Determinisztikus idősorelemzés, idősorok szűrése

Ferenci Tamás
tamas.ferenci@medstat.hu

Tartalom

Tartalomjegyzék

1. Determinisztikus idősorelemzés **1**
1.1. Alapgondolat 1
1.2. Determinisztikus idősormodellezés regresszióval 1
1.3. Trend és szezonalitás 5

2. Idősorok szűrése **6**

1. Determinisztikus idősorelemzés

1.1. Alapgondolat

A determinisztikus idősorelemzés

- Az idősor alakulása *elvileg* függvényszerűen felírható bizonyos tényezők alapján
- Csak azért nem tudjuk tökéletesen megtenni, mert nem ismerjük e tényezőket, nem tudjuk milyen függvényformával hatnak, nem tudjuk pontosan mérni stb. ezért fogunk hibázni
- De pont: a hibának *csak* ennyi szerepe van...
- ...beállítja az aktuális időszaki értéket, *és kész*

Dekompozíciós idősormodellek

- Minderre a legtipikusabb – és egyben legklasszikusabb – példát a **dekompozíciós idősormodellek** jelentik
- A legismertebb additív modell:

$$Y_t = R_t + C_t + S_t + u_t,$$

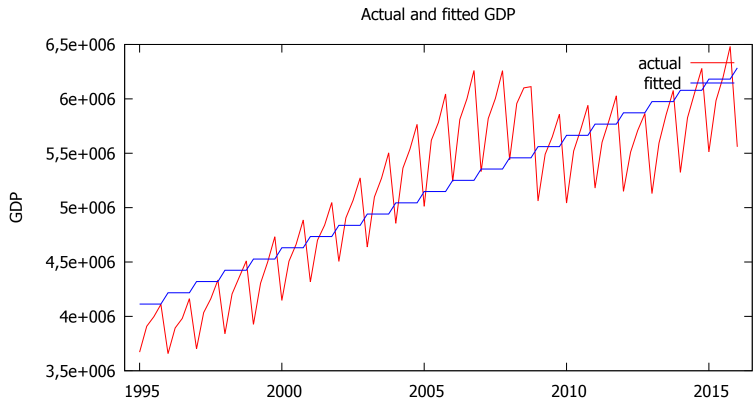
  ahol R_t, C_t és S_t a trend, a ciklus és a szezonalitás t-edik időszakbeli értéke rendre, u_t pedig a már említett eltérésváltozó
- Becslés?

1.2. Determinisztikus idősormodellezés regresszióval

Regresszió alkalmazása

- Az előbbi modell teljesen természetesen becsülhető regresszióval, ha R_t, C_t és S_t helyébe beírjuk a feltételezett – paraméteres – függvényformákat
- (Most tehát mindvégig paraméteres regressziót fogunk használni)
- Legegyszerűbb eset: $R_t = \alpha + \beta t$, $C_t = 0$ és $S_t = 0$ (egyszerű lineáris trend)
- Az így kapott modell OLS-sel becsülhető

Negyedéves GDP (éves) lineáris trenddel I.



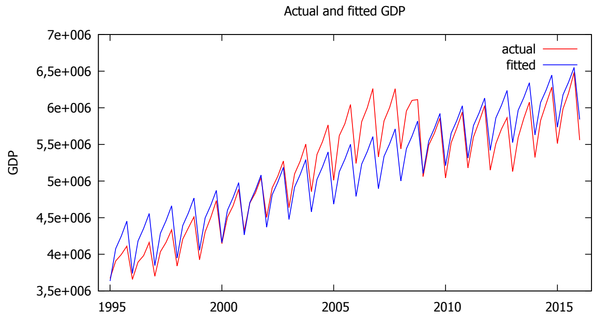
	Coefficient	Std. Error	t-ratio	p-value
const	−2,02165e+008	1,50629e+007	−13,4214	0,0000
EV	103398,	7512,17	13,7641	0,0000

Mean dependent var	5161052	S.D. dependent var	765270,3
Sum squared resid	1,50e+13	S.E. of regression	424924,3
R^2	0,695356	Adjusted R^2	0,691686
$F(1, 83)$	189,4494	P-value(F)	3,94e−23
Log-likelihood	−1221,169	Akaike criterion	2446,339
Schwarz criterion	2451,224	Hannan–Quinn	2448,304
$\hat{\rho}$	0,315141	Durbin–Watson	1,343686

Negyedéves GDP (éves) lineáris trenddel II.

