

Többdimenziós normális eloszláshoz közeli sűrűségfüggvények becslése és képi megjelenítése 3D-ben

Estimation of multivariate densities beyond Gaussian distributions and visualization in 3D

Dr. TERDIK György

Debreceni Egyetem Informatikai Kar
4028 Debrecen, Kassai út 26, Magyarország,
<http://www.inf.unideb.hu>, Terdik.Gyorgy@inf.unideb.hu

Abstract

This paper discusses the visualization of multivariate density functions close to the Gaussian model. It briefly reviews the Gram-Charlier approximation and the relationship between Gaussian and skew-normal distributions. The main emphasis is on graphical representations of three-dimensional densities using spiral and cone-based constructions. The proposed visualizations provide an intuitive interpretation of departures from normality and help reveal positivity problems of Gram-Charlier density approximations.

Keywords: multivariate density, Gram-Charlier approximation, skew-normal distribution, visualization, shrinkage

Kivonat

A dolgozat a normális eloszláshoz közeli többdimenziós sűrűségfüggvények vizualizációjával foglalkozik. Röviden ismertetjük a Gram-Charlier-közelítés alap gondolatát, valamint a normális és a ferde normális eloszlások közötti kapcsolatot. Kiemelt figyelmet fordítunk a háromdimenziós sűrűségfüggvények szemléltetésére spirális és kúp felületi reprezentációk segítségével. Az ábrák alkalmasak a normálítástól való eltérések, valamint a Gram-Charlier-becslés pozitivitási problémáinak bemutatására.

Kulcsszavak: többváltozós sűrűség, Gram-Charlier-közelítés, ferde normális eloszlás, vizualizáció, zsugorítás

1. BEVEZETÉS

A többváltozós adatok vizsgálatakor gyakran a normális eloszlás szolgál kiindulópontként. A valóság azonban ritkán ilyen szabályos. Egy pénzügyi adathalmaz, egy kommunikációs rendszer mérési eredménye vagy akár egy gépi tanulási modell kimenete gyakran torzult, ferde alakú vagy szokatlanul nagy szélsőérték-gyakoriságú eloszlást mutat [1], [2]. A pénzügyi és biztosításmatematikai alkalmazásokban gyakran nem az átlagos, hanem a szélsőséges események viselkedése áll az érdeklődés középpontjában. Ezek vizsgálatára szolgál a nagy veszteségek tartományát leíró módszer (Multivariate Tail Conditional Expectation (MTCE)), amely a kockázatos, ritkán előforduló állapotokban egy többdimenziós rendszer komponenseinek várható értékét jellemzi ([7]). A szélsőséges értékek tartományának meghatározása és becslése ezért nagymértékben függ az alkalmazott többváltozós modelltől, ami indokolja a normális eloszláshoz közeli modellek vizsgálatát. A kérdés nemcsak az, hogyan írjuk le ezeket a jelenségeket matematikailag, hanem az is, hogyan tehetjük őket láthatóvá. E dolgozat a normális eloszláshoz közeli többdimenziós sűrűségfüggvények becslésének és háromdimenziós sűrűségfüggvények megjelenítésének néhány lehetőségét mutatja be.

2. NORMÁLIS ELOSZLÁSHOZ KÖZELI MODELL, A GRAM-CHARLIER MÓDSZER

A többdimenziós ritka események vizsgálatához olyan valószínűségi modellre van szükség, amely a normális eloszlásnál rugalmasabban képes leírni az adatok viselkedését. Erre alkalmas eszköz a Gram-Charlier-közelítés, amely a normális eloszlás geometriai szerkezetét megtartva figyelembe veszi a ferdeség és

a lapultság hatását és lehetőséget ad a normálistól való eltérések modellezésére és grafikus megjelenítésére is. Egy dimenzió esetén tekintsük egy X változó $f_X(x)$ sűrűség függvényének az Hermite polinomok általi sorfejtését

