

二段抽出因子分析モデルにおけるデザイン効果の検討

前田 忠彦
統計数理研究所*

2001 年 11 月 20 日

概要

二段抽出因子分析モデルにおけるデザイン効果について検討した。ここでいうデザイン効果とは、データが由来する二水準の階層構造を無視して、一水準の（単純無作為抽出を仮定した）データとして因子分析を行った場合に、二水準の因子分析モデルでの分析と大きく異なる結果を得てしまうケースが生じる、というほどの意味である。この問題について、母集団レベルでの因子分析の「モデル」において、および現実の社会調査データの解析において、という二つの文脈において、事例研究的な検討を行った。前者については3変数単因子というきわめて限られた条件下ではあるが簡単な数値実験に基づく検討も行った。モデルの文脈の検討では、相関比が小さくなるにつれてデザイン効果の生じる可能性が低まることと、第一次抽出単位内の因子分析モデルで変量が因子に一樣に高く負荷するような好条件下ではデザイン効果が生じにくいことを指摘した。現実の社会調査データでは、二段抽出に対応した分析を行っても、一水準のモデルで分析した場合と類似した構造を見いだす例と、二段抽出モデルならではの分析事例の双方を示し、類似した構造が得られた理由が相関比の低さにあるとの可能性を指摘した。

1 はじめに

1.1 一つの問

次の性質は線形代数のテキストなどでも取り上げられ、よく知られている（例えば Harville, 1997）。

性質 1 2つの n 次正方行列 P_1 と P_2 が共に非負定値のとき、その和 $P_1 + P_2$ も非負定値である。

(証明省略)

もちろん、上記の性質を繰り返し適用することによって容易に ℓ 個の非負定値行列の和の場合に拡張することができる：

性質 1' ℓ 個の n 次正方行列 $P_1, P_2, \dots, P_\ell$ が全て非負定値のとき、それらの和 $\sum_{i=1}^{\ell} P_i$ もまた非負定値である。 ■

さて、因子分析モデルとは、一般に p 個の変量の母共分散行列 Σ に対して k 個の潜在変数（共通因子）を仮定した場合に、次式のような「共通因子分解」が与えられるとするものである：

$$\Sigma = \Lambda_k^* \Lambda_k^{*'} + \Theta^* \quad (\text{ただし } \Theta^* \text{ は正定値対角行列})$$

* 著者連絡先 e-mail: maeda@ism.ac.jp

$p \times k$ の行列 Λ_k^* は p 個の変量の k 個の因子の上への回帰における係数行列であり、因子負荷行列と呼ばれる。添え字 k は因子数を明示する目的で用いる。 Θ は残差分散行列に相当し、独自性行列と呼ばれる。上記の共通因子分解が成立することを母共分散行列が k 因子モデルに従う、と呼ぶ。

ここで因子分析モデルに関連して、前記性質 1 に類似の形式を持つ次のような問題を考えてみる。

問題 1 2つの共分散行列 Σ_1 と Σ_2 がそれぞれ k 因子モデル $\Sigma_1 = \Lambda_{k1}^* \Lambda_{k1}^{*'} + \Theta_1^*$ と $\Sigma_2 = \Lambda_{k2}^* \Lambda_{k2}^{*'} + \Theta_2^*$ にしたがうとき、それらの和 $\Sigma_1 + \Sigma_2 = \Sigma_T$ もまた k 因子モデルにしたがうか？

性質 1 の場合と同様にこれは容易に ℓ 個の行列の場合の間に拡張されることは言うまでもない¹。

この問題は必ずしも有名な問ではないようであるが、有名でない理由は明白で、答えが「常に Yes (k 因子モデルにしたがう) とは限らない」からである。答えが否定的であるならば、これは面白い問とは言えない。となれば問題 1 は、「最低限どのような条件が加われば Σ_T が k 因子モデルに従うことが保証されるか」という趣旨のものに書き替えられねばならないが、このように書き替えた間に一般的な解答を得るのは容易ではなさそうである。

1.2 数値例

問題 1 の答が否定的であることを示すために、 $p = 3, k = 1$ とした因子分析モデルの数値例を与えよう。因子分析の「モデル」では、共通因子分解を共分散行列 Σ_T に対して与えても相関行列 P に対して与えても同じことであるから、相関行列 P の共通因子分解 $P = \Lambda \Lambda' + \Theta$ を中心に考える。ここで「同じことである」とは、 Σ の対角要素の正の平方根すなわち各変数の標準偏差を要素とする対角行列を D として次のように両者の分解を対応づけられるの意味である：

$$(D\Lambda)(D\Lambda)' + D\Theta D = D(\Lambda\Lambda' + \Theta)D = DPD = \Sigma = \Lambda^* \Lambda^{*'} + \Theta^*$$

ここで $\Lambda^* = D\Lambda, \Theta^* = D\Theta D$ とおく。

3 変数の相関行列が一意の単因子共通因子分解を持つための条件およびその確率を Sato(1992) が与えている。その条件は直感的な言語表現に頼ると、相関行列の 3 つの非対角要素（全て正であるとする）の中に一つだけ飛び抜けて絶対値の小さなものがあるとはならない、とでも記述できるものである。正確には Sato(1992: Proposition 2.5) 参照。

数値例 1 さて

$$\begin{aligned} \lambda_a &= \begin{bmatrix} \lambda_{a1} \\ \lambda_{a2} \\ \lambda_{a3} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.8 \\ 0.8 \end{bmatrix}, & \Theta_a &= \begin{bmatrix} 0.96 & 0.00 & 0.00 \\ 0.00 & 0.36 & 0.00 \\ 0.00 & 0.00 & 0.36 \end{bmatrix} \\ \lambda_b &= \begin{bmatrix} \lambda_{b1} \\ \lambda_{b2} \\ \lambda_{b3} \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.8 \\ 0.2 \end{bmatrix}, & \Theta_b &= \begin{bmatrix} 0.36 & 0.00 & 0.00 \\ 0.00 & 0.36 & 0.00 \\ 0.00 & 0.00 & 0.96 \end{bmatrix} \end{aligned}$$

¹また問題 1 の変形として、 Σ_T が k 因子モデルに限定されない因子分析モデルに従うかという（すなわち因子数が変わっても良いとする）型の問題を考えることもできるが、とりあえずここではこの拡張は考えない。

とおき、簡単のため $D_a = 10I_3, D_b = 10I_3$ とする。このとき

$$\Sigma_a = \begin{bmatrix} 10.0 & 1.6 & 1.6 \\ 1.6 & 10.0 & 6.4 \\ 1.6 & 6.4 & 10.0 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 10.0 & 6.4 & 1.6 \\ 6.4 & 10.0 & 1.6 \\ 1.6 & 1.6 & 10.0 \end{bmatrix}$$

より,

$$\Sigma_T = \Sigma_a + \Sigma_b = \begin{bmatrix} 20.0 & 8.0 & 3.2 \\ 8.0 & 20.0 & 8.0 \\ 3.2 & 8.0 & 20.0 \end{bmatrix} \rightarrow P_T = \begin{bmatrix} 1.00 & 0.40 & 0.16 \\ 0.40 & 1.00 & 0.40 \\ 0.16 & 0.40 & 1.00 \end{bmatrix}$$

