Technical Report of the 1st Workshop on Latent Dynamics
(Jun 22、 2011、 Tokyo)

多変量データとパネルデータの相関構造に関する注意と試み

椿　広計*

Abstract: 線形多変量相関構造については、潜在因子構造による記述と顕在因果構造による記述の２つの方針が排他的に用いられているが、その両者の統合については、ARMA モデリングや共和分解析などが定型化している時系列解析分野を除いてはあまり流布していない。偏残差プロットの意味や記述多変量解析としての主成分分析が必要な共分散構造近似戦略に対してもつ意味を明確にしなければならない。また、これらの同時モデリングが可能となる共分散構造モデリングにおいても、潜在因子の測定と潜在因子からの因果影響を区別していない状況がある。ここでは、これらの多変量モデリングにおける基本的問題について注意を喚起する。これら線形多変量モデリングにダイナミックスを導入する試みとしては、多変量繰り返し測定データ（パネルデータ）に対する水準と成長に対する潜在因子導入は、潜在成長曲線モデルとして知られているが、 これを因果構造モデルとリンクする試みも紹介する。

Keywords: Covariance Structure Analysis、 Latent Growth Curve Modeling

1　表線形と裏線形の可視化

量的変数間の連関性は、一般に散布図行列上に端的に表現されると考えられてきた。そして、散布図行列上の 2 変数間の関係が直線的であるならば、関係性の数値的表現として（全）相関係数行列 $\boldsymbol{R_T}$ を用いることが合理的と考えられてきた。この相関係数行列をできるだけ簡単な構造で近似するのが主成分分析や因子分析である。これに対して、偏相関係数行列 $\boldsymbol{R_P}$ が、Karl Pearson とその最初の弟子である Yule らによって、相関係数開発の直後(19 世紀末)に開発された。今日、それを更に積極的に連関の解釈に利用としたのがグラフィカルモデリングである。

p 次元観測変量ベクトル $\mathbf{Z}^{\mathbf{T}}=(X, Y, \mathbf{A}^{\mathbf{T}})$ に対して、通常の偏相関係数は、連関性評価の関心対象となっている X、Y 以外の全変数 **A** で調整した偏相関係数を指す。一方、X が Y に直接影響を与えているが、Y は X には影響は与えていないという、因果関係の方向性 $X \rightarrow Y$ が既知ならば、Y の予測値には、X の情報を使い、E[Y|X、 **A**]と考えるのが自然である。しかし、連関性の評価を行いたい一方の変数を調整に用いた瞬間に他の変数の残差からは、調整変数の情報は消去されてしまい、関連性は 0 となる。この種の関連性評価に拘るのならば、因果関係を無視して対称性を導入する、すなわち、X の予測値にも、Y の情報を使い、E[X|Y、 **A**]と考えれば良い。両偏残差間相関は、通常の偏相関係数の符号を逆転したものとなる（早稲田大学、永田靖氏指摘）。探索的状況では、他の全ての変数を眺めて当該変数の回帰関数（条件付き期待値）を評価することは、記述統計学的にも有用である。なぜならば、各変量毎に偏残差 X-E[X|Y、 **A**]が偏相関を計算する変量の組み合わせに依存せず、一意に定まり、偏残差散布図行列を表示できるからである。宮川[1]のデータに対して散布図行列と偏残差散布図行列を示したものが図 1a、図 1b である。

