elektronikától függ. Jelen esetben az adatokat korrelálatlannak és azonos szórásúaknak

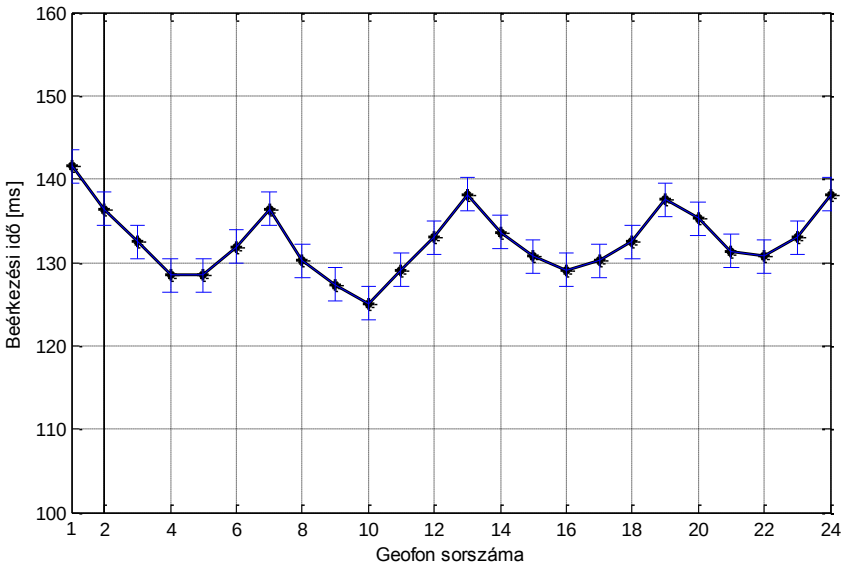
feltételeztük. A σ_d=3% választás mellett a mérési adatokhoz a 73. ábrán látható hibaintervallumok tartoznak. A kezdeti modellre (x_0=1m, y_0=5m, v=200ms^{-1}, t_0=300ms) számított és mért adatok távolsága 99% volt, mely a 4. iterációs lépés után 1.35%-ra csökkent (ld. 74. ábra). Kiszámítva a modell-kovariancia mátrix elemeit, a hibák ismeretében felállíthatjuk a modell-paraméterek megbízhatósági intervallumait. A 7. táblázatban az inverzióval becsült modellparaméterek értéke, valamint azok alsó (becsült érték és a hiba különbsége) és felső (becsült érték és a hiba összege) korlátai szerepelnek. A becsült paraméterek korrelációs mátrixa a 8. táblázatban található, melyre S=0.48 korrelációs átlag adódott. Látható, hogy az x_0 és y_0 koordináták függetlenek egymástól, ezért az eredmény elfogadható. A sebesség a többi paraméterrel kismértékben, ill. közepesen korrelál. A probléma a v és t_0 paraméterek együttes meghatározásával van. A teljes korreláció (r=0.99) oka az, hogy a 66. ábrán szereplő válaszegyenlet átrendezésével $(t-t_0)v=konstans$ kifejezés adódik, így a két paraméter kombinációjával végtelen számú ekvivalens megoldás lehetséges. Ez azt jelenti, hogy a két paraméter (inverzióval becsült) értéke nem megbízható. Ilyenkor azt lehet tenni, hogy az egyik paramétert más (jelen inverzión kívül) forrásból határozzuk meg (pl. a sebességet) és értékét nem változtatjuk (fix értéken tartjuk) az inverziós eljárás során. Ebben az esetben t_0-ra az inverz feladat már egyértelműen megoldható.

7. táblázat

Paraméter	Becsült	Minimum	Maximum
x_0	6.59	6.37	6.81
y_0	6.03	5.84	6.23
v	252.6	210.5	294.6
t_0	105.1	100.5	109.6

8. táblázat

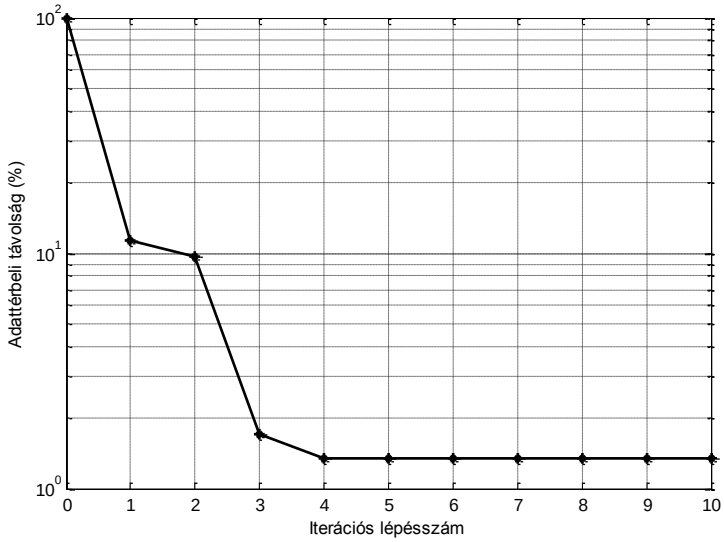
1.00	0.01	0.45	0.44
0.01	1.00	0.03	0.03
0.44	0.03	1.00	0.99
0.44	0.03	0.99	1.00



73. ábra *Terepi időadatok és azok hibája*

14. Globális szélsőérték-kereső eljárások

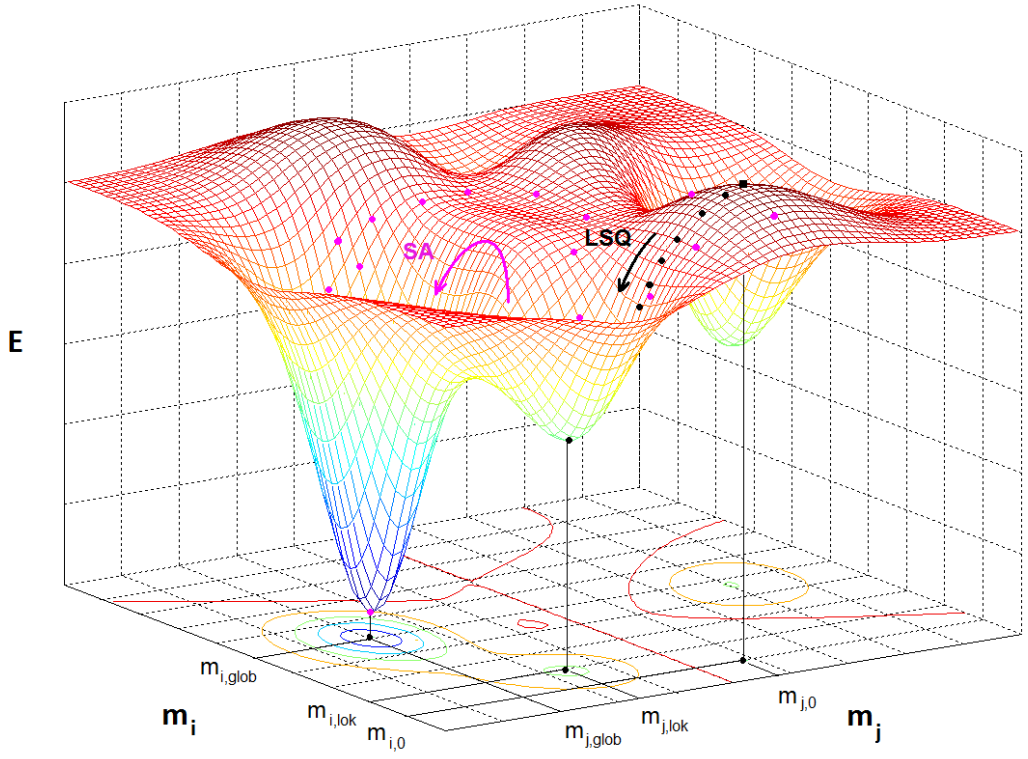
A lineáris inverziós eljárások (ld. 12. fejezet) kedvezően megválasztott kiindulási modell esetén gyors és kielégítő megoldást szolgáltatnak. Azonban e módszerek alkalmazása nagyméretű (sokváltozós) problémáknál nem mindig célravezető, mivel a (ekvivalens modellek miatt) nagyszámú lokális szélsőértékkel rendelkező (mért és számított adatok eltérését kifejező) célfüggvény optimalizálásakor gyakran egy helyi minimumban határozzák meg a megoldást. Az alapproblémát a 75. ábrán figyelhetjük meg. Az inverziós eljárást egy lokális minimumhoz közeli startmodelltől (ahol $m_{i,0}$ és $m_{j,0}$ az i-edik és j-edik modellparaméter kezdeti értéke) indítjuk. A legkisebb négyzetek módszere (LSQ) gradiens alapú keresést hajt végre, így az inverziós eljárás a célfüggvény legközelebb eső lokális minimumában (ahol $m_{i,lok}$ és $m_{j,lok}$ az i-edik és j-edik modellparaméter becsült értéke a helyi minimumban) stabilizálódik. Ez a mért és számított adatok egyezése szempontjából nem optimális megoldást jelent. Az ún. **globális optimalizációs** módszerek (Simulated Annealing, Genetikus Algoritmus) véletlen keresési mechanizmusa lehetővé teszi a lokális minimumból való kiszabadulást, így az inverziós eljárás képes megtalálni a célfüggvény abszolút (globális) minimumát (ahol $m_{i,glob}$ és $m_{j,glob}$ az i-edik és j-edik modellparaméter becsült értéke az abszolút minimumban). Az abszolút minimumhoz tartozó modellt tekintjük az inverz feladat optimális (legjobb) megoldásának.