となる。この相関行列 P_T は一因子の共通因子分解を持たない (Θ_T が正定値にならない)。ちなみに P_T の固有値は大きい順に 1.651, 0.840, 0.509 である。

この P_T は共通因子分解を持つ場合と持たない場合のいわば境界に当たるケースであり、前記因子負荷ベクトル λ_a, λ_b の要素 0.8 vs. 0.2 を 0.9 vs. 0.1 のようにより強いコントラストを持つものに変えるならば、境界から離れて共通因子分解を持たない範囲に入る。

数値例 2 足し合わせる 2 つの行列の相関構造は例 1 と同じでも、両変数群の分散が変化させれば、 P_T が共通因子分解を持つ場合があることを示す。

例 1 と同じ $\lambda_a, \Theta_a, \lambda_b, \Theta_b$ の組み合わせに対し、 $D_a = 10I_3, D_b = 20I_3$ とする。

$$\Sigma_T = \Sigma_a + \Sigma_b = \begin{bmatrix} 30.0 & 9.6 & 4.8 \\ 9.6 & 30.0 & 14.4 \\ 4.8 & 14.4 & 30.0 \end{bmatrix} \rightarrow P_T = \begin{bmatrix} 1.00 & 0.32 & 0.16 \\ 0.32 & 1.00 & 0.48 \\ 0.16 & 0.48 & 1.00 \end{bmatrix}$$

となり $P_T = \lambda_T \lambda_T' + \Theta_T$ なる共通因子分解は

$$\lambda_T = \begin{bmatrix} \sqrt{\frac{0.32-0.16}{0.48}} \\ \sqrt{\frac{0.32-0.48}{0.16}} \\ \sqrt{\frac{0.16-0.48}{0.32}} \end{bmatrix} \simeq \begin{bmatrix} \sqrt{0.1067} \\ \sqrt{0.9600} \\ \sqrt{0.2400} \end{bmatrix} \quad \Theta_T \simeq \begin{bmatrix} 0.8933 & 0.0000 & 0.0000 \\ 0.0000 & 0.0400 & 0.0000 \\ 0.0000 & 0.0000 & 0.7600 \end{bmatrix}$$

により一意に定まる。

1.3 何が問題か

本稿では、二段抽出因子分析モデルについて扱う。二段抽出因子分析モデルとは、データが第 1 次抽出単位 (PSU: Primary Sampling Unit) の下に第 2 次抽出単位がネストした形で得られた、二段階の階層構造を持つデータ (二段抽出によるデータ) に対する因子分析モデルである。端的に言って PSU 間共分散行列 Σ_B と PSU 内共分散行列 Σ_W がそれぞれ因子分析モデルにしたがうとするモデルと考えて良い。他方二段抽出のデータの基本性質として全共分散行列 Σ_T は $\Sigma_T = \Sigma_B + \Sigma_W$ のように分解される。このことは、すなわち共分散行列の和により表現される全共分散行列に対して因子分析モデルが当てはまるかという問 1 のような素朴な疑問に対する回答をまずは検討しておく必要が生じることを意味する。

問1に対する回答は否定的であることは既に述べたが、本稿では Σ_T に対する因子分析と Σ_T の分解成分である Σ_B および Σ_W に対して同時に因子分析モデルを考えるということの間の関係について、より詳細に検討する。

2 問題～研究の背景と目的～

2.1 社会調査における標本設計と分析法

社会調査で通常採用される標本設計は層化二段無作為抽出である。教育関係の調査では学級等に対するクラスター（集落）抽出が用いられることも多い。これらの抽出方法では通常、単純無作為抽出によるよりも調査精度が悪くなる（標本誤差が大きくなる）。しかしながら、現実の応用的研究では、調査データを標本設計まで考慮して分析することが多いとは言えない。特に多変量解析の文脈では標本設計を考慮した分析の方法が必ずしも十分広く知られておらず、多くの分析事例が単純無作為抽出 (S.R.) の仮定下での分析に止まっている。実際には標本設計を考慮した分析の手続きが煩雑であることが多いため、全ての分析で標本設計を考慮した分析法を採用すべきとの主張は現実味を欠くが、それでは標本設計を無視して (S.R. と見なした) 分析がどの程度の誤判断に結びつきうるのか、いわば広義のデザイン効果を、多様な文脈で検討しておくことが重要になると考える。

2.2 広義のデザイン効果

標本設計の複雑さを無視し単純無作為抽出を仮定した分析を行うことに伴う（悪）影響は、大まかに言えば標本誤差の過小評価と要約できよう。例えば検定の文脈で言えば第一種の誤りの増加の方向に働く。しかし、特に多変量解析の文脈ではこの解釈は必ずしも直感的な明瞭さを持たず、標本設計を無視した分析がどのような悪影響を持ち得るのかを一言で表すことは困難である。本研究では広義の「デザイン効果」を、漠然と以下のような内容と考えておく。

広義のデザイン効果 (仮の定義) 「標本設計を無視してあたかも単純無作為抽出によって得たデータであるかのように分析した結果として、標本設計に合わせた適切な分析手法によって得られるべき本来の結果から偏りが生じる、あるいは重要な構造を見落とす、などの（悪い）影響が出ること」

いわゆる Multilevel Analysis の文脈では上記で「標本設計」を「データの階層構造」などと置き換えることによりほぼ同様の概念を想定することができよう。

デザイン効果が無視しうる程度の物であるか否かの判断はもちろん考慮するデータの性質と分析の目的に依存するが、一般的な社会調査において、どのような形の効果が存在するかについて検討しておくことは、その後の分析手法選択に当たって重要な参考情報を提供する。

そこで、このデザイン効果について次に説明する二つの文脈で事例に基づく検討を行う。

2.3 本研究の目的

本研究では因子分析について、特に数学モデルとしての因子分析の「モデル」の文脈および、具体的な標本に基づく母数の推定という文脈の双方で、広義のデザイン効果について、その性質を検討する。特に標本からの母数の推定すなわちデータ解析の文脈については、実際の調査データ

に基づきデザイン効果を検討し、標本設計を考慮した分析手法の効用を考察する材料とする。ただし本研究は考察材料の準備といった位置づけであり、広義のデザイン効果について何らかの結論的な言明を用意することを意図していない。

具体的には因子分析モデル（および因子分析によるデータ解析）において、どのようなデザイン効果が生じ得るかを、実例に基づき考察する。実用面からは層化二段抽出下での考察が重要であるが、多変量の文脈ではやや問題が複雑になるので、層別を考えない二段抽出（又は集落抽出）の状況のみを考える。

3 デザイン効果の定義

デザイン効果の概念を考察する際に、まず単変量的集計の文脈での狭義のデザイン効果の概念を導入しておくことは、本稿で用いている広義のデザイン効果の定義との対比の面で有効と考える。ここでは狭義のデザイン効果を定義した上で、広義のデザイン効果の概念を因子分析の適用状況に即して導入する。

3.1 狭義のデザイン効果～一般的定義～

狭義のデザイン効果は標本調査法の文献では良く知られた概念である（例えば Kish(1965) 参照）。本研究での広義のデザイン効果がこの用語の拡大解釈になっている。