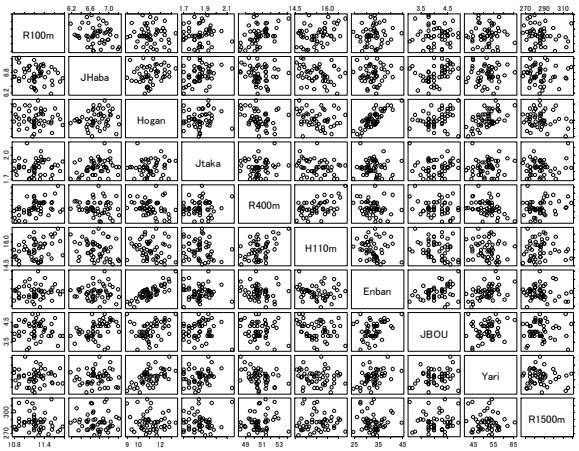

図 1a 宮川の近代 10 種競技データの散布図行列

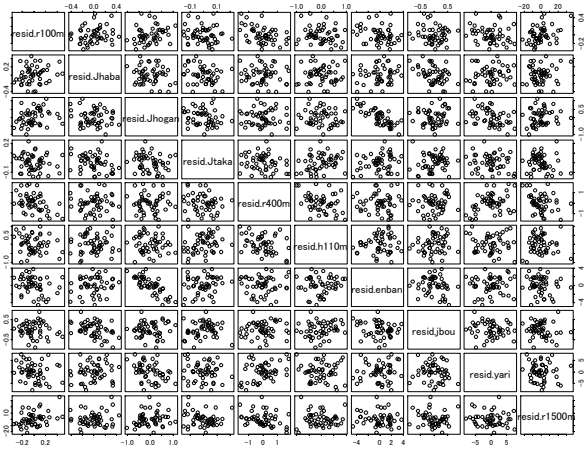

図 1b 近代 10 種競技データの偏残差散布図行列

Z が平均ベクトル **0**、共分散行列Σの最小 Fisher 情報量分布、すなわち多変量正規分布に従うとすると、対数確率密度関数は(1)となる。

$$\lambda(\mathbf{z}) = -\mathbf{z}^{\mathbf{T}}\Sigma^{-1}\mathbf{z}/2 + \log\left|\Sigma^{-1}\right|^{1/2} - \log(2\pi)^{1/2} \quad (1)$$

これから、確率分布としての「自然母数(Natural Parameter)」は、$\mathbf{zz^T}$ の期待値母数 $\mathbf{\Sigma}$ ではなくて、逆共分散行列-Σ^{-1}となり、この逆共分散行列の i 行 j 列要素を σ^{ij} と書くと、第 i 変数と第 j 変数の偏相関係数は、$-\sigma^{ij}/(\sigma^{ii}\sigma^{jj})^{1/2}$ となる。偏残差散布図行列は、この逆分散行列の可視化表現そのものである。なお、探索的共分散構造分析の第 2 段階は、散布図行列ないしは、偏残差散布図行列の世界で、外れ値の検討を含む直線性、等分散性をチェックすることである。

2 共分散の近似か逆行分散の近似か

主成分分析や主因子分析は、共分散行列あるいは相関係数行列を直接最良近似するために、「主成分」や「因子」といった潜在変量ベクトル **Q** を観測変量の背後に導入した。すなわち、**Z** を **Q** に回帰し、その残差ベクトルの共分散行列をできるだけ小さくしようとしたのである。q 次元潜在ベクトル **Q** の共分散行列をΩとし、**Z** と **Q** の共分散行列をBとすると、回帰残差の共分散行列は、$\Sigma - B\,\Omega^{-1}\,B^{T}$となる。例えば、Ωを単位行列と仮定し、通常の最小二乗基準で最小化すれば、Σは第 q 主成分までの固有空間を使って近似するのが最善と言う事になる。特に、「回転」という操作で達成される「単純構造(Simple structure)」とは、「検証的因子分析(Confirmatory Factor Analysis)」が興隆してきた 1980 年代後半以降、解釈の問題というよりは、潜在変量から観測変量へのパス係数のできるだけ多くを 0 にするという、ケチの原理の観点から考えるべきものとなった。

このように、古来主成分分析や因子分析が、この種の共分散構造あるいは相関構造を最小二乗近似する事に、強い意義があるかの如き心象をユーザーに植え付けてきた。これは、データの変動を要領よく近似するのが記述統計の果たすべき役割と考えれば自然な事である。従って、主成分分析などは、最小二乗的センスの累積寄与率評価が重要なものと認識されてきた。しかし、前節で述べたように共分散行列の逆行列である情報行列こそ自然なパラメータだとすれば、この情報行列を近似するのが望ましいと言う事になる。これは、大変な困惑を生み出す。なぜならば、Σ^{-1} の固有値はΣの固有値の逆数であり、固有ベクトルは共通である。従って、Σ^{-1}を効率的に近似しようとすれば、小さな固有値に対応する主成分の空間を重要視しなければならないからである。このように、偏相関分析あるいはそれを発展させたグラフィカルモデリングは、主成分分析や因子分析による連関分析とは全く異なった共分散構造近似に基づいていると考えられる。

