PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Jarosław Krajewski
Uniwersytet Mikołaja Kopernika w Toruniu

WPŁYW TRANSFORMACJI DANYCH NA WYNIKI ESTYMACJI DYNAMICZNEGO MODELU CZYNNIKOWEGO SZACOWANEGO METODĄ GŁÓWNYCH SKŁADOWYCH*

Streszczenie: Artykuł dotyczy dynamicznych modeli czynnikowych (DFM) − ich idei, zastosowania metody głównych składowych do ich estymacji oraz zastosowania kryteriów informacyjnych Baia i Ng do specyfikacji liczby czynników. Głównym celem artykułu jest ukazanie wpływu transformacji danych makroekonomicznych na ostateczną postać DFM. W badaniu wykorzystano 62 zmienne makroekonomiczne w postaci szeregów czasowych o częstotliwości miesięcznej z okresu od stycznia 2002 do marca 2009 roku. W wyniku szacowania DFM na różnych etapach transformacji danych uzyskano postacie DFM różniące się co do zawartych w nich zmiennych oraz co do wartości parametrów stojących przy poszczególnych zmiennych.

Słowa kluczowe: dynamiczny model czynnikowy, metoda głównych składowych, transformacja danych.

1. Wstęp

Dynamiczne modele czynnikowe (DFM – *Dynamic Factor Models*) stały się w ostatnim czasie bardzo popularne w analizach makroekonomicznych. Na ich popularność niewątpliwy wpływ mają banki centralne wielu krajów, które stymulują ich rozwój, widząc w tym szansę na odkrycie narzędzia dającego szybsze i bardziej trafne prognozy niż otrzymywane dotychczas ze standardowo stosowanych w tym celu narzędzi. Poza prognozowaniem DFM stosuje się do konstruowania głównych

* Praca naukowa współfinansowana ze środków Europejskiego Funduszu Społecznego i Budżetu Państwa w ramach Zintegrowanego Programu Operacyjnego Rozwoju Regionalnego, Działania 2.6 „Regionalne Strategie Innowacyjne i transfer wiedzy” projektu własnego Województwa Kujawsko-Pomorskiego „Stypendia dla doktorantów 2008/2009 – ZPORR”.

wskaźników koniunktury, analiz polityki monetarnej i badania międzynarodowych cykli koniunkturalnych.

Za pionierów w zakresie DFM uważa się Geweke'a [Geweke 1977] oraz Simsa i Sargenta [Sargent, Sims 1977], którzy zastosowali ten typ modeli do małych zbiorów danych. *Dynamic Factor Models* prezentują ateoretyczne podejście do modelowania ekonometrycznego [Sims 1980].

Głównym celem artykułu jest ukazanie wpływu transformacji danych makroekonomicznych na ostateczną postać empiryczną dynamicznego modelu czynnikowego oraz zweryfikowanie, czy standaryzacja zmiennych może pogarszać wyniki modelowania DFM i prognozowania na ich podstawie.

W artykule przybliżono rzadkie w polskiej literaturze zastosowanie metody głównych składowych do dynamicznej analizy makroekonomicznych szeregów czasowych.

2. Dynamiczny model czynnikowy

Koncepcja modeli czynnikowych opiera się na założeniu, że zachowanie się większości zmiennych makroekonomicznych może być dobrze opisane za pomocą małej liczby nieobserwowalnych wspólnych czynników. Czynniki te często są interpretowane jako wiodące siły w ekonomii. Poszczególne zmienne mogą wtedy zostać wyrażone jako liniowa kombinacja mniej niż 20 czynników, które wyjaśniają znaczącą część ich zmienności [Kotłowski 2008].

Niech y_t oznacza pewien szereg czasowy i $\mathbf{X_t}$ wyraża wektor N zmiennych w postaci szeregów czasowych zawierających informacje użyteczne w modelowaniu, a także prognozowaniu wartości y_t. W dynamicznym modelu czynnikowym zakładamy, że wszystkie zmienne x_{it} zawarte w wektorze $\mathbf{X_t}$ mogą zostać wyrażone jako liniowa kombinacja bieżących i opóźnionych nieobserwowalnych czynników f_{it}

$$x_{it} = \lambda_i(L)\mathbf{f_t} + e_{it}, \text{ dla } i = 1,...,N, \tag{1}$$

gdzie: $\mathbf{f_t} = [f_{1t}, f_{2t},..., f_{\bar{r}t}]'$ − wektor $\bar{r}$ nieobserwowalnych wspólnych czynników w momencie *t*;

$\lambda_i(L) = \lambda_{i0} + \lambda_{i1}L + \lambda_{i2}L^2 + ... + \lambda_{iq}L^q$ − operator opóźnień, natomiast e_{it} wyraża swoisty błąd zmiennej x_{it} [Stock, Watson 1998].

