Adatmodellezés, függvényillesztés

Bevezetés

Mérési (szimulációs) adatokhoz akarunk egy egyszerű alakú, általában valamilyen elméleti megfontolás alapján adott függvényt találni. A függvény paraméterek segítségével felírható. Cél olyan paraméterek találni, melyekkel a függvény legjobban visszaadja a mérési pontokat. Ezeknek a paramétereknek sokszor konkrét fizikai jelentése van. (Pl. ingamozgás kitérés-idő diagrammjára való illesztés segítségével akarjuk meghatározni a gravitációs állandót.)

Általában felírhatunk egy *költségfüggvényt* (hibafüggvény, energia) ami az illesztés jóságát jellemzi. Ennek minimumát keressük a paramétereket változtatva. Ez általánosan egy nemlineáris optimalizációs probléma melyről máshol szólunk. Bizonyos függvényillesztésekre vannak speciális, erre optimalizált megoldások.

Az illesztéskor figyelembe kell venni, hogy a mérési adatok hibát tartalmaznak, vagyis nem mindig a mérési pontokhoz legközelebbi függvényt keressük. Általánosan mindegyik koordináta tartalmazhat hibát.

Szükség lehet az illesztés során megtalált paraméterek pontosságára, hibájára is, amit befolyásol az adatok mérési hibája illetve az, hogy mennyire felel meg a modell a mérési adatoknak (szisztematikus hiba). Optimalizációnál gyakran probléma, hogy a költségfüggvénynek több lokális minimuma is lehet.

Az illesztési probléma lehet egy illetve több dimenziós, a paraméterek szerepelhetnek lineáris vagy nemlineáris alakban.

Gyakran előfordulnak *kiugró* (outlier) pontok melyek jelentősen torzíthatják az illesztést.

Megjegyzés: *A függvény illesztés és az interpoláció hasonló, de lényegében különböző probléma. Interpolációnál az ismert pontokon áthaladó, itt azokat közelítő függvényt keresünk.*

A legkisebb négyzetek módszere

Tegyük fel, hogy az (x_i, y_i), $i = 1...N$ mérési adatokra akarunk illeszteni egy függvényt melynek paraméterei a_j, $j = 1...M$.

$$y(x) = y(x; a_1, a_2...a_m)$$

A gyakran használt legkisebb négyzetek módszere a következő módon keresi a paramétereket:

$$\min_{a_1...a_m} \left(\sum_{i=1}^{N} [y_i - y(x_i; a_1...a_m)]^2 \right)$$

Nyilván nem ez az egyedüli lehetséges költségfüggvény, de bizonyos feltételezések mellett ez adja a paraméterek *leg-valószínűbb* halmazát. Pontosabban arra a kérdésre keressük a választ, hogy mely paramétervektor esetén maximális annak a valószínűsége, hogy az adott mérési eredményeket kapjuk. Ezt *maximális valószínűségű paraméterbecslésnek* (maximum likelihood estimator) hívjuk.

Ha *csak* az y_i adatok mérési hibáját vesszük figyelembe és ez a hiba *Gauss eloszlású*, valamint a hiba eloszlásának *szórása azonos* mindegyik mérési pontban, akkor a fent emlegetett valószínűség így írható fel:

$$P \propto \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{y_i - y(x_i)}{\sigma} \right)^2 \right] \Delta y \right\}$$

Ennek keressük a maximumát, vagy $-\log(P)$ minimumát.

$$-\log(P) = \left[\sum_{i=1}^{N} \frac{[y_i - y(x_i)]^2}{\sigma} \right] - N \log(\Delta y)$$

Mivel N, σ és Δy állandók, ez pont a legkisebb négyzetek módszerét adja vissza. P értéke megmondja, hogy mennyire jó az illesztés.

Fontos: *ha a fenti feltételek nem teljesülnek, a legkisebb négyzetek módszere ilyen formában nem optimális, de legalábbis nem a legnagyobb valószínűséghez tartozó paramétereket kapjuk.*

Khi-négyzet illesztés

Ha mérési pontok hibájának szórása nem egyforma, könnyen általánosítható a fenti módszer (ún. *khi-négyzet* illesztés) és a következő költségfüggvény kapható:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1 ... a_M)}{\sigma_i} \right)^2$$

Tekinthetjük úgy, hogy a σ_i szórások segítségével súlyozzuk az eltéréseket, vagy pedig, hogy egységnyi szórásúra normálunk minden pontnál.