主成分分析が共分散構造や、相関係数構造を近似するという事はどういう基準で行えば良いのかというのが、椿、椿[2]での問題意識であった。この答えは既に、CGGM などのグラフィカルモデルのソフトなどでは、実現していることだが、「逸脱度(Deviance)」すなわち、カルバックの擬距離に基づいて評価せよというものである。データの平方和積和行列を S とし、Σの推定量を $\hat{\Sigma}$ とすれば、推定量の逸脱度は(1)から、

$$\mathrm{Dev} = tr\left(S\hat{\Sigma}^{-1}\right) - n\log\left|\hat{\Sigma}^{-1}\right| + \mathrm{const}$$

となる。Σの対角要素の推定量として当該変量の標本分散 v_i を用い、

$$\hat{\mathrm{P}} = \mathrm{diag}(v_i)^{1/2}\,\hat{\Sigma}\,\mathrm{diag}(v_i)^{1/2}$$

として、相関係数行列の推定量に基づいて逸脱度を算出すれば、

$$\mathrm{Dev} = n\left\{tr\left(\mathbf{R}\hat{\mathrm{P}}^{-1}\right) - \left(\log\left|\hat{\mathrm{P}}^{-1}\right| + \sum_{i=1}^{p}\log v_i\right)\right\} + \mathrm{const} \quad (2)$$

となる。特に、ここで $\hat{\mathrm{P}}$ を相関係数行列 **R** の固有値 λ_i・固有ベクトル $\mathbf{p}_i$ による分解（相関係数行列起点の主成分分析、あるいはスペクトル分解）

$\mathbf{R} = \sum_{i=1}^{p}\lambda_i\mathbf{p}_i\mathbf{p}_i^{\mathbf{T}}$ に対して、$\Sigma W_i\mathbf{p}_i\mathbf{p}_i^{\mathbf{T}}$ の形に制約し、

(2)式に代入すると、$\mathrm{Dev} = \mathrm{const} + n\sum_{i=1}^{p}\frac{\lambda_i}{W_i} + \log W_i$

となり、$W_i = \lambda_i$ のときに、逸脱度は最小となる。一方、$W_i = (1+\delta_i)\lambda_i$ と逸脱度最小の値から、微小定数倍の摂動を考えると、このときの逸脱度の増大は、$n\{(1+\delta_i)^{-1} + \log(1+\delta_i) - 1\}$となり、固有値 λ_i の大きさには依存しない。すなわち、固有値の対数の値を一定値だけ、標本相関係数の固有値の対数から変化させる事は逸脱度に同等の変化を与えると言ってもよい。決して大きな固有値に対応する固有空間が重要であるということはないのである。

以上の準備の下で、宮川[1]のデータの相関係数行列の固有値の対数をプロット（対数スクリープロット）したものが図 2 である。

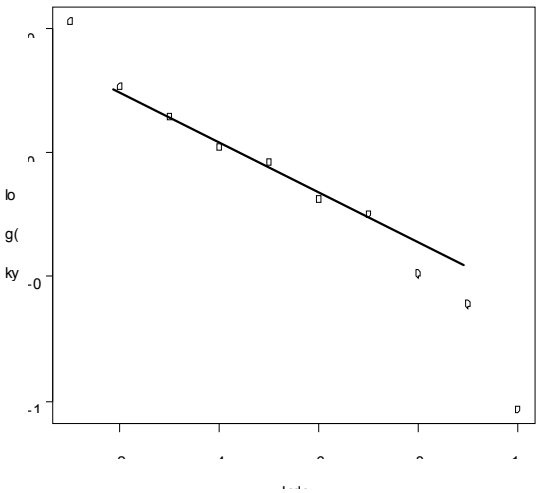

図 2　十種競技データの対数スクリープッロット

図 2 を眺めると、第 4 固有値から第 7 固有値までは、ほぼ同じスピードで減少しているが、第 7-10 固有値はそれより顕著に大きなスピードで、一方第一固有値から第 3 固有値までの減少も若干スピードが速いことに気付く。この種の現象は、広く多変量データ解析の現場で見られる。椿は、これを「大陸―大陸棚―海溝の構造」と呼んでいる。これまで主成分分析が注目してきた大陸の構造が、潜在因子構造に相当し、海溝の構造は、見過ごされがちであった共線性構造に相当するのである。ここで、第 4 から第 7 までの固有値は、減少のスピードが緩やかなので、これらの対数固有値が全て等しいと想定しても、逸脱度の増大はそれほどにはならない可能性がある。これから、相関係数行列を次のように近似することが考えられる。

$$\hat{\mathrm{P}} = \sum_{i=1}^{3} \lambda_i \mathbf{p_i} \mathbf{p_i}^{\mathrm{T}} + \bar{\lambda} \sum_{i=4}^{7} \mathbf{p_i} \mathbf{p_i}^{\mathrm{T}} + \sum_{i=8}^{10} \lambda_i \mathbf{p_i} \mathbf{p_i}^{\mathrm{T}} \qquad (3)$$

