Statisztikai Szemle

Közzététel: 2023. május 23.

A tanulmány címe:
Szezonális előrejelzési bizonytalanság a villamosenergia-piacon

Szerző:
MÁK FRUZSINA
a Budapesti Corvinus Egyetem Adatelemzés és Informatikai Intézete Statisztika Tanszékének egyetemi adjunktusa
E-mail: fruzsina.mak@uni-corvinus.hu

DOI: https://doi.org/10.20311/stat2023.05.hu0403

Mák Fruzsina

Szezonális előrejelzési bizonytalanság a villamosenergia-piacon

Seasonally varying prediction uncertainty in the electricity market

Mák Fruzsina, a Budapesti Corvinus Egyetem Adatelemzés és Informatikai Intézete Statisztika Tanszékének egyetemi adjunktusa
E-mail: fruzsina.mak@uni-corvinus.hu

A tanulmányban a hazai villamosenergia-rendszer terhelésének előrejelzési bizonytalanságát vizsgálom, Gauss-keverékregresszió *(Gaussian Mixture Regression*) felhasználásával. A rendszerterhelés heteroszkedasztikus viselkedésének leírása a Gauss-keverékmodellre (*Gaussian Mixture Model*) épülően kézenfekvő, azonban az energiaszektort érintő szakirodalomban elsősorban a nemlineáris (és interakciós) kapcsolatok modellezésére történő alkalmazása szerepel. Ez többek között azzal magyarázható, hogy a villamosenergia-fogyasztás bizonytalanságának explicit modellezése iránti igény az elmúlt években jelent meg igazán markánsan. A tanulmányban bemutatom, hogy a Gauss-keverékregresszió mennyire jól megragadja a villamosenergia-rendszer terhelésének előrejelzési bizonytalanságát, ami a várható értékhez hasonlóan szintén erős szezonális viselkedést mutat. A számítások implementálása a nyílt forráskódú Python programnyelvben történt.

Kulcsszavak: Gauss-keverékregresszió, villamosenergia-piac, előrejelzési bizonytalanság

In this paper the forecast uncertainty of the Hungarian electricity system load is investigated using Gaussian Mixture Regression. The description of the heteroskedastic behaviour of electricity load based on the Gaussian Mixture Model is very straightforward, but in the literature its application to the modelling of nonlinear (and interaction) relationships is first and foremost considered. This can be explained, among other things, by the fact that the need for explicit modelling of uncertainty in electricity load has become really pronounced in recent years. It is shown that Gaussian Mixture Regression captures well the prediction uncertainty of electricity load, which also exhibits strong seasonal behaviour similar to the variation of it. The calculations are implemented in the open source Python programming language.

Keywords: Gaussian mixture regression, electricity market, prediction uncertainty

A villamosenergia-piac az egyik legkomplexebb (energia)piac: mivel a keresletnek és a kínálatnak minden időpillanatban meg kell egyeznie, és a villamos energia nagymennyiségű tárolása nehezen megoldható, a potenciálisan felmerülő kockázatok mindkét oldalon könnyen eszkalálódhatnak. A nyersanyag- és kvótaárfüggőség, a megújuló energiák terjedésével a kínálati oldalon is megjelenő időjárásfüggőség, a keresleti oldalon jelentkező (gazdasági vagy szabályozási eredetű) hosszabb vagy rövidebb távú strukturális változások csak néhány példa azon tényezők közül, amelyekkel az energiapiaci szereplők szembesülnek.

A tanulmányban a keresleti oldalt, elsősorban annak az előrejelzési bizonytalanságát vizsgálom. Egyre több az olyan terület, ahol nemcsak a kereslet várható alakulásának, hanem a bizonytalanságának az ismerete is szükséges. Mindennek a módszertani és empirikus szakirodalmi háttere az elmúlt 10-15 évben vált egyre és egyre fontosabbá az energiapiacon, többek között például a verseny növekedése, a megújulók terjedése és egyéb infrastrukturális kihívások, vagy a kereslet oldali szabályozás *(demand side management)* miatt *(Hong–Fan, 2016).*

A tanulmány célja kettős: egyrészt a villamosenergia-rendszer terhelési bizonytalanságának empirikus eredményeit, másrészt a Gauss-keverékregresszió ilyen célú alkalmazása mögötti módszertani hátteret szeretném bemutatni, mivel a magyar nyelvű közgazdasági szakirodalomban kevésbé publikált módszertanról van szó. A tanulmány eredményei tehát nem csak az alkalmazott módszertan szempontjából fontosak: a magyar rendszerterhelési bizonytalanság szezonális mintázatának elemzése és magyarázata ugyanúgy releváns.

A tanulmányban elvégzett számítások a Python *sklearn* és *gmr* csomagjai felhasználásával készültek *(Pedregosa és szerzőtársai, 2011; Fabisch, 2021).*

1. Felhasznált adatok

A tanulmányban vizsgált országos, órás villamosenergiarendszer-terhelési adatok forrása a MAVIR Magyar Villamosenergiaipari Rendszerirányító Zrt.[1] A hőmérsékletadatok forrása az Iowa State University – Iowa Environmental Mesonet[2] ASOS-mérései, Budapest mérési pontra vonatkozóan.

A hőmérsékletadatok alapvetően félórás gyakoriságúak, ettől ritkán néhány perces, még ritkább esetben egy-egy órás eltérések is lehetnek. A hiányzó félórás időpontok (a teljes idősor 0,2%-a) értékeit a legközelebb eső korábbi időpont értékével,[3] az így létrejött – immár ekvidisztáns megfigyeléseket tartalmazó – adatbázisban a „ténylegesen" hiányzó értékeket (0,02%) pedig a félórával korábbi értékkel helyettesítettem.

Természetesen más imputációs technikák is alkalmazhatók, azonban a hiányzó adatok alacsony száma és aránya, illetve véletlenszerűen szórt jellege miatt a végső eredményeket valószínűleg nem befolyásolja a választott megoldás.

2. A villamosenergia-fogyasztás stilizált tényei

Elsősorban pénzügyi piacokon elterjedt és népszerű az ún. stilizált tények *(stylized facts)* megfogalmazása és vizsgálata. Ezek olyan, főként kvalitatív jellegű állítások, amelyeket a pénzügyi idősorok többsége esetén általánosan igaznak, érvényesnek fogadhatunk el. A modellekkel szemben jogos elvárásként fogalmazzuk meg, hogy ezeket a stilizált tényeket minél jobban képesek legyenek leképezni.

Ahogyan a pénzügyi piacokra *(Cont, 2001),* úgy a villamosenergia-árakra *(Marossy, 2010),* a pénzügyi és az energiapiacok kapcsolatára *(Leng és szerzőtársai*, *2014)*, vagy akár a fogyasztási idősorokra vonatkozóan is megfogalmazhatók ún. stilizált tények. Ezek közül – a teljesség igénye nélkül – a legfontosabbak az alábbiak (bővebben *Mák*, *2017*):

– magas időbeli függőség,
– szigorú értelemben vett (erős) stacionaritás hiánya,

[1] https://www.mavir.hu/web/mavir/rendszerterheles

[2] https://mesonet.agron.iastate.edu/request/download.phtml?network=HU__ASOS

[3] Pl. amennyiben 10:00, 10:28 és 11:00 időpontban vannak mért adataink, akkor a hiányzó 10:30-as időpont a 10:28 időpont méréseit örökli. Amennyiben a mérés hiányzik 10:28-kor (ún. *missing*), a 10:30-as értéket is hiányzóként kezelem, amit később imputálok (a 10:00-ás értékkel, amennyiben az nem hiányzó).

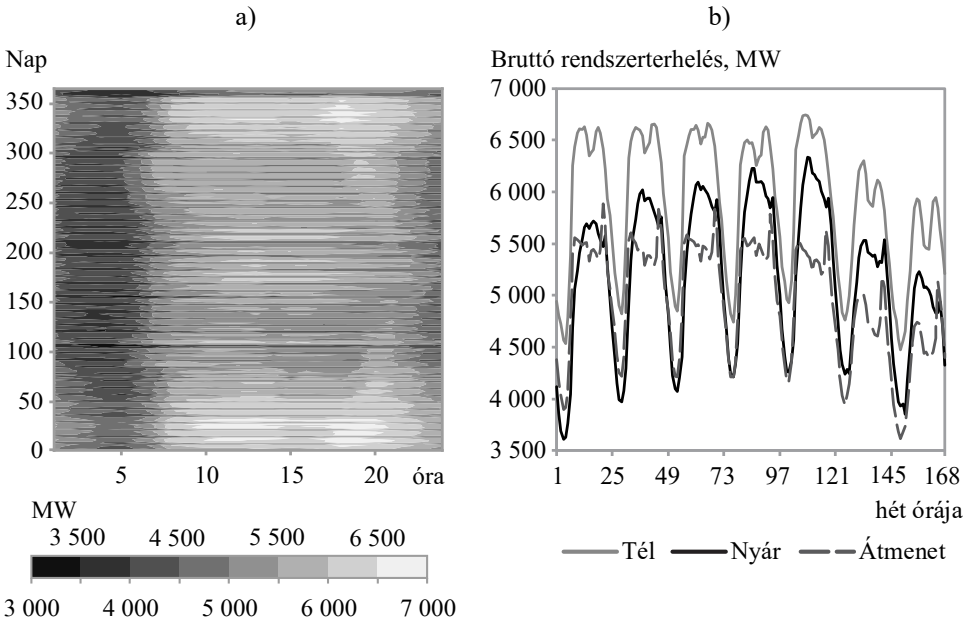
– többszintű (éves, heti, napon belüli) szezonalitás,
– időjárástól való függőség,
– heteroszkedaszticitás, azaz időben változó szóródás.

A tanulmányban elsősorban az utolsóra koncentrálok, de közvetetten természetesen a többi sem hagyható figyelmen kívül. Ehhez kapcsolódóan tekintsük a 2017-es év bruttó rendszerterhelésére készült ún. szintvonalábrát (*contour plot*, 1. (a) ábra), ahol a napon belüli órák (vízszintes tengely) és az év napjai (függőleges tengely) függvényében az órás bruttó rendszerterhelés értékeit ábrázolom az ábra alján szereplő osztályköz-színkód kombinációnak megfelelően. Emellett az 1. (b) ábrán egy-egy téli, átmeneti és nyári hét villamosenergia-terhelési görbéjének alakulása szerepel.[4]

1. ábra

A rendszerterhelés szintvonalábrája (a) és heti idősorai egy téli, egy átmeneti és egy nyári héten (b), 2017

Contour plot of the system load (a), weekly time series of the system load for a winter, a transition and a summer week (b), 2017



Forrás: saját számítás.

[4] A választott hetek az alábbiak: tél: 2017. január 23. – 2017. január 29., átmenet: 2017. március 27. – 2017. április 2., nyár: 2017. július 31. – 2017. augusztus 6.

Elsősorban olyan tendenciákra hívom fel a figyelmet az ábrák kapcsán, amelyek a tanulmány szempontjából fontosak, és kizárólag egy klasszikus idősoros ábrát vizsgálva nagyon könnyen elvesznek.

A szintvonalábráról nagyon sok, a villamosenergia-fogyasztást jellemző tendencia egyértelműen leolvasható:

– Télen az alacsony, nyáron a magas hőmérséklet miatt magasabb a rendszerterhelés értéke. Az ún. fűtési hatás különösen az év első és utolsó két hónapjában érvényesül a teljes nap során. Az ún. hűtési hatás nagyjából a 170. és a 220. nap között látszik, elsősorban a napközbeni órákban. Az 1. (a) ábra halványabb szürke színei jelölik ezeket a napokat, illetve napszakokat.
– A hétvégi napokon jellemzően alacsonyabb a rendszerterhelés, mint egyébként: ezt a hétvégi napoknak megfelelő, a környező napokhoz képest valamivel sötétebb sávok jelölik az 1. (a) ábrán.
– A naplemente miatti viszonylag erős ún. világítási hatást (vagy naplementehatást) az ábrán a 15. és a 20. óra környékén – az óraátállításkor kicsit megtörő – félkörív mutatja.

Az ábra számtalan további elemezni valót rejt magában, aminek a vizsgálatát az olvasóra bízom, azonban kiegészítésként két fontos megállapítást szeretnék kiemelni:

– A rendszerterhelés alakulásában napon belül jellemzően két csúcsidőszak van, ami a szintvonalábráról talán kevésbé tűnik annyira ki: az egyik napközben (elsősorban a hőmérséklet és az emberi aktivitásból fakadó tényezők miatt), a másik a nap végén (a naplementehatás miatt).
– A hőmérséklet nemlineáris hatása napon belül sem feltétlenül azonos, nem feltétlenül lehet „egyszerűen” csak fűtési, illetve hűtési hatásról beszélni. Az éjjeli, kora hajnali órák esetében ugyanis elsősorban a fűtési szezonbéli és az azon kívüli fogyasztásban van markáns eltérés – ez is jelent egyfajta hőmérsékletfüggőséget, azonban a kapcsolat korántsem olyan szoros, mint például a napközbeni csúcsórák esetén. Ugyanígy a napvégi órák esetén is más a hőmérséklethatás a napnyugta időpontjának (és az abból adódó ún. világítási hatásnak) a dominanciája végett.

A felsoroltakból is látható, hogy a villamosenergia-fogyasztás alakulása rendkívül összetett folyamat, nem véletlen, hogy számtalan kutatás foglalkozik vele, rengeteg szempontból.