Mi ezzel a baj? Hibatag jól specifikált? Aligha!

Negyedéves GDP (éves) lineáris trenddel és szezonalitással I.

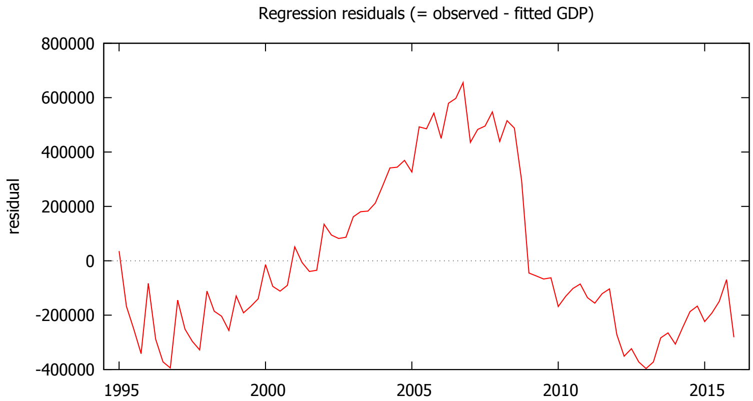


	Coefficient	Std. Error	t-ratio	p-value
const	−2,04994e+008	1,06300e+007	−19,2845	0,0000
EV	104985,	5301,64	19,8024	0,0000
DNEGYEDEV_1	−815807,	91469,3	−8,9189	0,0000
DNEGYEDEV_2	−375072,	92487,9	−4,0554	0,0001
DNEGYEDEV_3	−203380,	92487,9	−2,1990	0,0308

Mean dependent var	5161052	S.D. dependent var	765270,3
Sum squared resid	7,19e+12	S.E. of regression	299695,1
R^2	0,853937	Adjusted R^2	0,846634
$F(4, 80)$	116,9271	P-value(F)	1,34e−32
Log-likelihood	−1189,928	Akaike criterion	2389,855
Schwarz criterion	2402,068	Hannan–Quinn	2394,768
$\hat{\rho}$	0,946516	Durbin–Watson	0,116617

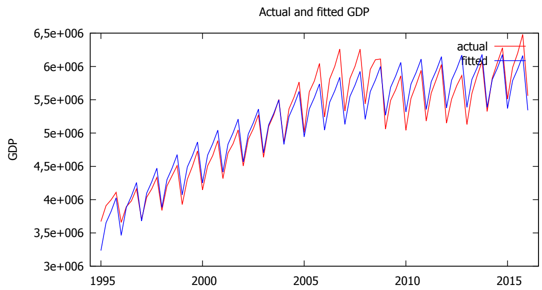
Negyedéves GDP (éves) lineáris trenddel és szezonalitással II.

A szezonalitás jónak tűnik, de az alaptrendet még mindig nem sikerült megragadni:



A szezonalitás azért tűnik jónak, mert nincs interakció az év és a szezon között, azaz minden évben hasonló a szezonalitás mintázata.

Negyedéves GDP (éves) kvadratikus trenddel és szezonalitással I.

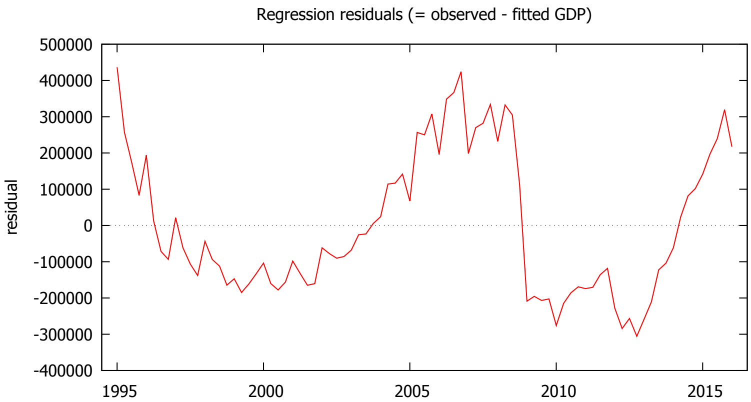


	Coefficient	Std. Error	t-ratio	p-value
const	−2,60273e+010	2,61365e+009	−9,9582	0,0000
EV	2,58613e+007	2,60697e+006	9,9201	0,0000
DNEGYEDEV_1	−792077,	61608,7	−12,8566	0,0000
DNEGYEDEV_2	−375072,	62247,4	−6,0255	0,0000
DNEGYEDEV_3	−203380,	62247,4	−3,2673	0,0016
sq_EV	−6422,59	650,072	−9,8798	0,0000