Stąd też y_t może być zapisane jako funkcja bieżących i opóźnionych wspólnych czynników zawartych w wektorze $\mathbf{f}_t$ oraz opóźnionych wartości y_t w następujący sposób:

$$y_t = \beta(L)\mathbf{f_t} + \gamma(L)y_t + e_t. \tag{2}$$

Zatem można powiedzieć, że dynamiczny model czynnikowy składa się z równań (1) i (2).

3. Estymacja parametrów DFM i specyfikacja liczby czynników

Jedną z najczęściej używanych metod estymacji parametrów i czynników w modelach czynnikowych jest metoda głównych składowych. W metodzie tej obie macierze, czynników i parametrów, są nieznane. Model przedstawiony jako równanie (1) może zostać zapisany w następującej formie macierzowej:

$$\mathbf{X} = \mathbf{FHH}^{-1}\mathbf{\Lambda}' + \mathbf{e}, \tag{3}$$

gdzie **H** to niejednostkowa macierz o wymiarach $r \times r$. Niezbędne jest wykonanie odpowiedniej normalizacji macierzy **H**. Stock i Watson [Stock, Watson 1998] zaproponowali warunek $(\mathbf{\Lambda}'\mathbf{\Lambda}/N) = \mathbf{I_r}$, który może zostać nałożony na parametry modelu i sprawi, że macierz **H** będzie ortonormalna.

Estymacja macierzy **F** i $\mathbf{\Lambda}$ przy użyciu metody głównych składowych polega na znalezieniu takich estymatorów macierzy $\hat{\mathbf{F}}$ i $\hat{\mathbf{\Lambda}}$, które będą minimalizować sumę kwadratów reszt równania (3) wyrażoną w następujący sposób:

$$V(\mathbf{F}, \mathbf{\Lambda}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \mathbf{\Lambda_i}'\mathbf{F_t})^2 \ . \tag{4}$$

W pierwszym kroku należy dokonać minimalizacji funkcji (4) w odniesieniu do macierzy czynników **F** przy założeniu, że macierz $\mathbf{\Lambda}$ jest znana i stała. W wyniku tego otrzymany zostanie estymator $\hat{\mathbf{F}}$ jako funkcja $\mathbf{\Lambda}$, który następnie zastępuje w powyższym równaniu prawdziwe wartości **F**. W drugim kroku minimalizowana jest funkcja (4) w odniesieniu do macierzy $\mathbf{\Lambda}$ z warunkiem normalizacji $(\mathbf{\Lambda}'\mathbf{\Lambda}/N) = \mathbf{I_r}$, w ten sposób uzyskuje się bezpośrednio estymator $\hat{\mathbf{\Lambda}}$. Warto zauważyć, że jest to równoznaczne z maksymalizacją wyrażenia $tr[\mathbf{\Lambda}'(\mathbf{X}'\mathbf{X})\mathbf{\Lambda}]$.

Kolejne kolumny macierzy $\hat{\mathbf{\Lambda}}$ są wektorami własnymi, macierzy **X’X** pomnożonej przez $\sqrt{N}$, odpowiadającymi największym wartościom własnym tej macierzy. Z kolei estymator macierzy **F** jest wyrażony jako

$$\hat{\mathbf{F}} = (\mathbf{X}\hat{\mathbf{\Lambda}})/N \ . \tag{5}$$

Stock i Watson podkreślają, że jeżeli liczba zmiennych jest wyższa od liczby obserwacji, tzn. $N > T$, wtedy z obliczeniowego punktu widzenia łatwiejsza do zastosowania jest procedura polegająca na oszacowaniu $\tilde{\mathbf{F}}$ przez minimalizację (4) z uwzględnieniem dla **F** warunku $\mathbf{F}'\mathbf{F}/T = \mathbf{I_r}$. Macierz $\tilde{\mathbf{F}}$ zawiera wówczas wektory własne z macierzy **X’X** odnoszące się do *r* największych wartości własnych z tej macierzy, przemnożonej przez $\sqrt{T}$. Z kolei estymator macierzy $\tilde{\mathbf{\Lambda}}$ przyjmie następującą formę:

$$\tilde{\mathbf{\Lambda}}' = (\tilde{\mathbf{F}}'\mathbf{X})/T\,. \tag{6}$$

Oba estymatory $\hat{\mathbf{F}}$ i $\tilde{\mathbf{F}}$ są równoważne.