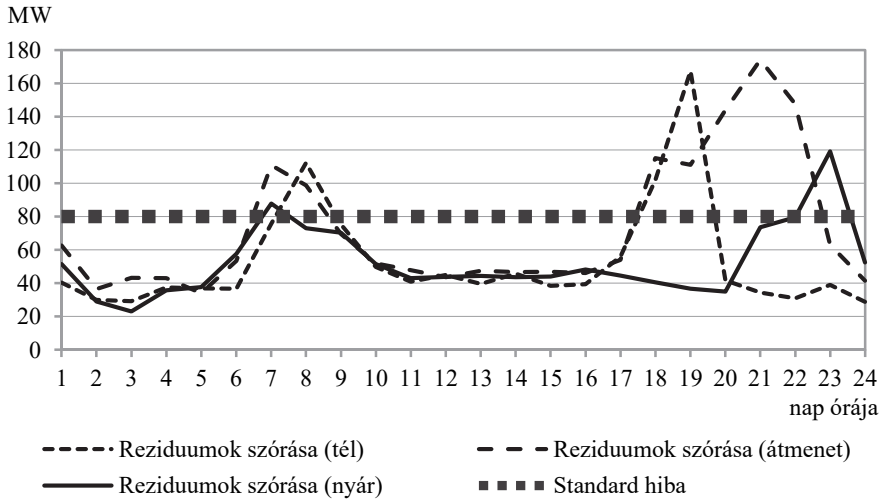
Annak érdekében, hogy az előző ábrákhoz hasonló egyszerűbb, de mégis intuitív eszközzel szolgáljunk a fogyasztás bizonytalanságáról is, vizsgáljuk meg a

2017-es év rendszerterhelési idősorára illesztett lineáris regresszió[5] reziduumainak alakulását! A 2. ábra[6] azt sugallja, hogy nemcsak a villamosenergia-rendszer terhelésének idősorát, hanem annak a bizonytalanságát is szezonális viselkedés jellemzi: jellemzően a reggeli felfutás óráiban, illetve az esti órákban a reziduumok szórása magasabb, mint egyébként. A reziduumok szórása azonban nyilvánvalóan rendkívül zajos, így ez alapján nehézkes a bizonytalanság pontos karakterisztikájáról érdemileg nyilatkozni. Az egyértelműen látszik viszont, hogy a konstans standard hiba és a reziduumok szórása nincsenek összhangban egymással.

2. ábra

A rendszerterhelésre illesztett lineáris regresszió standard hibája és reziduumainak szórása órás bontásban, 2017

Standard error and standard deviation of the residuals of the linear regression fitted to the system load, hourly breakdown, 2017



Forrás: saját számítás.

A fogyasztás bizonytalanságára – amely minden szempontból kapcsolódik az előzőekben felsorolt jellemzőkhöz – vonatkozóan is vannak természetesen tanulmányok. *Lo és Wu (2003)* egy olyan módszert mutat be a bizonytalanság mé-

[5] A lineáris regresszió illesztése a 2017. január 1. és 2017. december 31. közötti időszakra történt. A felhasznált magyarázóváltozók az alábbiak: a hét napjait, a nap óráit jelölő *dummy*változók, az ünnepnapokat és áthelyezett munkanapokat jelölő *dummy*változók, az órás rendszerterhelés 1, 2 és 24 órás késleltetése, valamint a hőmérsékletből számított hűtési, illetve fűtési napfokértékek. A napfokmódszerről röviden a 3. fejezetben lesz még szó, néhány további eredményről pedig a 4. fejezetben.

[6] A szezonális bontás az alábbiak szerint történt: téli hónapok: január, február, december; átmeneti hónapok: március, április, május, illetve szeptember, október, november, nyári hónapok: június, július, augusztus.

résére, amely nem igényel konkrét modellt: az órás változások szórásán alapul. *Subbarao és szerzőtársai (2011)* k-NN (k-legközelebbi szomszéd) módszere a becsült modell reziduumainak egyfajta simításán alapul, ami még szintén egy egyszerűbb megközelítésnek mondható.

Több olyan tanulmány is van, amely kvantilis regresszió használatára épül *(Bracale és szerzőtársai, 2019; Wang és szerzőtársai, 2018),* illetve ezen a területen is elterjedtek a különböző neurális hálózatok *(Taylor–Buizza, 2002; Goodfellow és szerzőtársai, 2014; Wang és szerzőtársai, 2020),* vagy ezek kombinációi *(Zhang és szerzőtársai, 2019).* Szintén elterjedt még a Gauss-folyamatregresszió alkalmazása *(Heo–Zavala, 2012; Manfren és szerzőtársai, 2013). Srivastav és szerzőtársai (2013)* Gauss-keverékregressziót használnak, ugyanúgy, mint *Mák (2017).*

A 3. fejezetben egyszerű példákkal kiegészítve bemutatom a Gauss-keverékmodell és a ráépülő Gauss-keverékregresszió módszertanát. A Mellékletben néhány további részlet is található az elméleti háttérről. A 4. fejezetben a rendszerterhelés 2011–2019 közötti idősorán bemutatom a keverékregresszió eredményeit: a számított (időfüggő) standard hibák (szezonális) alakulását, illetve azt, hogy a módszer a reziduumokkal konzisztens standard hibákat ad. Az 5. fejezetben összefoglalom a főbb konklúzióiókat.

3. Módszertani áttekintés

A keverékmodellek, azon belül is a Gauss-keverékmodell (*Gaussian Mixture Model,* GMM) alkalmazása egyre több helyen van jelen mind a szakirodalomban, mind a gyakorlati alkalmazások területén. Az utóbbi években talán leginkább a jelfelismerés, jelfeldolgozás területén lehet vele találkozni *(Bishop, 2006; Yuksel és szerzőtársai, 2018),* de nagyon sok közgazdasági alkalmazás is van *(Ausín–Galeano, 2007).* A bővebb elméleti háttérről számos helyen talál irodalmat az érdeklődő olvasó *(Dempster és szerzőtársai, 1977; McLachlan–Krishnan, 1977; McLachlan–Peel, 2000).*

A tanulmány lényegi eredményeit tartalmazó 4., empirikus fejezetet megelőlegezendően kisebb empirikus példák már ebben a módszertani fejezetben is szerepelnek: a nemzetközi szakirodalomhoz való kapcsolódásból fakadóan a GMM és az arra épülő Gauss-keverékregresszió (*Gaussian Mixtures Regression*, GMR) módszertanát a nemlineáris kapcsolatok miatti modellezési igény szempontjából kiindulva és ezen végigvezetve érdemes tárgyalni. Bemutatom azt is,

hogy a GMR milyen módon alkalmazható a heteroszkedaszticitás modellezésére, mivel a 4. fejezet fő fókusza az utóbbi.

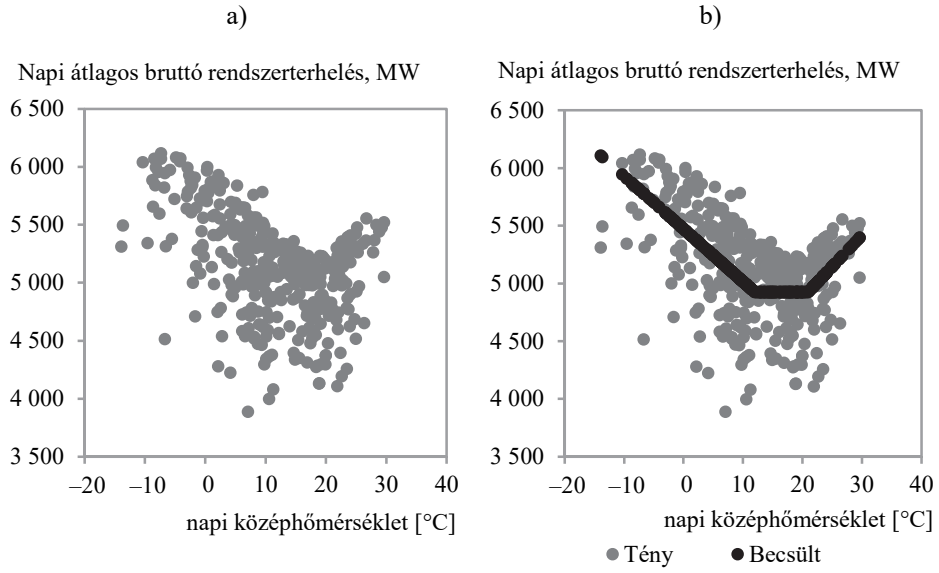
3.1. A linearitás hiányának kezelése a villamosenergiarendszer-terhelés vizsgálatában

Tekintsük bevezetésként – az energiapiaci szakirodalomban már közismertnek számító – 3. ábrát, amely a napi átlaghőmérséklet függvényében ábrázolja a napi átlagos villamosenergiarendszer-terhelést a 2017-es évre vonatkozóan. Tisztán leolvasható a két változó együttes eloszlásának, kapcsolatának az a bonyolultsága, amelyről a 2. fejezetben a szintvonalábra vizsgálata során is említést tettem már: a hőmérsékletfüggés iránya és erőssége sem állandó a teljes mintán.

3. ábra

Rendszerterhelés a hőmérséklet függvényében (a) és napfokmódszerrel becsült rendszerterhelés (b), 2017

System load as a function of temperature (a), predicted system load based on degree-day method (b), 2017



Forrás: saját számítás.

A linearitás hiányának ábrán megjelenő kezelését sok esetben szakaszosan lineáris függvények kezelésével oldjuk meg. Az egyik legalapvetőbb megoldás az ún. napfokmódszer, amely tapasztalati alapon vagy modellszelekció útján hatá-

roz meg küszöbértékeket a szakaszos linearitás biztosításához *(Sugár, 2011; Espinoza és szerzőtársai, 2005)*.[7, 8] Nagyon elterjedt megoldás az ún. MARS *(Multiple Adaptive Regression Splines)* is *(Friedman, 1991),* amely a nemlineáris (és interakciós) hatások hasonló szemléletű, de jóval általánosabb kezelésére is alkalmas (energiapiaci alkalmazásként pl. *Hiruta és szerzőtársai*, *2022*).

A linearitás hiányának ilyen módon történő kezelése természetesen csak egy lehetséges és rendkívül kézenfekvő megoldás a sok közül, a tanulmánynak nem célja ezeknek a részletes ismertetése.

A keverékmodellek alapötlete, hogy több változó bonyolult együttes eloszlását „egyszerűbb" eloszlások kombinációjaként (azaz súlyozásaként, keverékeként) fogalmazza meg – a GMM esetében ez az „egyszerűbb" eloszlás a normális eloszlás.

A GMR-ről elöljáróban annyit érdemes megjegyezni, hogy a linearitás hiányának modellezése itt is lineáris összefüggésekre van visszavezetve, azonban mind a kiindulópont (azaz a GMM), mind a megvalósítás eltérő az előzőekben említett módszerekhez, vagy ahhoz képest, ahogy a linearitás hiányáról klasszikusan gondolkozunk.

A formálisabb módszertani összefoglaló előtt néhány, a tanulmány témájához legközelebb álló energiapiaci témájú munkát még megemlítek. *Srivastav és szerzőtársai (2013)* bemutatják, hogy az épületi hűtési energiafelhasználás hogyan modellezhető *GMR*-rel a hőmérséklet, páratartalom és napsugárzás változók felhasználásával. *Hossain (2017)* a szélsebesség rövid távú előrejelzésére használ *GMR*-t. *Wang és szerzőtársai (2017)* szintén épületek energiafelhasználásának modellezésére használják a módszertant. Ezek a példák mind alapvetően regressziós célú alkalmazások, és *(Srivastav és szerzőtársai, 2013* munkáját leszámítva*)* közös bennük, hogy elsősorban a nemlineáris hatásokra és kevésbé a heteroszkedaszticitás modellezésére fókuszálnak. További GMR/GMM-alkalmazásokat, illetve a módszertan egyedi villamosenergia-fogyasztói görbék profilozására történő alkalmazását lásd *Mák, 2017.*

[7] *Sugár (2011)* alapján definiáljuk az ún. fűtési napfokot max(0,12°*C*–*T*) módon, a hűtési napfokot pedig max(0, *T*–21°*C*) módon (T a hőmérséklet °C-ban). Az ezen napfokváltozók felhasználásával készült regressziós becslés eredményei szerepelnek a 3. (b) ábrán. A küszöbértékek természetesen az idősor aggregáltsági szintjétől, ágazattól (pl. földgáz vagy villamos energia) is függenek (földgázfogyasztás esetén lásd pl. *Mák*, 2015).

[8] Érdemes megjegyezni, hogy az itt és a 3. fejezet további részeiben bemutatott példák csak a hőmérséklet-rendszerterhelés változópárra vonatkoznak, ezért bizonyos idősorelemzési szempontokat nem vesznek figyelembe (pl. hétköznapok és hétvégék megkülönböztetése, időbeli függőség kezelése stb.). A késleltetési struktúra figyelmen kívül hagyása miatt pl. az reziduumok jellemzően autokorreláltak lesznek, de a becsült paraméterek torzítatlansága ebben az esetben is megmarad, így a módszertan illusztrálására ez az egyszerű példa is megfelelő.

3.2. A Gauss-keverékmodell

A GMM megközelítése a paraméterek meghatározására a *maximum likelihood* elméletből ismerősen hangzik: az a célunk, hogy olyan paramétereket becsüljünk, amelyek mellett maximális a *likelihood*ja annak, hogy az éppen realizálódott minta keletkezik. A GMM alapötletének, az „egyszerűbb" eloszlások bevezetésének azonban természetesen ára van: az ún. keveréksúlyok nem megfigyelhetők. A becslés során a nehézséget ezeknek a keveréksúlyoknak a meghatározásához bevezetett ún. látens, rejtett változók kezelése jelenti.

A problémát formálisabban felírva jelölje x a megfigyelhető változó(ka)t, z pedig az előbb említett látens változót. Legyen az együttes eloszlás sűrűségfüggvénye az alábbi:

$$p(x,z) = p(z) \cdot p(x|z) \ . \qquad /1/$$

A z látens tulajdonságából adódóan azonban *likelihood*-függvényt ehhez kapcsolódóan nem tudunk felírni. A z szükségszerű megtartásával a sűrűségfüggvény felírható csak a megfigyelhető változókra:

$$p(x) = \sum_z p(x,z) = \sum_z p(z) \cdot p(x|z). \qquad /2/$$

A z látens mivolta azonban ezzel természetesen nem szűnik meg: a kezelésére az *expectation-maximization* (EM) módszer jelent majd megoldást.

3.2.1. A Gauss-keverékmodell felírása

A GMM kiindulópontja tehát az, hogy a megfigyelt változók együttes eloszlása egy K darab komponenst tartalmazó Gauss-keverékeloszlással írható le, amelynek sűrűségfüggvénye:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot p(x; \mu_k, \Sigma_k) \,, \qquad /3/$$

ahol:

– π_k a k-adik komponens (ún. keverék)súlya (k=1,2…K),
– p a normális eloszlás sűrűségfüggvénye μ_k átlagvektorral és Σ_k kovarianciamátrixszal (k=1,2…K),
– K a komponensek száma.

Az előbbiekben említett z látens kategóriaváltozó azt reprezentálja, hogy a különböző megfigyelések melyik normális eloszlású komponensből állnak elő. A kategoriális eloszlást követő z valószínűségi változó súlyfüggvénye legyen $p(z)$, ahol $\sum_{k=1}^{K} \pi_k = 1$ és $\pi_k \geq 0$ minden k-ra.