Mean dependent var	5161052	S.D. dependent var	765270,3
Sum squared resid	3,21e+12	S.E. of regression	201704,8
R^2	0,934664	Adjusted R^2	0,930529
$F(5, 79)$	226,0280	P-value(F)	2,80e−45
Log-likelihood	−1155,736	Akaike criterion	2323,473
Schwarz criterion	2338,128	Hannan–Quinn	2329,368
$\hat{\rho}$	0,889365	Durbin–Watson	0,173391

Negyedéves GDP (éves) kvadratikus trenddel és szezonalitással II.

Reziduumok kicsit jobbak:



Mindezek limitációi

- *Egyrészt* el kell találni a függvényformát
- Persze modelldiagnosztika (az előbb látott grafikus módszerek és tesztek is) ott van
- (Ez igazából már keresztmetszetnél is így volt)
- Pl. a kvadratikus nyilván csak erre az időszakra jó, az általánosítóképessége botrányos lenne
- *Másrészt* a hibatag diagnosztikája bonyolultabbá válik, egy új szempont is megjelenik (autokorreláció) → később még nagyon sokat fogunk róla beszélni

1.3. Trend és szezonalitás

A trend megadása

- **Trend**: „hosszú távú alapirányzat”
- A mostani trend (determinisztikus trend) bármi lehet, amit paraméteres függvényformában megadunk; például:
  - Lineáris trend: $a + bt$
  - Kvadratikus trend: $a + b_1 t + b_2 t^2$
  - Polinomiális trend: $a + b_1 t + b_2 t^2 + \ldots + b_k t^k$
  - Exponenciális trend: ae^{bt}
  - Aszimptotikus trend: $c + \frac{1}{a+bt}$
  - Logisztikus trend: $\frac{1}{c+e^{a+bt}}$
  - stb. stb.
- (Persze amelyik nem lineáris, ott vagy linearizálni kell vagy – ha ez nem lehetséges – akkor nem OLS-sel becsülni)
- Ezek mind paraméteres trendek voltak, elképzelhető nem-paraméteres trend is, a legismertebb a spline-ok használata (de ne feledjük, annak a becslése kevésbé hatásos, nem kapunk egyetlen vagy néhány számba sűrített – és jó esetben tárgyterületileg értelmezhető – eredményt, valamint az előrejelzés is problémásabb)

Szezonalitás megadása

- **Szezonalitás**: „éven belüli mintázat”, exogén módon rögzített hosszúságú, periodikus (vs. **ciklus**: „éven túli”, nem feltétlenül exogén módon adott, ismert hosszúságú)
- A szezonalitásnál viszont tipikusabb a nem-paraméteres megadás: minden negyedévnek (hónapnak, félévnek stb.) saját paramétere van
- (Dummy-kkal, ld. később, regressziós keretbe szintén szépen illeszkednek!)
- Persze itt is elképzelhető paraméteres megadás, a legismertebb a trigonometrikus (harmonikus) függvények használata

Dummy-kódolás szezonalitáshoz: referenciakódolás

- Az egyik szezon indikátorát elhagyjuk: **referenciakódolás**

	D_{Q1}	D_{Q2}	D_{Q3}
Q1	1	0	0
Q2	0	1	0
Q3	0	0	1
Q4	0	0	0

- Értelmezés: eltérés a referenciacsoporthoz képest (ami az elhagyott indikátorú csoport)

Dummy-kódolás szezonalitáshoz: kontrasztkódolás I.

- Egy másik népszerű megoldás a **kontrasztkódolás**: viszonyítsunk az *átlaghoz*!
- Ehhez hogyan kell kódolni...?