Mivel a mérési pontokról feltételeztük, hogy normális eloszlást követnek, χ^2 ilyen véletlen változók négyzetének összege. Az ilyen típusú valószínűségi változók nem a Gauss eloszlást, hanem az ún. $(N - M)$ *szabadsági fokú khi-négyzet eloszlás*t követik. Ha az a_j paraméterek lineárisan szerepelnek, akkor ez az eloszlás analitikusan megadható, így megmondható annak valószínűsége (Q), hogy az adott paraméterekkel jellemzett modellen végzett mérés χ^2 -nél nagyobb eltérést ad. ($Q \approx 0.1$ tipikus, $Q \approx 0.01$ elfogadható, $Q < 0.001$ rossz modellre, vagy hibabecslésre utal.) Fontos, hogy a mérési hibák becslése jó legyen, különben megtévesztő eredményre juthatunk.

Annak feltétele, hogy χ^2 -nek minimuma van az, hogy az a_j paraméterek szerinti deriváltja 0 legyen.

$$\frac{\partial \chi^2}{\partial a_j} = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i^2} \right) \left(\frac{\partial y(x_i; a_1 ... a_M)}{\partial a_k} \right) \quad j = 1...M$$

Ez általában M nemlineáris egyenletből álló rendszerre vezet, de ha az a_j paraméterek lineárisan szerepelnek az $y(x; a_1 ... a_M)$ kifejezésben, akkor az egyenletek is lineárisak lesznek.

Példa: egyenes illesztése

A legegyszerűbb példa, amikor egyenest szertnénk illeszteni (lineáris regresszió).

$$y(x) = y(x; a, b) = a + bx$$

A költségfüggvény:

$$\chi^2(a, b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

A minimumban a deriváltak eltűnnek:

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

A fenti kifejezésekben a szummákat szétbonthatjuk az alábbi jelölések segítségével:

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

Így a minimum feltétele:

$$aS + bS_x = S_y$$

$$aS_x + bS_{xx} = S_{xy}$$

Az egyenletrendszer megoldása:

$$\Delta \equiv SS_{xx} - S_x^2$$

$$a = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta}$$

$$b = \frac{SS_{xy} - S_x S_y}{\Delta}$$

Figyelembe véve a hibaterjedés törvényét:

$$\sigma_f^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial f}{\partial y_i} \right)^2$$

behelyettesítve a-t és b-t:

$$\sigma_a^2 = S_{xx}/\Delta$$

$$\sigma_b^2 = S/\Delta$$

Ezek a hibák csak a mérési hibákhatását fejezik ki, ettől a pontok még szórhatnak messze az egyenestől. Az illesztés jóságát (azt, hogy mennyire jól modellezi az egyenes a mérési adatok közti összefüggést) az $(N-2)$ szabadsági fokú khi-négyzet eloszlás adja meg a χ^2 helyen.

Ha a mérés hibája nem ismert, akkor a fenti képletek $\sigma_i = 1$ behelyettesítésével használhatók (úgy tekintjük hogy mindegyik pont hibája megegyezik).

Ha (mint a valÓSÁgban általában) nem csak az y_i-knek van hibája, hanem az x_i-knek is, akkor az alábbira módosul a költségfüggvény:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)}{\sigma_{y_i}^2 + b^2\sigma_{x_i}^2}$$

Ennek deriváltjaiban, amiből a minimumot szerenénk meghatározni a lineárisan, viszont b nem lineárisan szerepel, így egy nemlineáris egyenletrendszert kell megoldani, vagy pedig általános minimumkeresést kell végrehajtani, explicit analitikus formula nem adható meg.

Általános lineáris legkisebb négyzetes illesztés

Az egyenesillesztést könnyen általánosíthatjuk bármi olyan függvényre ahol a paraméterek lineárisan szerepelnek. Ilyen pl. egy polinom:

$$y(x) = a_1 + a_2 x + a_3 x^2 ... + a_M x^{M-1}$$

De a hatványok helyett vehetjük x-nek tetszőleges $X_k(x)$ függvényeit:

$$y(x) = \sum_{k=1}^{M} a_k X_k(x)$$

Az X_k függvények természetesen lehetnek nemlineárisak, a probléma az a_k-k szempontjából lineáris marad. A költségfüggvény:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right)^2$$

Ennek minimumát keressük, melynek az a feltétele, hogy minden a_k szerinti derivált 0 legyen:

$$0 = \frac{\partial \chi^2}{\partial a_k} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(i) \qquad k = 1...M$$