(3)の近似モデルは、第 3 項の海溝構造を除けば、T。W。 Anderson[3]の多変量解析の古典的テキストで主成分分析の固有値の数を決定するために考察した仮説検定問題と類似である。このとき、逸脱度の増加を最小にするには、$\bar{\lambda} = \dfrac{\sum_{i=2}^{7} \lambda_i}{6}$ ととれば良く、このときの逸脱度の増加量は、$n(4\log \bar{\lambda} - \Sigma \log \lambda_i)=4$。70 (自由度 3)に過ぎない。

一方、近似(3)の意味は、次のように考える事ができる。**Z***を **Z** の各要素を平均 0 分散 1 に標準化したものとする。このとき、

$$\{\mathrm{I} + \sum_{i=8}^{10} \left(\sqrt{\frac{\bar{\lambda}}{\lambda_i}} - 1 \right) \mathbf{p_i} \mathbf{p_i}^{\mathrm{T}} \} \mathbf{Z}^* = \sum_{i=1}^{3} \left(\lambda_i - \bar{\lambda} \right)^{1/2} \mathbf{p}_i f_i + \boldsymbol{\varepsilon}$$

といった構造がデータに存在すると考える事ができる。ここで、f_l は、標準化正規変量、$\boldsymbol{\varepsilon}$ は、平均 0、共分散 $\bar{\lambda}$ I の誤差変量である。この構造は、一般化すれば、

$$\mathbf{Z}^* = \mathrm{A}\mathbf{Z}^* + \mathrm{B}\mathbf{f} + \boldsymbol{\varepsilon}$$

といった構造を想定した事になり、時系列解析の ARMA モデルを多変量解析で実現したようなものである。

3　潜在因子の測定と影響

古典的計測工学では、原因系測定量が影響を与えている結果系変量の追加削減について、原因系測定構造の不変性を要求している。もちろん、これを物理測定の独自性ととらえることもできよう。しかし、潜在構造分析の多母集団モデルで測定構造不変性を要求するセンスと、測定対象の測定結果がその利用状況によって変動してはならないとする常識論の間は、それ程距離のある話ではないこの点は、偏相関分析、グラフィカルモデリングでは、解析変量群に因果関係の観点で半順序が付く場合、「因果連鎖分析」なる手順を実施するという形で実現している。そしてそこでは、結果系経由で生じる関連性は、偽偏相関として充分意識され、かつ適切な対処がなされている。

一方、線形潜在構造分析では、観測変量の因果関係による半順序の問題はどのように扱われているのだろうか？共分散構造分析の最大の貢献は、検証的研究対象となる「概念」の「測定モデル」と概念間の因果関係を記述する「構造モデル」を明示させることが研究の初動段階と位置づけたことである。

一方、このモデル化手続き自体は、因子得点として推定された潜在概念のレベルをあたかも観測変量と見なして行うパス解析と理念的には大差ないように考えられる。しかし、潜在概念を一度得点化し顕在尺度化する方法では、コンセプト間のパス係数は、対応する線形潜在構造分析に比べて「過小評価」、計量心理学者のいう「希薄化 (Attenuation)」がおきる。この理由付けとしてよく知られているのは、潜在因子自体を計測したのではなく、それを推定したために、因子得点間に成立する統計モデルが、「変数誤差モデル(Errors in Variables Model)」になっているためというものである。

しかし、「希薄化」問題には、この変数誤差側面以外に「原因系概念の結果系変数からの因果逆流による概念自体の変質」というより根源的な問題があ

る。すなわち、共分散構造分析では統計モデル作成に際しては、「測定モデル」と「構造モデル」との分離が明確に意識されるのに、モデル識別に際してはこの分離がなされていない。

