ASK, 1998, nr 7, strony 83-102

Jarosław Górniak*
Instytut Socjologii UJ

Analiza czynnikowa i analiza głównych składowych[1]

Analiza czynnikowa i analiza głównych składowych są bardzo popularnymi metodami analizy danych w badaniach społecznych. Często techniki te są traktowane jako warianty tej samej metody. W artykule dyskutowane są różnice między tymi metodami i typowe obszary analizy, w których są stosowane. Omówiono kolejne etapy analizy, kryteria wyboru metod wyodrębniania i rotacji czynników oraz podstawy interpretacji wyników.

W standardowych pakietach statystycznych pod nazwą procedury: *analiza czynnikowa* kryją się dwie metody, które różnią się pod względem założeń: analiza głównych składowych (*Principal Components Analysis* — **PCA**) i analiza czynnikowa (*Factor Analysis* — **FA**). Obie służą sprowadzaniu informacji zawartych w wielu zmiennych (wskaźnikach) do niedużej liczby zastępujących je/wyjaśniających wymiarów/czynników. Często traktowane są one jako warianty tej samej metody, chociaż w istocie nimi nie są. Dodajmy jednak od razu, że, w praktyce, wyniki uzyskiwane przy pomocy obu metod są zbliżone i rzadko prowadzą do odmiennych wniosków.

Niektórzy statystycy, zwłaszcza o orientacji pragmatycznej, postulują stosowanie w większości sytuacji PCA (ze względu na pewne zalety formalne, o których dalej), zwłaszcza w sytuacji, gdy uzyskane tą metodą skale czynnikowe chcemy stosować w innych analizach[2]. Inni, na odwrót, postulują używanie właściwej analizy czynnikowej (zwykle metodą osi/czynników głównych lub metodą największej wiarygodności), zwłaszcza w zastosowaniu do analizy testów psychologicznych lub przy konstruowaniu modeli przyczynowych obserwowanych zjawisk, a to ze

* Uwagi do autora lub prośby o nadbitki prosimy kierować do: Jarosław Górniak, Instytut Socjologii, Uniwersytet Jagielloński, ul. Grodzka 52, 30-044 Kraków, e-mail: usgorna@cyf-kr.edu.pl.

[1] Tekst ten zamieszczony jest równolegle w książce autora pt. „My i nasze pieniądze. Wyniki badań empirycznych nad postawami wobec pieniądza", w której pełni rolę aneksu przybliżającego czytelnikom jedną z zastosowanych w niej technik analizy. Książka ta została przygotowana na podstawie badań finansowanych z grantu KBN nr 1 P109 046 05.

[2] Np. Wilkinson i Stenson podkreślają, że — w przeciwieństwie do głównych składowych — model wspólnych czynników nie jest jednoznacznie zdefiniowany; i to nie ze względu na to, że może być dowolnie rotowany (tak jak i główne składowe), ale dlatego, że bazuje na większej liczbie nieobserwowanych parametrów od liczby obserwowanych danych, co jest „niezwykłą okolicznością w statystyce" (Wilkinson, Stenson 1996: 569). Dla niektórych rodzajów macierzy możliwa jest nieskończona liczba doskonale dopasowanych modeli czynnikowych. Ponadto w FA mamy do czynienia z problemem konieczności szacowania wartości czynnikowych, które nie mogą być bezpośrednio wyliczone z modelu, jak to ma miejsce w PCA.

względu na fakt, że analiza czynnikowa nie dąży do wyjaśnienia całej wariancji każdej zmiennej w baterii pytań, a więc i jej części wynikającej z błędu, lecz tylko tej jej części, która jest podzielana z innymi zmiennymi, a więc może być uznana za pozostającą pod wpływem wspólnego czynnika — ukrytej zmiennej/konstruktu. Inni wreszcie podają praktyczne reguły, dotyczące typowego zastosowania analizy czynnikowej do baterii pytań kwestionariuszowych lub testów.

(1) Jeśli bateria pytań obejmuje wiele pytań (ok. 15 lub więcej) poleca się wstawianie na głównej przekątnej macierzy korelacji wartości 1,0, czyli przeprowadzenie analizy metodą głównych składowych.

(2) Przy mniejszych bateriach pytań zaleca się wstawienie na główną przekątną macierzy korelacji oszacowanych zasobów zmienności wspólnej, np. podniesionego do kwadratu współczynnika korelacji wielokrotnej każdej ze zmiennych z pozostałymi zmiennymi z baterii — tzn. przeprowadzenie analizy czynnikowej metodą głównych czynników/osi głównych (por. Holm 1976: 24 i 27).

Podkreślmy jeszcze raz: w praktyce wyniki różnych metod wyodrębniania czynników nie prowadzą do odmiennych wniosków. Jednak należy rozumieć różnice pomiędzy analizą głównych składowych i analizą czynnikową, by metody te stosować świadomie, gdyż oparte są one na odmiennych założeniach.

ZAŁOŻENIA CO DO TYPU DANYCH, KTÓRE MOŻNA ANALIZOWAĆ

PCA i FA prowadzi się z założenia na zmiennych co najmniej interwałowych, a ponadto przyjmuje się, że między zmiennymi mamy do czynienia ze związkami liniowymi. Dobre rezultaty dają te analizy także w przypadku powszechnie stosowanych w badaniach społecznych i marketingowych skal semantycznych typu Likerta (najlepiej co najmniej pięciopunktowych), skal dyferencjału semantycznego, skal ratingowych itp., mimo że formalnie trudno uznać je za skale interwałowe. Problem stosowania tego rodzaju formatów pytań kwestionariuszowych jako źródła danych do analizy czynnikowej ciągle bywa żywo dyskutowany. Zwykle stosuje się w przypadku takich danych macierz współczynników korelacji Pearsona jako punkt wyjścia do dalszej analizy (co czynią domyślnie programy komputerowe), mimo dyskusyjności tego rozwiązania w przypadku skal porządkowych. Powszechność takiego podejścia wynika z jego praktycznych walorów, choć niezbędna jest ostrożność badacza, ze względu na możliwe zniekształcenia. Niekiedy w przypadku analizy na tego rodzaju danych postuluje się stosowanie współczynników tau-b Kendalla[3] (Arminger, 1979: 148–152). Są jednak przeciwnicy takiego stanowiska, którzy podkreślają fakt, że zmienne w analizie czynnikowej muszą być interwałowe i pozostawać w liniowym związku, a korelacja powinna być mierzona współczynnikiem r Pearsona, czyli być miarą kowariancji pomiędzy standaryzowanymi zmiennymi (por. Kim & Mueller 1994b). W wielu przypadkach r i tau-b prowadzi do takich samych rezultatów. Bacher (1990) podkreśla stosunkowo dużą odporność analizy czynnikowej na zniekształcenia spowodowane zastosowaniem danych z pomiaru na skalach porządkowych stosowanych jako typowy format

[3] Postuluje się w takich przypadkach ograniczanie analizy do samej struktury czynników i unikanie wyliczania wartości czynnikowych, których obliczanie wymaga skal w pełni interwałowych.

pytań w badaniach postaw. Jeśli w rzeczywistości mamy do czynienia ze zmiennymi ciągłymi, które są przez nas tylko mierzone przy pomocy skal porządkowych, to im silniejszy jest związek pomiędzy tymi „prawdziwymi", ciągłymi zmiennymi, tym bardziej jest on tłumiony przez zastosowanie skal porządkowych. Im większa liczba pozycji na skali, tym efekt tłumienia jest mniejszy. Z tego punktu widzenia skala dziewięciopunktowa jest lepsza od siedmiopunktowej, a ta ostatnia jest lepsza od pięciopunktowej. Ogólnie nie zaleca się stosowania skal krótszych niż pięciopunktowe (Bollen 1989). Ta wskazówka dotyczy zresztą w ogóle stosowania skal porządkowych reprezentujących zmienne ilościowe w modelach liniowych. W przypadku skal krótszych niż pięciopunktowe pewnym wyjściem, dostępnym w przypadku niektórych programów do modelowania strukturalnego (LISREL), jest stosowanie tzw. korelacji polichorycznych. Nie jest to jednak rozwiązanie bezdyskusyjne, zwłaszcza w przypadku prób o liczebności mniejszej od kilku tysięcy. W praktyce, w naukach społecznych, traktuje się pięcio- czy siedmiopunktowe skale odpowiedzi na pytania o postawy tak, jakby to były skale interwałowe.

