Mihályffy László:

Kalibrálás és konvex programozás – egy határterület áttekintése

Előadásom címe „Kalibrálás és konvex programozás – egy határterület áttekintése”. A téma a véges sokaságok mintavételes eljárással való megfigyeléséhez kapcsolódik; kalibráláson egy olyan módszert értünk, amelynek segítségével a mintából származó becslések pontosságát külső információ segítségével javíthatjuk. Matematikai szempontból a módszer konvex programozási feladatra vezet, éspedig egy konvex célfüggvényt kell minimalizálnunk egy lineáris egyenletrendszer mint feltételrendszer mellett; a változókra individuális korlátok is szerepelhetnek, bár ezt nem mindig követelik meg. Kézenfekvő lenne tehát a kalibrálás feladatát matematikai programozási módszerrel kezelni, a gyakorlatban azonban nem ez történik, az esetek többségében másfajta, gyakran heurisztikus módszereket alkalmaznak. Előadásomban arra a kérdésre próbálok választ adni, hogy milyen előnyökkel és esetleges hátrányokkal jár ez a gyakorlat, miszerint programozási algoritmus helyett másfajta, esetenként heurisztikus módszereket alkalmaznak.
Mindenek előtt néhány szót kell szólni a kalibrálásról Célszerű ehhez az előbbi programozási feladatot más jelölésekkel felírni, éspedig olyanokkal, amelyek az egyes mennyiségek tartalmára jobban utalnak.

$$F(w_1,...,w_n,w_1^o,...,w_n^o) = \sum_{j=1}^{n} G(w_j / w_j^o) = \min!$$

$$\sum_{j=1}^{n} x_{ij} w_j = X_i\,, \quad i = 1, 2, \ldots, m\ (< n)$$

$$(\ L \le w_j / w_j^o \le U\,, \quad j = 1, 2, \ldots, n\ )$$

A feladat egy n elemű valószínűségi mintához kapcsolódik, melyben a megfigyelésekhez tartozó, eredeti mintasúlyok $w_1^o,...,w_n^o$, ezeket design súlyoknak is nevezzük. A kalibrálás célja ezeknek a megváltoztatása, módosítása úgy, hogy az itt látható feltételek teljesüljenek; a módosítás eredményeként kapott kalibrált súlyokat jelöljük w_1-gyel, w_2-vel, …, w_n-nel. Az x-ek tehát itt nem ismeretlenek, hanem az ún. segédváltozóknak a mintán megfigyelt értékei; itt olyan változókról van szó, amelyek nem tartoznak a vizsgálat, a felvétel célját jelentő változók körébe, viszont megfigyelhetők, és segítségükkel a mintából származó becslések pontossága javítható. Az egyenletek jobb oldalán a nagy X-ek a segédváltozók sokaságbeli értékösszegei; ezeknek vagy a pontos érékét, vagy legalább is nagy pontosságú közelítését ismernünk kell. Az ismeretlen w_j-kre vonatkozó individuális korlátokat egyes szerzők w_j^o-től független alakban írják föl, illetve bizonyos esetekben nem is veszik figyelembe, ezért tettem zárójelbe ezeket. Mindenesetre ebben a felírásban $0 \le L < 1$ és $U > 1$.
A célfüggvényt itt távolságfüggvénynek nevezzük, mivel azt a törekvést fejezi ki, hogy a kalibrált w_j súlyok az eredeti w_j^o súlyokhoz lehetőleg közel legyenek Az egyenletrendszerre a következőképpen hivatkozunk mátrix-vektor írásmóddal:

$\mathbf{X}w = X$, ahol

$\boldsymbol{w} = (w_1, w_2, ..., w_n)^T$ és $X = (X_1, X_2, ..., X_m)^T$;

a félkövér **X** tehát itt *m*×*n*-es mátrix, a kurzív *X* pedig *m*-dimenziós oszlopvektor, a felső *T* a transzponálás jele. Az eddig elmondottak alapján világos, hogy az $\mathbf{X}w = X$ egyenletrendszer azt fejezi ki, hogy a kalibrálás eredményeként bármely segédváltozónak a mintából becsült értékösszege megegyezik a megfelelő sokaságbeli értékösszeggel. Itt jegyzem meg, hogy a kalibrálás filozófiája szerint mintavételes eljárásunk célváltozóira, tehát azokra, amelyekre a felvétel irányul, a közvetlen egyszerű becslésnél jobb becsléseket várhatunk, hacsak szoros statisztikai kapcsolat áll fenn a célváltozók és a segédváltozók között. A célfüggvényről, illetve most már távolságfüggvényről fel szokták tenni, hogy szigorúan konvex, folytonosan differenciálható, továbbá nem negatív és csak akkor nulla, ha $w_j = w_j^{\mathrm{o}}$ a minta bármely elemére.

Igen sokféle kalibrálási feladatot lehet felírni a távolságfüggvény választásának függvényében, meg aszerint, hogy a változókra, tehát a kalibrált súlyokra vonatkozóan kirovunk-e individuális korlátokat vagy sem. Mondanivalóm szempontjából elegendő lesz négyféle feladat-típusra szorítkozni a következők szerint

A *G* függvényre csak a következő két lehetőséget vesszük figyelembe:

Kvadratikus távolságfüggvény: $G(w, w^{\mathrm{o}}) = \tfrac{1}{2}(w - w^{\mathrm{o}})^2 / w^{\mathrm{o}}$

Információ-divergencia: $G(w, w^{\mathrm{o}}) = w \log \frac{w}{w^{\mathrm{o}}} - w + w^{\mathrm{o}}$

és ezek valamint a kalibrálási feltételek mellett vagy megköveteljük individuális korlátok meglétét is, vagy nem. Az így meghatározott feladat-típusokra a római I., II, III. és IV. sorszámokkal hivatkozom. Ez a négy feladat-típus egyébként lefedi a kalibrálás jelenlegi gyakorlatának jelentős részét.

I. kvadratikus távolságfüggvény, a változókra vonatkozó individuális korlátok nélkül
II. információ-divergencia, a változókra vonatkozó individuális korlátok nélkül
III. kvadratikus távolságfüggvény, a változókra vonatkozóan individuális korlátokkal
IV. információ-divergencia, a változókra vonatkozóan individuális korlátokkal.

Az I. típus esetén a feltételes szélsőérték-számítás klasszikus eljárása zárt alakban megadható megoldáshoz vezet. A Lagrange-függvény:

$$\Psi = \tfrac{1}{2} \sum_{j=1}^{n} (w_j - w_j^{\mathrm{o}})^2 / w_j^{\mathrm{o}} - \lambda^{\mathrm{T}} (\mathbf{X}\mathbf{w} - X)$$

$$\lambda = (\lambda_1, \ldots, \lambda_m)^T$$

w_j szerinti deriválásával, a deriváltak nullával egyenlővé tételével a kalibrált súlyokra a

$$\mathbf{w} = \mathbf{w}^{\mathrm{o}} + \Omega \mathbf{X}^T (\mathbf{X}\Omega\mathbf{X}^T)^{-1} (X - \hat{X})$$

kifejezés adódik, ahol Ω diagonális mátrix, átlójában az eredeti súlyokkal. A kifejezés a regressziószámítás összefüggéseire emlékeztet, nem véletlenül, ugyanis ezekkel a súlyokkal egy *y* célváltozó becsült értékösszege

$\hat{Y}^{\mathrm{kal}} = \hat{Y} + \mathbf{y}^T \Omega \mathbf{X}^T (\mathbf{X}\Omega\mathbf{X}^T)^{-1} (X - \hat{X})$, illetve

$$\hat{Y}^{\text{kal}} = \hat{Y} + \sum_{i=1}^{m} b_i (X_i - \hat{X}_i)$$

amelyet általánosított regressziós becslésnek nevezünk. $\hat{X} = (\hat{X}, ..., \hat{X}_m)^T$ a segédváltozók értékösszegének mintából számított közvetlen egyszerű becslésére a