狭義のデザイン効果は、ある標本設計下で母数の推定をした場合の推定量の分散の、同じ母数を同じ標本サイズの単純無作為抽出 (S.R.) の下で推定した場合の推定量の分散に対する比と定義される。すなわち、標本サイズ n 共通の条件下で、 θ を関心下の母数、 $\hat{\theta}_{DES}$ を当該 DESign の下での推定量、 $\hat{\theta}_{SR}$ を SR の下での推定量、 $V(\text{推定量})$ を推定量の分散として、

$$\text{デザイン効果} \quad \text{Deff}(\hat{\theta}_{DES}) \equiv \frac{V(\hat{\theta}_{DES})}{V(\hat{\theta}_{SR})}$$

により定義される。標本から推定する場合は、分母分子を適当な推定量で置き換える。

なお、社会調査データの単変量的集計の文脈における（狭義の）デザイン効果の検討事例は、前田・中村 (2000) を参照。

3.2 因子分析の文脈での広義のデザイン効果

以下話題を二段抽出のケースに限ろう。

先の狭義のデザイン効果では、二段抽出を採用したことにより推定したい母数（母平均など）の推定量の精度が（S.R. に比べて）どの程度落ちたかという“望ましくない効果の現出”の有無が、分散比によって評価されている。

本稿で考える広義のデザイン効果ではこの“望ましくない効果”の部分の推定量の分散の増大に限定せず、「データの階層構造を無視してあたかも単純無作為抽出によって得たデータであるかのように分析した結果として、（悪い）影響が出ること」を漠然と広く含めて考えるという主旨のことを先に述べた。

考えている分析手法は因子分析モデルであるから、このことは本稿では次のことを意味する：

二段抽出によるデータであることを無視して（全共分散行列 Σ_T に対して）因子分析を行うことによって、二段抽出に対応した適切なモデルに基づく分析結果で見出されたであろう重大な結果を見落とすことを「デザイン効果が生じた」と見なす。

ここで、S.R. を仮定した Σ_T の分析結果と対比すべき、二段抽出の構造を考慮した分析の結果とは何であるか？ また重大な結果を見落とすといったことの具体的な内実はどのようなことか？

これらについては、二段抽出に対応した因子分析のモデルをより具体的に導入した後に改めて取り上げる。

4 二段抽出因子分析モデル

本節では、二段抽出に対応した因子分析モデルの概略を述べる。本節の内容は、Muthén(1994) 流の二段抽出因子分析モデルを考えているが、もちろん共分散構造（一般）モデルを展開することが可能で、例えば Muthén & Muthén (1998) や、日本語ではほぼそれに依拠した豊田 (2000) を参照すればよい。本節の記述は、豊田 (2000) を参考にして、二段抽出に対応した一般共分散構造モデルの特別な場合としての二段抽出因子分析モデルを述べたものに過ぎない。

なお、Muthén (1994) は、変数としての因子にのみ階層構造を考えて因子負荷行列は PSU 間と PSU 内で共通と考える（階層構造に対応したパラメータ行列を考えない）ような、より制約的なモデルから論を起し本節で述べるモデルへ展開しているが、本稿では制約的なモデルについては一貫して考慮の外に置いている。

記号と用語 次の記号と用語を用いる。

PSU : 一次抽出単位（例：調査地点、学校）

($c = 1, \dots, C$)

SSU : 二次抽出単位（例：個人、生徒）

($j = 1, \dots, N_c$)

$N = \sum_c^C N_c$ は総標本サイズ

観測データの発生機構 x_{cj} を第 c 番目の PSU の j 番目の SSU に関する観測変数ベクトルとする。次のような観測モデルを考える。

$$x_{cj} = x_c^* + v_{cj}^* = \begin{bmatrix} x_c \\ y_c \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ v_{cj} \end{bmatrix} \quad (1)$$

x_c^* : PSU の特徴を記述する変数

v_{cj}^* : PSU 内の SSU の特徴を記述する変数

x_c : PSU に対し定義される変数

v_{cj} : SSU に対し定義される変数 ($E[v_{cj}] = \mathbf{0}$)

y_c : PSU に対し定義される v_{cj} の平均

n_W : v_{cj} , y_c のサイズ

n_B : x_c のサイズ

n : x_{cj} のサイズ ($n = n_W + n_B$)

ただし，本稿では $\mathbf{x}_c$ のような変数は具体的には登場せず，(1) 式の確率ベクトルの下側部分の要素のみが扱われる。

母平均 (1) 式の母平均は次の通りである：

$$\text{母平均 } \boldsymbol{\mu} = E[\mathbf{x}_{cj}] = E[\mathbf{x}_c^*] + \mathbf{0}.$$

共分散行列 PSU 間共分散行列と PSU 内共分散行列の定義，およびそれに基づく総共分散行列の分解が次のように与えられる。総共分散行列の分解は一変量の場合の二段抽出に於ける総分散の内分散と外分散への分解に対応するものである。

- PSU 間共分散行列 $\boldsymbol{\Sigma}_B$

$$\boldsymbol{\Sigma}_B = E[(\mathbf{x}_c^* - \boldsymbol{\mu})(\mathbf{x}_c^* - \boldsymbol{\mu})'] = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_c} & \boldsymbol{\Sigma}'_{\mathbf{y}_c \mathbf{x}_c} \\ \boldsymbol{\Sigma}_{\mathbf{y}_c \mathbf{x}_c} & \boldsymbol{\Sigma}_{\mathbf{y}_c} \end{bmatrix}$$

- PSU 内共分散行列 $\boldsymbol{\Sigma}_W$

$$\boldsymbol{\Sigma}_W = E[\mathbf{v}_{cj}^* \mathbf{v}_{cj}^{*'}] = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Sigma}_v \end{bmatrix}$$

$$\text{ただし, } \boldsymbol{\Sigma}_v = E[\mathbf{v}_{cj} \mathbf{v}_{cj}']$$

- 総共分散行列 $\boldsymbol{\Sigma}_T$ の分解

$$\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$$

4.1 標本共分散行列の期待値と級内相関の推定

母集団（モデル）での PSU 間，PSU 内共分散行列の標本における対応物は，それぞれ次のように定義される。標本に基づく $\boldsymbol{\Sigma}_W$ と $\boldsymbol{\Sigma}_B$ 等の推定量も合わせて与える。

- PSU 内の標本共分散行列

$$\mathbf{S}_{PW} = \frac{1}{N - C} \sum_c^C \sum_j^{N_c} (\mathbf{x}_{cj} - \overline{\mathbf{x}}_c)(\mathbf{x}_{cj} - \overline{\mathbf{x}}_c)'$$

$$E[\mathbf{S}_{PW}] = \boldsymbol{\Sigma}_W$$

- PSU(平均) 間の標本共分散行列

$$\mathbf{S}_B = \frac{1}{C - 1} \sum_c^C N_c (\overline{\mathbf{x}}_c - \overline{\mathbf{x}})(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})'$$

$$E[\mathbf{S}_B] = \boldsymbol{\Sigma}_B + \tilde{m} \boldsymbol{\Sigma}_W$$

$$\text{ただし, } \tilde{m} = \left[N^2 - \sum_c^C N_c^2 \right] [N(C - 1)]^{-1}$$

- Σ_W と Σ_B の標本に基づく推定

$$\begin{aligned}\widehat{\Sigma}_W &= S_{PW} \\ \widehat{\Sigma}_B &= (S_B - S_{PW})/\tilde{m}\end{aligned}$$