Prowadzi się także analizy na zmiennych dychotomicznych. Jeśli każda zmienna zawiera informację o innej cesze, analiza głównych składowych daje zwykle użyteczne wyniki. Jednak gdy mamy do czynienia ze zmiennymi dyskretnymi o więcej niż dwóch kategoriach, zakodowanymi przy pomocy zestawu zerojedynkowych zmiennych wskaźnikowych (*dummy variables*), stosowanie analizy głównych składowych lub analizy czynnikowej nie jest poprawnym podejściem. W przypadku wielowariantowych cech nominalnych należy stosować wieloraką analizę korespondencji lub tożsamą z nią analizę HOMALS z modułu SPSS Categories[4]. Także wówczas, gdy frakcje jedynek w poszczególnych zmiennych („trudność" kategorii) znacznie się różnią, analiza czynnikowa może być zwodnicza, gdyż korelacje między zmiennymi mogą wynikać z różnic w owej „trudności", a nie z merytorycznego związku cech[5]. Mimo to używa się analizy głównych składowych zmiennych zerojedynkowych w celu wyodrębnienia skupień zmiennych. Niekiedy lepsze wyniki może dać analiza skupień, w której mamy do dyspozycji wiele miar odległości pozwalających na identyfikację interesujących nas wzorów powiązań cech kodowanych binarnie.

Najczęściej analizie czynnikowej i analizie głównych składowych poddaje się zmienne w ich postaci standaryzowanej (tzn. analizuje się macierz korelacji, a nie macierz kowariancji); standaryzacja taka jest wykonywana domyślnie przez komputerowe programy analizy czynnikowej w takich popularnych pakietach jak SPSS, STATISTICA czy SYSTAT.

---

[4] SPSS daje nam możliwość eksperymentowania z narzucaniem naszym zmiennym różnego poziomu pomiarowego. Ogólniejszym modelem, który umożliwia takie analizy jest PRINCALS — nieliniowa analiza głównych składowych.

[5] Zniekształcenia mogą zresztą wystąpić również w przypadku wspominanych już pięcio- czy siedmiopunktowych zmiennych porządkowych, jeśli występują w nich bardzo duże różnice w częstości występowania poszczególnych kategorii. Duże rozbieżności w kształcie rozkładów mogą wpływać także na wyniki w przypadku analizy prowadzonej, zgodnie z zasadami, na zmiennych nie budzących wątpliwości co do interwałowego poziomu ich pomiaru; w związku z tym niekiedy przed wykonaniem PCA lub FA niezbędna jest transformacja zmiennych podobnie jak w analizie regresji.).

ANALIZA GŁÓWNYCH SKŁADOWYCH (PCA)

Główne składowe to liniowe kombinacje[6] zmiennych, które (1) są ortogonalne w stosunku do siebie, tzn. nie są wzajemnie skorelowane i mają taką właściwość, że (2) pierwsza główna składowa wyjaśnia największą ilość łącznej wariancji zmiennych, druga jest ortogonalna do pierwszej i wyjaśnia największą część łącznej wariancji zmiennych nie wyjaśnionej przez pierwszą główną składową itd. Maksymalna liczba głównych składowych potrzebna do wyjaśnienia całości wspólnej wariancji k zmiennych jest równa k.

Analiza głównych składowych (PCA) ma wielorakie zastosowania.

(1) Jest metodą redukcji przestrzeni danych, to znaczy jej celem jest przedstawienie informacji zawartej w zbiorze *k* zmiennych przy pomocy *j<k* głównych składowych przy zachowaniu jak największej ilości informacji z pierwotnego zbioru zmiennych. Korzystając z faktu, że kolejne składowe wyjaśniają malejący zakres łącznej wariancji zmiennych, dla celów prezentacji zależności w zbiorze danych wykorzystujemy j pierwszych składowych. W celu uzyskania interpretowalnych wyników główne składowe można poddać rotacji (o tym dalej).

(2) Jest metodą przekształcenia k skorelowanych zmiennych wyjściowych w k głównych składowych. Korzyścią z takiego przekształcenia zbioru zmiennych w zbiór głównych składowych jest możliwość ujęcia całości informacji zawartej w zmiennych (ich wariancji) w postaci zestawu ortogonalnych, a więc niezależnych, składowych. Takie składowe można użyć w wygodny sposób w analizie regresji lub analizie dyskryminacji, zwłaszcza w sytuacji, gdy pierwotny zbiór zmiennych niezależnych jest silnie skorelowany (występuje w nim zjawisko silnej przybliżonej współliniowości zmiennych niezależnych). W praktyce w dalszej analizie wykorzystuje tylko część wyodrębnionych składowych głównych. Niżej podaję kilka praktycznych reguł wykorzystania składowych głównych w modelach liniowych.

(3) Jest metodą prezentacji graficznej struktury wielowymiarowego zbioru danych na płaszczyźnie z jak najmniejszym zniekształceniem informacji.

Model analizy głównych składowych można wyrazić następująco:

główna składowa = liniowa kombinacja *obserwowanych zmiennych*

W analizie głównych składowych przedmiotem wyjaśnienia jest **całkowita** wariancja wszystkich zmiennych. Główne składowe, jako liniowe kombinacje obserwowalnych zmiennych, są jednoznacznie określone. Zatem dla każdej obserwacji w bazie danych można jednoznacznie wyliczyć wartości na głównej składowej (FACTOR SCORES), dodając do siebie wartości standaryzowane danego przypadku na poszczególnych zmiennych wymnożone przez odpowiednie wagi (*współczynniki wartości czynnikowych* — FACTOR SCORES COEFFICIENTS).

Matematyczną podstawą analizy głównych składowych jest dekompozycja pełnej macierzy korelacji zmiennych (z wartościami 1 na głównej przekątnej) na wektory własne i wartości własne.

---

[6] Kombinacja liniowa ma postać $y = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n$ ($a_0 = 0$ w przypadku modelu dla zmiennych standaryzowanych).

ANALIZA CZYNNIKOWA

Analiza czynnikowa (FA) jest metodą badania struktury leżącej u podstaw związków obserwowanych między zmiennymi. Celem tej metody jest sprowadzenie zaobserwowanych korelacji (kowariancji) między wieloma zmiennymi do niedużej liczby wyjaśniających je zmiennych nieobserwowalnych: wspólnych czynników, konstruktów. W modelu analizy czynnikowej przyjmuje się, że na każdą ze skorelowanych ze sobą zmiennych wpływają w różnym stopniu wspólne czynniki, które wyjaśniają zaobserwowaną korelację. Wariancja zmiennych dzieli się na:

(1) wariancję wspólną, podzielaną przez zmienne z zestawu (wyjaśnioną przez czynniki wspólne); część wariancji zmiennej wyjaśnioną przez wspólne czynniki nazywamy jej zasobem zmienności wspólnej (*communality*);

(2) wariancję swoistą każdej ze zmiennych, niesprowadzalną do współzmienności wywołanej oddziaływaniem wspólnych czynników.

Tę ostatnią dzieli się jeszcze na wariancję specyficzną zmiennej oraz wariancję wynikającą z błędu.

Celem analizy czynnikowej jest wyjaśnienie zasobu zmienności wspólnej mierzonych zmiennych. U podstaw analizy czynnikowej mamy więc model teoretyczny ukrytej struktury przyczynowej, wyjaśniającej zaobserwowaną strukturę korelacji wskaźników. Model analizy czynnikowej można więc wyrazić następująco:

obserwowana zmienna = liniowa kombinacja *czynników* + błąd

Matematycznie rzecz sprowadza się do analizy struktury tzw. zredukowanej macierzy korelacji, tzn. macierzy korelacji, w której na przekątnej umieszczone są wartości wskazujące **proporcję wariancji wspólnej** — wyjaśnianej przez wspólne czynniki — **w całkowitej wariancji każdej ze zmiennych (*zasoby zmienności wspólnej — communalities*)**. Najprostszym sposobem oszacowania tej proporcji (zasobów zmienności wspólnej) jest wykorzystanie kwadratu współczynnika korelacji wielorakiej każdej ze zmiennych z pozostałymi zmiennymi z analizowanego zestawu. Wielorakie R^2 jest dolną granicą zasobu zmienności wspólnej każdej ze zmiennych w modelu, granicą mającą tę zaletę, że jest ustalana empirycznie, a nie szacowana. Innym sposobem jest iteracyjne szacowanie wartości *communalities* poprzez wielokrotne prowadzenie analizy głównych składowych zredukowanej macierzy korelacji i podstawianie za każdym razem na główną przekątną nowo oszacowanych zasobów zmienności wspólnej, aż do osiągnięcia sytuacji, w której modele z dwóch kolejnych kroków nie różnią się znacząco (można manipulować tym kryterium zbieżności).

Odrębną metodą wyodrębniania czynników jest metoda największej wiarygodności: czynniki i zasoby zmienności wspólnej wyznaczone są w taki sposób, by z największym prawdopodobieństwem wytwarzały zaobserwowaną korelację między zmiennymi.