$$f_X(x) = \left(1 + \sum_{k=3}^{\infty} \frac{1}{k!} B_k(0,0, \kappa_{Y,3}, \dots, \kappa_{Y,k}) H_k(y) \right) \varphi(x|\mu_X, \sigma_X^2),$$

ahol B_k a Bell-polinomot, μ_X a várható értéket, σ_X^2 a szórásnégyzetet, φ pedig a normális eloszlás sűrűségfüggvényét jelöli. Továbbá az X standardizáltja $Y = (X - \mu)/\sigma$, amelynek értékei y , és $\kappa_{Y,k}$ a megfelelő kumulánsok. Az $f_X(x)$ becslése során természetesen nem a végtelen sorfejtést hanem csak ennek az első két tagját érdemes használni:

$$\tilde{f}_X(x) = \left(1 + \frac{1}{3!} B_3(0,0, \kappa_{Y,3}) H_3(y) + \frac{1}{4!} B_4(0,0, \kappa_{Y,3}, \kappa_{Y,4}) H_4(y) \right) \varphi(x|\mu_X, \sigma_X^2).$$

Ennek több oka is van:

1. a $\kappa_{Y,3}$ és $\kappa_{Y,4}$ a harmadik és a negyedik momentumon alapul és a szóban forgó eloszlás ferdeségét és lapultságát jellemzi,
2. már a harmadik és a negyedik momentum becsléséhez is nagy mintaelem szükséges,
3. a Bell polinomok egyszerű alakja $B_3(0,0, \kappa_{Y,3}) = \kappa_{Y,3}$, $B_4(0,0, \kappa_{Y,3}, \kappa_{Y,4}) = \kappa_{Y,4}$,
4. további tagok figyelembevételével a becslés nehezebbé válik, az interpretáció elveszíti egyszerűségét, a pozitivitási problémák is gyakoribbak.

A többdimenziós esetben a Gram–Charlier közelítés formailag analóg:

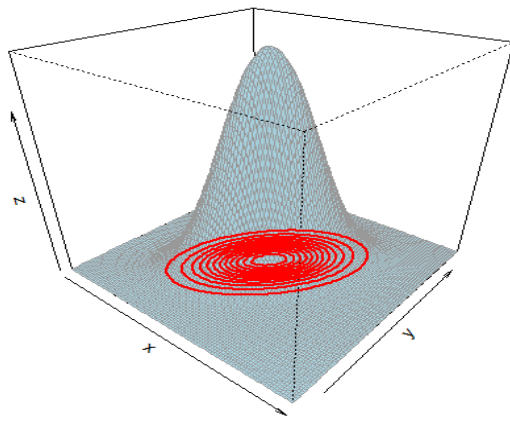
$$\tilde{f}_X(x) = \left(1 + \frac{1}{3!} H_3^T(y) \kappa_{Y,3}^{\otimes} + \frac{1}{4!} H_4^T(y) \kappa_{Y,4}^{\otimes} \right) \Big|_{y=\Sigma_X^{-1/2}(x-\mu)} \varphi(x|\mu_X, \Sigma_X),$$

de például $Y = \Sigma_X^{-1/2}(X - \mu_X)$, viszont a H_k Hermite tenzort jelöl vektor alakban, továbbá a $\kappa_{Y,3}^{\otimes}$ a ferdeség és a $\kappa_{Y,4}^{\otimes}$ lapultság szintén tenzor-vektorok. A Gram–Charlier-közelítés részletes kifejtése megtalálható a [3] könyvben, illetve a becslések elvégezhetők a MultiStat R-csomag segítségével, lásd [5], [6].