さて、共分散構造分析でモデリングのみならず、推論においても、測定モデルと構造モデルを分離において椿[4]が注目したのが Conditionality Principle である。この原理は数理統計学的には次のように抽象化される:「変量 **X**、**Y** の同時分布が、$f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})g(\mathbf{X}|\boldsymbol{\xi})$ のように分解される場合には、X は関心のある母数 θ の補助統計量（Ancillary Statistics）と呼ばれ、関心のある母数 **θ** に関する推論は、$\mathbf{X}=\mathbf{x}_{\mathbf{OBS}}$ と条件付けた分布を用いて行う。

簡単のため全ての変量（*X, Y, Z*）は簡単のため標準正規確率変量と想定し、図 3 のようなモデルを考える。

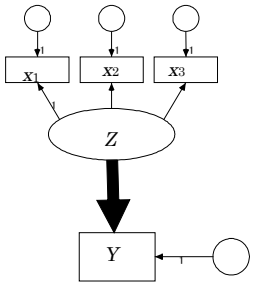

図 3　潜在測定モデルに付随する結果変数

図 3 で矢線の太さを変えているのは、細線は「測定モデル」、太線は指標変量により測定した概念が応答変量に及ぼす影響を表現した「構造モデル」のパラメータに対応することを明示するためである。この単純なモデルは Cox and Wermuth [5]も取り上げているが、形式的には、4 変量検証的 1 因子モデル[1]に過ぎない。しかし、実質的には 1 つの応答変量 (Response Variable)Y、3 つの指標変量(Indicator Variable)X_i からなっており、その区別は重要である。図 3 のモデルは、測定モデル(4)と構造モデル(5)で表現される。

$$X_i=\xi_i Z+\varepsilon_i \text{ 、 } i=1,2,3\text{、 } \varepsilon_i \sim N(0,1-\xi_i^2) \quad (4)$$

$$Y=\theta Z+\delta\text{、 } \delta \sim N(0, 1-\theta^2) \quad (5)$$

ここで *Z* は標準正規分布に従う潜在変量と想定（変量因子モデル）しているのだが、この値を母数（母数因子モデル）と想定した条件付き分布を計算すると、図 3 から示唆される条件付き独立性より、*Z* を与えた下での Y の条件付き分布は、ξ に依存せず、

$$f(Y|Z) = N(\theta Z, 1-\theta^2) \quad (6)$$

となる。従って、構造モデルの母数 θ を関心のある母数、測定モデルの母数ξを攪乱母数とする場合には、*Z* が θ の補助統計量の役割を果たすことになる。逆に、測定モデルの母数ξが関心のある母数、構造モデルの母数 θ が攪乱母数の場合には、*Z*、(及び *Y*) がξの補助統計量となる。

この補助統計量を所与とした条件付き推論は、仮に *Z* が観測変量の場合だとすれば Conditionality Principle に従ったことになる。問題は潜在変量に対しても同様の原理を適用して良いかということである。残念ながら図 3 のモデルでは実際に観測されている原因系変量 $\mathbf{X}=(X_1, X_2, X_3)$は、補助統計量ではなく、これを与えたときの *Y* の条件付分布は、

$$N(\theta\lambda\sum_{i=1}^{3}\frac{X_i\xi_i}{1-\xi_i^2}, 1-(1-\lambda)\theta^2) \quad (7)$$

となる。この条件付分布(7)を、

$$N(E[Z|\mathbf{X}]\theta, 1-\theta^2+Var[Z|\mathbf{X}]\theta^2) \quad (8)$$

と表現すれば、補助統計量が観測された場合の条件付分布(6)との対応も明らかである。すなわち、(8)式を **X** によって *Z* が不確かさ無く測定可能できる理想的状況で考えると(6)式になるのである。椿[4]は、第一段階としての尺度化、すなわち、$E[Z|\mathbf{X}]$の推定を行い、これをもとに構造母数を推定するのが自然となる状況があると主張するとともに、測定モデルとしての 3 変量 1 因子モデル当てはめと同等の結果を導く 4 変量飽和モデルを提示し、その適合度検定こそ、Cox and Wermuth[5]が外的適合度と呼んだものである。Cox らは、外的適合度を、指標変量による測定モデルが適合するという前提で、潜在変量 Z を与えたときに指標変量と応答変量とが条件付独立になることを検定している統計量と位置づけている。