- x_k に対する級内相関（相関比）の推定

$$\eta_k^2 = \frac{[\widehat{\Sigma}_B]_{kk}}{[\widehat{\Sigma}_B]_{kk} + [\widehat{\Sigma}_W]_{kk}}$$

以上の結果は因子分析に限らず、二段抽出における多変量の観測について一般に成り立つ性質である。以下の項で因子分析モデルが導入される。

4.2 因子分析の測定モデル

二段抽出に対応した因子分析モデルでは、 $\mathbf{x}_c^*$ と $\mathbf{v}_{cj}$ に対してそれぞれ因子分析的な測定モデルが導入される。測定モデルに対する仮定についても合わせて要約して示すと以下の通りである。

- $\mathbf{x}_c^*$ に対する測定モデル

$$\mathbf{x}_c^* = \boldsymbol{\nu}_B + \boldsymbol{\Lambda}_B \boldsymbol{\xi}_{B_c} + \boldsymbol{\epsilon}_{B_c} \quad (2)$$

ただし、

$$\begin{aligned}\boldsymbol{\nu}_B &: \text{切片} \\ \boldsymbol{\xi}_{B_c} &: \text{PSU 間変動に対する説明因子} \\ \boldsymbol{\Lambda}_B &: \text{因子負荷行列} \\ \boldsymbol{\epsilon}_{B_c} &: \text{誤差項}\end{aligned}$$

更に以下を導入:

$$E[\boldsymbol{\xi}_{B_c}] = \boldsymbol{\alpha}_B, \quad V[\boldsymbol{\xi}_{B_c}] = \boldsymbol{\Psi}_B, \quad V[\boldsymbol{\epsilon}_{B_c}] = \boldsymbol{\Theta}_B$$

- $\mathbf{v}_{cj}$ に対する測定モデル

$$\mathbf{v}_{cj} = \boldsymbol{\Lambda}_W \boldsymbol{\xi}_{W_{cj}} + \boldsymbol{\epsilon}_{W_{cj}} \quad (3)$$

ただし、

$$\begin{aligned}\boldsymbol{\xi}_{W_{cj}} &: \text{PSU 内変動に対する説明因子} \\ \boldsymbol{\Lambda}_W &: \text{因子負荷行列} \\ \boldsymbol{\epsilon}_{W_{cj}} &: \text{誤差項}\end{aligned}$$

更に以下を導入:

$$V[\boldsymbol{\xi}_{W_{cj}}] = \boldsymbol{\Psi}_W, \quad V[\boldsymbol{\epsilon}_{W_{cj}}] = \boldsymbol{\Theta}_W$$

4.3 期待値・共分散構造

前項のような測定モデルの仮定下で、PSU 間の期待値および共分散構造と PSU 内の共分散構造が下記のように導出される：

- PSU 間の期待値・共分散構造

x_c^* (あるいは x_c の) 期待値構造

$$\mu(\theta) = \nu_B + \Lambda_B \alpha_B \quad (4)$$

x_c^* の共分散構造

$$\Sigma_B(\theta) = \Lambda_B \Psi_B \Lambda_B' + \Theta_B \quad (5)$$

- PSU 内の共分散構造

$$\Sigma_W = \begin{bmatrix} O & O \\ O & \Sigma_v(\theta) \end{bmatrix} = \begin{bmatrix} O & O \\ O & \Lambda_W \Psi_W \Lambda_W' + \Theta_W \end{bmatrix} \quad (6)$$

4.4 パラメータ推定の方法

本稿ではデータに基づくパラメータの推定の方法について詳述する余裕はない。Full Information の最尤推定量 (FIMLE) も提案されているが、本稿では推定法として Muthén (1994) の提案による ML-Based の簡便推定量 (MUMLE) を利用することにする。これは PSU 内の SSU の数が全て等しいいわゆる balanced case では FIMLE に一致することが知られている。

MUMLE について MUMLE のアイデアの要点を示すと下記の通りである。

- 以下の二つの項（尤度関数）を結合利用する。
 - f_W : (6) 式の構造と S_{PW} に基づく適合度関数
 - f_B : (4) 式と $\bar{x}$, 及び S_B と $\Sigma_{B*} = \Sigma_W + \tilde{m} \Sigma_B$ の関係に基づく適合度関数
- 結局次の適合度関数を最小化することに等しい：

$$f_{\text{MUMLE}} = C f_B + (N - C) f_W$$

この尤度関数は、多母集団の同時分析モデルの枠組みで表現可能な形に構成されていることから、分析（最適化）の実装は、既存の共分散構造分析のソフトウェアによって可能である。

なお、本稿で示すデータ解析例ではいずれも MPlus(Muthén & Muthén, 1998) を利用した推定値を示している。

4.5 二段抽出因子分析の実際

Muthén (1994) は、二段抽出モデルは複雑な分析方法なので、データ解析に際して確固たる Strategy が必要であることを指摘し、共分散構造モデル（因子分析モデル）の場合は、実際の二段抽出分析を行う前に、以下のような4つのステップを踏む予備的な分析を行うことを推奨している。

1. 通常の因子分析モデルによる Σ_T の分析. ← 当てはまりそうなモデルの探索
2. 級間変動の推定 ← 二段抽出分析を行うに足るデータか？
本来は $\Sigma_B = \mathbf{O}$ の検定を行うべきだが、便法として、各変数の級内相関係数を推定してみる.
3. 級内の構造の推定 ← Σ_{PW} の通常の因子分析モデルによる分析.
これは二段抽出分析で Σ_B に構造を入れないモデルに相当する.
経験的には、二段抽出分析モデルと類似した結果を与える.
4. 級間の構造の推定: Σ_B に当てはまる構造の探索.
この部分が難しい.

4.6 因子分析の文脈での広義のデザイン効果 (続)

再び、デザイン効果をどう考え、定義するかという問題に戻ろう。

以下、因子分析の文脈で二段抽出のデザイン効果を主として次の意味で考えることにする。

二段抽出因子分析におけるデザイン効果 標本設計を無視して一段階の因子分析（全共分散行列

Σ_T に対する因子分析）を行った場合に、二段抽出に対応した分析を行った場合の PSU 内共分散行列 Σ_W に当てはまる因子分析モデルから大きな乖離が生じること

二段抽出因子分析（あるいはより広い文脈では二段階の階層構造を持つデータ、二水準データの分析）では PSU 内 (共) 分散と PSU 間 (共) 分散との分離に興味がある。上記の定義で PSU 内共分散行列 Σ_W の側の結果にだけ着目するのは、 Σ_T の分析が個人（第2次抽出単位）レベルの変動をモデル化したことに対応させたわけである。異論の余地があるかも知れない。本稿ではとりあえずそのように見なすことにする。

結果の乖離という場合に、例えば次のようないくつかの水準での乖離を考えるべきであろう：

1. 二段抽出モデルでは分析ができて、 Σ_T の分析ではそもそも因子分析モデルが当てはまらなくなるケース。
2. 見いだされるべき因子を見落とす、とか、あるいは併合されるべきであった因子を別ものと考えてしまうなどの影響がでるケース
3. 因子の数といったマクロな構造ではなく、たとえば因子に対する個々の変量の負荷の大小関係が変化する（つまりその因子に負荷する最も主要な項目が何であるかが変化する）といったややミクロな構造についての影響がでるケース
4. それらの複合