A fentiek mellett a minta *loglikelihood*-függvénye az alábbi:

$$l(\pi,\mu,\Sigma) = \ln p(X;\pi,\mu,\Sigma) = \sum_{i=1}^{n} \ln\big(p(x_i;\pi,\mu,\Sigma)\big) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i, z_i = k;\mu,\Sigma,\pi) = \\ = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i|z_i = k;\mu,\Sigma) \cdot \mathrm{p}(z_i = \mathrm{k};\pi) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i;\mu_k,\Sigma_k) \cdot \pi_k, \qquad /4/$$

ahol π, μ és Σ rendre a π_k, μ_k és Σ_k komponensenkénti paraméterek összevont, egyszerűsített jelölésére szolgálnak (azaz $\pi = (\pi_1, \pi_2 \dots \pi_K), \mu = (\mu_1, \mu_2 \dots \mu_K)$ és $\Sigma = (\Sigma_1, \Sigma_2 \dots \Sigma_K)$), x_i és z_i pedig az i-edik megfigyelés megfigyelhető és látens változóértékeit jelölik.

3.2.2. Az *expectation–maximization* (EM) becslési eljárás

A paraméterbecslés, azaz a π, μ és Σ értékek meghatározásának nehézségét alapvetően a *loglikelihood*-függvény logaritmuson belüli kifejezése adja, vagyis az, hogy a komponenstagságot jelölő *z* változó nem megfigyelhető és értéke függ az átlag- és kovariancia paraméterek értékétől.

A megoldást jelentő *expectation–maximization* eljárás alapvetően két fő lépés iteratív elvégzéséből áll. Az iteráció leállási feltétele az, hogy a paraméterek (vagy az ún. Q-célfüggvény) értékeiben nincs szignifikáns változás.

E-lépés (expectation step)

Az adott lépésben (vagy az inicializálás[9] eredményeként) rendelkezésre álló π, μ és Σ paraméterek mellett számítsuk ki minden megfigyelésre a *k*-adik komponenshez tartozás *posterior*-valószínűségeit (a *z posterior* eloszlását) a *Bayes*-formula alapján: vagyis azt, hogy az egyes mintaelemeket milyen eséllyel realizálódhattak a π, μ és Σ paraméterekkel leírható keveréksűrűség-függvény komponenseiből!

Az *i*-edik megfigyelés *k*-adik komponenshez való tartozásának *posterior*-valószínűsége:

$$p_{ik} = p(z_i = k|x_i) = \frac{\pi_k . p(x_i|z_i = k;\mu,\Sigma)}{\sum_{j=1}^{K} \pi_j \cdot p(x_i|z_i = j;\mu,\Sigma)} = \frac{\pi_k \cdot p(x_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot p(x_i;\mu_j,\Sigma_j)}. \qquad /5/$$

[9] A kezdő (ún. inicializált) paraméterek meghatározása történhet bármely klaszterezési eljárással, de a felhasználó által meghatározott (ún. szakértői) kezdő paraméterek is megadhatók. A tanulmányban a kezdő π_k, μ_k és Σ_k értékek meghatározása a *KMeans++* eljárással történt. A fejezet későbbi részében kitérek arra, hogy a GMM-et miért szokás (modellalapú) klaszterezési módszertanként emlegetni, annak ellenére, hogy a módszertan alapvetően változók együttes eloszlásának a becslésén alapul.

A *posterior*-valószínűségek felhasználásával felírható az ún. Q-függvény:

$$Q(\pi,\mu,\Sigma) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \cdot ln(p(x_i, z_i = k; \pi, \mu, \Sigma)) = \\ = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \cdot \ln(p(x_i; \mu_k, \Sigma_k)) + \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \cdot \ln(\pi_k) \quad /6/$$

M-lépés (maximization step)

Az E-lépésben felírt Q-függvény maximalizálása ekvivalens azzal, hogy a π, μ és Σ paraméterek értékét úgy változtatjuk meg, hogy az új (optimális) paraméterek mellett a megfigyelt minta (azaz a megfigyelések együttese) a legnagyobb esélylyel realizálódjon. Technikailag a maximalizálási feladatot az alábbi módon szokás felírni:

$$\hat{\pi}, \hat{\mu}, \hat{\Sigma} = argmin(Q(\pi, \mu, \Sigma)), \quad /7/$$

ahol $\hat{\pi}, \hat{\mu}$ és $\hat{\Sigma}$ az új paramétereket jelöli, amelyekkel a következő E-lépés *posterior*-valószínűségeit számoljuk (amennyiben a leállási feltétel nem teljesül). Viszonylag könnyen levezethető, hogy az M-lépés eredménye a becsülni kívánt paraméterekre zárt formulákat ad (bővebben ld. a Mellékletet).

Fontos megemlíteni, hogy a Q-függvényben – a p_{ik} *posterior*-valószínűségek bevezetése mellett – a logaritmus és a $\sum_{k=1}^{K} \ldots$ kifejezések lényegében „helyet cseréltek" a *loglikelihood*-függvény formulájához képest, és ez tette kezelhetővé az optimalizálási feladat hatékony megoldását – a Mellékletben bemutatom röviden ennek a módszertani hátterét.

3.2.3. A Gauss-keverékmodell alkalmazása a napi hőmérséklet-rendszerterhelés idősorra

Tekintsük a napi középhőmérséklet és villamosenergiarendszer-terhelés kapcsolatát a Gauss-keverékmodell szempontjából!

A 4. (b) ábrán az illesztés eredményeként előálló BIC-kritériumok értéke látható, 2 és 10 közötti komponensszám mellett. Mivel a BIC-kritérium minimuma jelzi a választandó modellt, 4 komponens illesztése mellett döntöttem. Az ezek felhasználásával előállítható Gauss-keveréksűrűség-függvény szintvonalábrája szerepel a 4. (c) ábrán, ahol a sötétedő színskála magasabb sűrűségfüggvény(*likelihood*)-értékeket jelöl. A becslést standardizált változók terében végez-

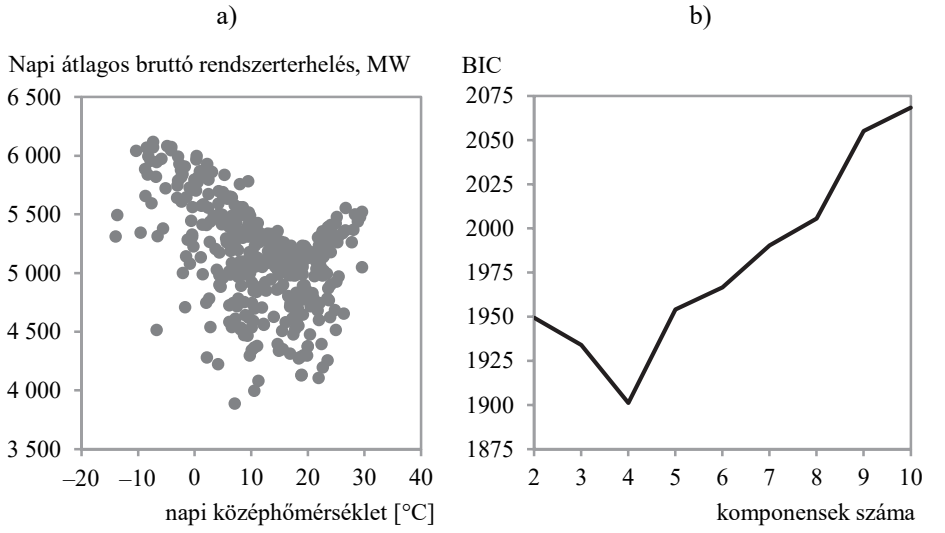
tem el, annak érdekében, hogy a változók nagyságrendje ne befolyásolja a klaszterezés eredményét.[10]

Ábrázoljuk külön-külön csak azokat a komponensenket, amelyek alapján a változók együttes eloszlásáról alkotott előzetes elképzeléseink és az eredmények könnyebben összehasonlíthatók! Az 5. (a) ábrán a GMM komponenseknek megfelelő kétdimenziós normális eloszlások 95%-os konfidenciaellipszisei szerepelnek, vagyis – amennyiben az eloszlásbeli feltételezés helyes – az ábrán szereplő ellipszisek területére komponensenként a megfigyelések 95%-a esik.[11] Az illesztett 4 komponens nagyjából megfelel az előzetes elképzeléseknek, hiszen van egy erős pozitív és két negatív irányú kapcsolatot reprezentáló komponens a magasabb, valamint egy nagyon enyhén negatív irányú kapcsolatot reprezentáló az alacsonyabb terhelési szinteken. Utóbbi jelenség azzal magyarázható, hogy az alacsonyabb fogyasztási szintek hőmérsékletfüggése jóval enyhébb, és elsősorban a téli vagy a hidegebb átmeneti hónapokban van jelen, ahogy az az 1. fejezetben szereplő szintvonalábra tárgyalása során is említettem.

4. ábra

Rendszerterhelés a hőmérséklet függvényében (a), az illesztett GMM-modellek BIC-értékei (b), GMM-keveréksűrűség-függvény 4 komponens esetében (c), 2017

System load as a function of temperature (a), BIC values of the GMM models (b), mixture density function for GMM model with 4 components (c), 2017

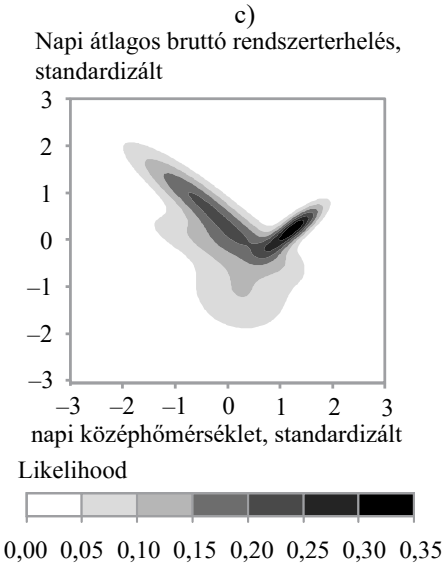


(Az ábra folytatása a következő oldalon.)

[10] A változók standardizálása hasonló megfontolásból a 4. fejezetben is megtörtént. A 4. fejezetben az ellentétes irányú transzformációt is elvégeztem a regressziós eredményeken, annak érdekében, hogy azok az eredeti mértékegységben értékelhetőek és a tény rendszerterhelés értékekkel összehasonlíthatóak legyenek.

[11] Az ellipszisek paraméterei a kovarianciamátrix sajátérték-sajátvektor felbontásából származtathatók.

(folytatás)



A GMM komponensei egyébként klaszterként is interpretálhatók: a módszert a szakirodalom sokszor Gauss-keverékklaszterezés néven *(Gaussian Mixture Clustering)* emlegeti. Ebben az értelemben a komponensekre gondolhatunk úgy, mint az x_i változók dimenziójában kialakított homogén, valamilyen szempontból (ebben az esetben a BIC-kritérium) optimális csoportokra. Az x_i változók esetünkben a hőmérséklet és a rendszerterhelés napi megfigyeléseit tartalmazzák, a z_i változók pedig a látens klasztertagságok.

A megfigyelések klaszterekhez történő hozzárendelése azonban nem bináris, hanem a p_{ik} *posterior*-valószínűségek alapján valószínűségi jelleggel történik: egy megfigyelés olyan mértékben járul hozzá egy klaszterhez, amilyen mértéket a becsült π_k, μ_k és Σ_k paraméterek és a saját x_i ismérvértékek alapján számolt p_{ik} *posterior*-valószínűség diktál. Másképpen fogalmazva: ahhoz a klaszterhez történő hozzájárulás lesz a legnagyobb, ahova a saját x_i ismérvértékek alapján a legnagyobb eséllyel tartozik.

A komponensek elhelyezkedéséből adódóan természetesen az egyes megfigyelések besorolási bizonytalansága nagyon eltérő lehet. Ezt hivatott szemléltetni az 5. (b) ábra, amely megőrzi a mellette szereplő ábra színeit, viszont az ábrán

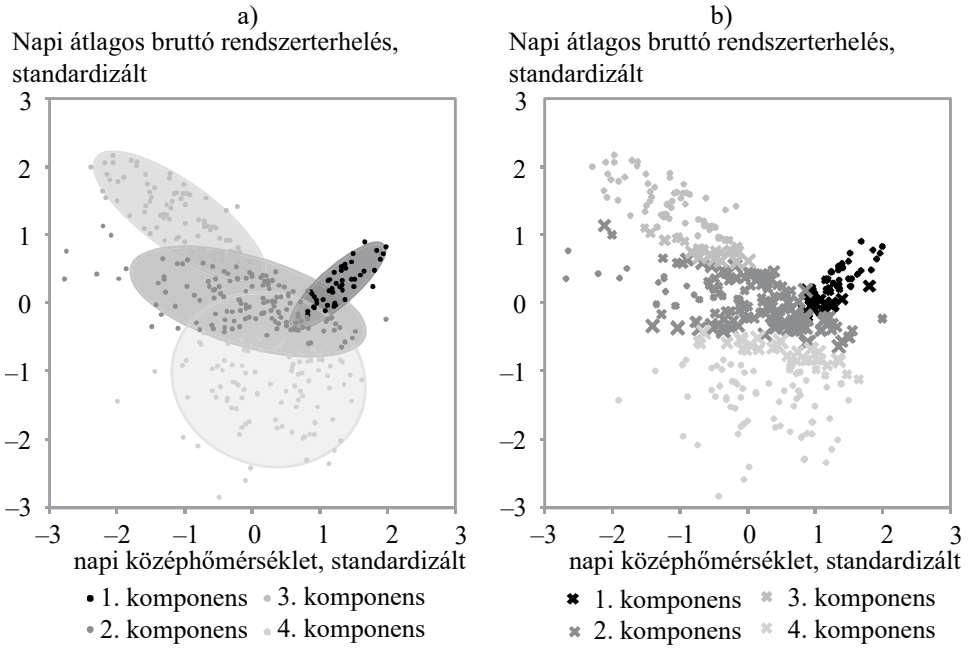
szereplő jelölő X-ek méretének konkrét tartalma is van: minél nagyobb egy besorolás bizonytalansága, annál nagyobb az ábrán szereplő jelölő X mérete.[12]

Elsőre némileg zavaró lehet az 5. ábrán, hogy az alacsonyabb fogyasztási szinteket túlnyomóan egy komponens képviseli. Valójában ez a képviselet a p_{ik} *posterior*-valószínűségek által vezérelve valószínűségi jellegű, ami azt jelenti, hogy az alacsonyabb fogyasztási szintek eloszlásának leírásához a többi komponens is hozzájárul, csak kisebb súllyal. Ez a szemlélet a GMM regressziós alkalmazásánál is fontos lesz.