W praktyce liczba czynników, niezbędna do pokazania związków pomiędzy zmiennymi, jest zazwyczaj nieznana. Istnieją jednak kryteria, których można użyć do wyznaczenia liczby czynników. Bai i Ng [2002] zaproponowali w tym celu następujące kryteria informacyjne:

$$IC_1(k) = \ln(\hat{V}(k)) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right), \tag{7}$$

$$IC_2(k) = \ln(\hat{V}(k)) + k\left(\frac{N+T}{NT}\right)\ln C_{NT}^2, \tag{8}$$

$$IC_2(k) = \ln(\hat{V}(k)) + k\left(\frac{\ln C_{NT}^2}{C_{NT}^2}\right). \tag{9}$$

We wzorach (7)-(9) $\hat{V}(k)$ oznacza sumę kwadratów reszt z *k*-czynnikowego modelu, a $C_{NT} = \min\left\{\sqrt{N}, \sqrt{T}\right\}$.

4. Transformacje danych i wyniki empiryczne

W badaniu zastosowanie znalazły miesięczne dane makroekonomiczne charakteryzujące polską gospodarkę. Zbiór danych zawiera 62 zmienne w postaci szeregów czasowych o częstotliwości miesięcznej. Dane dotyczą okresu od stycznia 2002 do marca 2009 roku, a więc każdy szereg składa się z 87 obserwacji. Jako zmienna objaśniana posłużyła inflacja, co ograniczyło pierwotny zbiór zmiennych objaśniających do 61. Wszystkie dane pochodzą z dwóch źródeł: Biuletynów Statystycznych Głównego Urzędu Statystycznego[1] i strony internetowej NBP (www.nbp.pl). Poddane zostały odpowiednim transformacjom. Po każdym etapie transformacji na podstawie otrzymanych danych wyznaczono czynniki i oszacowano najlepszy ze statystycznego punktu widzenia dynamiczny model czynnikowy. Zastosowane transformacje są analogiczne do stosowanych przez innych badaczy zarówno w klasycznej analizie czynnikowej, jak i w procedurze modelowania dynamicznych

[1] Wykorzystano Biuletyny Statystyczne z okresu styczeń 2002-kwiecień 2009.

modeli czynnikowych. Modele szacowane były na próbie skróconej do końca roku 2008. Otrzymano następujące dynamiczne modele czynnikowe:

a) DFM_sur – model oszacowany na podstawie danych surowych zebranych bezpośrednio ze wspomnianych już źródeł.

b) DFM_SA – model oszacowany na podstawie danych sprowadzonych do cen stałych ze stycznia 2002 roku i oczyszczonych z wahań sezonowych, które przeprowadzono za pomocą procedury X – 12 ARIMA.

c) DFM_STAC – model oszacowany na podstawie danych wykorzystanych do estymacji modelu DFM_SA oraz zlogarytmowanych i sprowadzonych do stacjonarności. Do wyznaczenia stopnia integracji poszczególnych zmiennych posłużył znany powszechnie test ADF.

d) DFM_STD – model oszacowany na podstawie danych wykorzystanych do estymacji modelu DFM_STAC, które poddano dodatkowo standaryzacji.

Ponadto dla przypadków c oraz d dwukrotnie dokonano również modyfikacji samego zbioru danych pierwotnych. W jego wyniku otrzymano modele:

a) DFM_STAC_D – model analogiczny do modelu c, z tym że część zmiennych surowych została uwzględniona w zbiorze tylko w postaci opóźnień o jeden okres. Zmienne, które zostały opóźnione, wybrano na podstawie wiedzy ekonomicznej.

b) DFM_STAC_DF – model oszacowany dla danych stacjonarnych. W modelu tym w zbiorze zmiennych pierwotnych znalazły się zarówno wartości bieżące wszystkich zmiennych, jak i wartości opóźnione części z nich.

c) DFM_STD_D – model oszacowany dla danych analogicznych z modelem DFM_STAC_D, ale dodatkowo poddanych standaryzacji.

d) DFM_STD_DF – model oszacowany dla danych analogicznych z modelem DFM_STAC_DF, ale dodatkowo poddanych standaryzacji.