Felcserélve a szummákat és bevezetve a következő jelöléseket:

$$\alpha_{kj} \equiv \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2}$$

$$\beta_k \equiv \frac{y_i X_k(x_i)}{\sigma_i^2}$$

az alábbi egyszerű kifejezést kapjuk:

$$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k$$

Ez egy lineáris egyenletrendszer, megoldható pl. a Gauss-Jordan módszerrel. Ez a módszer azért is jó, mert segítségével $C_{jk} \equiv [\alpha]_{jk}^{-1}$ inverzét is megkapjuk amire a paraméterek hibájának meghatározásakor szükség van.

$$a_j = \sum_{k=1}^{M} [\alpha]^{-1}_{jk} \beta_k = \sum_{k=1}^{M} C_{jk} \left[\sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \right]$$

Az a_j paraméterek hibáját pedig a hibaterjedés alapján kaphatjuk meg:

$$\sigma^2(a_j) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2$$

és mivel a C_{jk}-k nem függenek az y_i-ktől

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=1}^{M} C_{jk} X_k(x_i) / \sigma_i^2$$

Így a paraméterek hibája:

$$\sigma^2(a_j) = \sum_{k=1}^{M} \sum_{l=1}^{M} C_{jk} C_{jl} \left[\sum_{i=1}^{N} \frac{X_k(x_i) X_l(x_i)}{\sigma_i^2} \right]$$

A szögletes zárójelben lévő rész pont α_{kl} vagyis C inverze, vagyis:

$$\sigma^2(a_j) = C_{jj}$$

A diagonális elemek tehát megadják a paraméterek varianciáját.

Matematikailag a fent vázolt a legegyszerűbb megoldás, azonban bizonyos esetkben a lineáris egyenletrendszer megoldó algoritmusok a véges számábrázolás miatt elszállhatnak. Ilyenkor más minimumkereső algoritmus után kell nézni, pl. SVD (lásd a lineáris egyenlet megoldásáról szóló fejezetet, illetve [1]).

Megjegyzések:

1. Gyakran a paraméterekben nemlineáris problémák is átalakíthatók lineáris illesztéssé. Pl.:

$$y(x) = a \exp(-bx)$$

könnyen átírható a $z = log[y(x)]$ illetve $c = log(a)$ helyettesítéssel a

$$z(x) = c - bx$$

alakra, azaz a paraméterekben lineáris formára hozható.

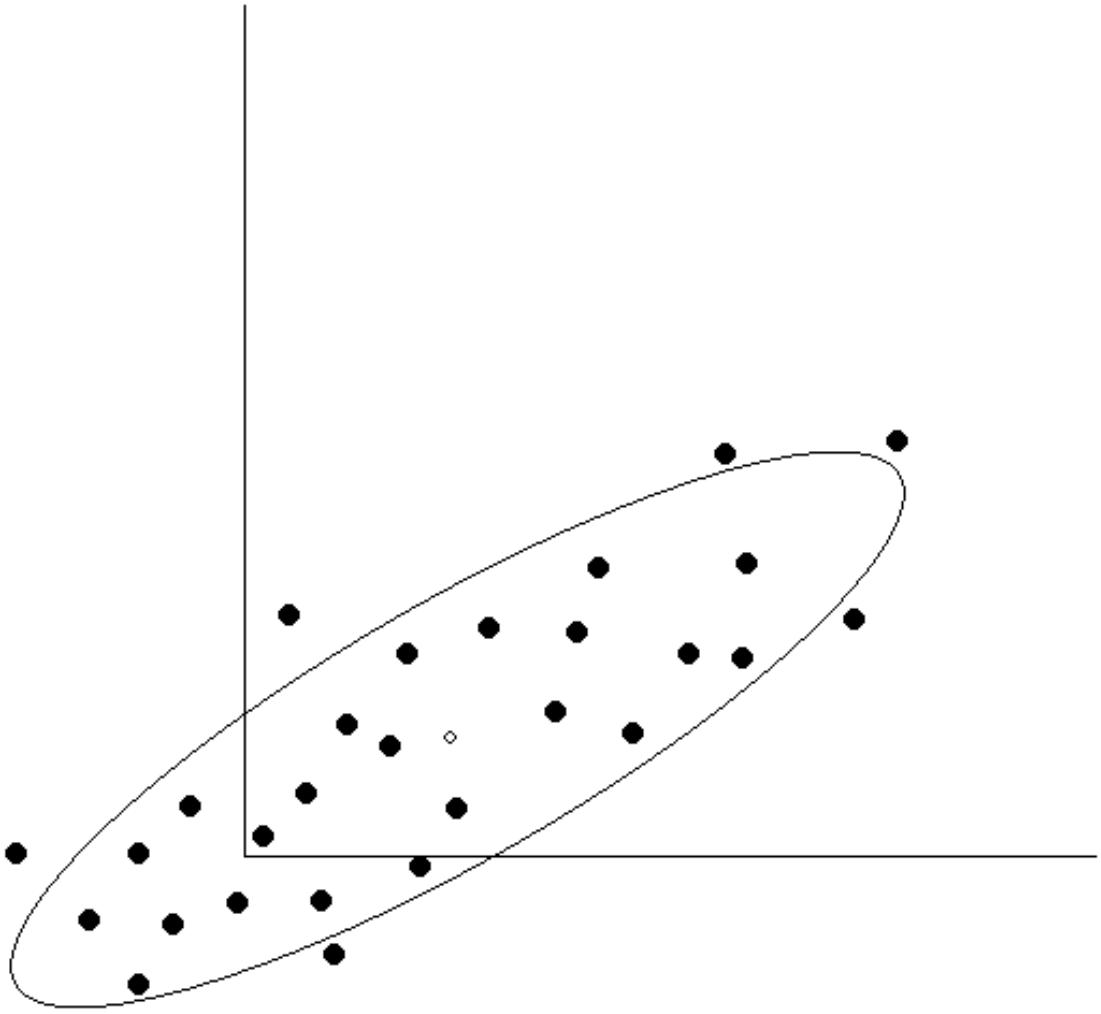
2. A fent vázolt általános legkisebb négyzetek módszerével való illesztés egyszeűen alakalmazható többdimenziós függvényillesztésnél is, csupán X_k az $\underline{x}_i$ vektorok függvényének kell tekinteni, a levezetés megegyezik az előzőekkel.