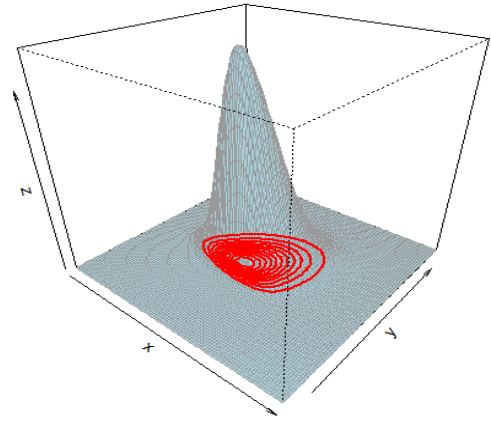
3. KÉPI MEGJELENÍTÉS

A Gram-Charlier-közelítés lehetővé teszi, hogy a normális eloszlástól való eltéréseket a ferdeség és a lapultság segítségével írjuk le. A ferde normális eloszlás a többváltozós normális eloszlás természetes általánosítása, amely egy további alakparaméter segítségével képes az aszimmetria modellezésére. Míg a többváltozós normális eloszlás szimmetrikus, a ferde normális eloszlás alakparamétere lehetővé teszi az aszimmetria figyelembevételét. A paraméter zérus értéke esetén visszakapjuk a klasszikus többváltozós normális eloszlást, ezért a ferde normális modell alkalmas a normalitástól való fokozatos eltérések leírására. A közelítés hatásának megértését jelentősen segíti a sűrűségfüggvények grafikus megjelenítése, különösen többdimenziós esetben, ahol a normálistól való eltérések a képletekből nehezen érzékelhetők. Mielőtt a Gram-Charlier-közelítés által létrehozott alakzatokat vizsgálnánk, célszerű áttekinteni a kétváltozós normális, illetve ferde normális eloszlások jellegzetes felületeit és szintvonalait. Ezek szolgálnak referenciamodellekként a későbbi összehasonlításokhoz.

Két dimenzióban a sűrűségfüggvény viszonylag könnyen ábrázolható felületként vagy szintvonalakkal (1. ábrák). Három változó esetén azonban a helyzet lényegesen bonyolultabb. A négydimenziós térben értelmezett sűrűségfüggvény közvetlen geometriai megjelenítése már nem lehetséges, ezért a szerkezet feltárásához speciális vizualizációs megoldásokra van szükség. Még a háromdimenziós normális eloszlás esetében sem egyszerű a valószínűségi tömeg elhelyezkedésének érzékeltetése. Az izosűrűségi felületek egymásba ágyazott ellipszoidok, amelyek fontos geometriai tulajdonságokat hordoznak, ugyanakkor hagyományos ábrázolásuk gyakran takarásba kerül vagy nehezen értelmezhető.



2D Normális



2D Ferde Normális

1. ábra. 2D normális és ferde normális sűrűségfüggvények és 2D kontúr ábrák

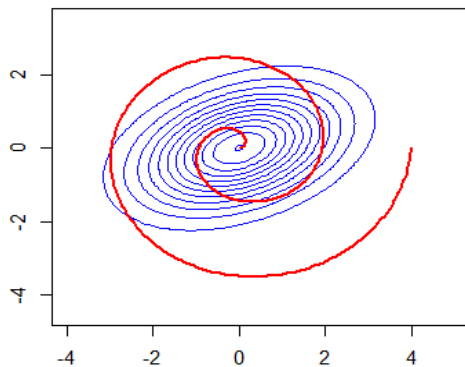
A háromdimenziós eloszlások vizualizációja nemcsak technikai, hanem szemléltetési probléma is. Míg a kétdimenziós esetben a szintvonalak és a felületi ábrák közvetlenül értelmezhetők, addig három dimenzióban a sűrűségfüggvény már egy négydimenziós objektum. Emiatt olyan reprezentációkra van szükség, amelyek a sűrűség szerkezetéről információt adnak anélkül, hogy a teljes négydimenziós objektum megjelenítését igényelnék. A spirál- és kúprezentációk ezt a célt szolgálják.