	C_{Q1}	C_{Q2}	C_{Q3}
Q1	1	0	0
Q2	0	1	0
Q3	0	0	1
Q4	-1	-1	-1

Dummy-kódolás szezonalitáshoz: kontrasztkódolás II.
Mert:

$$\alpha + \beta_{C_{Q1}} + 0 + 0 = \overline{y}_{Q1} \quad (1)$$
$$\alpha + 0 + \beta_{C_{Q2}} + 0 = \overline{y}_{Q2} \quad (2)$$
$$\alpha + 0 + 0 + \beta_{C_{Q3}} = \overline{y}_{Q3} \quad (3)$$
$$\alpha - \beta_{C_{Q1}} - \beta_{C_{Q2}} - \beta_{C_{Q3}} = \overline{y}_{Q4} \quad (4)$$

És így:

- (1)+(2)+(3)+(4) $\Rightarrow 4\alpha = \overline{y}_{Q1} + \overline{y}_{Q2} + \overline{y}_{Q3} + \overline{y}_{Q4} \Rightarrow \alpha$ tényleg a főátlag (mert azonosak voltak a csoportok elemszámai, különben ún. súlyozott kontraszt kellene)
- (2)+(3)+(4) $\Rightarrow 3\alpha - \beta_{C_{Q1}} = \overline{y}_{Q2} + \overline{y}_{Q3} + \overline{y}_{Q4} \Rightarrow \beta_{C_{Q1}} = 3\alpha - \left(\overline{y}_{Q2} + \overline{y}_{Q3} + \overline{y}_{Q4}\right) = 3\alpha - \left(4\alpha - \overline{y}_{Q1}\right) \Rightarrow \beta_{C_{Q1}} = \overline{y}_{Q1} - \alpha \Rightarrow$ tényleg az átlagtól való eltérés (és hasonlóan a másik kettő)

Dummy-kódolás szezonalitáshoz: egyebek

- Az angol irodalomban az általunk kontrasztkódolásnak nevezett módszert nagyon gyakran „effect coding”-nak nevezik...
- ... a kontraszt pedig az, amikor a csoportok tetszőleges – általunk meghatározott – lineáris kombinációját teszteljük

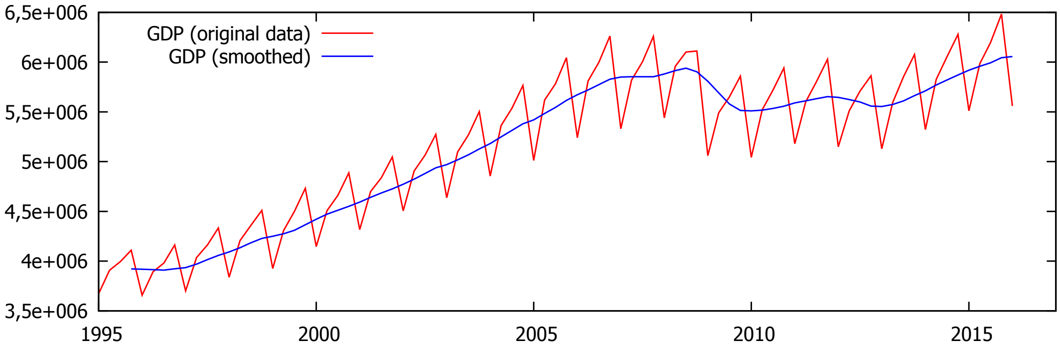
2. Idősorok szűrése

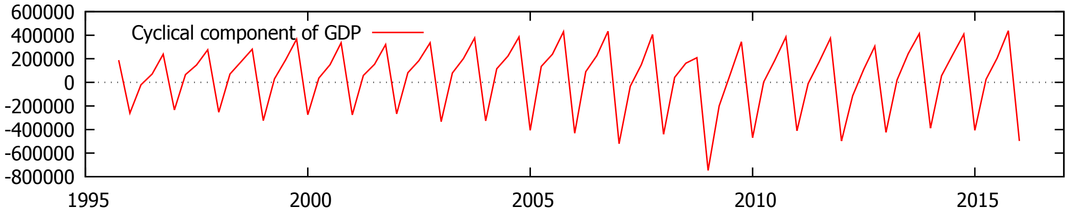
Célunk

- Szeretnénk elkülöníteni a trendet és a ciklikus komponenseket (szezon + ciklus); ez számos közgazdasági kérdésnél fontos feladat
- Egy lehetséges megoldás: „paraméteres szűrés”, azaz előírjuk a függvényformát és regresszióval megbecsüljük
- Igazából ezt megtettük az előbb is: megadtuk a trendet (lineáris vagy kvadratikus), megadtuk a szezonalitást (ezt nem-paraméteresen), és a látott reziduum a kiszűrt ciklus (+zaj) volt, ha van ilyen

- De ez függ a függvényforma-választás helyességén; nem lehetne ilyen feltevések nélkül is megtenni?
- Hogyne, például átlagoljunk ki évenként!
- Nem a legjobb, abrupt ugrások év végén, átlagoljunk inkább folyamatosan, mintegy csúszóablakot végigtolva (így is mindig négy különböző szezon lesz benne!)