74. ábra *Az adattávolság változása az inverziós eljárás során*

A **Simulated Annealing (SA)** eljárás egy a fémek speciális hőkezelési technikájának analógiája alapján tervezett hatékony globális optimalizációs módszer, mely az irányított Monte-Carlo módszerek családjába tartozik. A kohászatban a fémek lágyítását az olvadt állapothoz közeli hőmérsékletről történő lassú hűtéssel valósítják meg. Ennek hatására a nagyszámú atom fokozatosan veszít mozgási energiájából és a fém kristályosodni kezd. A kialakuló fémrács atomi összenergiája a hűtés időtartamának a függvénye. Elvileg végtelen lassú hűtés eredményezi a minimális energiájú (tökéletes) rácsszerkezetet, mely analóg az

inverz probléma E (mért és számított adatok illeszkedését jellemző) célfüggvény abszolút minimumban való stabilizálódásával. A gyakorlatban ilyen lassú hűtés nem valósítható meg, ezért gyorsabb hűtési eljárás szükséges. Gyorsabb hűtés következtében viszont a kristály-szerkezetben rácshibák alakulnak ki, és a fém egy magasabb energiaszinten fagy (tökéletlen) rácsba. Ez megfelel az inverziós eljárás lokális minimumban való stabilizálódásának. Az atomok speciális hőkezelés hatására azonban kiszabadulnak a magasabb energiaszintű kristályrácsból, és az ezt követő megfelelően lassú hűtés mellett képesek elérni az abszolút minimális energiájú rácsszerkezetet. Ez analóg az inverz feladat globális minimumának megtalálásával. Az SA eljárás a fenti folyamatot algoritmizálja az inverziós eljárás célfüggvénye globális minimumának meghatározása céljából.



75. ábra *A lokális (LSQ) és globális (SA) optimumkeresés*

Termikus egyensúly esetén az optimális modellek az ún. **Gibbs-féle eloszlás**sal írhatók le, melynek valószínűség-sűrűség függvénye a következő

$$P\left(m^{(i)}\right)=\frac{e^{-\frac{E\left(m^{(i)}\right)}{T}}}{\sum_{j=1}^{S} e^{-\frac{E\left(m^{(j)}\right)}{T}}}$$

ahol $P(m^{(i)})$ az i-edik modell előfordulási valószínűsége, S az összes lehetséges modell száma és T az **általánosított hőmérséklet**, melynek az SA algoritmusban nincs fizikai jelentése, viszont fontos folyamatjellemző (kontroll) paraméter.

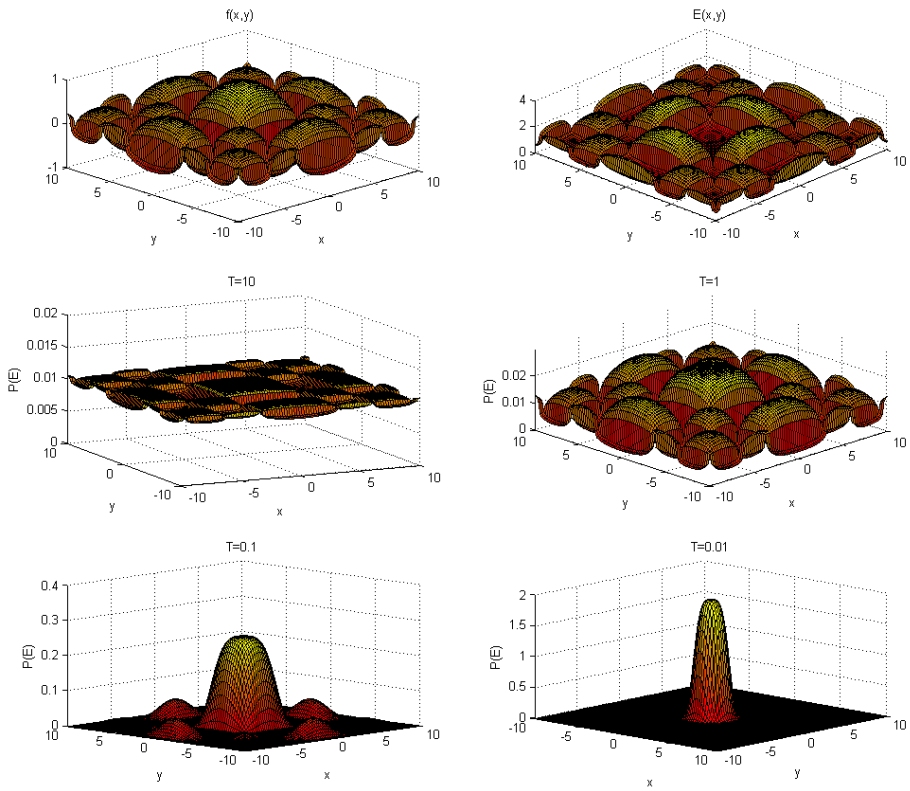
Példa. *Sen és Stoffa (1995)* alapján bemutatjuk a Gibbs-féle sűrűségfüggvény hőmérséklet függését. Legyen adott az *f(x,y)* kétváltozós függvény, melynek számos lokális maximuma, viszont az x=0 és y=0 helyen (egyetlen) abszolút minimuma van

$$f(x,y) = \operatorname{sgn}\left(\frac{\sin x}{x}\right)\left(\left|\frac{\sin x}{x}\right|\right)^{\frac{1}{4}} \operatorname{sgn}\left(\frac{\sin y}{y}\right)\left(\left|\frac{\sin y}{y}\right|\right)^{\frac{1}{4}}.$$

Képezzük az *E(x,y)* célfüggvényt, melynek globális minimuma ugyanazon a helyen van, ahol *f(x,y)* abszolút maximuma

$$E(x,y) = \left(1 - f(x,y)\right)^2 = \min.$$

Vizsgáljuk meg, hogy mely T értékek esetén adja meg a Gibbs-féle valószínűség-sűrűségfüggvény egyértelműen az *E(x,y)* célfüggvény globális minimumát! A feladat szoftveres megoldását *Szabó (2006)* segédlete tartalmazza. Az 76. ábra eredményéből látható, hogy minél kisebb a T értéke, annál határozottabban rajzolódik ki a globális szélsőértékhely. Az eredmény jól mutatja az SA eljárás keresési technikáját. A globális optimumkeresés elején nagy hőmérséklet értékeket célszerű alkalmaznunk (a modellek P elfogadási valószínűsége is nagy), annak érdekében, hogy kezdetben sok modellt (majdnem azonos valószínűséggel) kipróbáljunk. Amint a hőmérsékletet fokozatosan csökkentjük, egyre kisebb lesz az elfogadási valószínűség. Ez biztosítja, hogy ne történjenek nagy kiugrások a paramétertérben, és a globális optimum felé konvergáljon az eljárás.