$$\lambda = (\mathbf{X}\Omega\mathbf{X}^T)^{-1}(\mathbf{X}w - \mathbf{X}w^{\text{o}}) \; = \; (\mathbf{X}\Omega\mathbf{X}^T)^{-1}(X - \hat{X})$$

kifejezés adódik, ahol Ω a $w_1^{\text{o}},...,w_n^{\text{o}}$ diagonális elemekkel meghatározott diagonális mátrix:

$$\Omega = \begin{pmatrix} w_1^{\text{o}} & 0 & ... & 0 \\ 0 & w_2^{\text{o}} & & 0 \\ \vdots & & & \\ 0 & 0 & & w_n^{\text{o}} \end{pmatrix}$$

$b_1, b_2,..., b_m$ regressziós együtthatók. A w_j kalibrált súlyok között negatívok is lehetnek, illetve néhány súly irreálisan magas értéket is felvehet.

A klasszikus eljárás a II. feladat-típusnál is alkalmazható, itt azonban egy másik megoldás is kínálkozik, amelyet iteratív arányos közelítések módszerének vagy általánosított iteratív skálázásnak neveznek, angol neve raking ratio estimation vagy egyszerűen raking. Induló értékként a design súlyokat választva, az eljárás a következő két lépés ismétléséből áll:

(1) a w_j súlyok aktuális értéke mellett válasszunk olyan $r_1, r_2, ..., r_m$ szorzókat, hogy $r_i \sum_{j=1}^{n} x_{ij} w_j = X_i$ teljesüljön, i = 1, 2, …, m;

(2) j = 1, 2, …, n esetén legyen w_j új értéke $w'_j = w_j \prod_{i=1}^{m} r_i^{x_{ij} / x_{.j}}$, ahol $x_{.j} = \sum_{i=1}^{m} x_{ij}$,

tehát w_j szorzója $r_1, r_2, ..., r_m$ súlyozott mértani közepe, és a súlyok minden iterációban az $\mathbf{X} = (x_{ij})$ mátrix j-edik oszlopának elemeitől függ.

Amennyiben a kalibrálási feltételek konzisztensek, egyebek mellett a segédváltozók nem negatívok és az X_i értékösszegek pozitívok, akkor a w_j súlyok sorozata a kalibrálási feltételeknek egy megoldásához konvergál, és ez a megoldás minimalizálja a távolság-függvényt. Ezen túlmenően, a kapott súlyok nem negatívok, azonban 0-hoz közeli értékek valamint igen magas értékek előfordulhatnak közöttük.

A Központi Statisztikai Hivatalban ezt az eljárást használják kalibrálás céljára, pontosabban ennek egy egyszerűsített változatát, amelynél a (2) lépésben a mértani átlagok szerepét számtani átlagok veszik át. Az eljárás konvergenciája megmarad, a távolságfüggvény minimalizálása szempontjából azonban csak közelítő megoldáshoz jutunk.

Ellentétben az I-II. feladat-típussal, a III. és a IV: feladat-típusnál kézenfekvőnek látszik a konvex programozás alkalmazása. A jelenlegi gyakorlatban ezzel szemben a Newton módszer a legelterjedtebb eljárás, a következőképpen alkalmazzák. A

$$\Psi = \sum_{j=1}^{n} G(w_j, w_j^{o}) - \lambda^T (\mathbf{X}\mathbf{w} - X)$$

Lagrange-függvényt w_j szerint deriválva, a deriváltat nullával egyenlővé téve

$$\partial G / \partial w_j = \mathbf{X}_{.j}^T \lambda$$

adódik, ahol $\mathbf{X}_{.j}$ az **X** mátrix *j*-edik oszlopa, ebből pedig a

$$w_j = h(\mathbf{X}_{.j}^T \lambda),$$

összefüggéshez jutunk, ahol *h* a $\partial G / \partial w_j$ függvény inverze. Mivel *G* szigorúan konvex, $\partial G / \partial w_j$ monoton növekvő, tehát *h* létezik. A változókra kirótt individuális korlátok miatt a *G* függvényt most úgy kell módosítani, hogy amennyiben eszerint az összefüggés szerint w_j kilépne a megengedett tartományból, akkor *h* értéke legyen a megfelelő alsó- vagy felső korlát. Belátható ezután, hogy a Newton módszerrel az

$$X - \hat{X} - \phi(\lambda) = 0$$

egyenletet kell megoldani, ahol $\phi(\lambda) = \sum_{j=1}^{n} h(\mathbf{X}_{.j}^T \lambda)\mathbf{X}_{.j}$ - $\hat{X}$.