Żeby lepiej uświadomić sobie różnicę pomiędzy PCA i FA zwróćmy uwagę, że do wyjaśnienia **całkowitej wariancji** dwóch zmiennych skorelowanych np. na poziomie 0,81 potrzeba dwóch głównych składowych (wyznaczony zostanie po prostu nowy układ współrzędnych), podczas gdy do zupełnego wyjaśnienia **korelacji** między nimi (cel analizy czynnikowej) wystarczy jeden czynnik skorelowany z każdą z tych zmiennych na poziomie 0,9 (pomijając problem z identyfikacją takiego prostego modelu).

KIEDY STOSOWAĆ ANALIZĘ GŁÓWNYCH SKŁADOWYCH A KIEDY ANALIZĘ CZYNNIKOWĄ

Zastosowanie **analizy czynnikowej** jest preferowane w sześciu sytuacjach opisanych poniżej.

(1) Gdy chcemy wyjaśnić zaobserwowaną korelację między zmiennymi przy pomocy modelu przyczynowego opartego na strukturze związków zmiennych obserwowalnych z ukrytymi czynnikami.

(2) Gdy dysponujemy modelem teoretycznym struktury takiego związku[7] lub będziemy uzyskane wyniki interpretować w kategoriach teoretycznego modelu przyczynowego.

(3) Gdy koncentrujemy się na wyjaśnieniu korelacji między zmiennymi i dlatego chcemy wyłączyć z analizy wariancję swoistą zmiennych.

(4) Gdy zmienne są obciążone względnie dużym błędem pomiarowym, który badacz chce wyłączyć z analizy.

(5) Gdy celem analizy jest selekcja pozycji/wskaźników do skali sumarycznej Likerta (choć w tym przypadku, zwłaszcza przy dużej liczbie pozycji, stosuje się też analizę głównych składowych).

(6) Gdy celem analizy jest klasyfikacja zmiennych we względnie jednorodne grupy, w gruncie rzeczy będące właśnie wskaźnikami pewnych konstruktów.

Niektórzy statystycy (np. Wilkinson i Stenson 1996) zalecają porównanie rezultatów uzyskanych za pomocą analizy czynnikowej (np. metodą największej wiarygodności, osi głównych czy najmniejszych kwadratów) z wynikami analizy głównych składowych, żeby „uniknąć oszukania" przez degeneracje wynikające z niejednoznaczności modelu czynnikowego (por. przypis 2).

Inne są przesłanki stosowania **analizy głównych składowych**.

(1) Stosujemy ją wówczas, gdy nie dysponujemy potencjalnym modelem „głębokiej" struktury czynników wyjaśniających związki pomiędzy zmiennymi, taki model nie jest celem naszej analizy lub nie chcemy „wtłaczać" w taki model posiadanych danych empirycznych.

(2) Wybierać ją będziemy, gdy celem jest eksploracja, rozpoznanie struktury zbioru danych: gdy wyszukujemy przypadki osobliwe, chcemy przedstawić graficznie strukturę zbioru danych w przestrzeni dwu- lub trójwymiarowej przy możliwie najmniejszym zniekształceniu relacji zachodzących pomiędzy obserwacjami, szukamy skupień obiektów ze względu na podobieństwo w zakresie analizowanych cech, określamy minimalną liczbę wymiarów, za pomocą których jesteśmy w stanie wyjaśnić założoną część wariancji zbioru zmiennych.

(3) Jeśli wiemy, że wariancja specyficzna i wariancja wynikająca z błędu jest niewielka, a także, gdy analizujemy dużo (np. więcej niż 15) skorelowanych zmiennych lub gdy korelacja między zmiennymi jest względnie wysoka, lepiej jest stosować analizę głównych składowych: główne składowe są jednoznacznie określone — są kombinacjami liniowymi zmiennych i mogą być wprost wyliczone, podczas gdy wartości czynników głównych mogą być tylko szacowane, nie są jednoznacznie określone i przy zastosowaniu są źródłem pewnych kłopo-

[7] W tym wypadku nawet właściwsze będzie zastosowanie konfirmacyjnej analizy czynnikowej, dostępnej m.in. w programach SPSS AMOS, LISREL, EQS, RAMONA (SYSTAT) i SEPATH (STATISTICA).

tów (np. oszacowane zmienne z wartościami czynnikowymi mogą być skorelowane nawet wtedy, gdy czynniki nie są skorelowane lub mogą nie być doskonale skorelowane z rzeczywistymi czynnikami).

(4) PCA stosujemy, gdy chcemy wyliczyć nieskorelowane główne składowe w celu zastosowania ich w dalszych analizach wielowymiarowych (np. regresji lub dyskryminacji).

(5) Wybieramy tę metodę także wówczas, gdy chcemy wyliczyć jednoznacznie wartości skal reprezentujących wymiary mierzone przez zestaw zmiennych. Alternatywą dla PCA jest proste sumowanie dla każdego przypadku wartości z poszczególnych zmiennych, zaklasyfikowanych do skali na podstawie analizy czynnikowej („skala oparta na czynniku", a nie „skala czynnikowa"). Zastosowanie wartości czynnikowych wyliczonych w analizie czynnikowej (FA) jest problematyczne[8], choć też stosowane (por. podręcznikowy przykład w Backhaus i in. 1990).

Etapy analizy czynnikowej i analizy składowych głównych oraz zasady interpretacja wyników tych dwóch metod (przy świadomości różnic pomiędzy nimi) są takie same, dlatego też potraktujemy je łącznie.

KILKA UŻYTECZNYCH DEFINICJI

Wzorem Haira i in. (1984) warto podać słowniczek pojęć najczęściej spotykanych przy okazji analizy czynnikowej i analizy głównych składowych.

Zasób zmienności wspólnej — część wariancji oryginalnej zmiennej podzielana z wszystkimi pozostałymi zmiennymi włączonymi do analizy; w modelu ortogonalnym jest równa podniesionym do kwadratu ładunkom czynnikowym danej zmiennej. W przypadku wstępnej ekstrakcji czynników w analizie głównych składowych zasób zmienności wspólnej każdej ze zmiennych jest równy 1, co oznacza, że analizie poddana jest cała wariancja zmiennych. Po odrzuceniu części „najmniejszych" składowych zasób zmienności wspólnej mówi nam, jak dobrze reprezentowana jest dana zmienna przez model o zredukowanej przez nas liczbie wymiarów. W analizie czynnikowej szacowanie zasobu zmienności wspólnej jest jednym z kluczowych elementów procesu budowania modelu czynnikowego. Ostateczny zasób zmienności wspólnej informuje nas o tym, jaki zakres wariancji zmiennej jest sprowadzalny do ukrytych czynników ujętych w modelu.

Wartość własna — matematyczna własność macierzy kwadratowej; reprezentuje zakres wariancji wyjaśnianej przez dany czynnik. We wstępnej fazie analizy, przed rotacją, czynniki wyodrębniane są w taki sposób, że kolejno wyjaśniają największą możliwą część wariancji, spełniając jednocześnie warunek braku wzajemnej korelacji. Prowadzi to do tego, że kolejne czynniki (wektory własne) mają coraz mniejszą wartość własną. W PCA suma wartości własnej wszystkich składowych głównych (czyli ich wariancji) równa się liczbie zmiennych, gdyż każda zmienna standaryzowana ma wariancję równą 1. W analizie czynnikowej zredukowanej macierzy korelacji suma wartości własnych równa się sumie wartości umieszczonych na przekątnej tej macierzy (tzw. ślad

[8] Np. zmienne z wartościami oszacowanymi metodą regresyjną mogą być skorelowane nawet wtedy, gdy czynniki są ortogonalne.

macierzy). Procent wariancji wyjaśnionej przez czynnik obliczamy jako stosunek wartości własnej czynnika do sumy wszystkich wartości własnych (w PCA procentuje się do sumy równej iiczbie zmiennych, gdyż na przekątnej pełnej macierzy korelacji są jedynki — całkowite wariancje zmiennych standaryzowanych).

Ładunek czynnikowy — ogólne określenie współczynników umieszczanych w macierzy ładunków czynnikowych; w węższym znaczeniu: współczynniki regresji pomiędzy zmienną (standaryzowaną) a zestawem czynników wspólnych. W przypadku nierotowanych głównych składowych (które są nieskorelowane) i w przypadku rotacji ortogonalnej w obu opisywanych metodach są to jednocześnie współczynniki korelacji pomiędzy zmienną i każdym czynnikiem z osobna, jak i współczynniki regresji pomiędzy zmienną a zestawem czynników wspólnych. W przypadku rotacji skośnej mamy do czynienia z dwiema macierzami ładunków czynnikowych: **macierzą wzoru czynników** (*factor pattern matrix*) zawierającą ładunki czynnikowe, czyli współczynniki regresji pomiędzy zmienną (standaryzowaną) a zestawem czynników wspólnych oraz **macierzą struktury czynników** (*factor structure matrix*) zawierającą współczynniki korelacji pomiędzy każdą zmienną i każdym czynnikiem z osobna. W przypadku rotacji skośnej wartości współczynników w obu rodzajach macierzy nie są już sobie równe.