3.1 Spirál

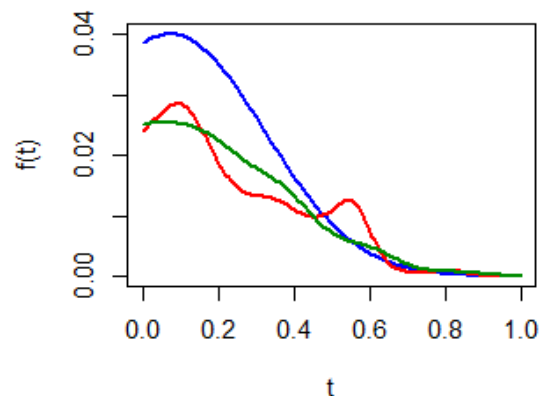
Egy viszonylag egyszerű módszer a sűrűségfüggvény értékeinek megjelenítésére 3D-s pontokban például a spirál [7]. Defináljuk az $(x, y, z) \in \mathbb{R}^3$ pontokat a következőképp

$$x(t) = r(t)\cos(p_s\pi t), \quad y(t) = r(t)\sin(p_s\pi t), \quad z(t) = 1 - t,$$

ahol $t \in (0,1)$. Látható, hogy a spirál $(0,0,1)$ pontból indul, és a sugarat $r(t) = p_s t^p$ (ahol $p = 1.05$) mértékben növelve éri el az x, y síkot, a $p_s = 4$ a fordulatok számát jelzi, lásd 2. ábra.



Spirál a 2D normális kontúr felett



3D normális sűrűségfüggvények spirális görbái

2. ábra. 3D spirál kontúrja és sűrűségfüggvények spirális görbái

3.2 Kúp

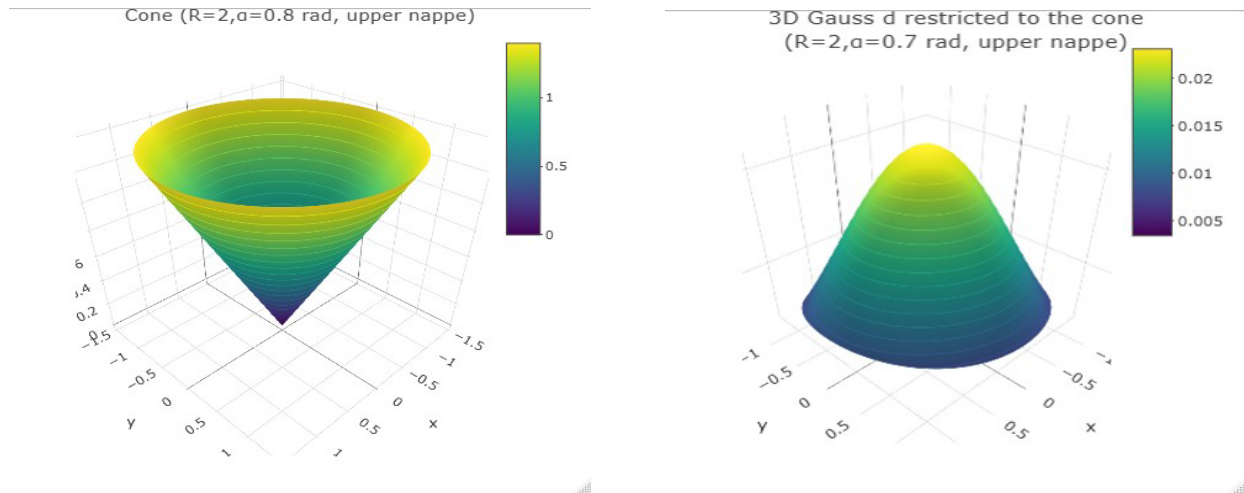
A spirál térbeli alakjának érzékeltetését egy kúpfelület segítségével egészítjük ki. A kúp egyszerű geometriai keretet ad a háromdimenziós ábrázoláshoz és jól szemlélteti a görbe térbeli elhelyezkedését. A 3D-s pontok $\mathbf{v} = (x, y, z)^T \in \mathbb{R}^3$ most egy kúp felületén helyezkednek el. Legyen most $\alpha \in (0, \pi/2)$ a kúp félnyílásszöge. Az origóból induló a pozitív z -tengely irányú kúp pontjai

$$\mathcal{C}_+ = \{(x, y, z) \in \mathbb{R}^3: z \geq 0, x^2 + y^2 = z^2 \tan^2 \alpha\}.$$