4　同時潜在成長曲線構造

共分散構造モデリング、線形潜在構造モデリングを動的構造に柔軟に拡張したのが状態空間モデリング（例えば、Durbin and Koopman[6]）である。これについては多くの実証的研究もすでに行われている。統計数理研究所に専門家集団も形成されているので、専門家でない筆者が紹介するのは不適切であろう。勿論、ここまで述べた相関構造の探索に関わる話題を動的因子構造においても議論する事は重要だが、以下では、共分散構造分析における代表的な動的構造モデリングの方法論としての潜在成長曲線モデルの筆者周辺の事例を２つ紹介する。一つは、データを仮想的に等間隔時系列化し、実質的には 99%以上が欠測という状況でモデリングを行ったアルツハイマー病の自然経過に与える影響の要因分析の事例(Arai, Tsubaki et al. [7])である。詳細な説明は省くが、この分析に用いたのが図 4 のバス図である。

もう一方は、東京工科大学の角埜恭央教授との共同研究で、企業のパフォーマンス計測に関する多変量不完全パネルデータに対する潜在成長曲線モデルあてはめを行い。様々な変数群の潜在水準因子。潜在

成長因子を探索的に導き、さらにその構造モデリングを行った事例である。

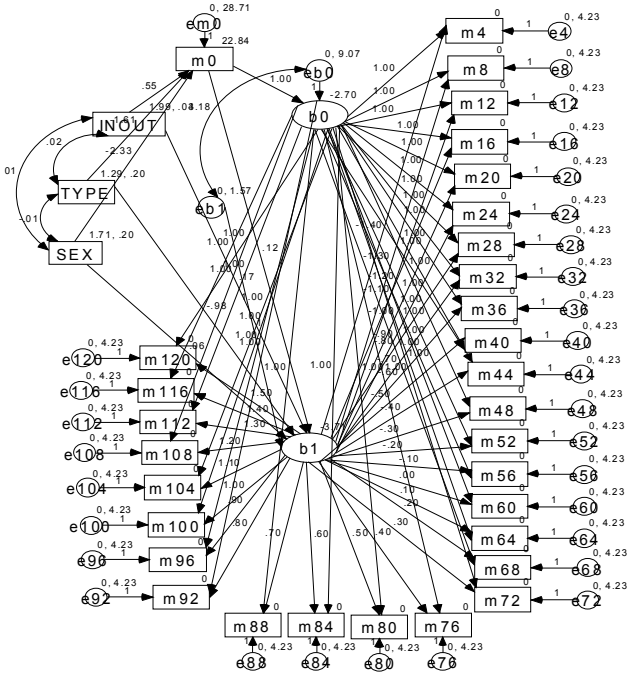

図 4　アルツハイマー病の 120 週仮想的定点観測に対する潜在曲線モデリング(b0:初期重症度潜在変数、b1:進行速度潜在変数、mt:観測開始後 t 週の重症度観測値、SEX: 性、TYPE:アルツハイマー病の確定診断の有無、INOUT:入院患者か外来患者かのダミー変数)

これらの詳細については当日時間の許す範囲で報告する。

参考文献

[1] 宮川雅巳, グラフィカル・モデリング, 朝倉書店、1997.

[2] 椿広計, 椿美智子, グラフィカルモデリングからの既成モデルの見直し, 日本統計学会第 65 回発表要旨集. pp.256－257, 1997.

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, 1984.

[4] 椿広計, 狩野論文へのコメント－「尺度化＋回帰分析」の問題点に関する注意, 行動計量学、Vol.29, No.2, pp.167-173, 2002.

[5] D. R. Cox and N. Wermuth, *Multivariate Dependencies – Models, analysis and interpretation*, Chapman and Hall, 1996.

[6] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, 2001.

[7] H. Arai, H. Tsubaki, Y. Mitsuyama, N. Fujimoto, Y. Urata and A. Homma, Early Onset Alzheimer Type Dementia More Rapidly Deteriorates than Late Onset Type: A Follow-up Study on MMSE Scores in Japanese Patients, Psychogeriatrics, Vol.1, pp.303-308, 2001.

*統計数理研究所、〒190-8562 、立川市緑町 10-3、tel. 050-5533-8562、e-mail: tsubaki@ism.ac.jp the Institute of Statistical Mathematics、10-3 Midori-cho、Tachikawa、Tokyo、Japan