5. ábra

Hőmérséklet-rendszerterhelés változópárra illesztett GMM komponsek (a) és besorolási bizonytalanságok (b), 2017

GMM components fitted to temperature and system load variables (a), component uncertainties (b), 2017



Forrás: saját számítás.

A komponensenként számolt kovarianciamátrix praktikus tartalma az, hogy a változók kovarianciastruktúrája a mintában nem állandó, hanem a magyarázóváltozók értékétől függ. Tekintve a klaszterezés tanító nélküli (ún. *unsupervised)*

[12] Feltételezve, hogy az *i*-edik megfigyelés oda lett besorolva, ahol a p_{ik} maximális (k=1,2,3,4), a bizonytalanság mértékének egy lehetséges számítási módja az alábbi kifejezés: $1 - \max_k p_{ik}$. Ahogy az várható, a klaszterhatárokon a bizonytalanság magasabb, hiszen az ott szereplő megfigyelések nagyobb eséllyel tartozhatnak két (vagy annál több) klaszterhez.

tulajdonságát, ez egyben azt is jelenti, hogy a változók közötti nemlineáris és interakciós kapcsolatokat úgy tudjuk figyelembe venni, hogy explicit módon nem kell őket definiálni, hiszen az együttes sűrűségfüggvény illesztése eredményeként megkeletkeznek. A regresszió módszertani hátteréről lásd bővebben a 3.3. fejezetet, illetve a Mellékletet.

3.3. A Gauss-keverékregresszió

A GMR mögötti elv a következő: mivel a GMM-ből adódóan a változók együttes eloszlása rendelkezésre áll, a feladat az, hogy a változók közül egy (a majdani eredményváltozó) feltételes eloszlását meghatározzuk a többi változó (mint magyarázóváltozók) feltétele mellett.

A Mellékletben bemutatjuk, hogy a GMR komponensenként illesztett regressziók keverékeként definiálható *(Sung, 2004).* Mivel a GMM kovarianciamátrixai az egyedi megfigyelések p_{ik} *posterior*-valószínűségekkel történő súlyozással keletkeztek meg, ennélfogva technikailag a komponensenként illesztett regressziók a megfelelő p_{ik} *posterior*-valószínűségekkel mint súlyokkal számolt súlyozott regressziónak tekinthetők. Mindebből fakadóan az eredményváltozó feltételes eloszlása is normális eloszlások keverékeként áll elő.

A legfontosabb konklúzió módszertani szempontból az, hogy mivel az eredményváltozó feltételes eloszlása az x_i magyarázóváltozók függvénye, a GMR jó alternatíva a heteroszkedaszticitás kezelésére (a lineáris és interakciós hatások kezelése mellett). Ebben a fejezetben az eredményváltozó feltételes eloszlásával kapcsolatos összefüggéseket közlöm, mivel a 4. fejezet eredményeihez, illetve azok megértéséhez ez kapcsolódik szorosabban. További részletek a Mellékletben szerepelnek.

A könnyebb követhetőség kedvéért a következő fejezetben használt jelölésrendszert a regresszióban szokásos jelölésrendszerhez igazítva módosítom. A továbbiakban y_i a választott eredményváltozót jelöli, x_i pedig y_i kivételével az összes többi változót, amelyek a GMM-ben megfigyelhető (nem látens) változóként voltak jelen. (Lásd még a Melléklet M2. táblázatát a fontosabb jelölésekről.)

3.3.1. Az eredményváltozó feltételes eloszlása és fontosabb momentumai

Az eredményváltozónak (y_i-nek az adott x_i magyarázóváltozók feltétele melletti) feltételes várható értékének és feltételes varianciájának (standard hiba négyzetének) a számítása a komponensenkénti feltételes várható értékek és varianciák alapján a p_{ik} *posterior*-valószínűségek, mint súlyok felhasználásával történik.

Jelölje m_{ik} az i-edik megfigyelés k-adik komponensbéli feltételes várható értékét, amelynek számítása az alábbi képlet szerint történik:

$$m_{ik} = \mu_k^y + (x_i - \mu_k^x)^T \cdot \widehat{\beta_k} \ . \qquad /8/$$

Az összefüggés kihasználja, hogy a $\widehat{\beta_k}$ paramétervektor nem tartalmaz konstans tagot, így az $(m_{ik} - \mu_k^y)$ és $(x_i - \mu_k^x)$ átlagtól vett eltérésekre felírt összefüggés értelemszerű átalakításaként adódik (lásd a Mellékletet).

Az s_{ik}^2, azaz az *i*-edik megfigyelés *k*-adik komponensbéli feltételes varianciája természetesen azonos a komponensenkénti reziduális varianciával (azaz $\widehat{\sigma_k^2}$-nel, lásd a Mellékletet).

Az eredményváltozó feltételes várható értéke és varianciája az alábbiak szerint írható fel:

$$\widehat{y_i} = \sum_{k=1}^{K} p_{ik} \cdot m_{ik}, \text{ illetve } var(\widehat{y_i}) = \sum_{k=1}^{K} p_{ik} \cdot (s_{ik}^2 + m_{ik}^2) - (\sum_{k=1}^{K} p_{ik} \cdot m_{ik})^2. \qquad /9/$$

Utóbbi összefüggésben (kétszer is) kihasználtam, hogy a variancia a négyzetes átlag négyzetének és a számtani átlag négyzetének a különbsége.

Mivel a komponensenkénti eloszlás normális, az eredményváltozó feltételes sűrűségfüggvénye is normális eloszlású sűrűségfüggvények keverékeként írható fel, az alábbi módon:

$$\Phi(y_i; p_{ik}, m_{ik}, s_{ik}) = \sum_{k=1}^{K} p_{ik} \cdot \frac{1}{\sqrt{2\pi} s_{ik}} \cdot exp\left(-\frac{1}{2}\left(\frac{y_i - m_{ik}}{s_{ik}}\right)^2\right), \qquad /10/$$

ahol az eredményváltozó feltételes eloszlását leíró p_{ik}, m_{ik} és s_{ik} paraméterhármas értéke függ az x_i magyarázóváltozók konkrét értékétől.

Az eredményváltozó feltételes eloszlása Gauss-keverékeloszlás, ezért a konfidenciaintervallum számítása is bonyolultabb: az α megbízhatósági szintű konfidenciaintervallum alsó, illetve felső határa az $\alpha/2$, illetve az $(1-\alpha/2)$ percentilisek értékei az előbb felírt keverékeloszlásból. Utóbbi a publikusan elérhető nyílt forráskódú programcsomagoknak nem része, így ennek implementálása is a tanulmány eredményének tekinthető.

3.3.2. A Gauss-keverékregresszió alkalmazása a napi hőmérséklet-rendszerterhelés idősorra

Visszatérve 2017-es év napi középhőmérséklet és a villamosenergiarendszer-terhelés kapcsolatára, a 6. ábra a GMR becslési eredményeit mutatja.

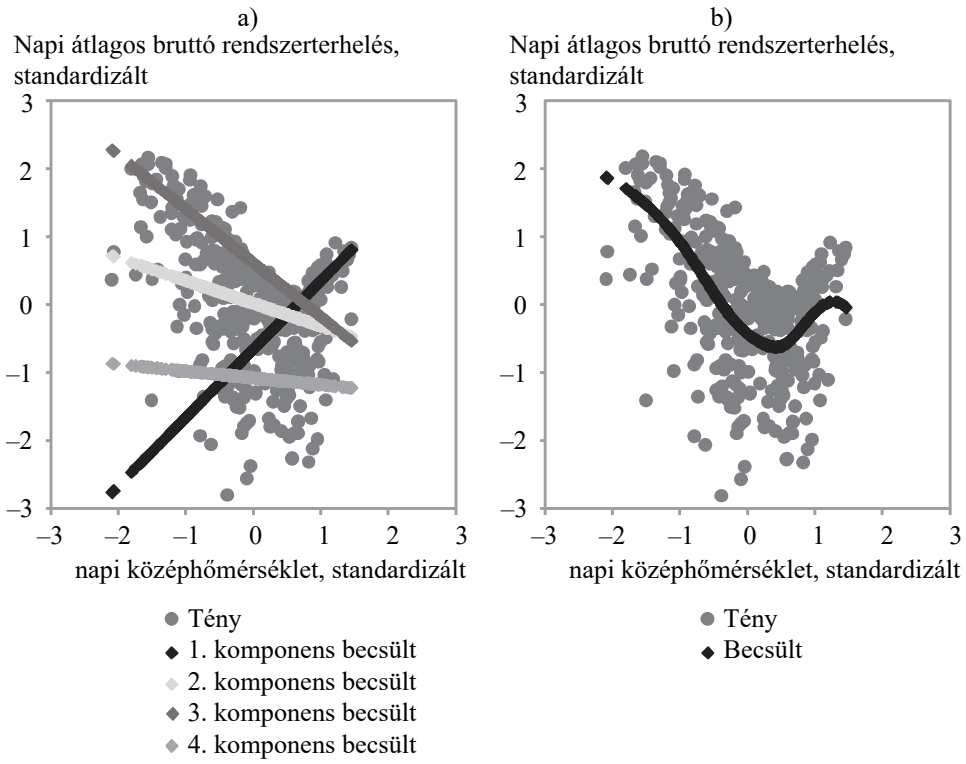
Mivel a változók közötti nemlineáris kapcsolatok leképezése a GMR esetében is lineáris összefüggésekre van visszavezetve, minden komponenshez tartozóan súlyozott (lineáris) regressziót becsülünk (6. (a) ábra).

Természetesen nem mindegyik regressziós egyenes érvényes a magyarázóváltozók teljes tartományán: a p_{ik} *posterior*-valószínűségek determinálják azt, hogy melyik komponenst milyen x_i értékek mellett kell nagyobb súllyal figyelembe venni. Így egy-egy megfigyelés adott k komponens melletti várható értékét (m_{ik}) a p_{ik} *posterior*-valószínűségekkel súlyozva kapjuk a GMR-becslést ($\hat{y}_i$, lásd 6. (b) ábra).

6. ábra

Komponensenkénti regressziós becslések (a) és GMR-becslés (b) a rendszerterhelés várható értékére, 2017

Component-wise regression predictions (a) and GMR prediction (b) of system load, 2017



Forrás: saját számítás.

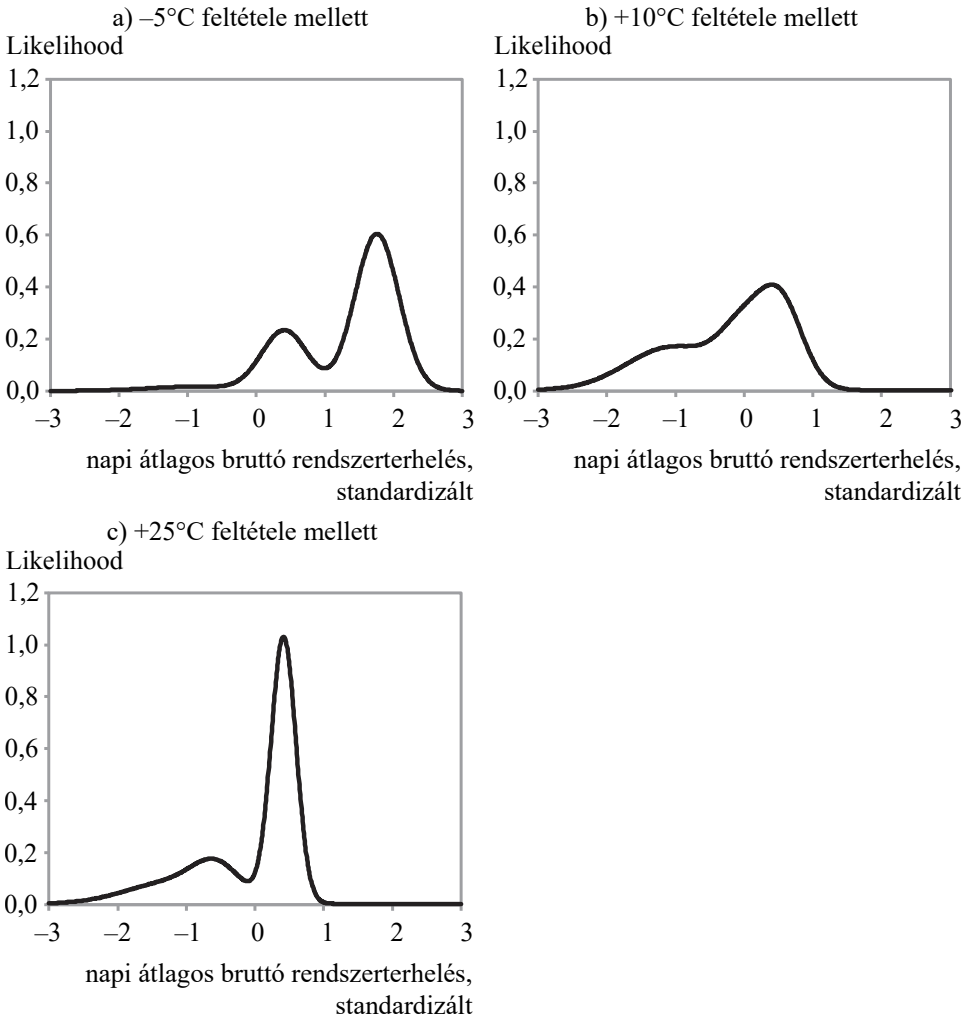
A GMR konstrukciójából adódóan nemcsak az eredményváltozó feltételes várható értéke, hanem a teljes feltételes eloszlása is származtatható. Egy-egy ilyen feltételes eloszlást mutat az 7. ábra (a), (b) és (c) része, választott –5°C, +10°C és +25°C, azaz egy-egy téli, átmeneti és nyári időszakban jellemző napi középhőmérséklet mellett (az ábrákon az eredmények a standardizált rendszerterhelésre vonatkozóan szerepelnek).

Az (a) és (c) (téli és nyári) ábrán a két móduszúság oka az, hogy mindössze két változóval dolgoztam, így adott hőmérséklet mellett a feltételes eloszlás alakulásában két komponens súlya lesz magas: az egyik az alacsonyabb (hétvégi) fogyasztási szintekhez kapcsolódik, a másik a magasabb (hétköznapi) fogyasztási szintekhez. A (b) ábrán ez jóval enyhébben jelenik meg, hiszen az átmeneti időszakban a hétvégék és a hétköznapok átlagos szintjei között nincs olyan jelentős különbség, mint télen vagy nyáron.