A paramétereket a következőképpen definiáljuk:

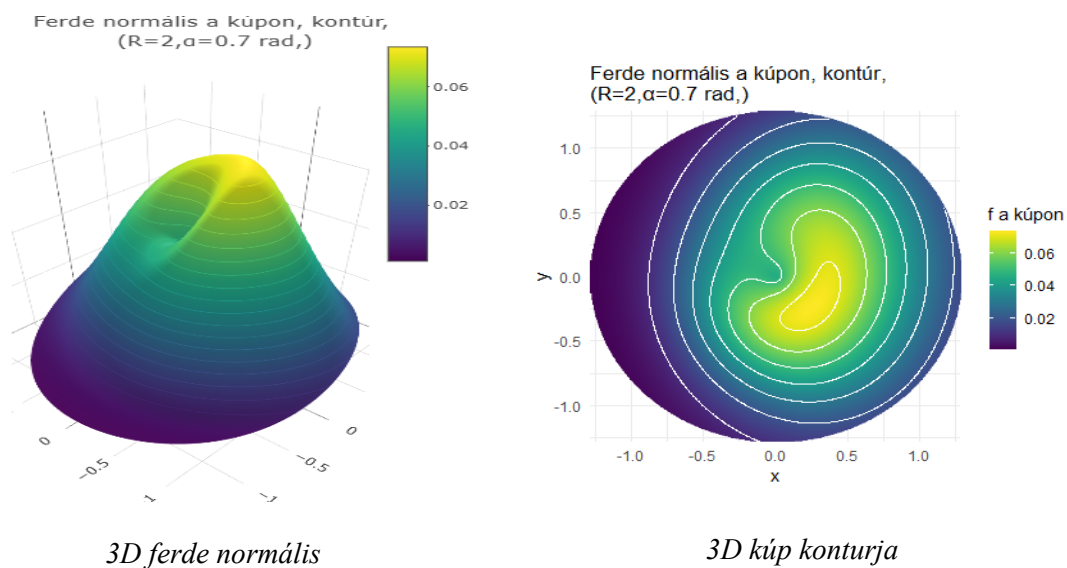
$$v(r, \theta) = \left[r \cos \theta, r \sin \theta, \frac{r}{\tan \alpha} \right], \quad r \geq 0, \quad 0 \leq \theta < 2\pi.$$

Bemutatjuk a normális és a ferde normális sűrűségfüggvények 2D-s ábrázolását a kúp segítségével 3. és 4. ábrákon.



3. ábra. A kúpfelület és a 3D normális sűrűségfüggvény megjelenítése

A kúpfelület előnye, hogy a sűrűség szintek egyetlen könnyen értelmezhető geometriai objektumon jeleníthetők meg.