76. ábra *Kétváltozós Gibbs sűrűségfüggvények különböző T értékek esetén*

Az aktuális modellparaméterekkel számított és mért adatok eltérését az E **energiafüggvény** jellemzi. Ha az adatok Gauss eloszlást követnek, akkor az $\vec{e}$ eltérésvektor L_2-normanégyzetének (az N adatszámmal történő normálás elhagyható) alkalmazása vezet optimális megoldásra

$$E_2 = \sum_{k=1}^{N} \left(d_k^{(m)} - d_k^{(sz)} \right)^2 = \min .$$

Kiugró adatok esetén az L_1-norma alkalmazása rezisztens megoldást biztosít

$$E_1 = \sum_{k=1}^{N} \left| d_k^{(m)} - d_k^{(sz)} \right| = \min .$$

Az SA eljárás véletlen keresést hajt végre a modelltérben, miközben a modellparamétereket iterációról-iterációra változtatja

$$m_i^{(új)} = m_i^{(régi)} + b \quad \text{ahol} \quad 0 \le b \le b_{max}$$

ahol b a paraméter változtatás mértéke ($bmax^{(új)} = \varepsilon \cdot bmax^{(régi)}$ ahol $0 < \varepsilon \le 1$). A régi (előző iterációs lépésbeli) és az új (aktuális iterációs lépésbeli) modellparaméterekkel számított energiafüggvény értékek különbsége

$$\Delta E = E\left(\vec{m}^{(új)}\right) - E\left(\vec{m}^{(régi)}\right).$$

Ha $\Delta E < 0$, akkor a mért és számított adatok illeszkedése javult, ellenkező esetben romlott. Az új modell elfogadására vonatkozó valószínűségi szabályt **Metropolis kritérium**nak nevezzük *(Metropolis, 1953)*

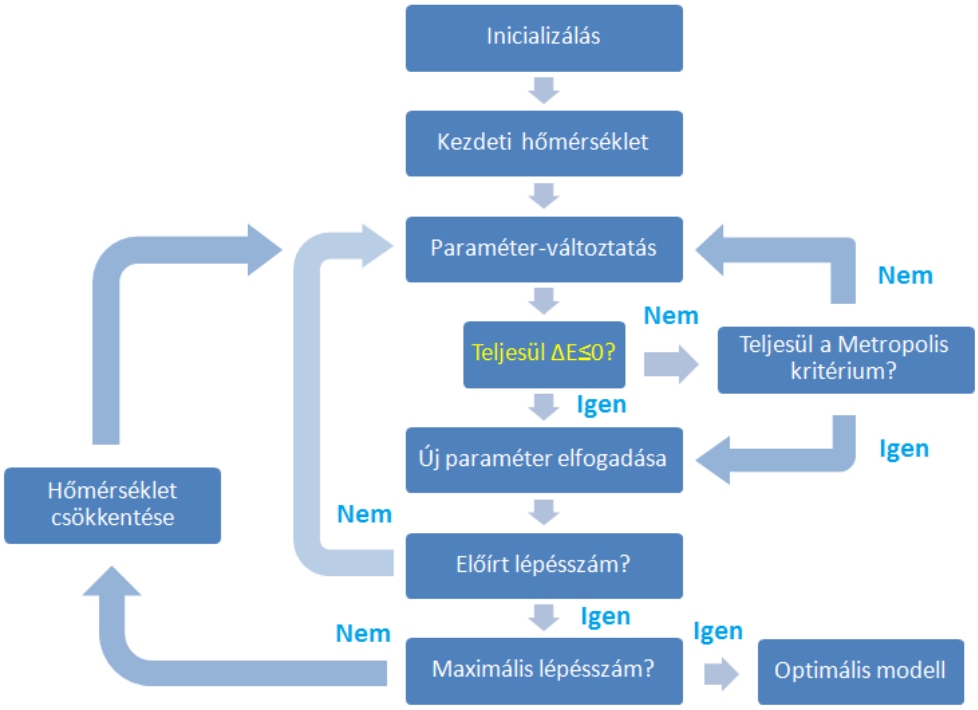
$$P = \begin{Bmatrix} 1 & \text{ha} \quad \Delta E \le 0 \\ e^{-\frac{\Delta E}{T}} & \text{ha} \quad \Delta E > 0 \end{Bmatrix}.$$

Ha a P elfogadási valószínűség nagyobb vagy egyenlő, mint egy α érték (0 és 1 tartományban egyenletes valószínűséggel generált véletlen szám), akkor az új modellt elfogadjuk, ellenkező esetben elvetjük. Ez tehát azt jelenti, hogy $\Delta E > 0$ esetben is van elfogadás, mely lehetővé teszi a lokális minimumból való kiszabadulást. A **hűtési ütem**, azaz a T általánosított hőmérséklet iterációs eljárásban történő csökkentése, nagymértékben befolyásolja az SA eljárás konvergenciáját. A globális minimumhoz történő konvergencia szükséges és elégséges feltétele a következő hűtési mechanizmus alkalmazása

$$T(q) = \frac{T_0}{\ln q} \quad \text{ahol} \quad q > 1$$

ahol q az iterációs lépésszám. A T_0 kezdeti hőmérséklet megadása empirikusan vagy próbafuttatásokkal történik. Például kiszámítjuk az azonos hőmérsékletnél elfogadott modellekhez tartozó energiafüggvény értékek számtani átlagát, az így kapott „átlaghibát" ábrázoljuk a hőmérséklet függvényében, majd a függvény minimumához tartozó hőmérsékletet fogadjuk el T_0-nak (energiaátlagok módszere).

Az SA eljárás folyamatábráját a 77. ábrán láthatjuk. Kezdetben az input adatok (mért adatok, startmodell, válaszegyenlet konstansok) és a folyamatjellemző paraméterek (kezdeti hőmérséklet, paraméter-változtatás mértéke, maximális paraméter-változtatás) beállítása után a modellvektor elemeit véletlenszerűen megváltoztatjuk. Ha ezzel csökken az energia-függvény értéke (a mért és számított adatok távolsága), akkor az új modellt elfogadjuk. Ellenkező esetben (az energia növekedése esetén) az új modell elfogadását a Metropolis kritériumhoz kötjük. Ha a megadott feltétel nem teljesül, akkor újra indítjuk a keresést. Ekkor még a paramétertér részletes vizsgálata céljából állandó hőmérsékleten folyik a keresés. Az eljárást előre megadott iterációs lépésszám elérése után (a belső ciklusból kilépve) kisebb hőmérsékleten és kisebb paraméterváltoztatást engedve folytatjuk. A fenti lépéssorozatot általában több ezerszer megismételjük. A kilépés feltételét a stopkritériumban határozzuk meg (előírt maximális lépésszám vagy ΔE küszöbérték), melynek teljesülése esetén a legutolsó lépésben elfogadott modellt tekintjük az inverz feladat megoldásának.



77. ábra *Az Simulated Annealing eljárás folyamatábrája*

Példa. A globális optimalizációs módszerek a kezdeti modell megválasztásától függetlenül konvergens és megbízható megoldást szolgáltatnak. Ezt a tulajdonságot **startmodell-függetlenség**nek nevezzük, melyet egy mélyfúrási geofizikai inverziós példán keresztül mutatunk meg. A 9. táblázatban szereplő modellparaméterek (ld. 12. fejezet) inverziós meghatározása (ahol H a rétegvastagságot jelöli, mely nem volt inverziós ismeretlen) céljából szintetikus *SP*, *GR*, *DEN*, *CN*, *AT*, *RMLL*, *RT* szelvényadatokat generáltunk (ld. 6. táblázat), majd azokhoz különböző mértékű zajt adtunk. Ily módon három különböző (quasi mért) adatrendszert hoztunk létre. Az első adatrendszerhez nem adtunk zajt (hibátlan adatok). A másodikat 2% Gauss eloszlásból származó zajjal terheltük. Az utolsót

6% Gauss zajjal szórtuk meg. A globális inverziós futtatások eredményeit a 10. táblázatban foglaltuk össze, ahol $D_{a,0}$ és $D_{m,0}$ a kiindulási, valamint D_a és D_m az inverzióval becsült modellhez tartozó adat- és modelltávolságot jelöli. Látható, hogy mindhárom adatrendszer esetén, a különböző helyről indított inverziós futtatások ugyanarra az eredményre vezettek. A maximális eltérés az illeszkedési értékek között 0.01% volt, ami elhanyagolható. Bizonyossággal azt mondhatjuk, hogy az adott inverziós problémánál az ismert modell környezetében 182%-os adattávolságon belül az SA eljárás startmodell-független. *Szabó (2006)* kimutatta, hogy még ennél is nagyobb adattávolságok (>1000%) esetén, ahol a lineáris módszerek már működésképtelenek (divergensek), az SA eljárás konvergens marad.