7. ábra

A rendszerterhelés feltételes eloszlásának sűrűségfüggvénye, 2017

Conditional density function of system load, 2017



Forrás: saját számítás.

Az említett jelenség lényegében arra utal, hogy fontos, releváns változó hiányzik a modellből. A tanulmány 4. fejezetében természetesen kezelem ezt a problémát. Itt a cél elsősorban az illusztráció volt és ebben az értelemben egy „hiányos" modell a módszertan lényegét még jobban tudta prezentálni.

4. Empirikus eredmények

A tanulmány fő empirikus eredményeit két részre bontva mutatom be: először a paraméterek keresztvalidációval történő becslését ismertetem, majd értékelem a modell teljesítményét a teljes előrejelzési intervallumon, kiemelten a hibák szóródására, az előrejelzési bizonytalanságra fókuszálva.

4.1. Paraméterbecslés keresztvalidációval

Mivel kellően hosszú idősor áll rendelkezésre, a komponensek optimális számának megválasztásához és a paraméterek becsléséhez ún. keresztvalidációs *(cross-validation)* technikát használtam. A becslési, validációs és előrejelzési *(train*, *validation* és *prediction)* intervallumokat az alábbi módon definiáltam:

- egy évet becslési és az ezt követő három hónapot validációs időszakként használva határoztam meg a komponensek optimális számát;
- az (egy és egynegyed évet követő) egy hónapot mint előrejelzési intervallumot használtam a modell teljesítményének kiértékelésére;
- a fenti két lépést egy hónapos eltolásokkal újra és újra elvégeztem a 2011. január 10. és 2019. december 29. közötti időszakra vonatkozóan.[13]

Az időintervallum-hosszok megválasztása alapvetően modellezői döntés, amelynek során figyelembe vettem, hogy sem a túl rövid, sem túl a hosszú intervallumok nem előnyösek. A villamosenergia-fogyasztás éves szezonalitásából adódóan legalább egyéves becslési időszakot mindenképpen érdemes választani. A becslési időszakhoz képest arányaiban túl rövid validációs időszak sem szerencsés, annak érdekében, hogy egy viszonylag stabil performanciát értékelni tudjunk. A GMM-módszertan egyik előnye a konstrukciójából adódóan éppen az, hogy akár alacsony mintaelemszám mellett is képes az új struktúrák felisme-

[13] Az időintervallumok tényleges hossza 52, 13 és 4 hét (azaz rendre 364, 91 és 28 nap) volt, annak érdekében, hogy minden intervallum teljes heteket fedjen le (2011. január 10. hétfőre esett, 2019. december 29. pedig vasárnapra). Így a csúszóablakos becslést könnyebben lehet implementálni, és az előrejelzési intervallumok eredményei is praktikusan összefűzhetők.

résére (azaz új komponens beazonosítására), ami igényt támaszt arra, hogy az eljárás teljesítményét havonta eltolt időperiódusokkal vizsgáljuk meg.[14]

A komponensek optimális számának meghatározása jórészt manuálisan történt:

– minden becslési/validációs intervallumpár esetén meghatároztam a GMM paramétereit 2 és 25 komponensszám között;
– azt a komponensszámot (és paraméterszettet) tekintettem optimálisnak, ahol a validációs mintán számított *score*[15]-érték már csak elhanyagolható mértékben növekedett, vagy a komponensek számának további emelésével ellaposodott, esetleg zajossá vált;
– a fenti eljárás eredményeként átlagosan 9-10 komponenst határoztam meg.

A felhasznált változók az alábbiak: bruttó rendszerterhelés, annak 1., 2., 8., 16., 24., 48. és 168. késleltetettjei, valamint órás hőmérséklet adatok, azaz összesen 9 változó dimenziójában dolgoztam. Az eredmények alapján ez a késleltetési struktúra a (napon belüli és heti) szezonális jellemzőket jól rögzítette, az optimális késleltetési struktúra modellszelekció útján történő meghatározásától eltekintettem.

4.2. Szezonális előrejelzési bizonytalanság

A fejezetben csak az előrejelzési intervallumok eredményeit mutatom be, hiszen a modell teljesítményéről leginkább ez nyilatkozik jól. Mivel a mozgóablakos becslésből adódóan az előrejelzési időszakok egy átfedésmentes időintervallumot alkotnak, a teljes, 2012 és 2019 közötti mintán kívüli teljesítményt értékelni tudom.

A GMR-eredmények mellett ebben a fejezetben a legkisebb négyzetek módszerével becsült lineáris regresszió (OLS) eredményeit is közlöm – elsősorban nagyságrendi összehasonlítás miatt, hiszen utóbbi sem a linearitás hiányát, sem a heteroszkedaszticitást nem kezeli. A lineáris regresszió illesztése a GMR előrejelzési időszakával azonos időszakra, a 2012 és 2019 közötti évekre történt.[16] Mivel a standard hiba konstans (egy csúszóablakos megoldással ezt a tulajdonságot például egyből el is veszíteném), valamint a lineáris jellegből fakadóan nem merül fel a túlilleszkedés problémája, nagyságrendi összehasonlításhoz ez a

[14] Természetesen más becslési, validációs és előrejelzési intervallumok is választhatók, de ezek észszerű megválasztása a tanulmány lényegi következtetéseire valószínűleg nincs érdemi hatással. A leginkább megfelelő kombináció ellenőrizhető egyébként akár keresztvalidációval is, de az jelentősen megnövelné a számítások komplexitását és időigényét. Ugyan a számítási kapacitásigény egy ilyen méretű problémánál ma már nem feltétlenül megoldhatatlan, az optimális(nak mondható) komponensszámoknak a keresztvalidáció jellegéből fakadó, jórészt manuális meghatározása mindenképpen kényelmetlen.

[15] Az ún. *score* az egy megfigyelésre jutó átlagos *loglikelihood*.

[16] A felhasznált magyarázóváltozók ugyanazok, mint a 2. fejezetben: a hét napjait, a nap óráit jelölő *dummy*változók, az ünnepnapokat és áthelyezett munkanapokat jelölő *dummy*változók, az órás rendszerterhelés 1, 2 és 24 órás késleltetése, valamint a hőmérsékletből számított hűtési, illetve fűtési napfokértékek.

megoldás megfelelő: elmondható, hogy a GMR mind a reziduumok, mind az abszolút reziduumok tekintetében valamivel jobb eredményeket produkál.

1. táblázat

A lineáris regresszió (OLS) és a GMR reziduumainak leíró statisztikái

Descriptive statistics of linear regression (OLS) and GMR residuals

(MW)

Jellemző	Reziduumok		Abszolút reziduumok	
	OLS	GMR	OLS	GMR
Átlag	0,00	–0,37	67,02	65,57
Szórás	92,20	92,06	63,31	64,62
25. percentilis	–51,37	–47,67	22,55	21,43
50. percentilis	–3,36	0,42	48,64	46,87
75. percentilis	45,70	46,08	89,97	87,69

Forrás: saját számítás.

A 3. fejezetben ismertetett módszertannak megfelelően a GMR esetében az eredményváltozó feltételes eloszlása minden időpontban más és más – a magyarázóváltozók függvényében kalkulált p_{ik}, m_{ik} és s_{ik} paraméterekkel leírható – keverékeloszlás. Az 8. ábrán a GMR standard hibák (azaz a feltételes eloszlások szórásának az) átlaga és a reziduumok szórása látható órás bontásban.

A szezonális bontás (nyár, tél, átmenet[17]) már önmagában véve okoz némi tökéletlenséget a megjelenítésben (különböző – például időjárási – okokból kifolyólag nem minden évben „viselkedik" ugyanúgy minden hónap). Emellett a standard hibák különböző (alakú) feltételes eloszlások szórásai, így az átlaguk is inkább csak nagyságrendileg releváns, valamint a reziduumok szórásának számolása is implicit módon feltételez egyféle homogenitást, ami csak részben teljesül. Mindezen korlátozó tényezők szem előtt tartásával megállapítható, hogy a két mennyiség együttmozgása a reziduumokkal konzisztens standard hibákra utal. Ezt a fejezet későbbi részében még ellenőrizni fogom.

Érdemi megállapítások fogalmazhatók meg a bizonytalanság szezonális viselkedéséről többek között az alábbi dimenziókban:

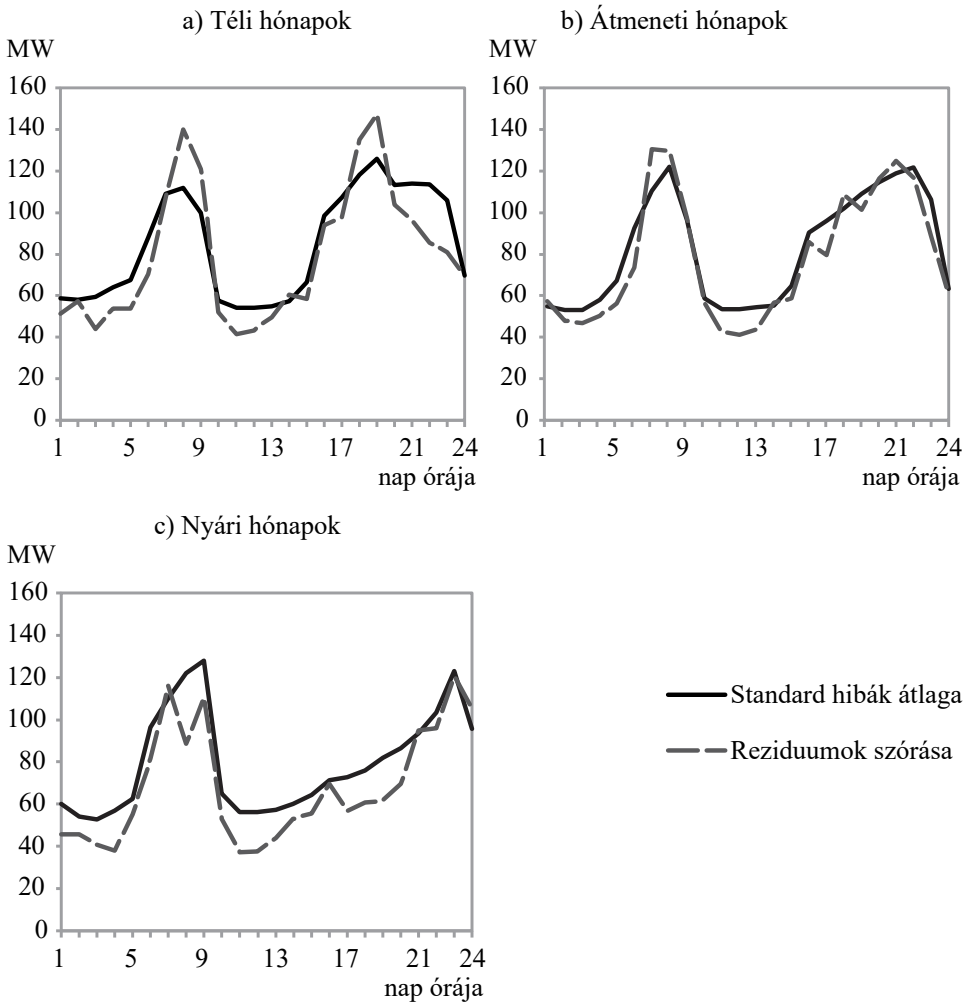
– a reggeli felfutás (ún. *ramp-up*) és a naplementét követő órák *versus* az ezen kívül eső időszakok bizonytalansága;
– a szezonálisan változó esti (naplemente utáni ún.) bizonytalansági csúcs időpontja;
– a nyári *versus* téli és átmeneti hónapok délutáni csúcsidőszaki óráinak bizonytalansága.

[17] A szezonális bontás a 2. fejezethez hasonlóan az alábbiak szerint történt: téli hónapok: január, február, december; átmeneti hónapok: március, április, május, illetve szeptember, október, november; nyári hónapok: június, július, augusztus.

8. ábra

A GMR standard hibák átlaga és a reziduumok szórása, 2012–2019

Mean standard errors and standard deviations of residuals in GMR, 2012–2019



Forrás: saját számítás.

Részletesebben kifejtve a fentieket, általánosságban elmondható, hogy évszaktól függetlenül a reggeli felfutások bizonytalansága relatíve nagyobb. Hasonló bizonytalanság jellemzi az esti időszakot is, de a nyári hónapokban összességében ez a bizonytalanság némileg alacsonyabb. Utóbbinak feltehetően az az oka, hogy a téli, illetve az átmeneti hónapokban a naplemente miatti világítási hatás

(és ennél fogva a hozzá kapcsolódó bizonytalanság is) sokkal erősebb és a naplemente időpontjával szinkronban sokkal korábban is kezdődik.

Elsőre talán meglepő eredmény, hogy a bizonytalanság mértéke nem teljesen arányos a rendszerterhelés szintjével: például a kora délutáni időszak bizonytalanságának mértéke nagyságrendileg hasonló az éjjeli, korai hajnali időszakhoz. Ennek az a magyarázata, hogy a csúcsidőszakban a fogyasztás sokkal szabályosabb, így sokkal nagyobb megbízhatósággal előrejelezhető, mint például a szintén aktív esti órákban.

A nyári időszakban a (kora) délutáni órák magasabb bizonytalansága a téli vagy az átmeneti hónapokhoz képest szintén fundamentálisan magyarázható: többek között feltehetően a magasabb hőmérséklet miatti kevésbé kiszámítható légkondicionáló-használat áll mögötte.

Ugyan a teljes előrejelzési időszakra vonatkozóan a GMR által becsült konfidenciaintervallumon kívül eső megfigyelések aránya minimálisan meghaladja az elméletileg várt 20, 10, illetve 5%-ot (2. táblázat), a meghaladás mértéke gyakorlatilag minimálisnak tekinthető évszaktól függetlenül.

2. táblázat

Konfidenciaintervallumon kívül eső megfigyelések aránya, 2012–2019

Percentage of observations outside confidence interval, 2012–2019

(%)

Megbízhatósági szint	Téli	Átmeneti	Nyári	Összesen
	hónapok			
80%	23,3	21,6	21,8	22,1
90%	12,8	11,3	11,4	11,7
95%	7,4	5,9	6,1	6,3

Forrás: saját számítás.

A standard hibák reziduumokkal való konzisztenciájának ellenőrzésére egy sokkal egzaktabb, de egyben sokkal szigorúbb elvárásoknak történő megfelelés tükröződik a 9. és a 10. ábráról. Ezek azt mutatják, hogy a választott 80, 90, illetve 95%-os konfidenciaintervallumon kívül eső megfigyelések aránya hogyan alakul a GMR és az OLS esetében különböző szezonális bontásokban.