Egyes szerzők – pl. Wayne Fuller – véleménye szerint ez az eljárás kvadratikus célfüggvény esetén ekvivalens a kvadratikus programozással, bizonyítással vagy arra való hivatkozással azonban nem találkoztam.

Érdemes megjegyezni, hogy a IV. feladat-típusnál, amely a II. feladat-típus megfelelője, a távolság-függvény szellemes módosításával sikerült elérni, hogy a *h*(.) függvény az argumentum minden értékénél a w_j -re vonatkozó korlátok között marad. A módosított távolságfüggvény a következő:

$$G(v) = \frac{1}{A}\left[(v-L)\log\frac{v-L}{1-L} + (U-v)\log\frac{U-v}{U-1}\right],$$

ahol $A = (U-L)/((1-L)(U-1))$ és $v = w/w^{o}$. Kimutatható, hogy *L*=0 és *U*→∞ $G(v)$ a $\frac{w}{w^{o}}\log\frac{w}{w^{o}} - \frac{w}{w^{o}} + 1$ függvénybe megy át, ami az információ-divergencia függvény $1/w^{o}$ -szorosa.

Mint említettem, a Központi Statisztikai Hivatalban a Darroch-Ratcliff-féle iteratív arányos közelítések módszerét, illetve ennek egy módosítását használják kalibrálási feladatok megoldására. Ezt is könnyű tovább módosítani úgy, hogy individuális korlátokat is kezeljen, éspedig úgy, hogy amennyiben a mintasúlynak a

$$w'_j = w_j \prod_{i=1}^{m} r_i^{x_{ij} / x_{.j}}$$

képlettel meghatározott új értéke elhagyná a megengedett tartományt, az új értéket a megfelelő alsó- vagy felső korláttal vesszük egyenlőnek Tapasztalat szerint az így módosított eljárás jól működik, az eredménynek a távolságfüggvény minimumával való kapcsolatáról azonban semmit sem állíthatunk.

Előadásom befejezéséül a következő négy eljárást hasonlítom össze egymással és részben a konvex programozással, a sorszámok a következő kis táblázatok fejrovatát is azonosítják:

1. Iteratív arányos közelítések módszere individuális korlátok kezelésének megfelelős módosítással, a változók új értékénél mértani közepek alkalmazásával;
2. Ugyanez, a változók új értékénél számtani közepek alkalmazásával;
3. Newton módszer, kvadratikus távolságfüggvény esetén (III. típusú feladat);
4. Newton módszer, információ divergencia távolságfüggvény esetén (IV. típusú feladat).

Az összehasonlítás szempontjai a következők:

– az egyes módszerekkel meghatározott megoldások – tehát kalibrált súlyok – esetén a távolságfüggvény értéke hogyan viszonyul a megfelelő konvex programozási eljárással elért minimumhoz;
– hogyan befolyásolja a módszer megválasztása a felvétel néhány célváltozójának értékét;
– milyen az egyes módszerek gépidő-igénye.

A tekintett módszerek mindegyikét fel lehet fogni, mint egy konvex programozási feladatnak megfelelő közelítő eljárást, ahol a megfeleltetést a távolságfüggvény-célfüggvény azonossága biztosítja. Az egyes módszerek a megfelelő konvex programozási feladat szempontjából kezdő megengedett megoldást szolgáltatnak, tehát azt lehet vizsgálni, hogy a konvex programozási eljárás ezt hogyan tudja javítani a távolságfüggvény szempontjából.