Oszacowano również modele dla niestacjonarnych danych zlogarytmowanych. Niestety żaden z nich nie posiadał parametrów istotnych na poziomie istotności 0,1 wobec czego w tym przypadku pominięto prezentację.

W obliczeniach wykorzystano dane dotyczące wielkości sprzedaży produkcji przemysłowej ogółem oraz jej części składowych, budownictwa w różnych aspektach, handlu krajowego i zagranicznego, inflacji i rynku pracy w różnych ujęciach, sfery budżetowej, a także charakterystyki szeroko rozumianej sfery polityki pieniężnej.

Po wstępnym przygotowaniu danych zastosowana została metoda głównych składowych w celu wyznaczenia czynników. Następnie wyznaczone zostały wartości kryteriów informacyjnych Baia i Ng w celu specyfikacji ich liczby. W poszczególnych modelach uwzględniona została taka liczba czynników, na jaką w danym przypadku wskazały kryteria informacyjne. W ostatnim etapie procedury za pomocą kryterium BIC wybrano opóźnienia zarówno dla zmiennej zależnej, jak i dla poszczególnych czynników występujących w danym modelu.

Po estymacji wszystkich wymienionych modeli wyznaczono na ich podstawie prognozy na 3 okresy w przód. Dla prognoz wyznaczone zostały odpowiednie błędy. Wyniki przeprowadzonych obliczeń zawierają tab. 1 i 2.

Tabela 1. Wyniki estymacji dynamicznych modeli czynnikowych

Model	*CPI*(-1)	*CPI*(-2)	*CPI*(-3)	*f* 1	*f* 1(-1)	*f* 2	*f* 2(-1)
DFM_sur				0,0043 (0,00)	–0,0045 (0,00)		
DFM_SA	0,9218 (0,00)	–0,3382 (0,01)	0,4171 (0,00)	–0,00001 (0,06)	0,00001 (0,06)		
DFM_STAC	–0,2907 (0,01)			-0,0076 (0,06)	–0,0077 (0,06)	0,0504 (0,00)	0,0276 (0,00)
DFM_STD					0,0803 (0,05)		
DFM_STAC_D	–0,3021 (0,05)	–0,3136 (0,00)	–0,2372 (0,03)	–0,0081 (0,07)	–0,0083 (0,07)		
DFM_STAC_DF	–0,3022 (0,01)	–0,3084 (0,00)	–0,2372 (0,03)	–0,009 (0,04)	–0,009 (0,05)		
DFM_STD_D				0,0961 (0,02)			
DFM_STD_DF				0,0594 (0,07)		–0,1188 (0,00)	

Źródło: opracowanie własne.

Tabela 2. Miary dobroci modelu i dokładności prognoz otrzymanych na podstawie dynamicznych modeli czynnikowych

Model	R2	Autokorelacja	RMSE	MAE	MAPE	Theil
DFM_sur	**0,57**	Tak	184,9600	161,0100	160,1	0,8800
DFM_SA	0,99	Tak	0,4329	0,4149	**0,4**	**0,0022**
DFM_STAC	**0,41**	**Nie**	**0,1500**	**0,1490**	**16,6**	**0,1790**
DFM_STD	0,05	**Nie**	1,4662	1,2445	**89,8**	0,7870
DFM_STAC_D	0,21	**Nie**	0,3830	0,3317	112,4	0,9072
DFM_STAC_DF	**0,22**	**Nie**	**0,1251**	**0,1090**	59 073,1	0,7313
DFM_STD_D	0,07	**Nie**	1,1839	0,8852	189,8	0,6556
DFM_STD_DF	0,14	**Nie**	0,9833	0,7667	102,6	0,5548

Źródło: opracowanie własne.

W tabeli 1 zawarto wartości parametrów stojących przy poszczególnych zmiennych. Brak wartości oznacza, że określona zmienna nie występuje w danym modelu. W nawiasach podano graniczne poziomy istotności parametrów (*p – value*).