Rotacja czynników — proces lokowania (transformacji) czynników ostatecznie zachowanych w analizie (także głównych składowych) w przestrzeni zmiennych tak, by uzyskać możliwie najprostszą, interpretowalną strukturę czynników.

Ortogonalne czynniki — czynniki nie pozostające ze sobą w korelacji; w przestrzeni: prostopadłe do siebie.

Rotacja ortogonalna — rotacja z zachowaniem niezależności (braku korelacji, prostopadłości) czynników.

Skośne czynniki — czynniki skorelowane ze sobą, nie tworzące w przestrzeni kąta prostego.

Rotacja skośna — rotacja czynników dopuszczająca korelację pomiędzy nimi, reprezentowaną przez odejście od prostopadłości czynników w przestrzeni.

Zredukowana macierz korelacji — macierz korelacji, w której na głównej przekątnej zamiast 1 umieszczone zostały oszacowane wartości zasobu zmienności wspólnej każdej zmiennej, zazwyczaj wartości współczynnika determinacji R^2 (wielokrotnego) danej zmiennej w jej regresji na wszystkie pozostałe zmienne ujęte w macierzy.

ETAPY ANALIZY

W analizie czynnikowej i analizie głównych składowych mamy do czynienia z pewną sekwencją czynności analitycznych.

(1) Podjęcie przez analityka decyzji o sposobie postępowania z brakiem danych: eliminacja parami, przypadkami czy zastępowanie średnią? A może należy podstawić w miejsce braków danych wartości na podstawie któregoś ze statystycznych modeli imputacji? Odpowiedź na te pytania wymaga uprzedniej analizy konfiguracji braków danych. Pomocny może być w tym np. moduł pro-

gramu SPSS: *Missing Value Analysis*. Ignorowanie problemów wynikających z braków danych może prowadzić do zniekształcenia wyników analizy.

(2) Obliczenie macierzy korelacji (program wykonuje to automatycznie).

(3) Wstępny ogląd macierzy korelacji i usunięcie z analizy zmiennych nie skorelowanych z pozostałymi (ewentualny test oceniający przydatność macierzy korelacji do zastosowania modelu czynnikowego) — w praktyce często jest jednak łatwiej przeprowadzić wstępne analizy metodą głównych składowych i „wyłapać" zmienne, które pojedynczo budują odrębne czynniki lub nisko ładują wszystkie czynniki zachowane w analizie.

(4) Wyodrębnienie czynników — wybór metody wyodrębnienia i określenie liczby czynników pozostawionych do dalszej analizy.

(5) Rotacja czynników w celu uzyskania klarownej interpretacji.

(6) Interpretacja znaczenia uzyskanych czynników na podstawie sensu zmiennych, które mają wysokie ładunki czynnikowe w przypadku danego czynnika (na ogół bierze się pod uwagę ładunki czynnikowe o wartościach bezwzględnych wynoszących co najmniej 0,6, choć nie jest to sztywna zasada i wiele zależy od konkretnych danych)

(7) Wyliczenie (w razie potrzeby) wartości czynnikowych i użycie ich do sporządzenia wykresów lub dalszych analiz.

METODA WYODRĘBNIANIA CZYNNIKÓW

Wybór pomiędzy analizą głównych składowych a właściwą analizą czynnikową został przedyskutowany wyżej. W ramach właściwej analizy czynnikowej stosujemy zazwyczaj jedną z poniższych metod.

(1) Analiza metodą głównych osi (*Principal Axis Factoring: PAF*) lub metoda najmniejszych reszt (*unweighted least-squares method: ULS* — metoda nieważonych najmniejszych kwadratów, znana również w literaturze jako metoda MINRES), które zasadniczo dają identyczne rezultaty[9]. Są to techniki iteracyjne korzystające z analizy głównych składowych jako punktu wyjścia w analizie zredukowanej macierzy korelacji, w wyniku których następuje wyodrębnienie czynników i oszacowanie zasobu zmienności wspólnej zmiennych użytych w modelu. Są to techniki eksploracyjne, opisowe, dla których nie mamy testu dopasowania modelu do danych.

(2) Metoda największej wiarygodności (*maximum likelihood: ML*) jest także metodą iteracyjną: czynniki wyznaczone są w taki sposób, by z największą wiarygodnością wywoływały zaobserwowaną korelację między zmiennymi, wszakże przy założeniu, że próba pochodzi z populacji, w której analizowane zmienne podlegają wielowymiarowemu rozkładowi normalnemu (co nakłada postulat normalności rozkładu także na każdą z nich z osobna — zjawisko rzadko spotykane w badaniach społecznych). Metoda ta daje możliwość przeprowadzenia testu dopasowania modelu opartego na określonej liczbie czynników do obserwowanej macierzy korelacji w warunkach dużej próby (test oparty na rozkładzie chi^2). Paradoksalnie, w warunkach dużej próby nawet niewielkie odchylenia odtworzonej na podstawie modelu czynnikowego macierzy korelacji od

[9] „W warunkach istnienia rozwiązania kanonicznego metoda MINRES jest identyczna z iteracyjną metodą czynników głównych dla R" (Arminger 1979: 52).

macierzy obserwowanej łatwo prowadzą do odrzucenia hipotezy o dopasowaniu modelu; chęć uzyskania potwierdzonego testem dopasowania prowadzi zwykle do zachowania zbyt dużej liczby czynników. Jeśli posłużymy się innymi kryteriami określania liczby czynników, zwłaszcza metodą merytorycznej interpretowalności czynników, metoda ta daje dobre rezultaty w analizie eksploracyjnej i jest często polecana. W procesie iteracyjnego wyodrębniania czynników tą metodą, w każdym kolejnym kroku, większa waga przypisywana jest tym zmiennym, które mają większy oszacowany zasób zmienności wspólnej. Z nazwy „metoda największej wiarygodności" nie wynika ocena tej metody, a jedynie wskazany jest przez nią model matematyczny, który stoi u podstaw tej techniki. Metoda ta nie usuwa problemu niejednoznaczności modelu czynnikowego. Podobne właściwości ma metoda uogólnionych najmniejszych kwadratów (*Generalised least squares — GLS*).

To, którą opcję wybrać, jeżeli już zdecydujemy się na analizę czynnikową, a nie głównych składowych, zależy od tego, czy chcemy testować jakość dopasowania modelu do danych w populacji i czy mamy podstawy ku temu (rozkład normalny, duża próba) — wówczas ML jest odpowiednia. Jeżeli prowadzimy analizę eksploracyjną zwykle używamy PAF. Wszystkie metody w praktyce badawczej dają zwykle takie same (merytorycznie, nie matematycznie) rezultaty.

OKREŚLANIE LICZBY CZYNNIKÓW

W analizie głównych składowych i analizie czynnikowej stosuje się różne kryteria pomocne przy podejmowaniu decyzji o liczbie czynników/głównych składowych pozostawionych do dalszej analizy:

(1) Kryterium wartości własnej Keisera w analizie głównych składowych: wartość własna każdego czynnika/głównej składowej (= jego wariancji) pozostawionego w dalszej analizie powinna być większa od 1 (a więc od wariancji każdej pojedynczej zmiennej standaryzowanej, która jest podstawą analizy). Program SPSS także w przypadku analizy czynnikowej przeprowadza najpierw analizę głównych składowych i kryteria selekcji odnoszą się do wartości własnych wyliczonych na tym etapie. Jako domyślne kryterium selekcji czynników stosuje się w tym programie kryterium Keisera.

(2) Kryterium Jolliffe (kryterium to sformułowane zostało dla PCA): w warunkach badania na próbie losowej błąd losowy może prowadzić do zaniżenia wartości własnej głównej składowej. W związku z tym, należy zachować w analizie te składowe, których wartość własna jest większa od 0,7.

(3) W przypadku analizy zredukowanej macierzy korelacji (np. analiza czynnikowa metodą osi głównych) stosuje się też w niektórych programach statystycznych (nie dotyczy to SPSS) kryterium Guttmana: wartość własna 0, które prowadzi jednak często do zachowania zbyt dużej liczby czynników.