Az ábrákon szereplő felbontáshoz szeretnék néhány megjegyzést fűzni. A 9. ábrán a napkelte/naplemente időpontjához igazodó napfényes/napfény nélkül időszak szerinti csoportosítás szerepel, mivel ez a bontás jobban diszkriminál, mint az állandó „kereskedelmi" jellegű csúcs/völgy bontás[18] (utóbbi esetben a lineáris regresszió esetében is viszonylag egyenletesek a konfidenciaintervallu-

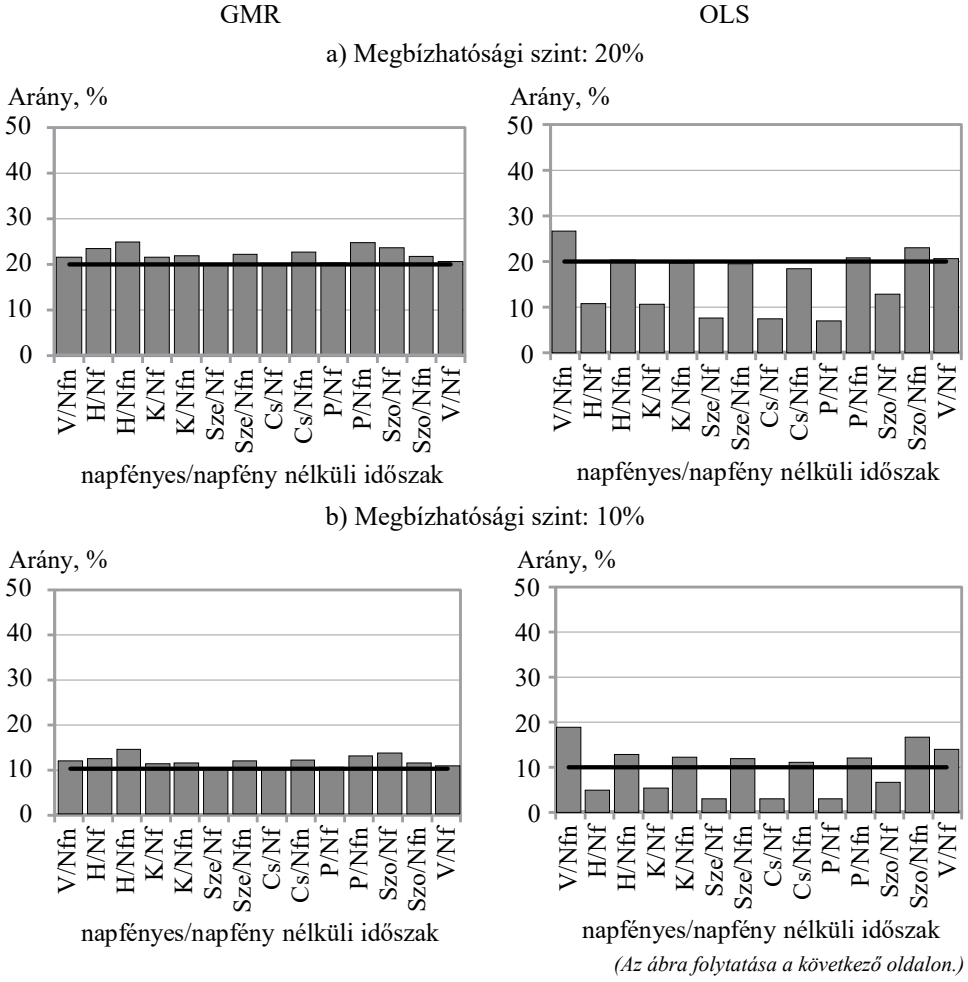
[18] A villamosenergia-kereskedelemben használt csúcs/völgyidőszak definíciója szerint csúcsidőszaknak számítanak a hétköznapok reggel 8 és este 8 óra között, völgyidőszaknak az ezen kívül eső órák, illetve a hétvégék.

mon kívüli arányok). A könnyebb ábrázolás érdekében egy adott naptári nap nap végi napfény nélküli időszakát és a következő naptári nap nap eleji, napfény nélküli időszakát összevontam, azaz pl. a vasárnapi napfény nélküli időszak (a 9. ábrán az első kategóriák) a vasárnapi nap végét és a hétfői nap elejét foglalja magában, a hétfői napfény nélküli időszak (a 9. ábrán a harmadik kategóriák) pedig a hétfői nap végét és a keddi nap elejét stb. A 10. ábra hónapok szerinti csoportosítást tartalmaz.

9. ábra

Konfidenciaintervallumon kívül eső megfigyelések aránya napfényes/napfény nélküli időszak bontásban, 2012–2019

Percentage of observations outside confidence intervals by daylight/no daylight period, 2012–2019

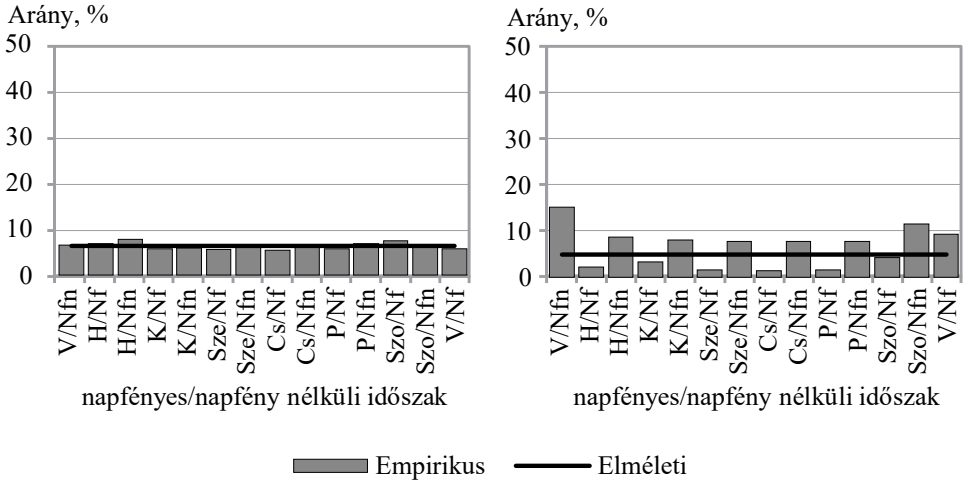


(Az ábra folytatása a következő oldalon.)

(folytatás)

GMR OLS

c) Megbízhatósági szint: 5%



Forrás: saját számítás.

Jól látható, hogy a GMR esetében a konfidenciaintervallumon kívül eső megfigyelések aránya jóval egyenletesebben oszlik el a 20, a 10, illetve az 5%-os értékek mentén, mind a napfényes/napfény nélküli, mind a havi bontás esetében. Szembetűnő javulás a napfényes/napfény nélküli időszak szerinti bontásban van, hiszen a napfény nélküli időszakok lefedik a naplemente utáni órákat és a reggeli felfutás óráit is, amikor magasabb a bizonytalanság (különösen télen és az átmeneti hónapokban, amelyek során később kel a nap).

Érdemes még megjegyezni, hogy abból adódóan, hogy a reziduumokra vonatkozó feltételek (normalitás, heteroszkedaszticitás) nem teljesülnek a lineáris regresszió esetében, a konfidenciaintervallumon kívüli arányok itt sem felelnek meg az elméletileg várt 20, 10, illetve 5%-os értékeknek. A reziduumok eloszlása a normálisnál csúcsosabb, ami azt eredményezi, hogy magasabb megbízhatósági szinteken (pl. 20%) inkább az elméletileg várt alatt, alacsonyabb megbízhatósági szinteken (pl. 5%) pedig inkább afelett van a vizsgált arány.

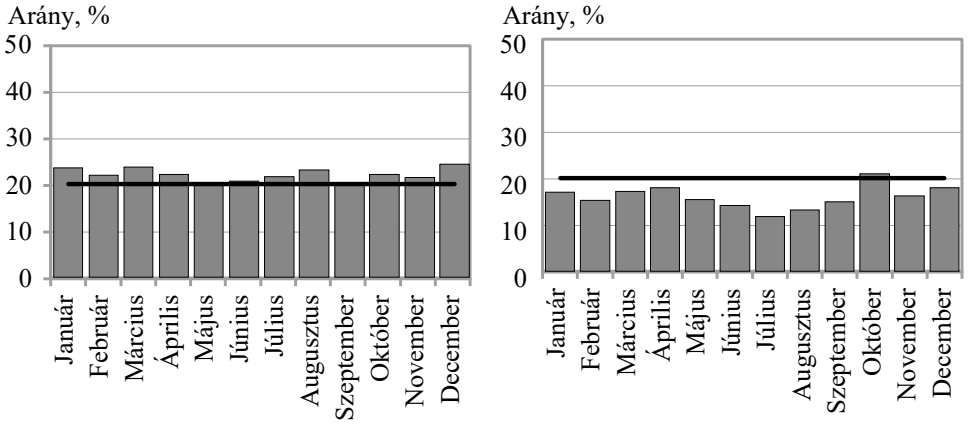
10. ábra

Konfidenciaintervallumon kívül eső megfigyelések aránya havi bontásban, 2012–2019

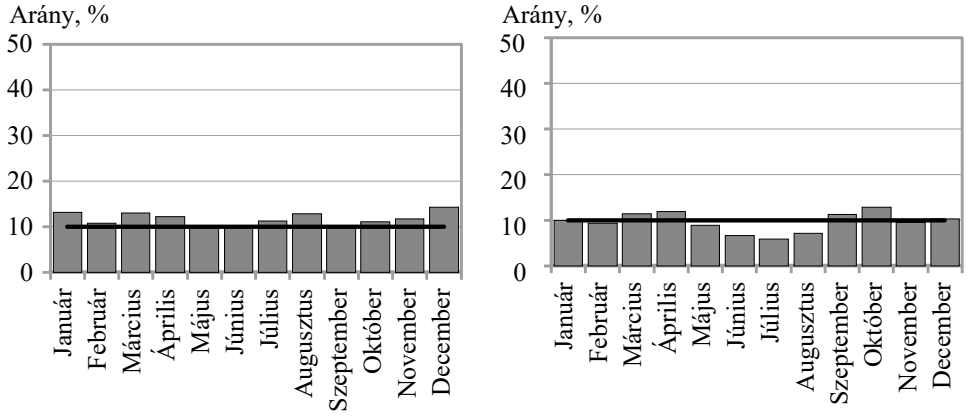
Percentage of observations outside confidence intervals by month, 2012–2019

GMR OLS

a) Megbízhatósági szint: 20%

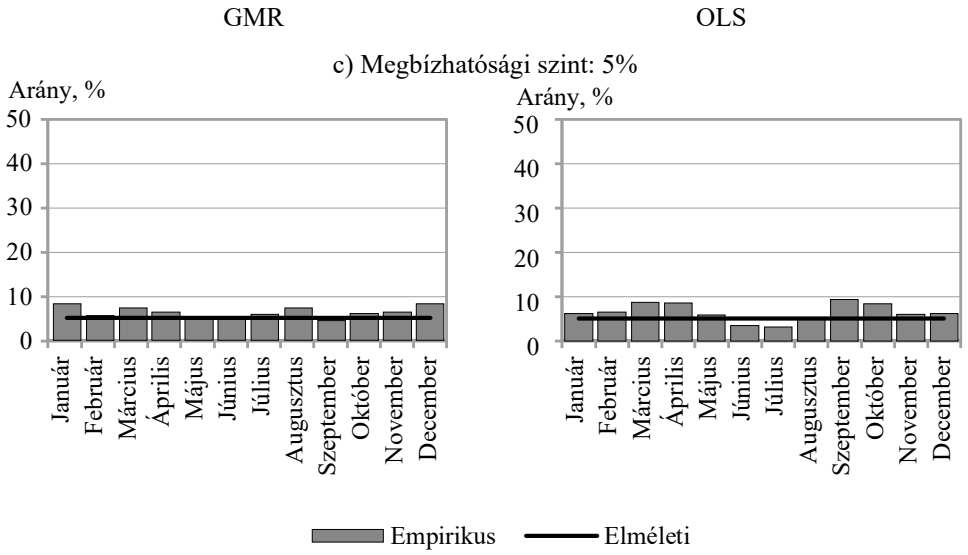


b) Megbízhatósági szint: 10%



(Az ábra folytatása a következő oldalon.)

(folytatás)



Forrás: saját számítás.

5. Összefoglalás

A fogyasztási/terhelési idősorok megbízhatóságának részletesebb vizsgálata elsősorban az utóbbi években kapott igazán nagy figyelmet. A tanulmányban a hazai bruttó villamosenergiarendszer-terhelés előrejelzési kockázatát vizsgálva bemutattam, hogy a villamosenergia-rendszer terhelésének alakulásához hasonlóan a bizonytalanságot is erős szezonális viselkedés jellemzi. Ennek a heteroszkedasztikus viselkedésnek a leírására az ún. Gauss-keverékregressziót *(GMR)* használtam. A módszer több helyen megjelenik a nemzetközi energiapiaci szakirodalomban, de elsősorban a nemlineáris és az interakciós hatások kezelésére alkalmazzák, és a mögöttes módszertani hátteret is sok esetben nagyon röviden tárgyalják.

Az empirikus eredmények az alábbiak szerint foglalhatók össze: az országos rendszerterhelés bizonytalansága évszaktól függetlenül a reggeli felfutás óráiban, illetve a naplementét követő órákban jóval magasabb, mint egyébként. Előbbi nem meglepő, hiszen a fogyasztás gradiense ilyenkor a legnagyobb. Utóbbi pedig leginkább azzal magyarázható, hogy a napközbeni aktív időszak sokkal kiszá-

míthatóbb, mint a nap végi, sok szempontból még szintén aktív, de már természetes napfény nélküli időszak. Mindemellett az eredmények alapján az is elmondható, hogy a nyári hónapok délutáni időszakaiban az előrejelzési kockázat valamivel magasabb, feltehetően a légkondicionáló-használat miatt.

Mindezek mellett bemutattam, hogy a GMR a reziduumokkal konzisztens standard hibákat becsül. Az ehhez szükséges GMR-konfidenciaintervallumok számítása a publikusan elérhető programcsomagnak nem része, és hasonló elemzés (reziduumok vs. standard hibák) a (GMR-t használó, de gyakran más) nyilvánosan elérhető publikációkban sincs. Megállapításaim minőségét erősíti, hogy kizárólag mintán kívüli (előrejelzési) eredményekről beszélek.