Analizując całościowo tab. 2, należy wyciągnąć wniosek, że najlepszy ze statystycznego punktu widzenia okazuje się model oparty na danych stacjonarnych, ale niepoddanych standaryzacji (DFM_STAC). W zbiorze danych pierwotnych tego modelu znajdują się wyłącznie zmienne bieżące. DFM_STAC charakteryzuje się najwyższym współczynnikiem R-kwadrat spośród modeli, w których brak jest autokorelacji, i najniższymi wartościami błędów MAPE i Theil oraz drugimi pod względem wielkości błędami RMSE i MAE. W swojej budowie uwzględnia on dwa czynniki i po jednym opóźnieniu zarówno dla każdego czynnika, jak i dla zmiennej zależnej. Niestety wskazania tego modelu nie można uznać za jednoznaczne ze względu na fakt, że DFM_SA ma wyższą wartość współczynnika R-kwadrat, która jest bliska 1, co może wynikać z wystąpienia w tym modelu przeważającego wpływu opóźnień zmiennej zależnej. DFM_STAC_DF ma natomiast niższe wartości błędów RMSE i MAE.

Na skomentowanie zasługują również wartości MAPE dla modeli DFM_SA oraz DFM_STAC_DF. Ich skrajność wynika z odpowiednich różnic, które wystąpiły między prognozami a empirycznymi wartościami prognozowanych zmiennych.

Na rysunku 1 zaprezentowano wartości empiryczne i wartości wyznaczone na podstawie danych stacjonarnych.

Rys. 1. Dynamiczny model czynnikowy inflacji w Polsce w latach 2002-2008, oszacowany na podstawie stacjonarnych danych źródłowych

Źródło: opracowanie własne.

5. Podsumowanie

W wyniku szacowania dynamicznych modeli czynnikowych na różnych etapach transformacji danych uzyskano różne postacie DFM, różniące się co do zawartych w nich zmiennych oraz co do wartości parametrów stojących przy poszczególnych zmiennych, a także pod względem współczynnika R-kwadrat i wybranego błędu prognozy.

Najczęściej występującą postacią modelu był dynamiczny model czynnikowy z trzema opóźnieniami dla zmiennej zależnej i jednym czynnikiem z jednym opóźnieniem. Przy czym modele DFM_STAC_D i DFM_STAC_DF okazały się bardzo zbliżone.

Najlepiej opisuje rzeczywistość model oszacowany na podstawie danych stacjonarnych: DFM_STAC. Okazało się, że model oszacowany na podstawie danych poddanych standaryzacji dał w tym konkretnym przypadku gorsze wyniki estymacji i wyższe wartości błędów prognoz.

Literatura

Bai J., Ng S., *Determining the number of factors in approximate factor models*, „Econometrica” 2002, 70, s. 191-221.

Geweke J., *The Dynamic Factor Analysis of Economic Time Series*, [w:] D.J. Aigner, A.S. Goldberger (red.), Latent Variables in Socio-Economic Models, North Holland, Amsterdam 1977.

Greene W.H., *Econometric Analysis*, Pearson Education, New Jersey 2003.

Kotłowski J., *Forecasting Inflation with Dynamic Factor Model – The Case of Poland*, Working Papers, 2-08, SGH, Warszawa 2008.

Marcellino M., Stock J. H., Watson M. W., *Macroeconomic Forecasting in the Euro Area: Country Specific versus Area – Wide Information*, Working Paper, 201, Innocenzo Gasparini Institute for Economic Research, 2001.

Sargent T., Sims C., *Business Cycle Modelling without Pretending to Have Too Much a Priori Economic Theory*, [w:] C. Sims (red.), *New Methods in Business Cycle Research*, Federal Reserve Bank of Minneapolis, Minneapolis 1977.

Sims C.A., *Macroeconomics and reality*, „Econometrica” 1980, 48, s. 1-48.

Stock J., Watson M.W., *Diffusion Indexes*, Working Paper, 6702, National Bureau of Economic Research, 1998.

DATA TRANSFORMATION INFLUENCE ON DYNAMIC FACTOR MODEL PRINCIPAL COMPONENTS ESTIMATION PERFORMANCE

Summary: The paper concerns the dynamic factor models (DFM). It presents the concept of factor models, principal components method of estimation and Bai & Ng information criteria which are used to estimate the number of factors. The main purpose of the article is to show macroeconomic data transformation influence on the final result of model estimation. In the analysis we used 62 monthly time series from the Polish economy from the period from January 2002 to March 2009. Data transformation influenced the final results of model estimation. It gave us models with different number of variables and with different values of parameters.