(4) Kryterium wystarczającej proporcji wyjaśnionej wariancji (popularne w ramach PCA): należy pozostawić tyle składowych, by wyjaśniały założony procent wariancji, np. 80% lub 95% czynników, które w świetle wstępnej analizy wyjaśniają w sumie określony zakres wariancji.

(5) Liczba czynników powinna być mniejsza od połowy liczby zmiennych (najbardziej „płynne" kryterium ze spotykanych w literaturze, obok kryterium, że naj-

mniejszy czynnik powinien wyjaśniać co najmniej 1%, 5% lub 10% całkowitej wariancji w PCA, a całkowitej wspólnej wariancji w FA).

(6) Kryterium interpretowalności czynników: badacz zachowuje taką liczbę czynników, która ma sens, da się zinterpretować w ramach jego modelu teoretycznego. Jest to ważne kryterium, choć jest subiektywne. Dane obciążone są błędami wynikającymi z losowania i samego pomiaru. Może to prowadzić do zniekształceń i wyodrębniania czynników reprezentujących przypadkowe konfiguracje zmiennych. Z drugiej strony, ważny jest walor „heurystyczny" analizy czynnikowej, jej zdolność ujawniania konfiguracji, których nie oczekiwaliśmy i podważania tych, z wizją których przystępujemy do badania. Odrzucenie czynnika, ze względu na jego „nieinterpretowalność", musi być więc poprzedzone stosownym namysłem.

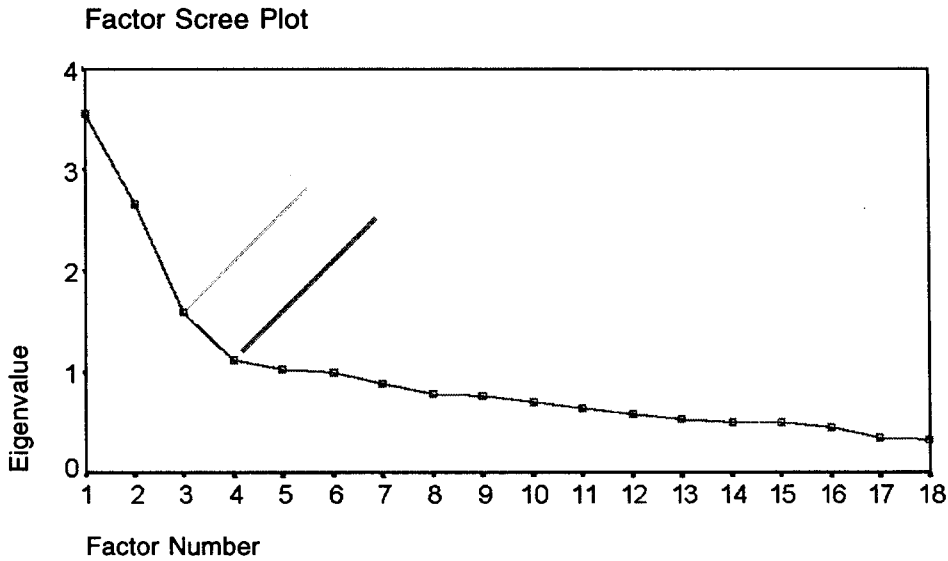
(7) Kryterium istotności statystycznego testu dopasowania odtworzonej macierzy korelacji do macierzy obserwowanej (tylko dla metody największej wiarygodności i GLS). Jak już wskazałem, stosowanie tego kryterium prowadzi często do pozostawienia dużej liczby „małych" czynników. Test ten stawia wymóg wspólnej wielowymiarowej normalności rozkładów zmiennych w populacji, z której pobrana jest próba. Testujemy kolejne modele zwiększając liczbę czynników o 1, aż do uzyskania wartości $p > 0{,}05$ w teście chi^2.

(8) Analiza odchyleń (reszt, *residuals*) obserwowanych współczynników korelacji od współczynników odtworzonych. Duże odchylenia odtworzonych współczynników korelacji świadczą o słabym dopasowaniu naszego modelu czynnikowego do danych i każe go zweryfikować. Musimy jednak pamiętać, że **nawet bardzo dobre dopasowanie modelu do danych nie gwarantuje jego prawdziwości**.

(9) Kryterium osypiska (Cattella): należy zachować tyle czynników, ile tworzy „zbocze", natomiast zignorować te, które tworzą „osypisko", „rumowisko" u podnóża na wykresie sporządzonym przez połączenie punktów opisujących wielkość wartości własnej (wariancji) kolejnych czynników[10]. Czasami trudno jest zdecydować, które miejsce stanowi rzeczywiście początek osypiska i wybór bywa nieco subiektywny. Metoda ta daje jednak często dobre rezultaty. Prowadzi zwykle do pozostawienia mniejszej liczby czynników niż kryterium Keisera i jest skuteczna zwłaszcza w przypadku analizy koncentrującej się na najważniejszych czynnikach i ignorującej mniej ważne. A oto przykładowy wykres ilustrujący kryterium „osypiska". Osypisko wyraźnie zaczyna się w przypadku 4 czynników, taką więc ich liczbę należałoby pozostawić w analizie. Można jednak dopatrywać się początku osypiska już przy 3 czynnikach, należy więc odwołać się dodatkowo do kryterium merytorycznej interpretowalności. Kryterium Keisera sugeruje rozwiązanie oparte na 4 czynnikach.

W analizie czynnikowej dużą rolę odgrywa doświadczenie i sztuka interpretacji, stąd badacz powinien elastycznie kierować się powyższymi wskazówkami, by dotrzeć do ostatecznego modelu.

[10] W literaturze spotyka się dwa stanowiska: jedno każe pozostawić tyle czynników, ile znajduje się na „zboczu" wraz z tym, od którego zaczyna się „osypisko"; inne stanowisko każe ignorować ten ostatni czynnik.



ROTACJA CZYNNIKÓW I INTERPRETACJA WYNIKÓW

Celem rotacji jest uproszczenie wzoru czynników tak, by (w idealnym przypadku) każda zmienna miała wysoki ładunek tylko na jednym czynniku i by każdy czynnik miał przynajmniej kilka ładunków bliskich 0 i kilka wysokich, bliskich 1 lub -1. Ułatwia to interpretację uzyskanego modelu. Taki ogólny cel może prowadzić do różnych szczegółowych kryteriów matematycznych, które kierują zmianą położenia czynników wobec zmiennych.

Aby uzyskać prostą strukturę macierzy ładunków czynnikowych, można dążyć do uproszczenia interpretacji każdej ze zmiennych przy pomocy minimum istotnych czynników, a więc do uproszczenia wierszy macierzy ładunków. Prowadzi to do rotacji QUARTIMAX[11], która w szczególnych sytuacjach może jednak skończyć się wyprodukowaniem wysokich ładunków dla wszystkich zmiennych na jednym czynniku.

Można też dążyć do uproszczenia interpretacji każdego z czynników, a więc doprowadzić do tego by względnie niewiele zmiennych miało wysokie ładunki na jednym czynniku, a pozostałe zmienne miały na tymże czynniku ładunki zerowe lub bliskie zero; oznacza to dążenie do uproszczenia kolumn macierzy ładunków. Prowadzi to do rotacji VARIMAX[12], która daje, ogólnie biorąc, klarowniejsze i bar-

[11] Kryterium rotacji jest w tym wypadku maksymalizacja wariancji podniesionych do kwadratu ładunków czynnikowych dla każdej zmiennej, przy danej liczbie czynników, danych zasobach zmienności wspólnej i zachowaniu ortogonalności czynników.

[12] Kryterium rotacji jest w tym wypadku maksymalizacja wariancji podniesionych do kwadratu ładunków czynnikowych dla każdego czynnika, przy danej liczbie czynników, danych zasobach zmienności wspólnej i zachowaniu ortogonalności czynników.

dziej stabilne wyniki. Jest to domyślna rotacja w programie SPSS. Kompromisem pomiędzy rotacją QUARTIMAX i VARIMAX jest rotacja EQUAMAX.

Najczęściej stosowaną w praktyce metodą rotacji jest ortogonalna rotacja VARIMAX (z normalizacją Keisera[13]). Powołując się na eksperymenty Keisera, Kim i Mueller (1994) piszą: „wzór czynników uzyskany poprzez rotację VARIMAX bywa bardziej stabilny (*invariant*) od uzyskanego w rotacji QUARTIMAX, gdy analizujemy różne podzbiory zmiennych" (s. 104). Z kolei Arminger (1979: 94–95) pisze, że w wielu wykonanych przez siebie analizach nie stwierdził większych różnic pomiędzy wynikami uzyskanymi przy pomocy tych rotacji, za wyjątkiem sytuacji, w których wśród zmiennych występowały duże różnice pomiędzy zasobami zmienności wspólnej. Konkludując: jeśli zasadne jest wykonanie rotacji ortogonalnej, nie dopuszczającej korelacji między czynnikami, używamy zazwyczaj rotacji VARIMAX.