9. táblázat

H(m)	POR	SX0	SW	VSH	VSD
6.0	0.2	0.8	0.4	0.3	0.5
2.0	0.1	1.0	1.0	0.8	0.1
8.0	0.3	0.8	0.3	0.1	0.6
4.0	0.1	1.0	1.0	0.6	0.3

10. táblázat

Zaj(%)	$D_{a,0}$(%)	$D_{m,0}$(%)	D_a(%)	D_m(%)
0	24.47	20.00	0.001	0.002
0	63.90	50.01	0.001	0.002
0	182.53	150.00	0.001	0.002
2	24.47	20.00	2.37	3.44
2	63.90	50.01	2.36	3.44
2	182.53	150.00	2.37	3.44
6	24.47	20.00	6.95	9.95
6	63.90	50.01	6.96	9.96
6	182.53	150.00	6.96	9.96

A klasszikus Metropolis kritériumon alapuló SA eljárás igen hatékony és robusztus, azonban komoly hátrányaként említhető, hogy akár több nagyságrenddel nagyobb futási idő jellemzi, mint a lineáris módszereket. Ezért a módszert a kisebb gépidő elérése céljából többen továbbfejlesztették. Az új módszerek közül a legfejlettebb a **VFSA eljárás** *(Very Fast Simulated Annealing)*, mely exponenciális-hűtést alkalmaz úgy, hogy a globális optimum megtalálása biztosított marad *(Ingber, 1989)*. A VFSA módszer véletlen keresést hajt végre a modelltérben, azonban a modellvektor komponenseit a klasszikus SA módszertől eltérő módon változtatja meg

$$m_i^{(új)} = m_i^{(régi)} + y_i\left(m_i^{(max)} - m_i^{(min)}\right)$$

ahol y_i egy [-1,1] intervallumba eső véletlen szám (u_i pedig 0 és 1 intervallumból egyenletes valószínűséggel generált véletlen szám)

$$y_i = \mathrm{sgn}\left(u_i - 0.5\right)T_i\left[\left(1+\frac{1}{T_i}\right)^{|2u_i-1|} - 1\right].$$

A fenti kifejezésben T_i az i-edik modellparaméter egyedi hőmérsékletét jelöli, mely különbözik a T általánosított hőmérséklettől. Kimutatható, hogy az abszolút minimum az alábbi (a logaritmikus hűtéstől jóval gyorsabb) hűtési ütem esetén is garantálható

$$T_i^{(q)} = T_i^{(0)} e^{-c_i \sqrt[M]{q}}$$

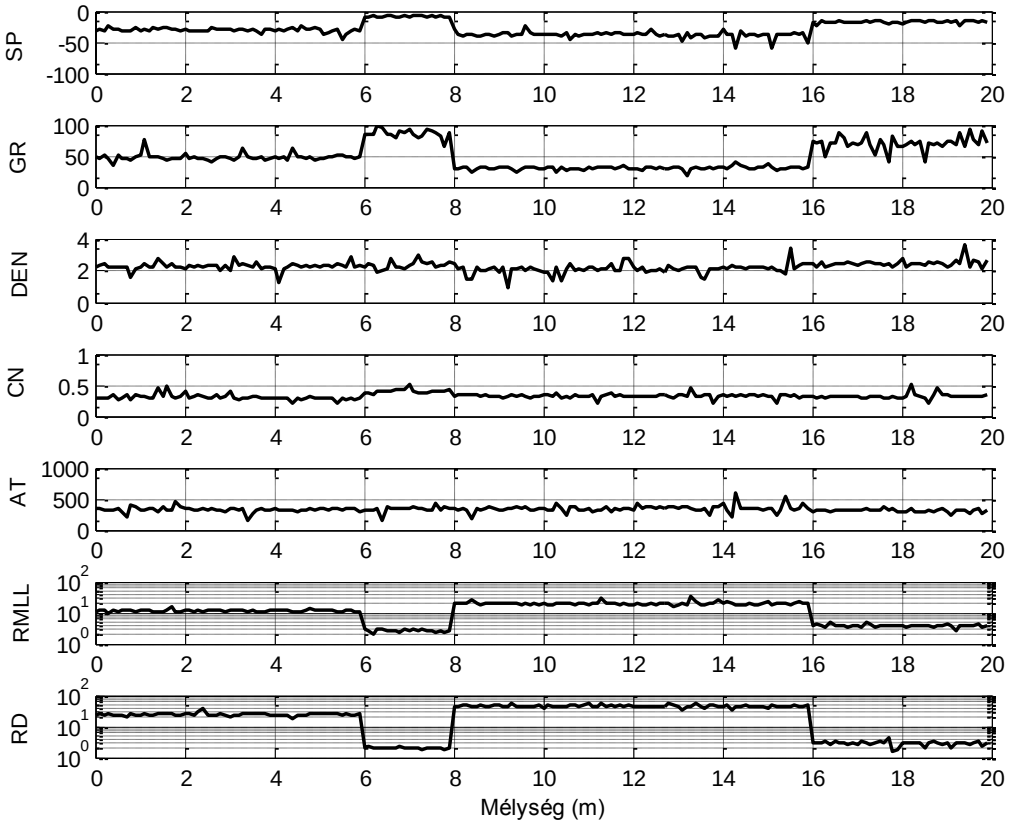
ahol $T_i^{(0)}$ az i-edik kezdeti modell-hőmérséklet, c_i az ehhez tartozó folyamatszabályozó konstans és M a modellparaméterek száma. A VFSA algoritmus a modellparaméterek megváltoztatásának módjától eltekintve hasonlóan épül fel, mint a klasszikus SA algoritmus (az új modellek elfogadását a Metropolis kritériumhoz köti).

A globális optimalizációs módszerek hátrányaként említhető az, hogy a becslés pontosságáról és megbízhatóságáról egyetlen futtatásból nem szolgáltatnak információt. Lehetőség van nagyszámú (ismételt) futtatásból a modellparaméterekre hibajellemzőket számítani, azonban ez irreálisan nagy gépidőt igényelne. Ehelyett alkalmazhatjuk az ún. **kombinált inverzió**s eljárást, mely relatíve kisszámú globális optimalizációs lépést követően átvált lineáris keresésre. Ennek lényege, hogy kezdetben a globális inverzió startmodell-függetlenségét kihasználva eljutunk az abszolút minimum környezetébe, ahonnan már a lokális módszerek is jól működnek és sokkal (legalább egy nagyságrenddel) gyorsabban megtalálják az optimumot. A gyorsaság mellett további előnyt jelent, hogy az utolsó lépésben számított $\underline{\underline{G}}$ érzékenységi mátrix felhasználásával a becsült paraméterek hibája és megbízhatósága a lineáris fázisban meghatározható *(Dobróka és Szabó, 2005)*.

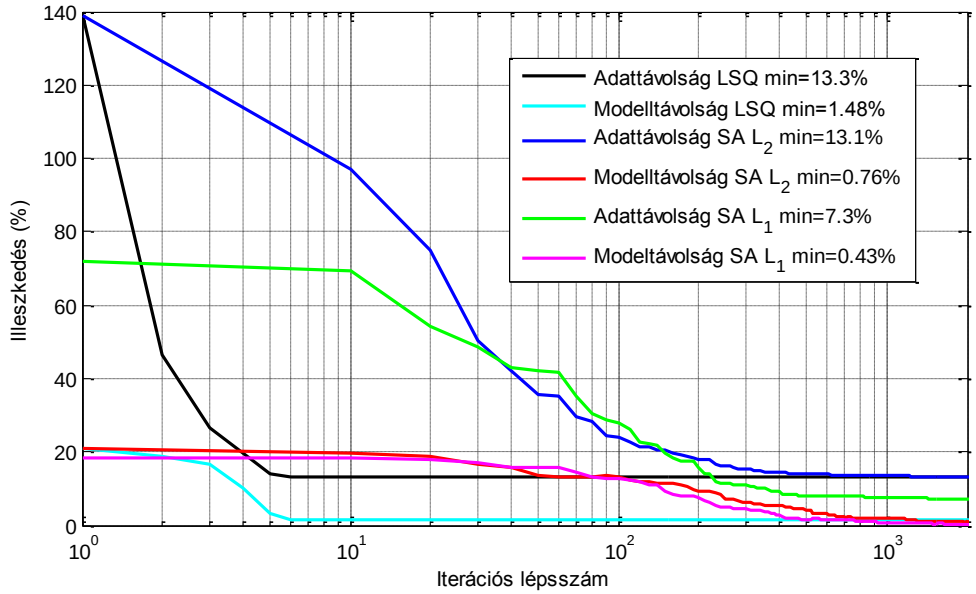
11. táblázat

H (m)	POR	SX0	SW	VSH	VSD
6.0	0.20	0.81	0.40	0.29	0.51
2.0	0.10	0.98	1.00	0.81	0.09
8.0	0.30	0.80	0.30	0.10	0.60
4.0	0.10	1.00	1.00	0.60	0.30