Motiváló példa





Mozgóátlagolás

- Ez volt az (egyszerű) **mozgóátlag**:

$$y'_t = \frac{y_t + y_{t-1} + y_{t-2} + \ldots + y_{t-p}}{p+1}$$

- (Néha nem visszafelé átlagolnak, hanem az eredmény az átlagolt ablak közepén van („centrált” mozgóátlag), a dolognak nincs nagy jelentősége: ez talán kicsit jobban néz ki, viszont jövőbeli megfigyeléseket is igényel)
- Ez azt jelenti, hogy minden megfigyelés ugyanolyan súlyú; kézenfekvő gondolat, hogy a régebbiek kevésbé számítsanak
- Például:

$$y'_t = \frac{py_t + (p-1)\,y_{t-1} + (p-2)\,y_{t-2} + \ldots + y_{t-(p-1)}}{p + (p-1) + (p-2) + \ldots + 1}$$

- Ez az ún. súlyozott mozgóátlag

Exponenciális mozgóátlag

- Ökonometriában gyakoribb az exponenciális súlyozás:

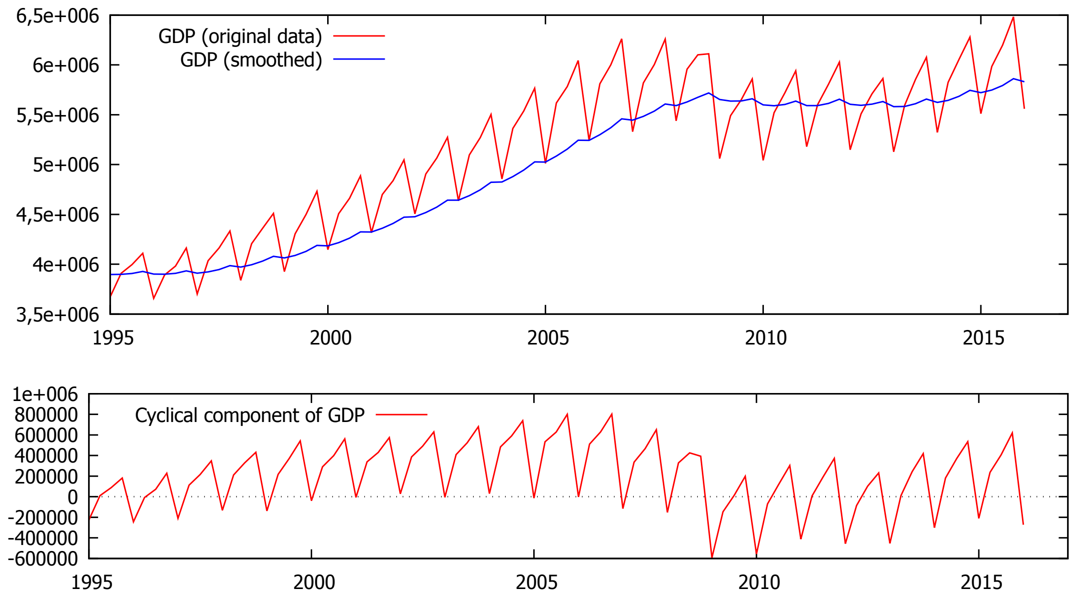
$$y'_t = \alpha y_t + (1-\alpha)\, y'_{t-1},$$