W wielu przypadkach nie mamy powodu zakładać ortogonalności czynników, należy dopuścić do korelacji między czynnikami, gdyż oczekujemy, że są one w rzeczywistości skorelowane. W takiej sytuacji przeprowadzamy nieortogonalną rotację prowadzącą do czynników skośnych (*oblique factors*). W analizie czynnikowej wypracowano kilka takich metod. W SPSS dostępny jest skośny odpowiednik rotacji VARIMAX — rotacja DIRECT OBLIMIN. Dopuszczalny poziom korelacji między czynnikami reguluje się w niej przy pomocy parametru Delta: wartość 0 lub nieco większa dopuszcza największe skorelowanie; im bardziej ujemna wartość, tym rozwiązanie bliższe jest uzyskanemu w rotacji VARIMAX.

Nie ma doskonałej recepty na ustawianie parametru DELTA. W analizie eksploracyjnej G. Arminger poleca następujący sposób postępowania (1979: 112–113).

(1) Najpierw zdefiniować konstrukty i zoperacjonalizować je przy pomocy mierzalnych zmiennych.
(2) Wykonać analizę bez rotacji i sporządzić wykres ładunków czynnikowych (problem przy większej liczbie czynników). Zmienne definiujące konstrukt powinny tworzyć zwartą chmurę punktów. Zmienne odosobnione należy wyłączyć z analizy.
(3) Jeśli przeprowadzimy osie przez chmury punktów, możemy mniej więcej ocenić kąt pomiędzy nimi. Cosinus tego kąta umożliwia ocenę korelacji pomiędzy czynnikami. Jeśli korelacja jest wysoka, ustawiamy DELTA > 0, jeśli niska — DELTA < 0.
(4) Zarówno przy eliminacji zmiennych, jak i przy wyborze DELTA ważne są rozstrzygnięcia merytoryczne: jeśli z teorii wynika, że nie powinno być korelacji, a my uzyskujemy niewysoką korelację przy DELTA=0, należy spróbować obniżyć wielkość DELTA.

Ustawienie parametru delta na 0, sprawdzenie uzyskanej korelacji między czynnikami i porównanie macierzy wzoru czynników z wynikami rotacji VARIMAX często pozwala na ostateczne podjęcie decyzji co do sposobu rotacji. Wielu badaczy sugeruje rotację skośną jako naturalne podejście w analizie czynnikowej i dopiero wówczas, gdy korelacja między czynnikami jest nieduża, rotowanie metodą VARIMAX. Trzeba jednak pamiętać, że skorelowane czynniki mogą być trudniejsze

[13] Polega ona na podzieleniu przed rotacją ładunków czynnikowych dla każdej zmiennej przez pierwiastek kwadratowy z zasobu zmienności wspólnej tej zmiennej, a to w celu wyrównania wpływu zmiennych na położenie rotowanych czynników niezależnie od ich zasobu zmienności wspólnej.

w interpretacji; wymagają często teorii wyjaśniającej zaobserwowaną korelację między czynnikami. Ponadto, możliwość manipulowania parametrem DELTA jest przez niektórych traktowana jako nadmiar arbitralności w modelowaniu rzeczywistości. Często też analizę czynnikową i głównych składowych prowadzi się po to, by uzyskać ortogonalny układ zmiennych do dalszych analiz. Wówczas rotacja nieortogonalna nie jest rozwiązaniem pożądanym.

Od wersji 7,5 pakietu SPSS dostępna jest również rotacja skośna PROMAX, która polega na potęgowaniu (zazwyczaj do 4 potęgi, co wyznacza parametr KAPPA), ładunków czynnikowych uzyskanych w rotacji VARIMAX, a następnie wyliczeniu kąta między czynnikami o uproszczonym przez potęgowanie wzorze czynników. W tym wypadku korelacja między czynnikami jest więc pochodną prostej struktury czynników: ich najlepszego dopasowania do poszczególnych skupień zmiennych. Rotacja PROMAX cieszy się sporym uznaniem w literaturze za jej efektywność przy odkrywaniu nieortogonalnej struktury czynników leżących u podstaw korelacji między wskaźnikami.

W wyniku rotacji nieortogonalnej uzyskujemy nie jedną, lecz dwie macierze współczynników, opisujących związki między czynnikami i zmiennymi.

(1) Macierz wzoru czynników (*factor pattern matrix*) — zawiera ładunki czynnikowe, czyli standaryzowane współczynniki regresji pomiędzy każdą zmienną (jako zmienną zależną) a czynnikami (jako zmiennymi niezależnymi).

(2) Macierz struktury czynników — zawiera współczynniki korelacji liniowej pomiędzy zmiennymi a czynnikami: w pierwszej kolumnie mamy współczynniki korelacji pomiędzy pierwszym czynnikiem i każdą zmienną z osobna, w drugiej — pomiędzy drugim czynnikiem i każdą zmienną z osobna itd.

W sytuacji, gdy czynniki są skorelowane, współczynniki korelacji pomiędzy zmienną a każdym z czynników nie są równe standaryzowanym współczynnikom regresji pomiędzy zmienną a tymi czynnikami jako zestawem zmiennych niezależnych, gdyż współczynniki regresji uwzględniają wzajemną korelację zmiennych niezależnych, a współczynniki korelacji — nie. W sytuacji, gdy czynniki są ortogonalne, współczynniki korelacji są równe standaryzowanym współczynnikom regresji pomiędzy zmiennymi i czynnikami (ładunkom czynnikowym) i dlatego mamy do czynienia z jedną macierzą ładunków czynnikowych.

W analizie czynnikowej rotowanej skośnie (OBLIMIN, PROMAX) interesuje nas zwykle macierz wzoru czynników (*Pattern*) — zawierająca ładunki czynnikowe/współczynniki regresji — co wiąże się z przyczynowym charakterem interpretacji modelu czynnikowego. Różnice struktury obu macierzy nie są jednak zwykle istotne dla interpretacji. Są one tym większe, im silniej skorelowane są czynniki. W przypadku bardzo wysokiej ich korelacji możliwa jest sytuacja, że ładunki czynnikowe (w *Pattern Matrix*) będą w pewnych przypadkach niskie, a współczynniki korelacji (w *Structure Matrix*) wysokie; np. zmienna V ma niski ładunek i wysoką korelację z czynnikiem X i wysoki ładunek i wysoką korelację z czynnikiem Y. Taką sytuację należy rozumieć następująco:

a) zmienność czynnika X pokrywa się w znacznym stopniu ze zmiennością czynnika Y, gdyż są one silnie skorelowane;

b) czynnik Y wyjaśnia większą część wariancji zmiennej V niż czynnik X, przy kontroli wpływu pozostałych czynników;

c) czynniki X i Y reprezentują pewien wspólny wymiar, a ich wyodrębnienie w analizie może być wynikiem niekompletnego doboru wskaźników lub np. część wskaźników ma ambiwalentny charakter; zawsze w takiej sytuacji pojawia się problem z kwalifikowaniem wskaźników do jednej lub drugiej skali/czynnika i konieczne jest włączenie kryterium merytorycznej interpretacji (problem trafności pomiaru).

Macierz struktury czynników ujawnia nam związki pomiędzy zmiennymi a czynnikami, które mogą być zacierane w macierzy wzorów, w której ładunki są wyliczane przy charakterystycznym dla regresji wyłączaniu (kontroli) wpływu innych skorelowanych czynników. Musimy jednak brać pod uwagę to, że proste współczynniki korelacji mogą reprezentować związki pozorne, właśnie dlatego, że w ich przypadku nie jest kontrolowany wpływ pozostałych zmiennych (czynników) w modelu.

Zwykle w przypadku badań kwestionariuszowych zakładamy, że **czynniki przez nas uzyskane powinny być dobrze rozróżnione, powinny posiadać swoją specyfikę, dlatego też nie powinny być one zbyt silnie ze sobą skorelowane.** Sposobem na zaobserwowaną wysoką korelację nie jest jednak wymuszanie ortogonalności, lecz przemyślenie modelu teoretycznego i doboru wskaźników.