第1のようなケースが生じる可能性がゼロでないことは、本稿第1節の問題1に関する検討により明らかである。

以下の節では上記のように定義した二段抽出におけるデザイン効果が、どのような場面で生じ

るかについて、「モデル」と具体的な「データ解析」という二つの水準で検討する。

5 因子分析の「モデル」におけるデザイン効果の性質

5.1 検討の目的

因子分析においてはデータ解析（すなわち標本データに基づくパラメータ推定）の文脈で理論的な問題を検討するのがしばしば難しく、本稿の問題も同様である。そのような場合に、まずは母共分散行列に対する共通因子分解の問題で—つまり「因子分析モデル」そのもので—二段抽出のデザイン効果を考えることから始めることは、問題の見通しを与える意味でも有効である。

更に、二段抽出モデルの問題の複雑さを緩和するために、最も単純なケースすなわち 3 変数単因子のモデル（第 1 節の表記で $p = 3, k = 1$ ケース）で検討を始めることがふさわしい。

重複するが、 $p = 3, k = 1$ として、PSU 間共分散構造 (5) と PSU 内共分散構造 (6) に基づき、それらの相関構造を書き直しておこう：

$$\text{PSU 間相関} : \mathbf{P}_B = \boldsymbol{\lambda}_B \boldsymbol{\lambda}_B' + \boldsymbol{\Theta}_B \quad (7)$$

$$\text{PSU 内相関} : \mathbf{P}_W = \boldsymbol{\lambda}_W \boldsymbol{\lambda}_W' + \boldsymbol{\Theta}_W \quad (8)$$

ただし PSU 内の構造は、 $\mathbf{v}_{cj}$ に対する部分のみを取り出して改めて $\mathbf{P}_W$ と表記したものである。また第 1 節の表記では共分散行列の分解に際して因子負荷行列 $\boldsymbol{\Lambda}$ 等に * を付して相関行列の分解の場合と区別したが、ここでは * による区別を省略している点に注意されたい。

5.2 若干の考察

まず、以下の点を指摘し得る。

- $\boldsymbol{\Sigma}_B$ と $\boldsymbol{\Sigma}_W$ が相関構造として全く同じ因子分析モデルにしたがう場合は、デザイン効果は皆無である、すなわち $\boldsymbol{\Sigma}_T$ の分析も $\boldsymbol{\Sigma}_W$ の分析と同じ相関構造に導かれることは自明であろう。

なおこの性質は 3 変数単因子のケースに限定されない。

- PSU 間分散が PSU 内分散に比べて小さければ、すなわち相関比の値が小さい場合にも、やはりデザイン効果は生じにくいことが容易に予想される。

後者は、1 節で提示した例題 2 に端的な例を示したが、次項の数値実験により積極的に裏付けることができる。

5.3 数値実験：3 変数単因子のモデルでの検討

本項では、3 変数単因子のケースでの簡単な数値実験の結果を示し、どのようなケースでデザイン効果が生じるかを考える一つの材料とする。

実験概要 次のような計算を行う（実験というよりも単なるモデルに具体的な数値を当てはめた計算である）。

一般性を失うことなく、(8) 式の $\boldsymbol{\lambda}_W$ において $0 \leq \lambda_{W1} \leq \lambda_{W2} \leq \lambda_{W3}$ を仮定する。また $\boldsymbol{\lambda}_B$ の各要素も正とする。

また簡単のために、分散行列（共分散行列の対角部分）は PSU 間、PSU 内ともに単位行列の定数倍であると仮定し、PSU 内分散と PSU 間分散の比は 3 変数の間で共通とする。これにより、相関比 $[= \text{PSU 間分散} / (\text{PSU 内分散} + \text{PSU 間分散})]$ が 3 変数で一定と仮定したことになる。

具体的に λ_W と λ_B に数値を与え、かつ関連する PSU 内分散と PSU 間分散を与えることにより、 P_W から Σ_W が、 P_B から Σ_B が構成される。これらから $\Sigma_T = \Sigma_B + \Sigma_W$ および P_T を計算し、 P_T が一因子の共通因子分解 $P_T = \lambda_T \lambda'_T + \Theta_T$ を持つか否かを検討する。

この時次の二つのケースを「デザイン効果が生じた」ケースと見ることにする。

Case(1) P_T が一因子の共通因子分解を持たない

Case(2) P_T は一因子の共通因子分解を持つが、 λ_W と λ_T の差が大きい。

ここで Case(2) の「差が大きい」とは、仮に、少なくとも一組の対応する要素 λ_{Wi} と λ_{Ti} について、 $\lambda_{Ti} < 2/3\lambda_{Wi}$ あるいは $\lambda_{Ti} > \lambda_{Wi} + 1/3(1 - \lambda_{Wi})$ となる場合、と規約することにしよう。Case(3) は Case(1)(2) 以外であるが、上記の「デザイン効果」の定義はやや狭いものである。

次の計算を行う。

1. PSU 内分散と PSU 間分散の比を定める（この条件は相関比を定めることに等しい）。
2. λ_W を与える。ここでは $\lambda_{W3} = 0.9, \lambda_{W2} = (\lambda_{W1} + \lambda_{W3})/2$ と規約し、 λ_{W1} を 0.1~0.9 まで 0.1 刻みに変化させた 9 条件を用いる。
3. λ_B の 3 の個要素については、0.01~0.99 まで 0.01 刻みで数値を変化させ、 99^3 個の格子点に対して Case(1)~(3) までのいずれかを判定する。

結果 図 1~図 3 は、相関比を 1/2, 1/3, 1/5 とし、所与の λ_W の値ごとに上記の計算を行った結果の Case(1), Case(2) の発生率を λ_{B1} の値をグラフの横軸にとり、 $\lambda_{B2}, \lambda_{B3}$ に関する 99^2 の組み合わせ中で Case(1)(2) が生じた割合をプロットしたものである。