  hiszen ez nyilván annak felel meg, hogy

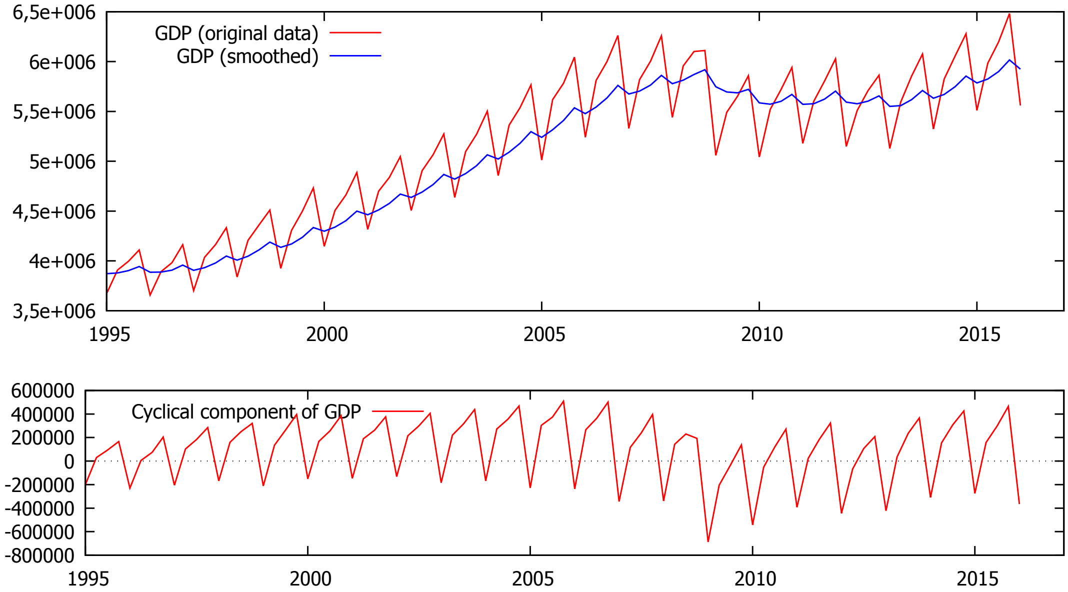
$$y'_t = \alpha y_t + (1-\alpha)\,\alpha y_{t-1} + (1-\alpha)^2\,\alpha y_{t-2} + \ldots + \\ + (1-\alpha)^{t-2}\,\alpha y_2 + (1-\alpha)^{t-1}\,\alpha y_1 + (1-\alpha)^t\,\alpha y_0$$

- (Az y_0 kezdőértéket nekünk kell megadni, a tipikus választások: y_1, az első néhány megfigyelés átlaga, az egész idősor átlaga; a következő példákban az első 4 megfigyelés átlaga került alkalmazásra)
- Tehát az ablak egyre nyúlik (végig az egész tartomány), és a súlyok exponenciálisan csengenek le

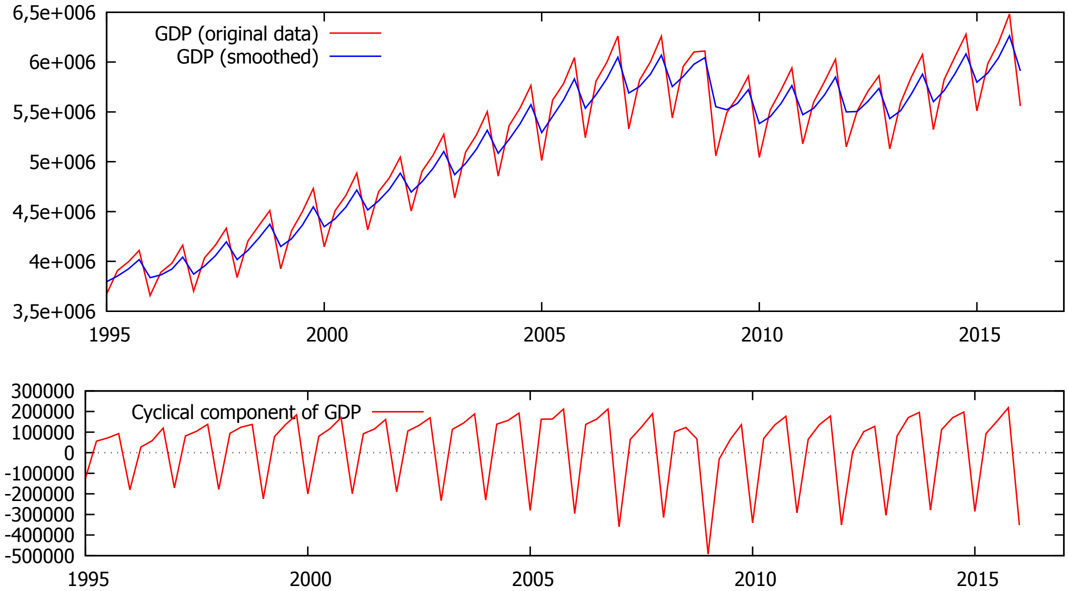
Negyedéves GDP exponenciális mozgóátlagolása, $\alpha = 0{,}1$