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{k=1}^{M} a_k X_k(\underline{x}_i)}{\sigma_i} \right)^2$$

3. Ha a paraméterek nemlineárisan szerepelnek a függvényben (pl. $y(x) = a\exp(-bx^2)$), akkor is hasonlóan írható fel a költségfüggvény, vizsont a minimumkeresés nem vezet lineáris egyenletrendszerre és csak iteratív nemlineáris optimalizációs módszerekkel kaphatjuk meg az eredményt. Egy elterjedt algoritmus a Levenberg-Marquardt módszer.

Konfidencia határok

A valódi jelenség paraméterit nem ismerjük. Azokat próbáltuk becsülni egy méréssorozat alapján. A mérések hibákat tartalmaztak. Azt szeretnénk megtudni, hogy az illesztésből származó paraméterek mennyire megbízhatóak. Elképzelhetjük ezt úgy is, hogy a paraméterek terében a valódi jelenség egy pont. Ha mérésteket végzünk, és arra való illesztéssel határozzuk meg a paramétereinket, akkor ezek a valódi pont körül szóródnak valamilyen eloszlásnak megfelelően. Jó lenne megmondani azt a tartományt pl. amin belül mondjuk 68% valószínűséggel vannak a paraméterek. A pontok eloszlása ebben a térben gyakran igen bonyolult összefüggésben van a mért paraméterek hibájával, azokból közvetlenül nem számítható ki. A probléma az, hogy az eloszlás valódi centrumát nem ismerjük, illetve, hogy nem tudunk sok mérést végezni, hogy így rajzoljuk fel az eloszlást.