A fenti eredmények relevanciája és a további kutatási lehetőségek céljából érdemes néhány további szempontot megemlíteni. A tanulmányban nem vettem figyelembe minden változót, ami elméletileg rendelkezésre állhat és a villamosenergia-fogyasztást szignifikánsan befolyásolja. Ilyenek például a szél, a levegő páratartalma vagy az égbolt felhőtakaróval való fedettsége. *Miller és Nam (2022)* szerint például magas hőmérséklet esetén a páratartalom figyelmen kívül hagyása a hőmérséklet hatását szignifikánsan alul-, míg alacsony hőmérséklet esetén a felhőtakaró figyelmen kívül hagyása a hőmérséklet hatását szignifikánsan felülbecsüli. A hőmérséklethez képest a többi időjárási változó esetében viszont sokkal nehezebb megbízható, a teljes országra jellemzőnek mondható adatokat beszerezni vagy konstruálni.

További fontos elemzési lehetőség a GMR kiegészítése, illetve kiterjesztése az ún. kevésbé tipikus időszakok irányítottabb figyelembevételével – önmagában vagy más módszerekkel kombinálva. Ez jelenti egyrészt az ún. speciális napok (ünnepnapok, munkanap-áthelyezések) kezelését, másrészt az elemzés elmúlt három évre történő kiterjesztését. Klaszterezés jellegű módszertanról lévén szó, a GMM-nek elsősorban a tipikus mintázatok felismerésében van előnye; ilyen tekintetben a speciális napok – alacsony számuk miatt – kevésbé nevezhetők tipikusnak. Ezen túlmenően pedig a bizonytalanság vizsgálata a Covid19-járvánnyal vagy a 2022 második felében megjelenő rezsiválsággal kapcsolatban ugyanúgy külön elemzés tárgyát kell, hogy képezze, mint magának a villamosenergia-fogyasztás várható szintjének a vizsgálata (*Hortay–Szőke, 2020*). Olyan alapvető strukturális változásokról beszélünk ugyanis (trendtörés, a fogyasztás napon belüli alakjának megváltozása stb.), amelyeknek a vizsgálata és a részletes bemutatása meghaladja ennek a tanulmánynak a kereteit. A bizonytalanságot illetően a nemzetközi szakirodalom sem tartalmaz még erre vonatkozóan sok publikált eredményt. Abból adódóan, hogy a GMM/GMR kis mintaelemszám mellett is képes az új struktúrák beazonosítására, a módszertan alkalmazása erre az időszakra is ígéretes.

Melléklet – Módszertani kiegészítések

M.1. A Gauss-keverékmodell felírása

Tegyük fel, hogy a megfigyeléseink egy K darab komponenst tartalmazó Gauss-keverékeloszlással írhatók le, aminek sűrűségfüggvénye az alábbi:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot p(x; \mu_k, \Sigma_k)\,, \qquad \text{/M1/}$$

ahol:

– π_k a k-adik komponens (ún. keverék)súlya (k=1,2…K),
– p a normális eloszlás sűrűségfüggvénye μ_k átlagvektorral és Σ_k kovarianciamátrixszal (k=1,2…K),
– K a komponensek száma.

Feltételezve, hogy a z látens változó a π_k paraméterekkel leírható kategoriális eloszlást követi (jelöljük ezt a valószínűségi súlyfüggvényt $p(z)$-vel, ahol $\sum_{k=1}^{K} \pi_k = 1$ és $\pi_k \geq 0$ minden k-ra) és követve a 3. fejezetből a $p(x) = \sum_z p(z) \cdot p(x|z)$ összefüggést, a *loglikelihood*-függvény (feltételezve, hogy a megfigyeléseink egymástól függetlenek) az alábbiak szerint írható fel:

$$l(\pi, \mu, \Sigma) = \ln p(X; \pi, \mu, \Sigma) = \sum_{i=1}^{n} \ln\big(p(x_i; \pi, \mu, \Sigma)\big) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i, z_i = k; \mu, \Sigma, \pi) = \qquad \text{/M2/}$$
$$= \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i | z_i = k; \mu, \Sigma) \cdot \mathrm{p}(\mathrm{z}_i = \mathrm{k}; \pi) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} p(x_i; \mu_k, \Sigma_k) \cdot \pi_k,$$

ahol π, μ és Σ rendre a π_k, μ_k és Σ_k komponensenkénti paraméterek összevont, egyszerűsített jelölésére szolgál ($\pi = (\pi_1, \pi_2 \ldots \pi_K), \mu = (\mu_1, \mu_2 \ldots \mu_K)$ és $\Sigma = (\Sigma_1, \Sigma_2 \ldots \Sigma_K)$). A becslési feladat nehézségét alapvetően az adja, hogy a z értékek látensek és értékük függ a többi paraméter értékétől, így a *loglikelihood*-függvény maximalizálása a *maximum likelihood* (ML-) módszerrel a szokásos módon nem hatékony (minden megfigyelésre a z változó K lehetséges értékét figyelembe véve kellene kiszámolni a *loglikelihood*-függvény értékét).

A z ismeretének hiányában az *expectation-maximization (EM-)* eljárás megoldása a problémára az, hogy a komponenstagságokat – konkrét ismeretük hiányában – megbecsüli iteratív módon, egészen addig, amíg az eredmények nem kon-

vergálnak. Technikailag annyi történik, hogy a *loglikelihood*-függvény helyett annak egy jól megválasztott *alsó korlátját* optimalizáljuk. A választás biztosítani fogja, hogy a logaritmus és a $\sum_{k=1}^{K}$ … kifejezések lényegében helyet cseréljenek, így az optimális paraméterekre jól interpretálható zárt formulákat fogunk kapni.

A könnyebb érthetőség kedvéért egyet lépjünk vissza: amennyiben a *z* értékek megfigyelhetők lennének, a *loglikelihood*-függvény az alábbi módon lenne felírható:

$$l(\pi,\mu,\Sigma)=\sum_{i=1}^{n}\ln\big(p(x_i,z_i;\pi,\mu,\Sigma)\big)=\sum_{i=1}^{n}\ln\big(p(x_i|z_i;\mu,\Sigma)\cdot p(z_i|\pi)\big) \qquad /M3/$$

Ahogy a jelölés is mutatja, ez nemcsak a paraméterek, hanem a megfigyelhető *x* és *z* változók függvénye is. A kifejezés átalakítható az alábbi módon:

$$\begin{aligned} l(\pi,\mu,\Sigma)&=\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\ln\Big(p\big(x_{i_k}\big|z_{i_k}=k;\mu,\Sigma\big)\Big)+\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\ln(p(z_{i_k}=k;\pi))\\ &=\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\ln\big(p(x_{i_k};\mu_k,\Sigma_k)\big)+\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\ln(\pi_k)\end{aligned}, \qquad /M4/$$

ahol n_k a *k*-adik komponensbe eső megfigyelések száma, i_k pedig a *k*-adik komponensbe eső *i*-edik megfigyelés (i_k=1,2…n_k), illetve $\sum_{k=1}^{K} n_k = n$). A fenti összefüggésben kihasználtuk, hogy a szorzat logaritmusa a tagok logaritmusának összegeként írható fel. Innen már látható, hogy a maximalizálási feladat jóval könnyebb, hiszen *K* darab normális eloszlás és egy kategoriális eloszlás paramétereinek ML-becslésére vezethető vissza.

Az EM-eljárás célfüggvénye (optimalizálási szempontból) nagyon hasonlít majd ehhez a *z* megfigyelhetőségét feltételező *likelihood*-függvényhez.

M.2. Az *expectation–maximization* (EM) becslési eljárás

A 3. fejezetben az EM-eljárás két legfontosabb lépését nagyvonalakban már ismertettem. Az alábbi táblázatban csak összefoglalom ezeket, célom elsősorban az *EM*-eljárás hátterének a bemutatása.

M1. táblázat

Az *expectation-maximization* eljárás fő lépései
Main steps of Expectation-Maximization algorithm

Inicializálás	A kezdő π_k, μ_k és Σ_k értékek meghatározása
iterálás	
E-lépés	– z posterior-eloszlásának (p_{ik} posterior-valószínűségeknek) a meghatározása a Bayes-formula alapján, az aktuális π_k, μ_k és Σ_k paraméterek mellett – Q-függvény felírása a p_{ik} posterior-valószínűségek felhasználásával
M-lépés	Q-függvény maximalizálása a π_k, μ_k és Σ_k paraméterek mentén
amíg: a π_k, μ_k és Σ_k paraméterek konvergálnak	

Érdemes megemlíteni, hogy az EM-eljárás lényege alapvetően az, hogy a *likelihood*-függvényt a látens z változó miatt nem tudjuk közvetlenül optimalizálni, viszont egy jól definiálható alsó korlátját igen. Az E- és M-lépések iteratív elvégzésével ez az alsó korlát folyamatosan javul; amennyiben nem, akkor elértük az alsó korlát (és egyben a *likelihood*-függvény) optimumát. Kicsit formálisabban, legyen a *loglikelihood*-függvényünk az alábbi:

$$l(\pi,\mu,\Sigma)=\sum_{i=1}^{n}\ln(p(x_i;\pi,\mu,\Sigma))=\sum_{i=1}^{n}\ln\sum_{k=1}^{K}p(x_i,z_i=k;\pi,\mu,\Sigma)\,. \qquad /M5/$$

Bevezetve egy új, q_k eloszlást, a *loglikelihood*-függvény az alábbi formában írható át:

$$l(\pi,\mu,\Sigma)=\sum_{i=1}^{n}\ln(p(x_i;\pi,\mu,\Sigma))=\sum_{i=1}^{n}ln\sum_{k=1}^{K}q_k\cdot\frac{p(x_i,\ z_i=k;\pi,\mu,\Sigma)}{q_k}, \qquad /M6/$$

amelynek az alsó korlátja az ún. Jensen-egyenlőtlenséget[19] felhasználva könnyen származtatható:

$$l(\pi,\mu,\Sigma)=\sum_{i=1}^{n}ln\sum_{k=1}^{K}q_k\cdot\frac{p(x_i,z_i=k;\pi,\mu,\Sigma)}{q_k}\geq\sum_{i=1}^{n}\sum_{k=1}^{K}q_k\cdot\ln\left(\frac{p(x_i,z_i=k;\pi,\mu,\Sigma)}{q_k}\right). \quad /M7/$$

Amennyiben a q_k eloszlást a p_{ik} *posterior*-eloszlásként határozzuk meg,[20] rövid átalakítások után megkapható az ún. Q-függvény vagy várható *loglikelihood*:

[19] A Jensen-egyenlőtlenség szerint a logaritmus mint konkáv függvény esetén $log(E(x))\geq E(log(x))$.
[20] Az is szemléletesen levezethető, hogy miért éppen a p_{ik} *posterior*-valószínűségeket célszerű választani, ennek ismertetésétől azonban most eltekintek.

$$Q(\pi,\mu,\Sigma)=\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}\cdot\ln(p(x_i,z_i=k;\pi,\mu,\Sigma))\,. \qquad /M8/$$

Ennek vagy az alsó korlátnak az optimalizálása ekvivalens, hiszen a $\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}\cdot\ln(p_{ik})$ nem függ az optimalizálni kívánt π, μ, és Σ paraméterektől.[21]

A könnyebb érthetőség végett az említett lépéseket kicsit formálisabban a konkrét GMM-problémára megfogalmazva az alábbiak történnek:

Legyenek az *r*-edik lépésbeli paramétereink $\pi_k^{(r)},\mu_k^{(r)}$ és $\Sigma_k^{(r)}$[22]. Ekkor a $p_{ik}^{(r+1)}$, azaz az *i*-edik megfigyelés *k*-adik komponenshez tartozási (x_i értékek feltétele melletti, *posterior*) valószínűsége az (*r*+1)-edik iterációban az alábbi:

$$p_{ik}^{(r+1)}=p(z_i=k|x_i)=$$
$$=\frac{p(z_i=k)\cdot p(x_i|z_i=k)}{p(x_i)}=\frac{p(z_i=k)\cdot p(x_i|z_i=k)}{\sum_{j=1}^{K}p(z_i=j)\cdot p(x_i|z_i=j)}=\frac{\pi_k^{(r)}\cdot p(x_i;\mu_k^{(r)},\Sigma_k^{(r)})}{\sum_{j=1}^{K}\pi_j^{(r)}\cdot p(x_i;\mu_j^{(r)},\Sigma_j^{(r)})} \qquad /M9/$$

Ezen *posterior*-valószínűségek felhasználásával felírva a Q-függvényt:

$$\begin{aligned}Q(\pi,\mu,\Sigma)&=\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}^{(r+1)}\cdot\ln(p(x_i,z_i=k;\pi,\mu,\Sigma))=\\&=\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}^{(r+1)}\cdot\ln(p(x_i|z_i=k;\mu,\Sigma)\cdot p(z_i=\mathrm{k};\pi)) \qquad , \qquad /M10/\\&=\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}^{(r+1)}\cdot\ln(p(x_i;\mu_k,\Sigma_k))+\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}^{(r+1)}\cdot\ln(\pi_k)\end{aligned}$$

amely a p_{ik} valószínűségektől eltekintve nagyon hasonló ahhoz, amit akkor kaptam, amikor a *loglikelihood*-függvényt úgy írtam fel, hogy a *z* változóról feltételeztem, hogy ismert.

[21] Amennyiben az alsó korlátot tovább bontjuk, az alábbi kifejezést kapjuk:
$\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}\cdot\ln\left(\frac{p(x_i,z_i=k;\pi,\mu,\Sigma)}{p_{ik}}\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}\cdot\ln(p(x_i,z_i=k;\pi,\mu,\Sigma))-\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ik}\cdot\ln(p_{ik})$.

Ebből a második tagban – tekintve, hogy a q_k eloszlást a p_{ik} *posterior* eloszlásként rögzítettem – semmi sem függ a π, a μ és a Σ paraméterektől, ezért a Q-függvényben csak az első tagot szerepeltetem: az optimális paraméterek ettől függetlenül ugyanazok lesznek.

[22] Az *r*=0 a kezdő inicializált paramétereket jelöli.