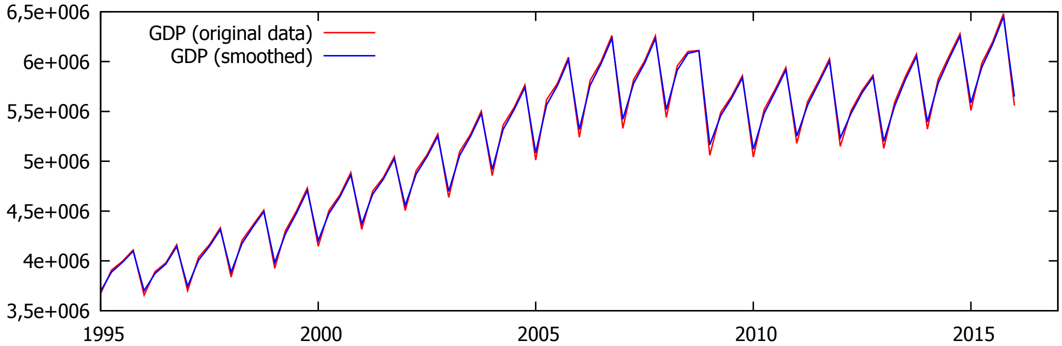
Negyedéves GDP exponenciális mozgóátlagolása, $\alpha = 0{,}2$

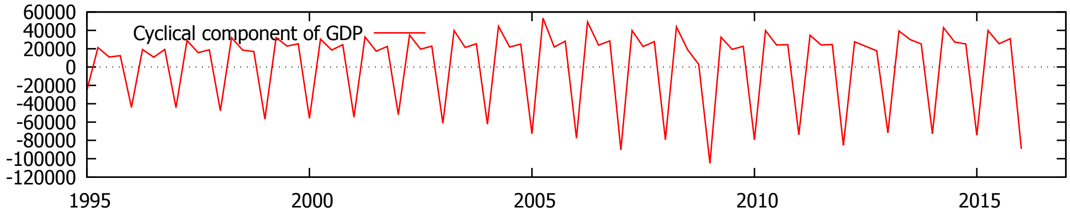


Negyedéves GDP exponenciális mozgóátlagolása, $\alpha = 0{,}5$



Negyedéves GDP exponenciális mozgóátlagolása, $\alpha = 0{,}9$





Lineáris szűrő

- Az egyszerű és a súlyozott mozgóátlag speciális esete annak, hogy

$$y'_t = a_0 y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} = \sum_{i=0}^{p} a_i y_{t-i}$$

- Például egyszerű mozgóátlagra $a_i = 1/(p+1)$
- Lényegében diszkrét konvolúció
- Ezt hívjuk lineáris szűrőnek, a tulajdonságait értelemszerűen teljes mértékben meghatározzák az $a_0, a_1, \ldots, a_p$ szűrőegyütthatók
- Roppant érdekes kérdés, hogy a szűrt idősor hogyan néz ki az eredetihez képest a szűrőegyütthatók függvényében, lehet otthon kísérletezgetni...

Hodrick–Prescott-szűrő

- Különösen makroökonómiában népszerű
- Alapgondolat: a trend (g_t) követi az idősort (azaz az $y_t - g_t$ kicsi), de nem nagyon ugrándozva (azaz g_t sima)
- A megoldandó feladat:

$$\min_{\{g_t\}_{t=1}^{T}} \sum_{t=1}^{T} (y_t - g_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(g_{t+1} - g_t) - (g_t - g_{t-1})\right]^2$$

- Első tag: mennyire követi jól az idősort a trend, második tag: mennyire „rángatózik” a trend

- A λ együttható határozza meg a két szempont egymáshoz viszonyított fontosságát ($\lambda = 0$: a trend nem kell, hogy sima legyen $\rightarrow$ pontosan az idősor lesz; $\lambda \rightarrow \infty$: trend legyen tökéletesen sima, nem érdekes, hogy mennyire követi az idősort $\rightarrow$ egyenes lesz)
- Negyedéves adatokra a tipikus választás $\lambda = 1600$

A második tagra azt írtuk, hogy „mennyire rángatózik a trend"; később majd pontosabban is fogjuk látni, hogy ez micsoda, mert van eg nagyon konkrét tartalma.

A HP-szűrő matematikája I.