3D ferde normális

3D kúp kontúrja

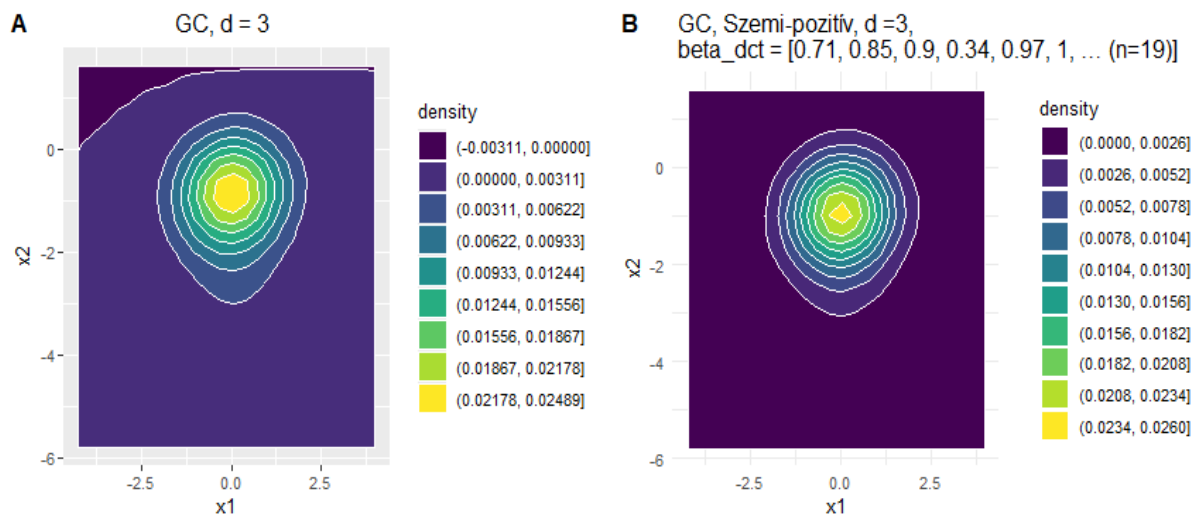
4. ábra. A 3D ferde normális sűrűségfüggvény ábrázolása a kúpfelületen és annak kontúrja

A szimmetrikus és aszimmetrikus eloszlások közötti különbségek a kúpon sok esetben szemléletesebben felismerhetők, mint a szokásos szintvonal-ábrázolásban.

4. POZITIVITÁSI PROBLÉMA

A sűrűségfüggvények egyik alapvető tulajdonsága, hogy pozitív. Aligha kell indokolni, hogy a Gram–Charlier-becslés egyik gyenge pontja az, hogy semmi nem garantálja a becsült sűrűségfüggvény pozitivitását az adott tartományban. Erről nagyon könnyű meggyőződni a spirál vagy a 3D-kúp kontúrjának ábrázolásával. Abban az esetben, ha negatív értékeket találunk, az egyik módszer a Hermite-polinomok együtthatóinak zsugorítása (shrinkage). Bemutatunk egy példát, amikor a ferde 3D-s sűrűségfüggvény Gram–Charlier-becslése

negatív értékeket tartalmaz, és a zsugorítási eljárással megváltoztatott együtthatókkal javított változatot, 5. ábra.



5. ábra. Negatív tartományokat tartalmazó és zsugorított becslések összehasonlítása, 3D-kúpok kontúrjai

5. ÖSSZEFOGLALÁS

A többváltozós eloszlások vizualizációja különösen nehéz feladat, mivel már három változó esetén is négydimenziós objektumként kellene a sűrűségfüggvényt ábrázolni. A megfelelő geometriai reprezentáció ezért nem pusztán esztétikai kérdés: befolyásolja azt is, hogy mennyire tudjuk felismerni az aszimmetriát, a szélsőséges tartományok szerkezetét vagy a becslési hibákból adódó torzulásokat. A bemutatott spirális és kúpfelületi ábrázolások ezt a vizuális értelmezést kívánják segíteni. A dolgozatban bemutattuk a normális eloszláshoz közeli többdimenziós sűrűségfüggvények modellezésének és vizualizációjának néhány lehetőségét. Röviden ismertettük a Gram-Charlier-közelítés alap gondolatát, amely a ferdeség és a lapultság figyelembevételével lehetővé teszi a normális modelltől való eltérések leírását. A hangsúlyt a képi megjelenítésre tettük. A hagyományos kétváltozós felületi és kontúrbrák mellett spirális és kúpfelületi reprezentációkat alkalmaztunk háromdimenziós sűrűségek szemléltetésére. Az ábrák rámutatnak arra, hogy a geometriai megközelítés hasznos segítséget nyújthat a többváltozós valószínűségi modellek szerkezetének megértésében, valamint a Gram-Charlier-becslések pozitivitási problémáinak feltárásában. A bemutatott módszerek további lehetőséget kínálnak a ritka eseményekhez kapcsolódó kockázati mutatók és sűrűségbecslések vizsgálatához. A bemutatott ábrák nemcsak kutatási célokra használhatók, hanem a többváltozós eloszlások oktatásában is. A vizuális reprezentációk segítségével a szimmetria, aszimmetria, valamint a magasabb rendű momentumok hatása intuitív módon szemléltethető.