Niekiedy spotyka się opinię, że o ile rotacja jest naturalnym elementem analizy czynnikowej, o tyle w analizie głównych składowych rotacja nie jest zasadna. Nie jest to podejście słuszne. Zarówno doświadczenie badawcze, jak i studia symulacyjne pokazują, że **rotowanie głównych składowych w celu uzyskania klarownej ich interpretacji jest uzasadnione**. Główne składowe są po rotacji, podobnie jak czynniki, często łatwiejsze do interpretacji — a celem analizy danych jest przecież zrozumienie danych, a nie ich matematyczne przetworzenie. Także wówczas, gdy główne składowe obliczamy w celu zastosowania w dalszych analizach, rotacja często jest lepszym rozwiązaniem. Tak więc w analizie skupień (*cluster analysis*) użycie rotowanych „istotnych" składowych głównych (np. o wartościach własnych powyżej 1) prowadzi do lepszego odtworzenia struktury danych niż stosowanie wszystkich wyodrębnionych głównych składowych (Bacher 1996: 194–198). Rotacja głównych składowych może też poprzedzać ich użycie w analizie regresji[14]. Takie podejście zbliża analizę głównych składowych do analizy czynnikowej, nie zacierając jednak ich formalnych różnic między tymi technikami.

Po rotacji można przystąpić do interpretacji uzyskanego modelu. W przypadku właściwej analizy czynnikowej nie powinno się interpretować czynników nierotowanych, wobec niejednoznaczności uzyskiwanego rozwiązania. W przypadku PCA interpretacja nierotowanych składowych jest możliwa i niekiedy właściwsza, rotacja zwykle jednak przynosi rozwiązanie łatwiejsze do interpretacji.

WYLICZANIE WARTOŚCI CZYNNIKOWYCH

Po wykonaniu rotacji możemy wyliczyć **wartości czynnikowe** — *factor scores* (w przypadku PCA można pominąć fazę rotacji). W pakiecie SPSS do szacowania wartości czynnikowych służą trzy metody: regresyjna, Bartletta i Andersona-Rubi-

[14] „Jeżeli główne składowe są nieinterpretowalne, wówczas możemy rotować zatrzymane składowe przed użyciem ich w regresji" (Dunteman 1994: 215).

na. W przypadku PCA wszystkie trzy metody obliczania wartości czynnikowych prowadzą do tych samych rezultatów, w przypadku FA — wszystkie prowadzą do pewnych kłopotów, gdyż wartości czynnikowe nie są jednoznacznie zdefiniowane. W wyniku wyboru opcji obliczania wartości tworzone są nowe zmienne, dodawane na końcu zbioru. Odpowiadają one poszczególnym czynnikom/głównym składowym. Zawierają (dla każdej obserwacji, w której nie ma braków danych) oszacowania wartości, które każda obserwacja uzyskała na wymiarze (skali) reprezentującej czynnik. Wartości czynnikowe wyliczane są przez pomnożenie wyliczonych przez program **współczynników wartości czynnikowych** (*factor score coefficients*) dla poszczególnych zmiennych przez te (standaryzowane) zmienne i dodanie do siebie wyników. Nowa zmienna jest więc kombinacją liniową wartości zmiennych, ważonych współczynnikami, określającymi wpływ poszczególnych zmiennych na wartość danego czynnika. Musimy pamiętać, że w przypadku właściwej analizy czynnikowej (FA) wartości czynnikowe są tylko oszacowaniem „prawdziwych" wartości czynników i, ze względu na właściwości tego modelu analizy, mogą być problematyczne. Dlatego w sytuacji, gdy chcemy używać wartości czynnikowych w dalszej analizie, lepiej jest skorzystać z analizy głównych składowych. W PCA wartości czynnikowe są wyliczane jednoznacznie, a nie szacowane. Składowe główne są liniowymi kombinacjami obserwowanych zmiennych, jednoznacznie określonymi[15].

WYKRESY ŁADUNKÓW CZYNNIKOWYCH I WARTOŚCI CZYNNIKOWYCH

Ładunki czynnikowe można przedstawić na wykresie rozrzutu (2W lub 3W). Osie układu współrzędnych reprezentują czynniki. Współrzędne punktów reprezentujących zmienne wyznaczone są przez ładunki czynnikowe. Skupienia zmiennych na wykresie wskazują na ich relatywnie silniejsze związki pomiędzy sobą. Często używa się strzałek, by połączyć punkty oznaczające zmienne z początkiem układu współrzędnych. Musimy zawsze pamiętać, że oglądamy obraz uproszczony, w którym sąsiedztwo punktów na wykresie 2W może być wynikiem „uproszczenia rzeczywistości" i zrzutowania punktu leżącego daleko, na niewidocznym wymiarze, na analizowaną płaszczyznę. Dotyczy to zwłaszcza punktów leżących bliżej centrum, czyli początku układu współrzędnych. Pewność naszego wnioskowania zależy od jakości modelu, mierzonej odsetkiem wyjaśnionej wariancji lub testem dobroci dopasowania. Jakość reprezentacji każdej zmiennej na dwuwymiarowym wykresie, opartym na dwóch pierwszych czynnikach/składowych opisana jest jej zasobem zmienności wspólnej (*communality*) oszacowanym (jednoznacznie wyliczonym w PCA) dla modelu opartego na dwóch pierwszych czynnikach.

Wykresy można również sporządzać korzystając z wartości czynnikowych. Umieszczamy wówczas na wykresie rozrzutu, którego osie reprezentują czynniki, punkty reprezentujące poszczególne obiekty (obserwacje). Punkty leżące blisko siebie stanowią skupienia podobnych obiektów. Jest to stwierdzenie tym bardziej prawdziwe, im większy odsetek wariancji wyjaśniają dwie pierwsze składowe, które de-

[15] W przypadku ortogonalnych głównych składowych współczynniki wartości czynnikowych otrzymuje się przez podzielenie ładunków czynnikowych przez wartość własną czynnika; to dzielenie wykonuje się po to, by uzyskać wartości czynnikowe znormalizowane tak, żeby wariancja wyliczonej zmiennej była równa 1.

finiują nasz wykres. W przypadku bazy danych złożonych z dużej liczby obserwacji, trudno przedstawić je w komplecie na wykresie. Wylicza się więc średnie z wartości czynnikowych dla wybranych segmentów (np. wykształcenia) i lokuje na wykresie te segmenty, posługując się średnimi wartościami czynnikowymi jako współrzędnymi. Jest to standardowa technika pozycjonowania.

Można ładunki czynnikowe zmiennych i wartości czynnikowe obiektów umieścić na jednym wykresie. Wymaga to wykonania uprzednio dość prostych zabiegów związanych z przygotowaniem wspólnej bazy danych zawierającej ładunki i wartości czynnikowe na dwóch pierwszych czynnikach oraz zmiennej odróżniającej jedne od drugich. Następnie wykonuje się wspólny wykres rozrzutu. Należy jednak pamiętać, że interpretacja odległości pomiędzy punktami na tym wykresie jest uprawniona tylko odrębnie w zbiorze zmiennych i odrębnie w zbiorze przypadków. Oba te zbiory należą do odrębnych przestrzeni: ładunków i wartości czynnikowych, których wspólnym elementem są osie układu reprezentujące czynniki. Dlatego też używamy punktów (strzałek) reprezentujących zmienne do interpretacji znaczenia wymiarów/osi układu współrzędnych, a następnie interpretujemy położenie punktów oznaczających przypadki (segmenty) względem tych zinterpretowanych wymiarów. Jest to technika powszechnie używana w pozycjonowaniu i eksploracyjnej analizie danych.

ROZMIARY ZBIORU DANYCH I DOBÓR ZMIENNYCH

Ile przypadków musi być w bazie danych, żeby przeprowadzić analizę czynnikową i składowych głównych

Minimum musimy mieć o jeden przypadek więcej niż wynosi liczba zmiennych. Analizę głównych składowych prowadzi się dla takich niedużych macierzy danych, by odkryć ich strukturę i zredukować do minimum (2 lub 3) wymiarów, w celu prezentacji graficznej. Analizie czynnikowej zasadniczo nie powinno się poddawać prób mniejszych niż 50 przypadków, a jeszcze lepiej, by miały 100 lub więcej przypadków. Reprezentanci bardziej ostrożnego podejścia mówią, że powinniśmy mieć cztery do pięciu razy więcej przypadków niż zmiennych, mniej konserwatywni zadowalają się stosunkiem 2:1. Dyskusje dotyczące wielkości próby dotyczą zwłaszcza metody największej wiarygodności; w tym wypadku sugeruje się, że liczba przypadków powinna być o 51 większa od liczby zmiennych. Można podać wzór:

$$N - n - 1 >= 50$$

gdzie: N - wielkość próby
n - liczba zmiennych

Niektórzy badacze (np. Thurstone) sugerują, że powinniśmy mieć przynajmniej po trzy zmienne na każdy czynnik, tzn. ładujące istotnie tylko ten czynnik. Jest to formułowane jako wystarczający warunek identyfikacji czynnika (Bacher 1990: 120). Dość powszechna zgoda panuje co do tego, że powinniśmy mieć co najmniej dwa razy więcej zmiennych niż czynników (por. Kim & Mueller 1994b: 144-145; Hair, Anderson & Tatham 1984: 237).