- Az érdekes az, hogy a fenti minimalizációs feladatnak van zárt alakú megoldása!
- Legyen $\mathbf{y}$ az y_t-k, $\mathbf{g}$ a g_t-k vektora és legyen

$$\mathbf{Q}_{(T-2)\times T} = \begin{pmatrix} 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -2 & 1 & 0 \\ 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix},$$

ekkor a megoldandó feladat

$$\min_{\mathbf{g}} (\mathbf{y} - \mathbf{g})^T (\mathbf{y} - \mathbf{g}) + \lambda (\mathbf{Q}\mathbf{g})^T (\mathbf{Q}\mathbf{g})$$

A HP-szűrő matematikája II.

- Deriválva $\mathbf{g}$ szerint:

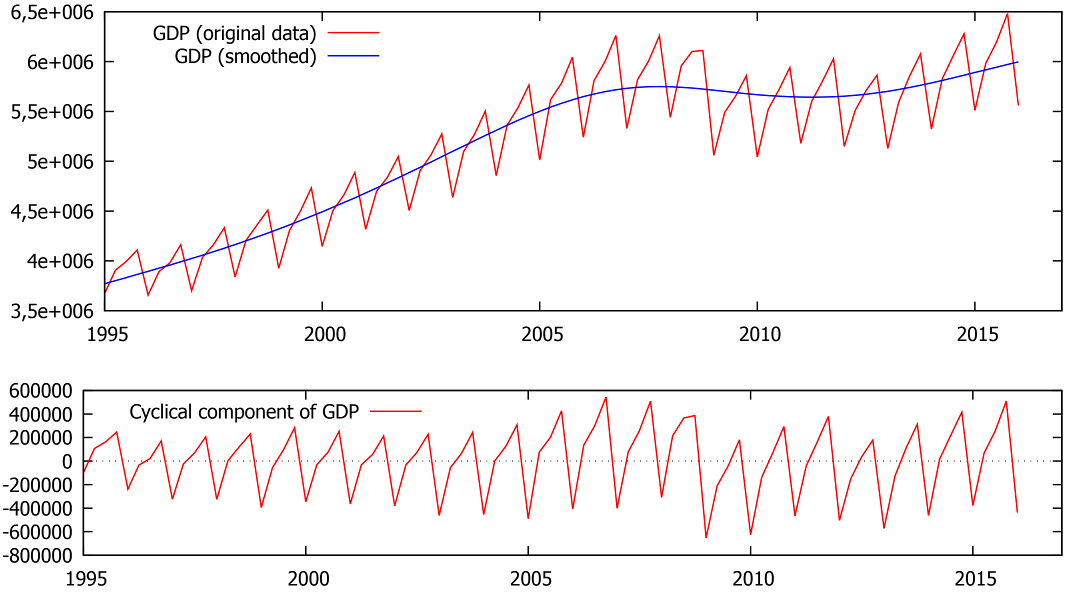
$$-2\mathbf{y} + 2\mathbf{g} + 2\lambda\mathbf{Q}^T\mathbf{Q}\mathbf{g}$$

- Egyenlővé téve nullával és megoldva (a másodrendű feltételek teljesülnek ahhoz, hogy ez tényleg minimum legyen):

$$-2\mathbf{y} + 2\mathbf{g} + 2\lambda\mathbf{Q}^T\mathbf{Q}\mathbf{g} = 0 \Rightarrow \widehat{\mathbf{g}} = \left(\mathbf{I} + \lambda\mathbf{Q}^T\mathbf{Q}\right)^{-1}\mathbf{y}$$

- (A megoldás mindegyik megfigyeléstől függ)

Negyedéves GDP HP-szűrése, $\lambda = 1600$



A HP-szűrő kritikája

Érdekes olvasmány: Hamilton, James D. „Why You Should Never Use the Hodrick-Prescott Filter." (2016). `http://econweb.ucsd.edu/~jhamilto/hp.pdf`.

Egyéb szűrők

- Több szűrő viselkedése alapvetően frekvencia-tartományon érthető meg
- Trend/ciklus szétválasztás: aluláteresztő szűrés (mert a trend az, ami lassan változik, a ciklus az, ami gyorsabban, persze kérdés, hogy hol a határ)
- Igazából már az egyszerű mozgóátlag is egy aluláteresztő szűrő volt!
- Vagy: LOESS-használata (STL-dekompozíció)