Az 1. és 2. módszer esetén saját fejlesztésű programot használtam, ez a SAS programrendszer IML, azaz, Interactive Matrix Language modulja alapján készült. A 3. és a 4.módszernél a Francia Statisztikai Hivatal, az INSEE által kifejlesztett CALMAR nevű programot alkalmaztam, ez egyébként az I-IV. feladat-típusok kezelésére készült.

Konvex programozási eljárásként a Wolfe-féle redukált gradiens eljárást választottam. Mivel a rendelkezésemre álló hardware eszközökön nincs FORTRAN fordító program, végeredményben ezt az eljárást is saját fejlesztésű SAS/IML programként használtam. A lebonyolított számítások azonban azt is jelezték, hogy ez a program is megfelelően működik.

A számításokat a KSH munkaerő-felmérésének 2006. februári és 2009. januári mintáján végeztem, melyek tíz-tízezer háztartás adatait tartalmazták. Mivel a munkaerő-felmérésben a kalibrálást megyei szinten végzik, összesen 40 kalibrálási feladatot kellett megismételni a

négy különböző módszerrel, majd a kapott eredményt meg kellett fejelni a redukált gradiens módszer alkalmazásával. Az individuális korlátokat nem számítva, minden egyes feladathoz 22 kalibrálási feltétel tartozott, éspedig 2x10 korcsoportos létszámadat, továbbá a háztartások valamint a megyei jogú városokban lakók száma a megyében.

Noha még nagyobb mennyiségű számítás elvégzése nyilván még több információt szolgáltatott volna, le kell szögezni, hogy bizonyos következtetések az említett két időszak adatainak alapján is levonhatók. A tekintett kalibrálási feladat struktúrája miatt ugyanis elég nagy bizonyossággal kijelenthető, hogy nagyjából ugyanazt a képet kaptuk volna, ha pl. 2005 januárjától 2009 márciusáig az 51 hónap munkaerő-felmérésének mintáját hasonló módon dolgoztuk volna fel, mint ahogy azt tettük a 2006. februári és 2009. januári felvétel mintájának esetében.

Tekintsük mármost a számítás eredményeit. Az 1. táblázat 1. oszlopában a nagyságrend érzékeltetésére a távolságfüggvénynek a 2. módszerhez tartozó értékét tüntettük fel megyénként, majd a 2., 3., 4. és 5. oszlopban azt az információt közöltük, hogy az 1., 2., 3., illetve 4. módszer eredményét milyen mértékben sikerült javítani a távolságfüggvény értelmében a redukált gradiens módszerrel. A 2. oszlopban egy 111,5591 adat például azt jelenti, hogy az 1. módszer eredménye 11.5591 százalékkal rosszabb a redukált gradiens módszerrel meghatározott minimumnál. Az elvégzett számítások szerint az 1. és a 2. módszer esetében ez a minimumtól való eltérés 8-9 százalékos, és a két módszer eredménye meglepően közel van egymáshoz. A 3. és 4. módszer, tehát a Newton módszer esetén viszont a minimumtól való eltérés jelentéktelen, csupán egy-két esetben éri el a 2-3 százalékot, és valószínűsíti azt a sejtést, hogy ezt a majdnem pontos megegyezést elméletileg is alá lehet támasztani.

A 2. táblázat a foglalkoztatott- és a munkanélküli létszám valamint a munkanélküliségi ráta adatait tartalmazza. Itt is a KSH gyakorlatának megfelelő 2. módszer az alap, itt az adatok abszolút értékét találjuk, míg az azt követő négy oszlopban az 1-4. módszernek a redukált gradiens módszerrel javított eredményei találhatók, pontosabban azoknak az 1. oszlop adataitól való relatív eltérései százalékban. Az eltérések átlagosan fél százalék alatt vannak a foglalkoztatott létszám esetén, míg a munkanélküli létszámnál egy-két esetben még az 5 százalékot is megközelítik. Az eltérések azonban minden esetben a mintavételi hiba határain belül vannak, összhangban Deville és Särndal egy híres eredményével (1992), miszerint bármely két kalibrálási módszer aszimptotikusan ekvivalens, a távolságfüggvény választása nem befolyásolja lényegesen a becsléseket. Az eltérések különösen kicsik a ráták esetében, itt az abszolút eltérések sehol sem haladják meg a fél százalékpontot, s többnyire 1/10 százalékpont alatt vannak.