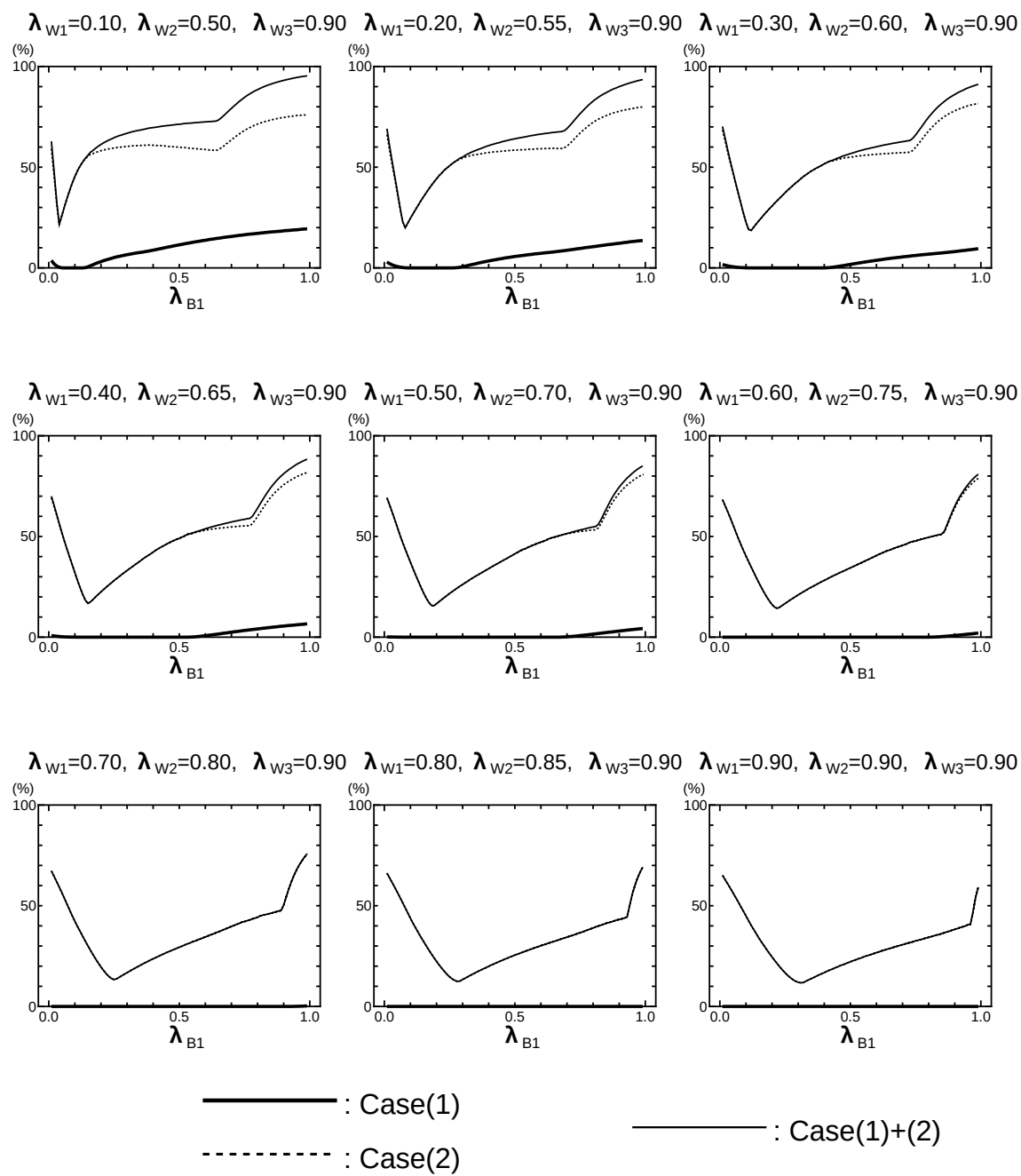
考察 上記結果について次の点を指摘できよう。

- 相関比が小さくなると、（予想通り）Case(1) や Case(2) のようにデザイン効果が生じる可能性が低まる（両ケースともに発生率は低くなっていく）。
- Case(1) は λ_{W1} の値が λ_{W3} に近づき、3 変数が PSU 内構造において一様に高い因子負荷を持つ傾向が強くなるにつれて、発生率が低まる。

後者の結果は、実用場面では PSU 内構造がはっきりとした一因子モデルを示す場合には、総共分散行列での分析でも一因子モデルによる分析は可能である場合が多くなることを示し、その意味で階層構造を無視した分析をした際にも当該因子を取り出せないというような「重要な構造の見落とし」が生じにくくなることを示唆している。それでもなお Case(2) のような推定値の大きな差が生じうることはある点にも留意は必要である。

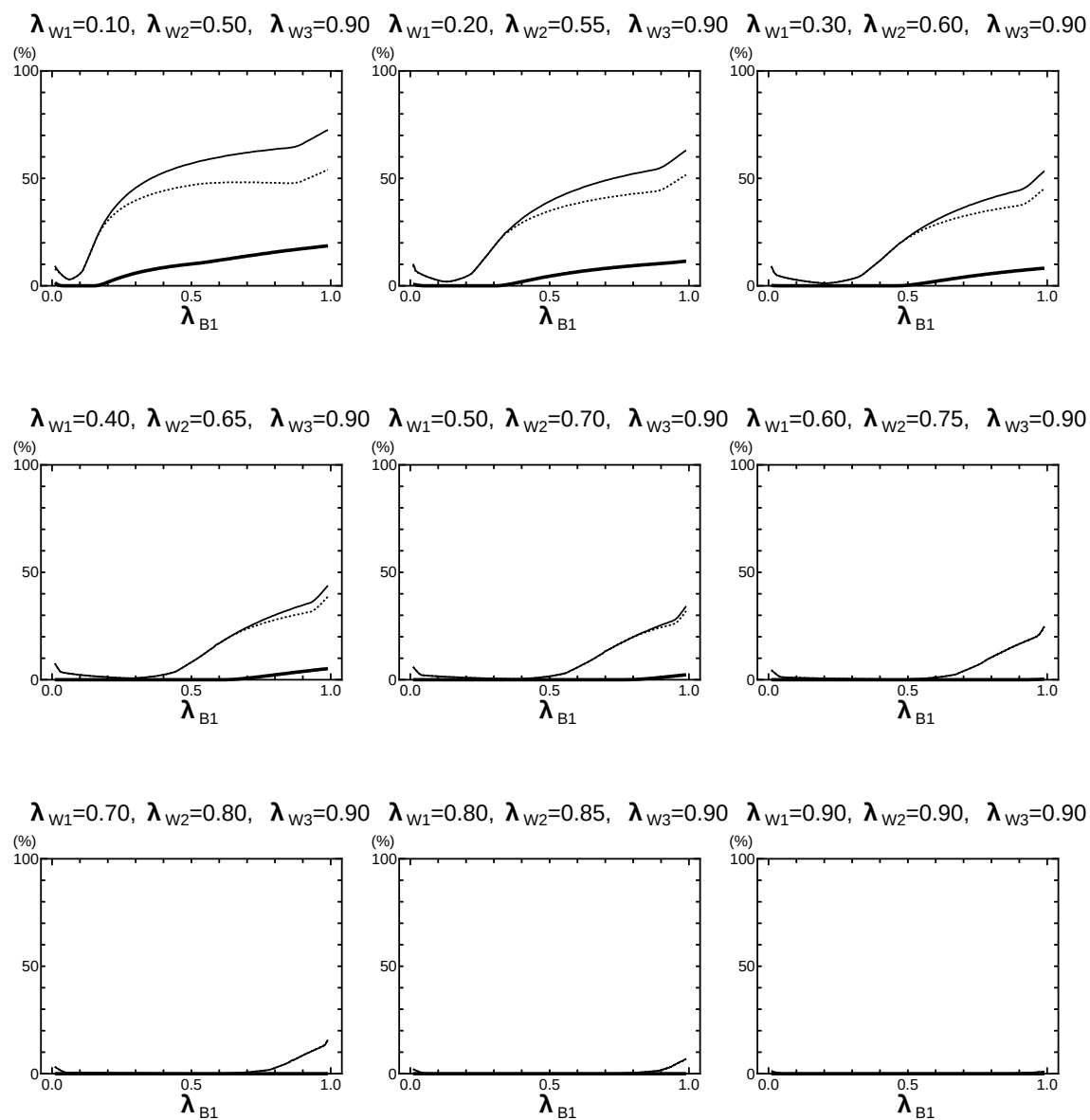
Case(2) については、本項の実験が現実のデータではあまり想定し得ないような値域についてまで因子負荷行列 λ_B の各要素の格子点上での振る舞いを調べている点は割引いて考える必要はある。