Példa. A 9. táblázatban szereplő kőzetfizikai modellt alapul véve új szintetikus adatrendszert generáltunk. Az adatokat az ún. intervallum inverziós módszerrel dolgoztuk fel *(Dobróka, 2001)*, melyeket 5% Gauss eloszlásból származó, továbbá kiugró adatokat tartalmazó véletlen zaj terhelte. Ez utóbbit úgy képeztük, hogy az adatok 1/5 részéhez az 5%-on felül további 25%-os Gauss zajt adtunk. A 78. ábrán az inverzió bemenő szelvényeit láthatjuk, ahol kiugró adatként jelenik meg pl. a természetes gamma (*GR*) szelvényen az 1.1m mélységben rögzített 75.3API, vagy a sűrűségszelvényen (*DEN*) 15.5m mélységben szereplő 3.42g/cm^3 érték. Először a lineáris LSQ módszert próbáltuk ki, mely köztudottan rosszul működik kiugró adatok jelenlétében. A 79. ábrán látható, hogy a becslésre vonatkozó adat- és modelltávolság ekkor a legnagyobb. Ezután az E_2 energiafüggvényen (L_2-normán) alapuló SA módszert alkalmaztuk. Ez sem volt rezisztens, viszont a globális optimalizáció kismértékben javított az eredményen. Az áttörést az E_1 energiafüggvényen (L_1-normán) alapuló SA módszer hozta. Ez gyakorlatilag kétszer jobb eredményt szolgáltatott a mért és számított adatok illeszkedése, és háromszor jobbat a becsült és az ismert modell illeszkedése szempontjából. Ez utóbbi inverziós eljárással kapott eredményt a 11. táblázatban láthatjuk.



78. ábra *Fúrási geofizikai szelvények (5% Gauss + 25% véletlen zaj)*



79. ábra *Lineáris (LSQ) és globális (SA) inverziós eljárások konvergenciája*

A **Genetikus Algoritmus (GA)** mint globális szélsőérték kereső eljárás a véletlen keresést használó módszerek csoportján belül az evolúciós algoritmusok között kap helyet. A GA működése biológiai analógián alapul, mely egyrészt az evolúciós fejlődést meghatározó természetes szelekciót, másrészt a genetikát használja fel. Köztudott, hogy a darwinisták szerint a természetben elsősorban azok az élőlények maradnak fenn és szaporodnak, melyek az adott körülmények között erre a legalkalmasabbnak bizonyulnak. Ezt az alapgondolatot felhasználva az első GA-okat az öröklődési mechanizmus leírására, később mesterséges rendszerek vizsgálatára alkalmazták. A GA robusztus és rendkívül jó adaptációs képességgel rendelkező globális optimalizációs eljárás, mely a geofizikai inverzió területén is *„változó körülmények között elfogadható teljesítmény”*-t nyújt a modell megválasztásától és az adatok eloszlástípusától függetlenül. A mesterséges populációk egyedeinek genetikai információit a DNS-lánc analógiája alapján kódolt számsorozatok (kromoszóma) hordozzák, melyek egyértelműen meghatározzák az optimalizációs probléma paramétereit. Mesterséges öröklődéskor a GA egy véletlen populációból választja ki a legalkalmasabb egyedeket (modelleket), azok között genetikus információcserét és mutációt hajt végre egy alkalmasabb generáció létrehozása érdekében. A GA a populációt a fenti genetikus operátorok (véletlen műveletek) alkalmazásával iteratív úton javítja. Alapvető különbség az SA módszerrel szemben, hogy a véletlen keresés nem pontról-pontra történik a modelltérben, hanem több pontot szimultán módon megvizsgálunk (több modellt javítunk párhuzamosan), mellyel hatékonyan elkerülhetők a lokális szélsőérték helyek. A modellparaméterek lehetséges tartományát az optimalizációs eljárás kezdetén meg kell adnunk, így egyes modell-tartományokat azonnal kizárunk a keresésből.

A GA következőképpen használható fel az inverz feladat megoldására. Az inverz probléma modellvektorát egy adott modellpopuláció egyedeként azonosítjuk. A populáció minden egyedéhez hozzárendelünk egy F alkalmassági (fitness) értéket, mely az egyed túlélési képességeit számszerűen jellemzi. Minél nagyobb ez az érték, az egyed annál nagyobb valószínűséggel és nagyobb számban szaporodik. Lényegében ez a fitness-érték határozza meg, hogy az egyedek bekerülnek-e a következő generációba vagy elpusztulnak. A GA az optimalizációs eljárás során a fitness, mint célfüggvény maximalizálására törekszik a legalkalmasabb modell megtartása érdekében. A fitness függvényt úgy kell megválasztanunk, hogy azzal a mért és a számított adatok eltérése mérhető legyen, és annak globális maximumához tartozzon az inverz feladat optimális megoldása. Az inverziós módszerek elméletében az $E^{(i)} = E\left(\vec{d}^{(m)} - \vec{g}\left(\vec{m}^{(i)}\right)\right)$ skaláris függvény (vektornorma) jellemzi a mérési és az i-edik modell (a $g^{(i)}$-edik direkt feladat keretében) alapján számított adatok eltérését. A GA eljárásban az E célfüggvény minimumát keressük, ezért a $F(\vec{m}) = max$ szélsőérték-feltétel biztosítása érdekében az **alkalmassági (fitness) függvény**t többféleképpen képezhetjük

$$F(\vec{m}^{(i)}) = -E^{(i)} \quad \text{vagy} \quad F(\vec{m}^{(i)}) = \frac{1}{E^{(i)} + \varepsilon^2}$$

ahol ε^2 az alkalmassági értékeket felülről szabályozó konstans, ahol i=1,2,…,S (S a populációt alkotó modellek száma). Az iterációs eljárás során konvergenciáról akkor beszélünk, ha az egymást követő populációk átlagos alkalmassági értéke fokozatosan nő.

A GA eljárás első lépésében beolvassuk a mérési adatokat, és a problémához illeszkedő alkalmassági függvényt definiálunk. Az eljárás elején meghatározzuk a keresési teret, és egy véletlen populációt (startmodell együttest) hozunk létre, melynek minden eleméhez kiszámítjuk az alkalmassági értéket. Ezután a modell paramétereit kódolt számsorozatokká alakítjuk (kódolás), amelyen közvetlenül alkalmazzuk a genetikus operációkat (véletlen műveleteket). Ezek közül a keresztezés művelete információcserét hajt végre két (véletlenül kiválasztott) kiinduló egyed között, melynek eredménye két teljesen új egyed lesz. Egy-egy egyed génjét a mutációs operátor alkalmazásával megváltoztathatjuk, melynél lényeges a **mutációs arány** (mutált egyedszám/összes egyedszám) megadása. Túl kis mutációs arány esetén a populáció könnyen homogenizálódhat, míg túl nagynál a keresés sokkal tovább tart és a konvergencia sem biztosított. Az új generáció összetételét a **reprodukció** művelete alakítja ki. Általában az átmeneti (genetikus műveleteken átesett) populáció egyedeiből építjük fel az új generációt, azonban létezik olyan algoritmus is, melyben kicseréljük az átmeneti populáció legrosszabb (legkisebb alkalmasság) egyedét a régi populáció legjobb (legnagyobb alkalmasság) egyedére. Ezt a műveletet **elitizmus**nak nevezzük. A fenti lépéseket addig ismételjük, míg egy megadott stopkritérium nem teljesül. Az utolsó generáció legalkalmasabb egyedét fogadjuk el az inverz feladat megoldásának.