Végül, ami a módszerek sebességét illeti, a Newton módszer a CALMAR program realizálásában rendkívül gyors, kb. kétszer olyan gyors, mint az iteratív arányos közelítések módszere. A redukált gradiens módszer viszont, amelyet itt csak a 2. fázisban használtam, nagyságrenddel lassúbb az előzőeknél. Természetesen itt nem kereskedelmi szoftverről van szó, hanem „házilagos” készítésű programról, de a módszerek viszonylagos gyorsaságára vonatkozó megállapítás valószínűleg akkor is érvényes, ha a redukált gradiens módszert egy „áramvonalas” program segítségével alkalmazzuk.

Összegezve, a tekintett módszerek – iteratív arányos közelítések, Newton módszer – alkalmazása kalibrálási feladatokban a végzett számítások tükrében helyeselhető, tekintettel

arra, hogy itt a távolságfüggvény valódi minimumának elérése vagy el nem érése másodlagos jelentőségű.

1. táblázat

Távolság-függvény	Távolság-fgv. értéke a minimum %-ában			
	1.	2.	3.	4.
55910	106.9	106.8	100.0	100.0
17143	108.1	108.3	100.3	102.6
9871	111.6	111.9	100.0	101.1
5160	106.2	106.2	100.0	100.5
9537	109.9	110.0	100.0	100.9
7613	115.3	115.3	100.0	102.3
6859	106.3	105.9	100.0	100.0
4914	110.3	110.9	100.0	100.0
3803	108.9	108.9	100.0	100.0
2735	110.2	109.9	100.0	100.1
4756	109.4	109.5	100.0	100.0
8034	108.5	108.5	100.0	100.1
7921	112.8	113.0	100.2	100.5

2a. táblázat

M	Foglalkoztatott					Munkanélküli				
	Létszám	1.	2.	3.	4.	Létszám	1.	2.	3.	4.
01	755618	0.00	-0.21	-0.22	-0.33	32183	0.02	0.57	1.00	0.73
02	145528	-0.07	-0.81	-0.41	0.05	14335	-0.03	-1.90	-1.60	-1.37
03	189137	0.02	-0.33	-0.33	-0.44	16732	-0.15	0.04	0.01	0.99
04	128268	0.03	-0.61	-0.66	-0.52	14398	0.01	3.41	3.13	2.50
17	91181	-0.00	-0.17	-0.15	-0.11	8641	0.03	0.50	0.90	0.04
18	114347	-0.01	0.03	0.02	0.04	9903	0.14	-0.24	-0.54	-1.20
20	131226	0.01	0.42	0.45	0.36	10420	-0.06	-2.12	-1.21	1.09
T	3865062	0.00	0.01	-0.06	-0.05	331249	-0.04	-0.44	0.52	0.27

2b. táblázat

M	Ráták, %						Rel. eltérések, %			
		1.	2.	3.	4.		1.	2.	3.	4.
01	4.09	4.09	4.12	4.13	4.13	.	0.02	0.75	1.17	1.02
09	9.16	9.15	9.26	9.29	9.23	.	-0.09	1.13	1.36	0.78
10	9.35	9.33	9.46	9.49	9.59	.	-0.25	1.14	1.47	2.48
11	10.81	10.81	10.40	10.41	10.34	.	0.06	-3.75	-3.64	-4.27
13	5.50	5.50	5.54	6.10	5.90	.	-0.01	0.86	10.88	7.27
14	9.43	9.40	9.42	9.41	9.52	.	-0.25	-0.13	-0.20	0.93
15	14.30	14.31	14.35	14.35	14.36	.	0.01	0.35	0.29	0.42
20	7.36	7.35	7.18	7.24	7.41	.	-0.07	-2.35	-1.54	0.67
T	7.89	7.89	7.86	7.94	7.92	.	-0.04	-0.42	0.53	0.30