現実の社会調査データで相関比が 0.5 にもなる変数の例は稀で、おおかたは 0.3 程度までと考えられるから、少なくとも Case(1) のような意味でのデザイン効果の発生はあまり多くないことが想像される。



相関比=1/2

図 1. 3 変数単因子の状況におけるデザイン効果の出現率 (1)



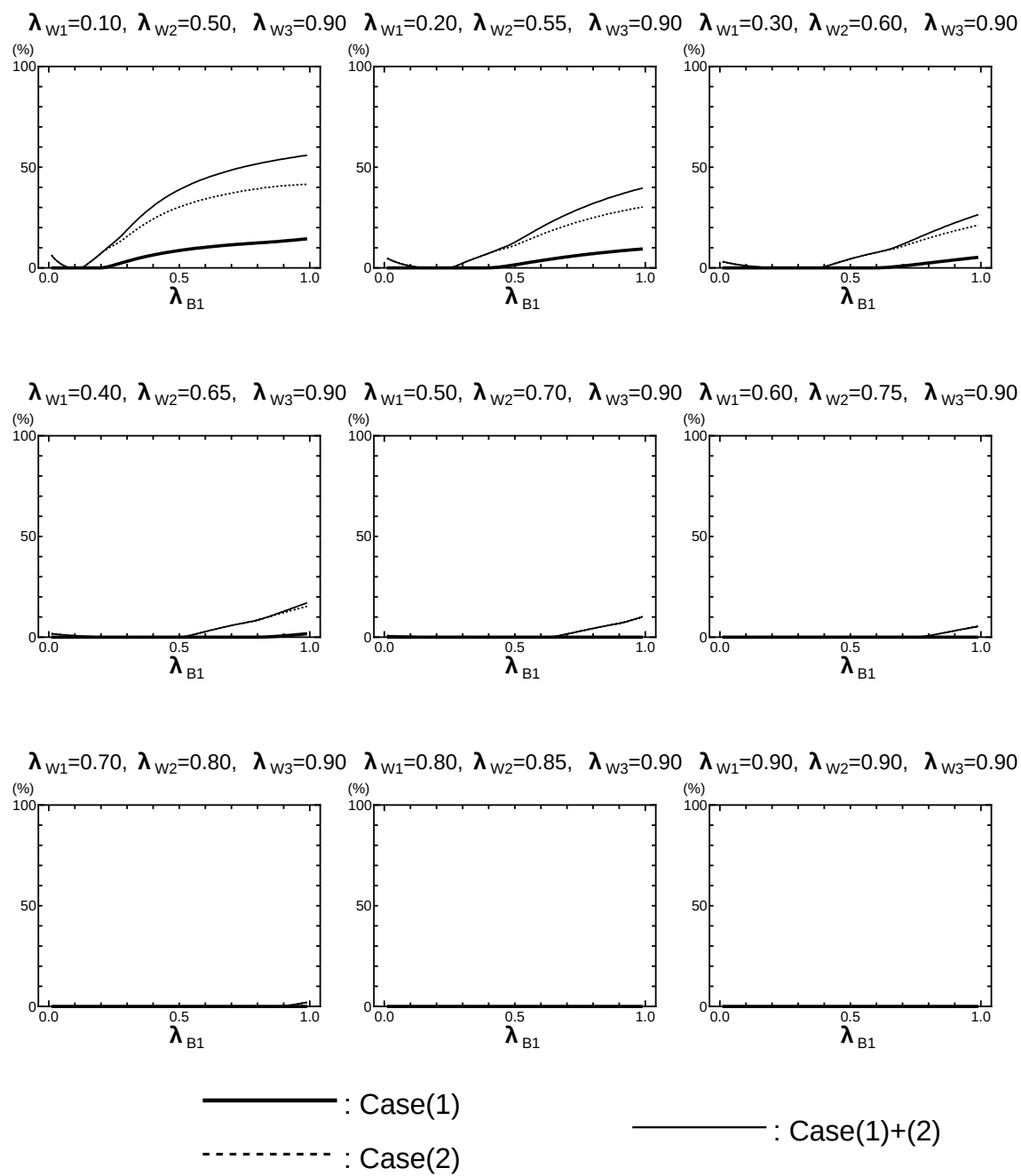
———— : Case(1)

----- : Case(2)

———— : Case(1)+(2)

相関比=1/3

図 2. 3変数単因子の状況におけるデザイン効果の出現率 (2)



相関比=1/5

図 3. 3変数単因子の状況におけるデザイン効果の出現率 (3)

6 データ解析における事例

本節では現実の調査データの解析例を示す。

6.1 二種類の使用データ

以下の二種類のデータを分析する。例示のために、変数数をともに3～4個に絞った小さなモデルでの結果のみ示す。 m を PSU の数, n を総サンプルサイズとする。

国民性調査データ 統計数理研究所による「日本人の国民性」調査（第9次全国調査）のデータから日本の水準を評価した3項目#9.12, #9.12c, #9.12d, 個人生活内での3領域での満足#2.3c, #2.3j, #2.3k と社会に対する満足感関連の2項目#2.3d, #7.40, を取り上げる。 $m = 300$, $n = 1905$ 。PSU はいわゆる二段抽出による第1次抽出単位（調査地点に相当）である。標本設計は便宜的に層別を考えない二段抽出と見なす。

進路指導用検査データ 水戸谷(1997)による。中学生の進路選択をめぐる意識や態度の成熟の程度を測る2種類計9尺度からなる検査データのうち、4つの尺度（尺度1, 8と尺度5, 9）を分析する。 $m = 90$, $n = 4235$ 。ここでのPSUは中学生のグループであるが、学級のような自然のまとまりではなく、データの都合上やや人為的に作成したクラスターである（その分相関比などは小さくなっているものと想像される）。具体的には15校の中学校内で性別×3学年で6グループを作成し、計90クラスターのデータと見なした。

6.2 単因子3変数のケース

国民性調査データの3変数の背後に単一因子が想定される分析事例を二つ示す。これは前節で「モデル」におけるデザイン効果の性質を考察した場合のデータ解析バージョンである。このケースはいわゆる飽和モデルであり、モデルの自由度が0になる。

表 1. 国民性調査データで3変数単因子分析の事例(1)

Between	推定値	標準誤差	標準解 1	標準解 2	決定係数	
#2.3c :家庭に満足か	1.000	0.000	0.155	0.941	0.885	
#2.3j :余暇に満足か	0.463	0.287	0.072	0.606	0.367	
#2.3k :健康状態に満足か	1.099	0.779	0.170	0.794	0.630	
Within	推定値	標準誤差	標準解 1	標準解 2	決定係数	Total に対する標準解 2
#2.3c :家庭に満足か	1.000	0.000	0.337	0.527	0.278	0.565
#2.3j :余暇に満足か	1.334	0.206	0.450	0.558	0.312	0.536
#2.3k :健康状態に満足か	0.987	0.136	0.333	0.406	0.165	0.438