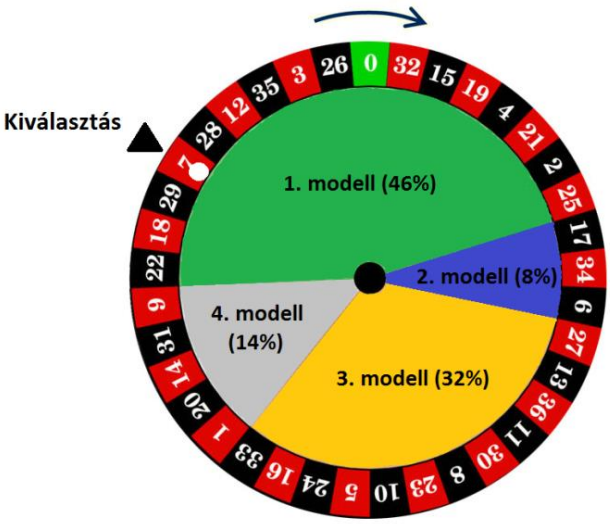
A **Klasszikus Genetikus Algoritmus** *(Classical Genetic Algorithm)* alapvető tulajdonsága, hogy a paramétereket kódolt formában kezeli, és a genetikus műveleteket közvetlenül a kódokon végzi el. Az optimumkeresés során az alkalmassági függvény kiértékeléséhez minden iterációs lépésben dekódolási művelet szükséges az elvi adatok számítása miatt. A **kódolás** *(coding)* legegyszerűbb módja a bináris átalakítás, mely a modellt egy véges hosszúságú számsorozattá konvertálja. Ennek szerkezete analóg a kromoszómával, mely a populáció minden egyedét genetikailag egyértelműen azonosítja. A modellt a modellparaméterek kódjaiból összefűzött bitfüzér reprezentálja, melynek alapelemeit, a géneket 0 és 1 értékű bitek alkotják. Például a 107-es érték kódolása 10 bit segítségével

$$0001101011 \quad \text{azaz} \quad 1\cdot 2^6 + 1\cdot 2^5 + 1\cdot 2^3 + 1\cdot 2^1 + 1\cdot 2^0 = 107.$$

A kódok a modellparaméterek értékeit csak korlátozott pontossággal bontják fel. Ha növeljük a bitek számát, a felbontás nő, viszont jelentősen lassul a GA eljárás. A bináris CGA algoritmus esetén a keresési tér pontjainak kódolását a bináris genetikus operátorok alkalmazása követi. Az első művelet a **szelekció** *(selection)*, mely véletlenszerűen kiválasztja és az alkalmassági értékük alapján sokszorosítja az egyedeket. Ez azt eredményezi, hogy a legalkalmasabb egyedek nagy számmal bekerülnek egy új populációba, melyben a legkisebb fitness értékkel rendelkező egyedek már nem vesznek részt. A leggyakrabban alkalmazott szelekciós művelet a Rulett-szelekció, mely a fitness-szel arányos kiválasztást valósít meg. A 80. ábrán látható, hogy a modellek kiválasztási valószínűsége a körcikkek területével arányosan növekszik. Az ábrán a legnagyobb túlélési képességekkel az 1. számú modell rendelkezik. A szelekció művelete kiválasztotta azt a következő populáció számára. Az ismétlés is engedélyezve van, tehát ezt a modellt egy következő próbálkozásnál újra versenyeztethetjük. A legkisebb esélye a túlélésre a 2. számú modellnek van. Ez a modell nagy valószínűséggel nem kerül kiválasztásra, mivel kis esélye van a túlélésre.

A **keresztezés** *(crossover)* műveletével a kiválasztott modelleket véletlenszerűen párosítjuk, majd részleges információcserét hajtunk végre közöttük. A legegyszerűbb keresztezési módszer az egypontos (egyszerű) keresztezés, ahol egy véletlen bitpozíciónál elmessük a kromoszómát és az attól jobbra elhelyezkedő géneket felcseréljük a két egyed között (a balra lévőket változatlanul hagyjuk). Például 10 bitből álló kromoszómák esetén a 8. bitpozíciónál a kiinduló modelleket elvágva az alábbi új modellpár jön létre

$$11000110|10 \qquad 1100011001$$
$$\rightarrow$$
$$10011100|01 \qquad 1001110010.$$



80. ábra *A Rulett-szelekció elve*

A bitfüzérek keresztezését modellparaméterenként egyszerre több bitpozíció mentén is elvégezhetjük többszörös metszéssel. A **mutáció** *(mutation)* művelete a CGA eljárás esetén egy (vagy több) bit értékének véletlenszerű megváltoztatásával hajtható végre. Például az alábbi modell 5. bitjének megváltoztatásával az eredmény

$$1100011010 \quad \rightarrow \quad 1100111010.$$

A fenti genetikus operátorok ismételt alkalmazásával a CGA eljárás a régi generációkból újabbakat generál (a generáció száma megegyezik az optimalizációs algoritmus iterációs lépésszámával). E folyamat végén az utolsó generáció egyedeit dekódolva az alkalmassági függvény maximumához tartozó modellt fogadjuk el megoldásként.

A klasszikus CGA meglehetősen időigényes eljárás. Elegendő, ha arra gondolunk, hogy az inverz probléma esetén iterációs lépésenként kódolás-dekódolás műveletet kell alkalmaznunk. A CGA módszert továbbfejlesztve létrehoztak olyan korszerű algoritmusokat, melynél a kódolás művelete elhagyható, és a bináris kódolás helyett természetesebb számábrázolással működnek. A **Valós Genetikus Algoritmus** *(Float-encoded Genetic Algorithm)* a gének lebegőpontos ábrázolásán alapul. Ez azt jelenti, hogy az FGA eljárás valós modell

paraméterekkel dolgozik, nem pedig kódokat választ ki, keresztez vagy mutál. Ez a módszer az inverz problémát is jobban reprezentálja. Az FGA esetén minden modell-paraméter egy-egy kijelölt (csak alulról és felülről korlátozott) valós intervallumból kerül ki, így a modelltér sokkal finomabban felbontható, mint bináris kódolással. Az FGA eljárás hatékonysága legfőképpen a futási idők tekintetében mutatkozik meg, mivel nagyság-rendekkel kisebb gépidőket produkál, mint a CGA eljárás.

Az FGA alapjában véve megegyezik a CGA-val, annyi különbséggel, hogy a genetikus műveleteket valós operátorokként definiálja és azokat közvetlenül a modellparamétereken (géneken) alkalmazza. A valós operátoroknak gazdag eszköztára van, melyből csak néhányat említünk meg ebben a jegyzetben. A **Rulett-szelekció** alkalmazása esetén az *i*-edik modell kiválasztásának valószínűségét a modell alkalmassági értékének és a populációban résztvevő összes modell fitness összegének hányadosa adja meg

$$P\left(\vec{m}^{(i)}\right)=\frac{F\left(\vec{m}^{(i)}\right)}{\sum_{j=1}^{S} F\left(\vec{m}^{(j)}\right)}.$$

Az *i*-edik egyed akkor kerül kiválasztásra, ha egy $\zeta \subset [0,1]$ egyenletes valószínűséggel generált számnál nagyobb az *i*-edik kumulatív valószínűség

$$C_{i-1} < \zeta_i < C_i \quad \text{ahol} \quad C_i = \sum_{k=1}^{i} P_k .$$

A fenti szelekciós eljárásban egyenként válogatjuk ki az egyedeket, addig, amíg a kiinduló modellek számának megfelelő új populációt nem kapunk. A kiválasztás következtében bizonyos egyedek elpusztulnak, mások pedig (akár többször is) kiválasztásra kerülnek. E művelet sémáját a 81. ábrán egy hat modellből álló populáció esetén mutatjuk meg. A fenti kiválasztást előíró feltételek alkalmazásán alapulnak a **rang szelekció**s műveletek is. E módszernél az egyedeket alkalmassági értékeik szerint sorba rendezzük, úgy hogy a legnagyobb alkalmassági értékű egyed rangja 1, a legkisebbé pedig *S* (populáció mérete) lesz. A normált geometriai rangszelekció esetén az *i*-edik egyed kiválasztási valószínűsége