Wpływ doboru zmiennych na wyniki analizy

Na wyniki uzyskane w analizie czynnikowej i analizie głównych składowych ma wpływ dobór zmiennych do analizy. W przypadku próby z szerszej populacji korelacja może wystąpić nawet pomiędzy tymi zmiennymi, które w populacji nie są skorelowane. Im więcej zmiennych używamy w analizie, tym większe jest prawdopodobieństwo, że w próbie losowej przypadkowo uzyskamy istotne korelacje nawet pomiędzy oryginalnie nieskorelowanymi zmiennymi, a to wpłynie na wyniki analizy czynnikowej i PCA. Należy więc dobierać do analizy takie zmienne, co do których mamy merytoryczne podstawy, by oczekiwać, że będą skorelowane z grupą innych zmiennych i będą wspólnie z nimi definiowały jakiś interpretowalny czynnik. Nawet przy takim podejściu zdarzają się różne niespodzianki (czasami o bardzo twórczych konsekwencjach), łatwiej jednak ustrzec się błędu interpretacji przypadkowych związków jako teoretycznie ważnych lub błędu nieuwzględnienia istotnych związków między zmiennymi. Analiza czynnikowa, jak cała statystyczna analiza danych, nie chroni automatycznie przed błędami i wymaga namysłu oraz starannej specyfikacji modelu. To skłania niektórych praktyków analizy czynnikowej do preferowania analizy konfirmacyjnej. Jednak ta ostatnia nie jest także wolna od problemów związanych z niejednoznacznością rozwiązania czynnikowego i możliwością dopasowania do danych wielu alternatywnych modeli.

Wstępna ocena przydatności danych do analizy czynnikowej

W analizie czynnikowej dostępne są także statystyczne techniki wspomagające wstępną selekcję zmiennych i ocenę przydatności macierzy korelacji do przeprowadzenia analizy czynnikowej. Takim narzędziem jest przede wszystkim *Keiser-Meyer-Olkin measure of sampling adequacy* — **KMO**. Służy on ocenie, na ile daną macierz korelacji można uznać za produkt oddziaływania wspólnych czynników, odnosząc współczynniki korelacji między zmiennymi (pożądane jest, by były wysokie, pomiędzy zmiennymi, na które działa wspólny czynnik) do cząstkowych współczynników korelacji między nimi (jeśli obserwowane korelacje między zmiennymi są wynikiem oddziaływania wspólnego czynników, wówczas korelacje cząstkowe pomiędzy tymi zmiennymi powinny być niskie). Współczynnik KMO można obliczyć dla całej macierzy korelacji. Im bliższa 1 jest jego wartość, tym lepiej model czynnikowy nadaje się do wyjaśnienia struktury danej macierzy korelacji. Keiser[16] wskazuje następujące dolne progi wartości KMO:

a) 0,9 — wspaniały
b) 0,8 — godny pochwały
c) 0,7 — niezły
d) 0,6 — przeciętny
e) 0,5 — nędzny
f) poniżej 0,5 — nie do przyjęcia.

Jeśli macierz korelacji ma niski współczynnik KMO, należy rozważyć sensowność użycia analizy czynnikowej. Współczynnik KMO może zostać wyliczony również dla każdej zmiennej. Jeśli zmienna uzyska niski KMO, należy rozważyć usunięcie jej z analizy. Współczynniki KMO dla zmiennych są umieszczone na przekątnej macie-

[16] Cyt. za: Maria Norusis, SPSS Professional Statistics 6.1, s. 52.

rzy ***Anti-image correlation matrix***. Nawiasem mówiąc, elementy tej macierzy, poza przekątną, to pomnożone przez -1 wartości korelacji cząstkowych pomiędzy zmiennymi[17]. Jeśli zmienne pozostają pod wpływem wspólnych czynników, wówczas ich korelacje cząstkowe powinny być bliskie 0. Duży odsetek wysokich wartości korelacji cząstkowych każe rozważyć sensowność modelu czynnikowego dla danej macierzy korelacji.

NA MARGINESIE: UŻYCIE GŁÓWNYCH SKŁADOWYCH W ANALIZIE REGRESJI

Głównych składowych używa się w analizie regresji w celu poradzenia sobie ze zjawiskiem wielowspółliniowości zmiennych niezależnych lub w celu uproszczenia analizy i interpretacji wyników. Możemy wprowadzić wszystkie nieskorelowane główne składowe — współczynniki korelacji między każdą z nich a zmienną zależną są równe standaryzowanym współczynnikom regresji (beta) pomiędzy każdą ze składowych a zmienną zależną. Możemy wprowadzić część głównych składowych, kierując się przy ich doborze poziomem korelacji ze zmienną zależną (zazwyczaj pierwsze składowe są najsilniejszymi predyktorami, zmiennej zależnej, ale nie zawsze. Przed użyciem w analizie regresji składowe główne można poddać rotacji w celu ułatwienia interpretacji wyników.

ZAKOŃCZENIE

Analiza czynnikowa i analiza głównych składowych to najpowszechniej stosowane techniki analizy wielowymiarowej. Są sprawdzonymi i dobrymi narzędziami, pod warunkiem dobrego zrozumienia, czego możemy od nich oczekiwać i jak je stosować. Wiele wyborów dokonywanych przez badacza ma charakter arbitralny. Z drugiej strony, jak to zauważyliśmy, analiza czynnikowa daje podobne rezultaty przy różnych metodach wyodrębniania czynników oraz podobne do analizy głównych składowych. W selekcji i interpretacji czynników ważne jest doświadczenie analityka i merytoryczna znajomość problemu. Najgorszym podejściem jest wkładanie do analizy czynnikowej danych „na ślepo" i następnie wiara w uzyskane rezultaty. W tej metodzie również obowiązuje święta zasada analizy danych: **włożysz śmieci — wyjmiesz śmieci**. Podkreślam to, niezależnie od przekonania o fundamentalnej roli eksploracyjnej analizy danych w poznaniu rzeczywistości i dobrych doświadczeń z użytkowaniem na tym polu analizy czynnikowej i głównych składowych.

LITERATURA

Arminger G. 1979. *Faktorenanalyse*. Stuttgart: Teubner.
Bacher J. 1996. *Clusteranalyse*. Muenchen: Oldenbourg.

[17] Tzn. wyliczonych pomiędzy resztami pozostałymi po wyodrębnieniu z każdej z tych zmiennych wpływu pozostałych zmiennych.

Bacher J. 1990. *Einfuehrung In die Logik der Skallerungsverfahren*, „Historical Social Research", Special Issue, Vol. 15, No. 3., Koeln: Center for Historical Social Research.

Backhaus K., Erichson B., Plinke W., Weiber R. 1990. *Multivariate Analysemethoden*. Berlin: Springer.

Duntemann G.H. 1994), *Principal Components Analysis*. W: M.S. Lewis-Back, *Factor Analysis an Related Technics*. London: Sage: 157-145.

Grabiński T. 1992. *Metody taksonometrii*. Kraków: Akademia Ekonomiczna.

Hair Jr. J.F., Anaderson R.E., Tatham R.L. 1984. *Multivariate data Analysis with Readings*. London: Macmillan.

Holm K. 1976. *Die Befragung 3: die Faktorenalyse*. Muenchen: Francke Verlag.

Jajuga K. 1993. *Statystyczna analiza wielowymiarowa*. Warszawa: PWN.

Kim J.-O., Mueller Ch.W. 1994. *Introduction to Factor Analysis: What It Is and How to Do It*. W: M.S. Lewis-Back, *Factor Analysis an Related Technics*. London: Sage: 1-73.

Kim J.-O., Mueller Ch.W. 1994. *Factor Analysis: Statistical Methods and Practical Issues*. W: M.S. Lewis-Back, *Factor Analysis an Related Technics*. London: Sage: 75-155.

Norusis M. 1994. *SPSS Professional Statistics 6.1*. Chicago: SPSS Inc.

Wilkinson L., Grant B., Gruber Ch. 1996. *Desktop Analysis with SYSTAT*. Upper Saddle River: Prentice Hall.

Wilkinson L., Stenson H. 1996. *Factor Analysis*. W: *SYSTAT 6.0 for Windows: Statistics*. Chicago: SPSS Inc.

FACTOR ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

Both factor analysis and principal component analysis are very popular among social researchers. They are often treated as variants of the same method. In the paper author discusses difference between the methods, and typical use of them made for data analysis. Other topics discussed here are steps of the analysis, criteria of choosing factor extraction and rotation methods, and basic interpretation of results.