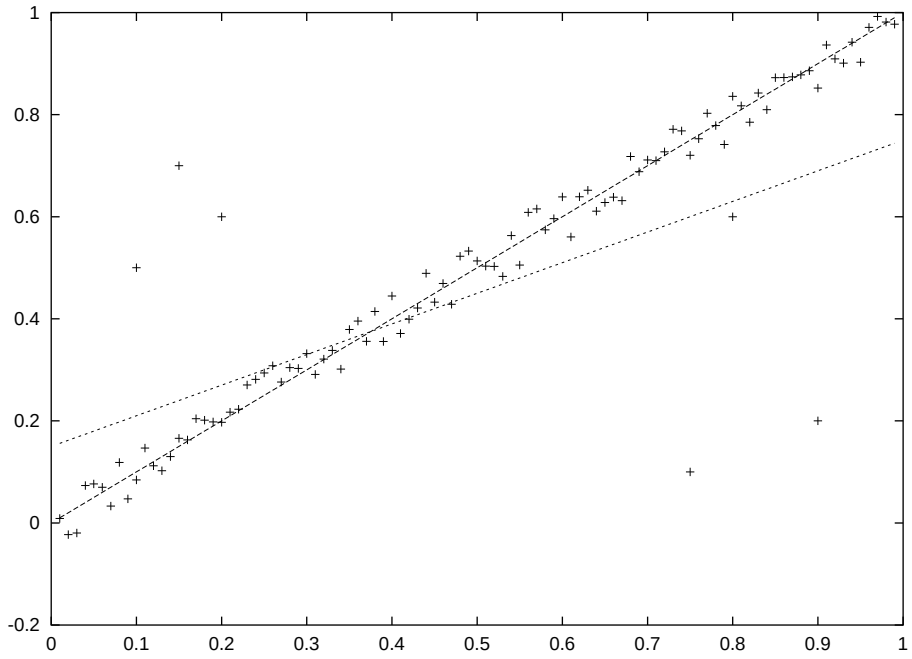
A sok mérést szimulálhatjuk Monte Carlo eljárással. Vegyük az illesztésből kijött paraméterekkel felírt modellt, és az alapján generáljunk mérési pontokat, melyhez adjunk hibákat. Ezekre a szimulált mérési adatokra ismét illesszünk modellt, és jegyezzük fel a kapott paramétereket. Sok ilyen szimuláció kirajzol egy eloszlást. Ez az eloszlás nem a valódi paramétervektor körül lesz, viszont feltételezhetjük, hogy az eloszlás alakja, paraméterei, megegyeznek az igazival.

Egy egyszerű, gyakran használt Monte Carlo módszer az ún. *bootstrap* módszer. Itt egyszerűen a meglévő mérési pontokat használjuk fel, úgy hogy úgy generálunk újabb és újabb szimulált méréssorozatokat, hogy *visszatevéssel* veszünk ki mérési pontokat. Ha pl. mindig N pontból álló sorozatokat csinálunk, azokban egy részben többszörösen szereplő pontok is lesznek. Ezekre a halmazokra aztán mindegyikre illesztünk, és az így kapott paramétervektorok eloszlását nézzük. Ez a módszer csak akkor ad értékelhető eredményt, ha a mérési pontok egymástól függetlenek és felcserélhetőek.

Ha a hibák Gauss eloszlásúak, akkor a költségfüggvény optimuma körül a költségfüggvény-(hiper)felület a paraméter-(hiper)síkkal párhuzamosan elmetszve, és azt levetítve ellipszoidokat kapunk. Ezek szintén használhatóak konfidencia határként. Nem Gauss hibák esetében nem ellipszoidokat, hanem általában bonyolultabb alakokat kapunk.

“Robusztus” becslés

A kiugró pontok (outliers) nagyon el tudják rontani az illesztést.



Ha a mérési hibának az eloszlását nézzük, akkor a Gauss-nál sokkal jobban kiterjedt, ún. hosszúfarkú eloszlást találunk. A probléma megoldása elvileg az, hogy visszamegyünk a kiindulóponthoz, amikoris felírtuk egy adott paramétervektorhoz tartozó valószínűséget:

$$P \propto \prod_{i=1}^{N} \{\exp\left[-\rho(y_i, y(x_i; a_1, ...a_M))\right] \Delta y\}$$

csak a Gauss eloszlás helyett az aktuális (hosszúfarkú) eloszlást írjuk be. Ha ρ a sűrűségfüggvény negatív logaritmusát

jelöli (Gaussnál ez egyszerűen a két argumentum különbségének négyzete osztva a szórásnégyzettel), akkor a fenti formula segítségével a legkisebb négyzetek módszeréhez hasonló kifejezéseket kapunk a fenti valószínűség maximalizálásakor, vagyis az alábbi kifejezés minimumát keressük az a_k paraméterek terében:

$$\sum_{i=1}^{N} \rho \left(\frac{y_i - y(x_i; a_1 ... a_M)}{\sigma_i} \right)$$

Egy másik lehetséges robusztus módszer nem a négyzetes eltérést minimalizálja, hanem az eltérések abszolút értékét. Mivel az eltérés nam négyzeten szrepel, a nagy eltérések relatíve kisebb súllyal adódnak a költségfüggvényhez, így kevésbé húzzák el az illesztést.

$$\sum_{i=1}^{N} \left| y_i - y(x_i; a_1 ... a_M) \right| / \sigma_i$$