$$P\left(\vec{m}^{(i)}\right)=\frac{q}{1-(1-q)^{S}}(1-q)^{r_i-1}$$

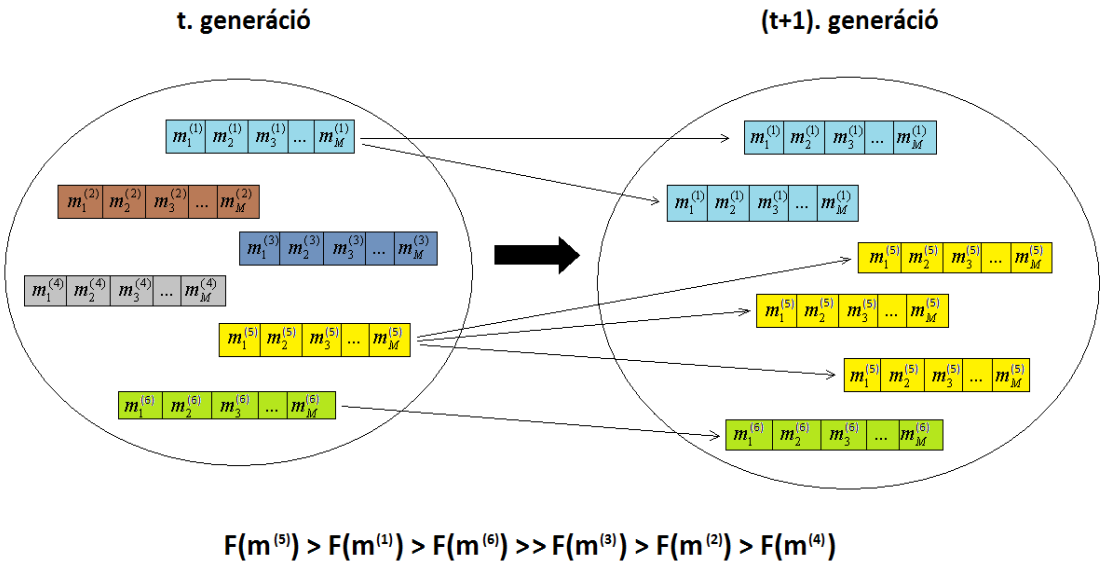
ahol r_i az *i*-edik modell rangja, *q* a maximális alkalmassági értékű modell kiválasztásának valószínűsége (előre megadott folyamatjellemző). E fenti két kiválasztási mechanizmus mellett érdemes említést tenni a **versenyszelekció**ról is, mely a legegyszerűbb, a kiválasztási valószínűség számítását nélkülöző eljárás. Alkalmazásakor egy bizonyos számú egyedet választunk ki (az ismétlés engedélyezése mellett) a populációból, majd ezek közül kiemeljük a legnagyobb alkalmassági értékkel rendelkező egyedet és áthelyezzük az új populációba. Ezt az eljárást egészen addig ismételjük, míg a kezdeti populációnak megfelelő számú egyedet nem kapunk.

A keresztezés legegyszerűbb módja az **egypontos (egyszerű) keresztezés**, melynek elvét a 82. ábrán láthatjuk. Valós esetben e művelet a következő

$$m_i^{(1,új)} = \begin{cases} m_i^{(1,régi)}, & \text{ha} \quad i < k \\ m_i^{(2,régi)}, & \text{ha} \quad i \geq k \end{cases}$$
$$m_i^{(2,új)} = \begin{cases} m_i^{(2,régi)}, & \text{ha} \quad i < k \\ m_i^{(1,régi)}, & \text{ha} \quad i \geq k \end{cases}$$

ahol $m_i^{(1,régi)}$ és $m_i^{(2,régi)}$ a kiinduló, és $m_i^{(1,új)}$ és $m_i^{(2,új)}$ a keresztezésen átesett modellek *i*-edik paramétere. A *k* véletlen egész számot a modellt felépítő paraméterek számának tartományából kell sorsolnunk ($k_{max}=M$), mely azt a paraméterpozíciót határozza meg, ahol a „metszést” végezzük. A következő gyakran alkalmazott művelet az **aritmetikus keresztezés**, mely a kiinduló modellek lineáris kombinációjával tér vissza. A keresztezéssel előálló modellpár $p \in [0,1]$ egyenletes valószínűséggel generált véletlen egész szám esetén

$$m_i^{(1,új)} = pm_i^{(1,régi)} + (1-p)m_i^{(2,régi)}$$
$$m_i^{(2,új)} = (1-p)m_i^{(1,régi)} + pm_i^{(2,régi)}.$$



81. ábra *Az FGA szelekció elve*

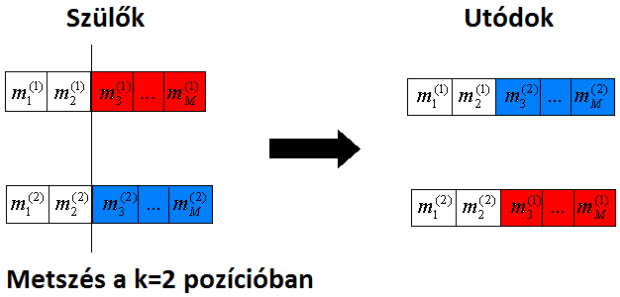
A **heurisztikus keresztezés** az aritmetikus keresztezés továbbfejlesztett változatának tekinthető, mely az $F(\vec{m}^{(2)}) < F(\vec{m}^{(1)})$ feltétel teljesülése esetén a következő módon származtatja a két új modellt

$$m_i^{(1,új)} = m_i^{(1,régi)} + p\left(m_i^{(1,régi)} - m_i^{(2,régi)}\right)$$
$$m_i^{(2,új)} = m_i^{(1,régi)}.$$

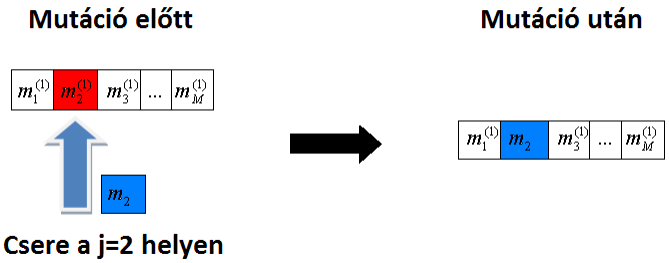
A harmadik genetikus operátor a mutáció, melynek sémája a 83. ábrán látható. Az ún. **uniform mutáció** az egyed *j*-edik paraméterét egy α véletlen számmal cseréli fel, mely a kiválasztott modellparaméter értéktartományából egyenletes valószínűséggel kerül ki

$$m_i^{(\text{új})} = \begin{cases} \alpha, & \text{ha } i = j \\ m_i^{(\text{régi})}, & \text{ha } i \neq j. \end{cases}$$

E művelet helyett alkalmazhatunk **határmutáció**t, mely a *j*-edik paramétert az *i*-edik paraméter minimális vagy maximális lehetséges értékére állítja be. A **nemuniform mutáció** pedig egy nem egyenletes eloszlásból származó véletlen számra cseréli fel a modellparamétert. Többszörös mutáció esetén egyszerre több modellparamétert is megváltoztathatunk.