MATEMATIKAI HÁTTÉR

A d -dimenziós normális eloszlás sűrűségfüggvénye:

$$\varphi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \mathbf{x} \in \mathbb{R}^d.$$

A d -dimenziós ferde normális eloszlás sűrűségfüggvénye:

$$f_{\mathbf{X}}(\mathbf{x}) = 2\varphi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}^\top(\mathbf{x} - \boldsymbol{\mu})),$$

ahol $\boldsymbol{\alpha}$ az alakparaméter és Φ az 1 dimenziós normális eloszlásfüggvény.

Zsugorítás (shrinkage). Vegyük a Gram-Charlier közelítés legegyszerűbb esetét

$$\tilde{f}(\mathbf{x}) = \left(1 + \frac{1}{3!}\kappa_{Y,3}H_3(y) + \frac{1}{4!}\kappa_{Y,4}H_4(y)\right)\varphi(\mathbf{x}|\mu_X, \sigma_X^2).$$

Az $\tilde{f}(x)$ pozitívitása azon múlik, hogy a φ együtthatója, jelöljük $1 + g(y)$ -al, pozitív-e. A feladat az, hogy találjunk olyan $\beta = [\beta_1, \beta_2] \in [0, 1]^2$ értékeket, amelyekkel teljesül a

$$g_{\beta}(y) = 1 + \beta_1 \frac{1}{3!} \kappa_{Y,3} H_3(y) + \beta_2 \frac{1}{4!} \kappa_{Y,4} H_4(y) \geq 0,$$

egyenlőtlenség. Mindezen közben a β koordinátái legyenek egyhez minél közelebb: $\|\mathbf{1} - \beta\|^2$ legyen minimális. Ezt a módszert nevezzük zsugorításnak (shrinkage). Érdekes megjegyezni, hogy a minimum kereshető ℓ_2 : $\|\mathbf{1} - \beta\|^2$ vagy ℓ_{∞} : $\|\mathbf{1} - \beta\|_{\infty}$ vagy más távolságok minimalizálásával is. Jelen esetben ℓ_2 a zsugorítást a koordinátákon egyenletesen végzi, míg ℓ_{∞} a legrosszabb torzítást egyenlíti. Mivel a mi esetünkben a sűrűségfüggvény alakjának megőrzése fontos, ezért az ℓ_2 -t alkalmazzuk.

IRODALMI HIVATKOZÁSOK

- [1] Azzalini, A. The Skew-Normal and Related Families. Cambridge University Press, 2014.
- [2] Kotz, S. N. Balakrishnan, and N. L. Johnson. Continuous Multivariate Distributions. 2000. 2nd ed. Vol. 1. Wiley.
- [3] Terdik, Gy. Multivariate Statistical Methods: Going Beyond the Linear. Frontiers in Probability and the Statistical Sciences. Springer International Publishing, 2021. <https://doi.org/10.1007/978-3-030-81392-5>.
- [4] Terdik, Gy. and E. Taufer. Beyond Gaussian Multivariate Statistics. Use R. Springer International Publishing,, 2026, to appear.
- [5] Terdik, Gy. and E. Taufer. “MultiStatM: Multivariate Statistical Methods in R.” The R Journal, 2024. 16 (4): 123–40.
- [6] Terdik, Gy. and E. Taufer. MultiStatM: Multivariate Statistical Methods. CRAN, 2025.
- [7] Terdik, Gy. and Taufer, E. Beyond Gaussian Multivariate Statistics. Use R. Springer International Publishing, 2026, to appear.