個人生活関係の満足感3項目に関する標本相関行列の分析により得られる因子負荷の値（表の標準解2に相当）は、0.565, 0.536, 0.438であり、二段抽出モデルのPSU内構造（表のWithin）の因子負荷行列と大差ないが、#2.3cと#2.3jに対する負荷量の順序が入れ替わる程度の違いがある。

表 2. 国民性調査データで 3 変数単因子分析の事例 (2)

Between	推定値	標準誤差	標準解 1	標準解 2	決定係数	
#9.12 :科学技術の水準	1.000	0.000	0.115	0.721	0.520	
#9.12c:経済力	2.075	0.515	0.238	0.814	0.663	
#9.12d:生活水準	1.532	0.406	0.176	0.830	0.688	
Within	推定値	標準誤差	標準解 1	標準解 2	決定係数	Total に対する標準解 2
#9.12 :科学技術の水準	1.000	0.000	0.221	0.396	0.157	0.428
#9.12c:経済力	2.298	0.279	0.509	0.741	0.549	0.753
#9.12d:生活水準	1.663	0.177	0.368	0.584	0.341	0.613

日本の 3 つの側面に対する評価（#9.12 等）についての分析結果では、Within の因子負荷と Total の相関構造の分析での因子負荷は、推定値としては異なるものの相対的な順位の差がなく、この例も明確なデザイン効果は見られない例である。

6.3 2 因子 4 変数のケース

4 変数に対し 2 因子の存在が想定されるようなケースでは、問題が複雑になりうる—Between と Within で因子数が異なるモデルが想定できるなどのような—事例を示そう。

表 3. 国民性調査データでの主なパラメータの推定結果

Between	推定値	標準誤差	標準解 1	標準解 2	決定係数	
#9.12c:日本の経済力	1.000	0.000	0.220	0.750	0.562	
#9.12d:日本の生活水準	0.871	0.388	0.191	0.901	0.812	
#2.3d :社会に満足か	1.000	0.000	0.135	0.954	0.911	
#7.40 :社会は公平か	1.076	0.518	0.146	0.642	0.412	
因子間共分散	0.015	0.009	0.517	0.517		
Within	推定値	標準誤差	標準解 1	標準解 2	決定係数	Σ_T に対する標準解 2
#9.12c:日本の経済力	1.000	0.000	0.392	0.571	0.327	0.598
#9.12d:日本の生活水準	1.218	0.163	0.478	0.757	0.573	0.772
#2.3d :社会に満足か	1.000	0.000	0.394	0.534	0.285	0.554
#7.40 :社会は公平か	1.054	0.184	0.415	0.594	0.353	0.599
因子間共分散	0.067	0.012	0.433	0.433		

本節で示すのは 2 因子 4 変数のケースによる分析である。

6.3.1 分析結果

国民性調査データには、 Σ_W と Σ_B に同じ 2 因子モデルが当てはまると仮定して分析した。表 3 は主なパラメータの推定結果である。表下半右端の欄は、全共分散行列 Σ_T に対して Σ_W と同じ 2 因子モデルを当てはめた場合の完全標準解（標準解 2）である。一方進路指導用検査データに対しては、 Σ_W に対し 2 因子モデルを、 Σ_B に対しては 1 因子モデルを仮定した分析結果を示す

表 4. 進路指導用検査データでの主なパラメータの推定結果

Between	推定値	標準誤差	標準解 1	標準解 2	決定係数	
尺度 1:自己実現的態度	1.000	0.000	1.033	0.808	0.652	
尺度 8:進路選択	0.437	0.102	0.452	0.483	0.233	
尺度 5:進路情報	0.888	0.099	0.918	0.916	0.839	
尺度 9:進路先への適応	0.889	0.099	0.918	0.990	0.981	
Within	推定値	標準誤差	標準解 1	標準解 2	決定係数	Σ_T に対する標準解 2
尺度 1:自己実現的態度	1.000	0.000	4.706	0.719	0.516	0.734
尺度 8:進路選択	0.813	0.024	3.825	0.857	0.734	0.833
尺度 5:進路情報	1.000	0.000	3.501	0.823	0.677	0.829
尺度 9:進路先への適応	0.867	0.023	3.036	0.761	0.578	0.773
因子間共分散	9.279	0.433	0.563	0.563		

(表 4)。

表 3 と表 4 の Within 側の標準解 2 を Σ_T の標準解 2 と比較すると、数値が大まかに似ており、この二つの分析事例では 3.2 の意味でのデザイン効果はあまり認められないと言えよう。潜在変数側の分散のみ 1 に基準化した標準解 1 の結果を見ると、Within 側に想定した因子と Between 側に想定した因子の、観測変数の変動に対する相対的な寄与の程度を比較することができる。両データのうち進路指導用検査データで Between 側因子の寄与の程度が低い、このデータの相関比の全般的な小ささによって説明がつく。

6.4 考察

本研究で検討した因子分析の事例では、あまり明確なデザイン効果が見られなかったが、これは相関比（級内相関）が 0.1 前後と小さい変数が多かったためであろう。経験的にはこの程度の値では二段抽出の効果が多変量データ解析に際し明確には現れにくい、ということを示唆しているように思える。例えば学級を集落抽出した教育調査データでは、学級間の風土の違いによりデザイン効果が大きくなるケースが考えられる。この点については実際のデータで更に分析経験を積むと共に、人工データでの検討が必要とされるところであろう。

7 いくつかの課題

残された課題がいくつかある。まず、多変量データ解析一般の文脈では、狭義のデザイン効果を表す指標を考案・導入することである。多変量のデザイン効果の導入に際しては、値の大小が標本設計を無視した多変量分析を行うことの危険の程度（広義のデザイン効果の有無）もある程度鋭敏に予測することができれば有用な指標となり得よう。例えば PSU 間共分散行列 Σ_B と総共分散行列 Σ_T の一般化分散の比を取ることも考えられようが、その後に想定する多変量解析の手法によって個別の指標を考える必要性を検討すべきかも知れない。

因子分析については、多因子モデルのケースを考察しなければならない。また 5 節の数値実験例は、現実のデータにより近い負荷量の値の範囲について結果を精査する必要があるだろう。

文献

- Harville, D.A. (1997) *Matrix Algebra from a Statistician's Perspective*, Springer.
- Kish, L. (1965) *Survey Sampling*, Wiley.
- 前田忠彦 (2001). 二段抽出におけるデザイン効果の検証, 日本行動計量学会第 29 回大会発表論文抄録集, 354–357.
- 前田忠彦・中村 隆 (2000). 近年 5 回の国民性調査の標本設計と標本精度について, 統計数理, **48**, 147–178.
- Muthén, B.O. (1994) Multilevel covariance structure analysis, *Sociological Methods and Research*, **22**, 376–398.
- Muthén, L.K. and Muthén, B.O. (1998) *Mplus User's Guide*, Muthén & Muthén
- 水戸谷貞夫 (1997) 中学校における進路指導の改善・充実に関する実証的研究—生涯学習の基礎を培う視点に立って—, 平成 6,7,8 年度科学研究費補助金基盤研究 B 研究成果報告書, 財団法人日本進路指導協会.
- Sato, M. (1992) A study of an identification problem and substitute use of principal component analysis in factor analysis. *Hiroshima Mathematical Journal*, Vol.**22**, No.3, 479–524.
- 豊田秀樹 (2000) 共分散構造分析 [応用編] —構造方程式モデリング— 朝倉書店.