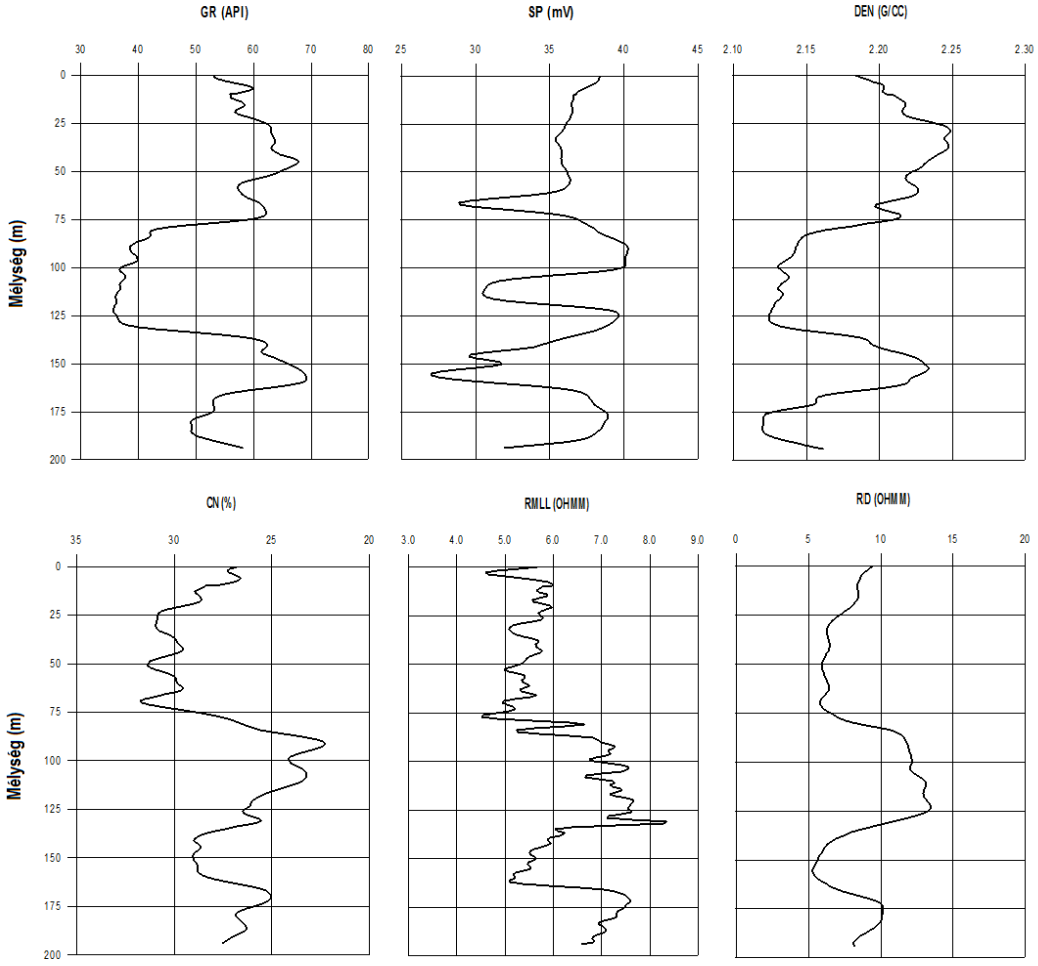
82. ábra *Az egyszerű keresztezés elve*



83. ábra *Az uniform mutáció elve*

Példa. A globális optimalizációs módszerek hatékonyan alkalmazhatók fúrási geofizikai adatok inverziójánál. Egy hazai szénhidrogén-kutató fúrás termálvizes rétegeiben mért karotázs szelvényanyag adatait dolgoztuk fel az FGA intervallum inverziós eljárással *(Dobróka 2001, Szabó 2004)*. Agyagos-homok rétegeket feltételezve ismeretlennek tekintettük a porozitást (*POR*), az agyagtartalmat (*VCL*) és a kvarc tartalmat (*VSD*). Mivel víztárolókban a pórusteret 100%-ban víz tölti ki, így a víztelítettségeket *SX0*=*SW*=1 konstansnak vettük. A négyréteges modellben rétegenként homogén kőzetfizikai jellemzőket (konstans modellparamétereket) feltételeztünk, ezért az ismeretlenek száma 12 volt. Az inverziós algoritmus alkalmas a réteghatárok helyzetét (réteghatár-koordináták) is automatikusan meghatározni, mellyel 15-re nőtt az ismeretlenek száma. Az inverziós eljárásba bemenő adatokat *SP*, *GR*, *DEN*, *RMLL*, *CN* és *RD* szelvények (ld. 6. táblázat)

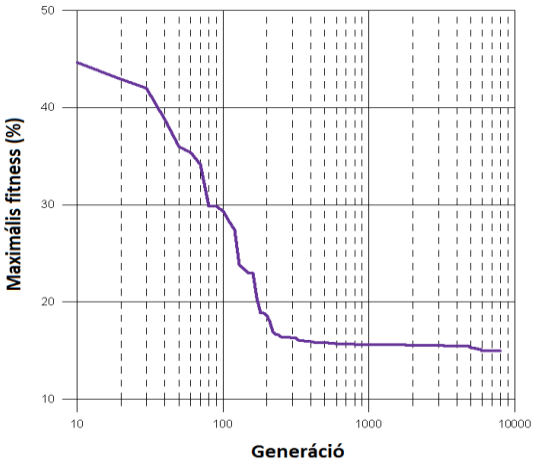
képezték (ld. 84. ábra). Az adatszám 1170 volt (195 mélységpontban), így az inverz feladat nagymértékben túlhatározott volt. Az FGA eljárást próbafuttatások alapján a 8000. iterációs lépésben megállítottuk. A 85. ábrán a generációk legjobb egyedének alkalmassági értékét (a mért és számított adatok eltérésének reciproka) ábrázoltuk az iterációs lépésszám (generációk száma) függvényében. Az inverziós eljárással becsült modell adattávolsága 6%-nak adódott (melyet a rétegben mért és számított adatok átlagának különbségeként definiáltunk). Az inverziós eredményeket a 86. ábra tartalmazza, melyen a pórustér, az agyag és a kvarc térfogatokat a mélység függvényében ábrázoltuk, valamint a becsült réteghatár-koordinátákat a mélységskála mentén piros színnel tüntettük fel.



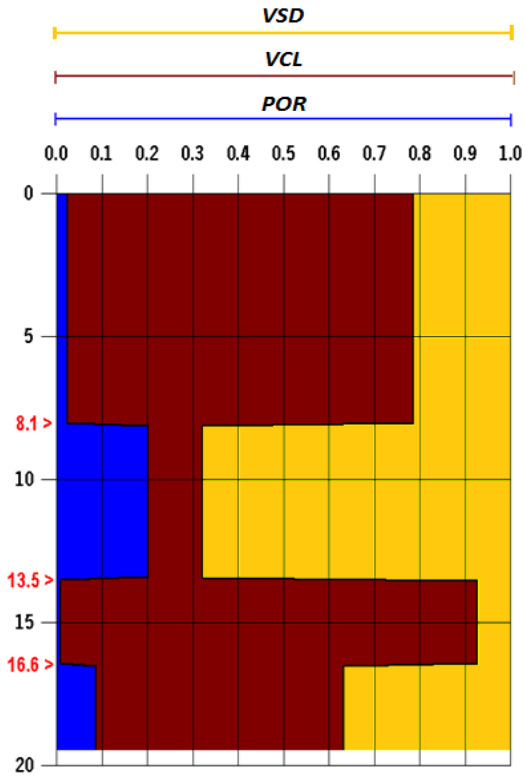
84. ábra *Az FGA inverziós eljárás bemenő szelvényei (MOL Nyrt. jóvoltából)*

A globális optimalizációs módszerek robusztusak és megbízhatóan működnek, viszont alkalmazásuk futási idő tekintetében hátrányos a lineáris eljárásokkal szemben. Ez a technika ugyanakkor rengeteg előnyös tulajdonságokat is mutat, pl. a pontosság, megbízhatóság, rezisztencia és jó adaptációs képesség tekintetében. Alkalmazásukat az is indokolja, hogy a lineáris optimalizációs módszerek gyakran kudarcot vallanak, mivel a Jacobi-mátrix megadásához elengedhetetlen a $\partial d/\partial m$ differenciálhányadosok numerikus számítása (rosszul

kondicionált lineáris egyenletrendszert kapunk), továbbá elegendő és megbízható előzetes információ szükséges a lokális szélsőértékhelyek elkerülése szempontjából (ami gyakran ezzel együtt sem sikerül). Mivel a globális eljárások nem alkalmaznak linearizálást, így azok nem igényelnek segédinformációkat (derivált-, és startmodell-függetlenek). E hatékony módszerek konvergenciáját azonban a folyamatjellemző paraméterek (GA-nál a genetikus operátorok paraméterei és azok kombinációja, SA-nál a kezdeti hőmérséklet, hűtési ütem és paraméterváltoztatás mértéke) megválasztása alapvetően befolyásolja.



85. ábra *Az FGA eljárás konvergenciája*



86. ábra *Az FGA inverziós eljárással kapott inverziós eredmények*

Irodalomjegyzék

Dobróka Mihály, 2001. Bevezetés a geofizikai inverzióba. Jegyzet, Miskolci Egyetem.

Dobróka M. and P. N. Szabó, 2005. Combined global/linear inversion of well-logging data in layer-wise homogeneous and inhomogeneous media. Acta Geodetica et Geophysica Hungarica, Vol. 40(2), pp. 203-214.

Edward H. Isaacs and R. Mohan Srivastava, 1989. An introduction to applied geostatistics. Oxford University Press.

Horvai György (szerk.), 2001. Sokváltozós adatelemzés (kemometria). Nemzeti Tankönyvkiadó, Budapest.

Kiss Bertalan és Ferenczy László, 1993. Szénhidrogén-tárolók mélyfúrási geofizikai értelmezése I. Nemzeti Tankönyvkiadó.

Lukács Ottó, 1987. Matematikai statisztika (Bolyai könyvek). Műszaki Könyvkiadó, Budapest.

Menke William, 1984. Geophysical data analysis – Discrete inverse theory. Academic Press, Inc. London Ltd.

Metropolis N., Rosenbluth M., Teller H. and Teller E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, pp. 1087-1092.

Móri Tamás, 1999. Főkomponens és faktoranalízis. ELTE Valószínőségelméleti és Statisztika Tanszék, jegyzet.

Sen M. and Stoffa P., 1995. Global Optimization Methods in Geophysical Inversion. Elsevier.

Steiner Ferenc, 1990. A geostatisztika alapjai. Tankönyvkiadó, Budapest.

Stoyan Gisbert, 2005. Matlab, frissített kiadás. Typotex.

Szabó Norbert Péter, 2004. Mélyfúrási geofizikai adatok modern inverziós módszerei. PhD értekezés. Miskolci Egyetem.

Szabó Norbert Péter, 2006. Geoinformatikai szoftverfejlesztés. Oktatási segédlet, Miskolci Egyetem, Geofizikai Intézeti Tanszék.

Köszönetnyilvánítás

A jegyzet a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként - az Új Magyarország Fejlesztési Terv keretében - az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Dr. Szabó Norbert Péter