Könnyen levezethető, hogy a fenti Q-függvényt maximalizálva a keresett π_k, μ_k és Σ_k paraméterek (r+1)-edik iterációbeli értékeire az alábbi zárt formulákat kapjuk[23]:

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{ik}^{(r+1)} \cdot x_i}{n_k^{(r+1)}}, \qquad \text{/M11/}$$

$$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{ik}^{(r+1)} \cdot (x_i - \mu_k^{(r+1)}) \cdot (x_i - \mu_k^{(r+1)})^T}{n_k^{(r+1)}},$$

$$\pi_k^{(r+1)} = \frac{n_k^{(r+1)}}{n}, \text{ ahol } n_k^{(r+1)} = \sum_{i=1}^{n} p_{ik}^{(r+1)} \text{ és } \quad n = \sum_{i=1}^{n} n_k^{(r+1)}$$

Ezekből már visszavezethető az ismert z értékek feltételezése melletti eset, ami – ahogy korábban is említettem – K darab normális eloszlás és egy kategoriális eloszlás paramétereinek *ML*-becslésével ekvivalens. Az M-lépés egyébként hasonló az ML-becsléshez, annyi különbséggel, hogy az M-lépésben várható *loglikelihood*ot maximalizálok (ezt és iteratív módon többször is elvégzem).

A gyakorlati alkalmazások során természetesen szembesülünk az optimális komponensszám meghatározásának problémájával. Ez történhet például modellszelekciós kritériumok alapján *(pl. AIC, BIC),* vagy keresztvalidációval. A tanulmányban az utóbbi megoldást választottam, ennek részleteiről az empirikus fejezetben írtam részletesebben.

M.3. A Gauss-keverékregresszió származtatása

A tanulmány 3. fejezetében már röviden ismertettem a GMR-nek a GMM-ből történő származtatásának az elvét. A Melléklet ezen fejezetében csak a komponensenkénti regressziós együtthatók formális származtatását részletezem, mivel a dolgozat lényegi tartalmához kapcsolódó összefüggések (az eredményváltozó feltételes eloszlásáról) a 3. fejezetben részletesen szerepelnek.

Végezzük el a GMM által becsült komponensenkénti μ_k és Σ_k átlagvektorok és kovarianciamátrixok partícionálását a regressziós eredmény-, illetve magyarázóváltozó felosztásnak megfelelően:

$$\mu_k = \begin{bmatrix} \mu_k^y \\ \mu_k^x \end{bmatrix} \text{ és } \Sigma_k = \begin{bmatrix} \Sigma_k^{yy} & \Sigma_k^{yx} \\ \Sigma_k^{xy} & \Sigma_k^{xx} \end{bmatrix}, \text{ az alábbi méretekkel: } \begin{bmatrix} 1 \times 1 \\ p \times 1 \end{bmatrix} \text{ és } \begin{bmatrix} 1 \times 1 & 1 \times p \\ p \times 1 & p \times p \end{bmatrix}. \text{ /M12/}$$

[23] A μ és a Σ paraméter meghatározásakor csak a Q-függvény első, a π meghatározásakor csak a Q-függvény második tagjával kell dolgoznunk, hiszen az ellenkező tagok nem tartalmazzák az ismeretlen – optimalizálni kívánt – paramétereket.

Mind az átlagvektorok, mind a kovarianciamátrixok a p_{ik} *posterior*-valószínűségekkel súlyozottak:

$$\mu_k^y = \frac{\sum_{i=1}^n p_{ik} \cdot y_i}{\sum_{i=1}^n p_{ik}} \qquad \Sigma_k^{yy} = \frac{\sum_{i=1}^n p_{ik} \cdot (y_i - \mu_k^y)^T \cdot (y_i - \mu_k^y)}{\sum_{i=1}^n p_{ik}}$$

$$\Sigma_k^{yx} = \frac{\sum_{i=1}^n p_{ik} \cdot (y_i - \mu_k^y)^T \cdot (x_i - \mu_k^x)}{\sum_{i=1}^n p_{ik}} \qquad \mu_k^x = \frac{\sum_{i=1}^n p_{ik} \cdot x_i}{\sum_{i=1}^n p_{ik}} \qquad \text{/M13/}$$

$$\Sigma_k^{xy} = \frac{\sum_{i=1}^n p_{ik} \cdot (x_i - \mu_k^x)^T \cdot (y_i - \mu_k^y)}{\sum_{i=1}^n p_{ik}} \qquad \Sigma_k^{xx} = \frac{\sum_{i=1}^n p_{ik} \cdot (x_i - \mu_k^x)^T \cdot (x_i - \mu_k^x)}{\sum_{i=1}^n p_{ik}}$$

Könnyen levezethető *(Sung, 2004),* hogy a komponensenkénti regressziós együtthatók a *posterior*-valószínűségekkel súlyozott regressziós együtthatókként állnak elő, azaz:

$$\widehat{\beta_k} = (\Sigma_k^{xx})^{-1} \cdot \Sigma_k^{xy} \qquad \text{/M14/}$$

A vázolt jelölésrendszerből adódóan a $\widehat{\beta_k}$ nem tartalmazza a regressziós konstans tagot, de ez összhangban van a 3.3. fejezetben bemutatott formulákkal.

Ehhez hasonlóan származtatható a komponensenkénti hibatag varianciája:

$$\widehat{\sigma_k^2} = \Sigma_k^{yy} - \Sigma_k^{yx} \cdot (\Sigma_k^{xx})^{-1} \cdot \Sigma_k^{xy} \qquad \text{/M15/}$$

A Melléklet ezen fejezetében bemutatott jelölések azonosak a 3.3 fejezetben használtakkal. A könnyebb átláthatóság érdekében ezeket összefoglalásképpen lásd az alábbi táblázatban.

M2. táblázat

Fontosabb jelölések a GMR-ben

Notation used in GMR

y_i	az eredményváltozó az *i*-edik megfigyelés esetén
x_i	a magyarázóváltozók (p×1) méretű vektora az *i*-edik megfigyelés esetén
p_{ik}	az *i*-edik megfigyelés *k*-adik komponensbe tartozásának *posterior*-valószínűsége
i	*i*-edik megfigyelés (i=1,2,…n)
k	*k*-adik komponens (k=1,2,…K)
K, n, p	komponensek száma, mintaelemszám, magyarázóváltozók száma
μ_k^y, μ_k^x	partícionált átlagvektorok a *k*-adik komponens esetén
Σ_k^{yy}, Σ_k^{yy}, Σ_k^{xy}, Σ_k^{xx}	partícionált kovarianciamátrixok a *k*-adik komponens esetén
$\widehat{\beta_k}$	regressziós együtthatók a *k*-adik komponens esetén
$\widehat{\sigma_k^2}$	reziduális variancia a *k*-adik komponens esetén
m_{ik}	az *i*-edik megfigyelés feltételes várható értéke a *k*-adik komponens esetén
s_{ik}^2	az *i*-edik megfigyelés feltételes varianciája a *k*-adik komponens esetén
$\widehat{y_i}$	az *i*-edik megfigyelés feltételes várható értéke
$var(\widehat{y_i})$	az *i*-edik megfigyelés feltételes varianciája (standard hibája)

Irodalom

Ausín, M. C. – Galeano, P. (2007): Bayesian estimation of the Gaussian mixture GARCH model. *Computational Statistics & Data Analysis*. Vol. 51. No. 5. pp. 2636–2652. https://doi.org/10.1016/j.csda.2006.01.006

Bishop, C. (2006): *Pattern recognition and machine learning.* Springer, New York.

Bracale, A. – Caramia, P. – De Falco, P. – Hong, T. (2019): Multivariate quantile regression for short-term probabilistic load forecasting. *IEEE Transactions on Power Systems.* Vol. 35. No. 1. pp. 628–638. https://doi.org/10.1109/TPWRS.2019.2924224

Cont, R. (2001): Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance.* Vol. 1. pp. 223–236. https://doi.org/10.1080/713665670

Dempster, A.P. – Laird, N.M. – Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1) pp. 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Espinoza, M. – Joye, C. – Belmans, R. – De Moor, B. (2005): Short-Term Load Forecasting, Profile Identification and Customer Segmentation: A Methodology based on Periodic Time Series. *IEEE Transactions on Power Systems*. Vol. 20. No. 30. pp. 1622–1630. https://doi.org/10.1109/TPWRS.2005.852123

Fabisch, A. (2021): GMR: Gaussian Mixture Regression. *Journal of Open Source Software*. Vol. 6. No. 62. p. 3054. https://doi.org/10.21105/joss.03054

Friedman, J. H. (1991): Multivariate Adaptive Regression Splines. *The Annals of Statistics*. Vol. 19. No. 1. pp. 1–67. https://doi.org/10.1214/aos/1176347963

Goodfellow, I. – Pouget-Abadie, J. – Mirza, M. – Xu, B. – Warde-Farley, D. – Ozair, S. – Courville, A. – Bengio, Y. (2014): *Generative adversarial nets. Advances in Neural Information Processing Systems.* pp. 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

Heo, Y. – Zavala, V. M. (2012): Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings*. Vol. 53. pp. 7–18. https://doi.org/10.1016/j.enbuild.2012.06.024

Hiruta, Y. – Gao, L. – Ashina, S. (2022): A novel method for acquiring rigorous temperature response functions for electricity demand at a regional scale. *Science of the Total Environment.* 819 p. 152893. https://doi.org/10.1016/j.scitotenv.2021.152893

Hong, T. – Fan, S. (2016): Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting.* Vol. 32. No. 3. pp. 914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011

Hortay O. – Szőke T. (2020): A kijárási korlátozás hatása a villamosenergia-rendszer terhelési és árgörbéire Magyarországon. *Statisztikai Szemle*. 98. évf. 10. sz. 1131–1150. o. https://doi.org/10.20311/stat2020.10.hu1131

Hossain, M. E. (2017): Application of Gaussian mixture regression model for short-term wind speed forecasting. *2017 North American Power Symposium (NAPS)*, Morgantown, WV, USA. pp. 1–6. https://doi.org/10.1109/NAPS.2017.8107222

Leng, T. K. – Cheong, C. W. – Hooi, T. S. (2014): Impact of global financial crisis on stylized facts bet-ween energy markets and stock markets. *Proceedings of the 3rd International Conference on Mathematical Sciences*. AIP Conf. Proc. Vol. 1602. pp. 994–1001. https://doi.org/10.1063/1.4882605

Lo, K. L. – Wu, Y. K. (2003): Risk assesment due to local demand forecast uncertainty in the competi-tive supply industry. *IEEE Proceedings – Generation, Transmission and Distribution.* Vol. 150. No. 5. pp. 573–581. https://doi.org/10.1049/ip-gtd:20030641

Mák, F. (2015): Az időjárás véletlen hatásának szerepe a szezonális kiigazítás során, a hazai földgázfogyasztás példáján. *Statisztikai Szemle*. 93. évf. 5. sz. 417–441. o.

Mák, F. (2017): *Fogyasztási kockázat a villamosenergia-piacon. Profilozás a fogyasztási bizonytalanság figyelembevétele mellett.* Budapesti Corvinus Egyetem. Budapest.

Manfren, M. – Aste, N. – Moshksar, R. (2013): Calibration and uncertainty analysis for computer models – A meta-model based approach for integrated building energy simulation. *Applied Energy.* Vol. 103 pp. 627–641. https://doi.org/10.1016/j.apenergy.2012.10.031

Marossy, Z. (2010): *A spot villamosenergia-árak elemzése statisztikai és ökonofizikai eszközökkel. PhD-értekezés.* Budapesti Corvinus Egyetem, Budapest.

McLachlan, G. – Krishnan, T. (1977): *The EM Algorithm and Extensions. Wiley Series in Probability and Statistics.* John Wiley & Sons. New York.

McLachlan, G. – Peel, D. (2000): *Finite Mixture Models. Wiley Series in Probability and Statistics.* John Wiley & Sons. New York.

Miller, J. I. – Nam, K. (2022): Modeling peak electricity demand: A semiparametric approach using weather-driven cross-temperature response functions. *Energy Economics.* Vol. 114 p. 106291. https://doi.org/10.1016/j.eneco.2022.106291.

Pedregosa, F. – Varoquaux, G. – Gramfort, A. – Michel, V. – Thirion, B. – Grisel, O. – Blondel, M. – Prettenhofer, P. – Weiss, R. – Dubourg, V. – Vanderplas, J. – Passos, A. – Cournapeau, D. – Brucher, M. – Perrot, M. – Duchesnay, É. (2011): Scikit-learn: Machine Learning in Python. *JMLR.* Vol. 12. pp. 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Srivastav, A. – Tewari, A. – Dong, B. (2013): Baseline building energy modeling and localized uncerta-inty quantification using Gaussian mixture models. *Energy and Buildings*. Vol. 65. pp. 438–447. https://doi.org/10.1016/j.enbuild.2013.05.037

Subbarao, K. – Lei, Y. – Reddy, T. A. (2011): The Nearest Neighborhood Method to Improve Uncertainty Estimates in Statistical Building Energy Models. *ASHRAE Transactions*. Vol. 117. No. 2. pp. 459–471.

Sugár, A. (2011): A hőmérséklet hatásáról a villamosenergia- és gázfogyasztás magyarországi példáján. *Statisztikai Szemle.* 89. évf. 4. sz. 379–398. old.

Sung, H. G. (2004): *Gaussian mixture regression and classification*. Rice University, Houston, TX.

Taylor, J. W. – Buizza, R. (2002): Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems.* Vol. 17 pp. 626–632. https://doi.org/10.1109/TPWRS.2002.800906

Wang, Y. – Chen, Q. – Zhang, N. – Wang, Y. (2018): Conditional residual modeling for probabilistic load forecasting. *IEEE Trans. Power Systems*. Vol. 33. No. 6. 7327–7330. https://doi.org/10.1109/TPWRS.2018.2868167

Wang, Y. – Hug, G. – Liu, Z. – Zhang, N. (2020): Modeling load forecast uncertainty using generative adversarial networks. *Electric Power Systems Research.* Vol. 189. No. 106 732. https://doi.org/10.1016/j.epsr.2020.106732

Wang, L. – Kubichek, R. – Zhou, X. (2017): Adaptive learning based data-driven models for predicting hourly building energy use. *Energy and Buildings*. Vol. 159. pp. 454–461. https://doi.org/10.1016/j.enbuild.2017.10.054

Yuksel, S.E. – Wilson, J.N. – Gader, P.D. (2012): Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems.* Vol. 23 No. 8. pp. 1177–1193. https://doi.org/10.1109/TNNLS.2012.2200299

Zhang, W. – Quan, H. – Srinivasan, D. (2019): An improved quantile regression neural network for probabilistic load forecasting. *IEEE Trans. Smart Grid*. Vol. 10. No. 4. pp. 4425–4434. https://doi.org/10.1109/TSG.2018.2859749