不完全データの分析

研究ノート

狩　野　　裕*

Analysis of incomplete data

Key Words：Full-information maximum likelihood, missing at random, missing indicator, missing mechanism

1　はじめに

本来データがあるべきなのに何らかの事情で手に入らないことがある．これをデータが欠測する(missing) という．欠測を含むデータ全体を不完全データ(incomplete data) という．いくつかの要因の効果を知るため実験をしたとしよう．もし，要因の組み合せのいくつかで実験を失敗しデータが得られなかったならば，その実験からは適切な結論を得ることは難しくなるだろう．予算とマンパワーがあれば再実験が可能かもしれないが，そのような場合でも，実験の順序や環境の変化など無作為化という実験の基礎的な仮定が崩れるという別の問題が生じる．一方，実験を失敗するには理由があるはずで，実験者の未熟さゆえの単純ミスであれば問題はないが，特定の条件下では実験を失敗しやすいということが暗示される場合は，欠測（実験失敗）という事実自体が意味を持つことになる．単にデータをそろえるために再実験することは，重要な発見を見落とす可能性がある．

本稿では不完全データの解析について基礎的な事項と最近の結果のいくつかを紹介する．

2　一変量の場合

ある対象物の寿命（生存時間，故障するまでの時間）をXと書き，寿命の期待値$E[X]$を推定したいとしよう．対象物をn個体用意し，時刻$t=0$から事前に定められた時刻$t=c$までそれらを観察する．いくつかの対象物は時刻$t=c$までに寿命を迎えるであろうが，他のいくつかは寿命を測定することなく観察が終了してしまうことがある[1]．いま，時刻cまでにm個体の寿命が測定され（$X_1,\cdots,X_m$とする），$n-m$個体の観察が打ち切られたとする．このとき，データは次のように並ぶことになる．

$$\underbrace{X_1,\cdots,X_m,}_{\text{測定された寿命（}m\text{個体）}}\quad\underbrace{??,\cdots,??,}_{\text{打ち切られた寿命（}n-m\text{個体）}}$$

測定できた寿命データだけを用いて$E[X]$を$\frac{1}{m}\sum_{k=1}^{m}X_k$によって推定するならば，これは$E[X]$を過小推定することは容易にわかる．打ち切られた個体の寿命を打ち切り時間cで置き換え，

$$\widehat{E[X]}_1=\frac{1}{n}\Big(\sum_{k=1}^{m}X_k+c(n-m)\Big)$$

としたとしても，過小の程度は緩和されるが依然として小さめに推定される．

この問題に対する一つの解答は次式で与えられる．

$$\widehat{E[X]}_2=\frac{1}{m}\Big(\sum_{k=1}^{m}X_k+c(n-m)\Big)$$

この推定量は$\widehat{E[X]}_1$の$n/m(\geq 1)$倍である．推定量$\widehat{E[X]}_2$はXが指数分布に従うとき正当化され，$n\to\infty$のとき$\widehat{E[X]}_2$は真の平均$E[X]$に概収束することが証明される．

この推定量はad hoc なものではなく，完全情報最尤法 (method of full-information maximum likeli

[1] このようなデータを時間打ち切り，または，タイプI センサリングという．

*Yutaka KANO
1958年11月生
大阪大学大学院基礎工学研究科 数理系専攻 博士前期課程修了（1983年）
現在、大阪大学 大学院基礎工学研究科 システム創成専攻 数理科学領域　教授
工学博士　統計科学・応用数学
TEL：06-6850-6485
FAX：06-6850-6485
E-mail：kano@sigmath.es.osaka-u.ac.jp

hood; FIML）という統一的な理論の下で導出することができる[2]．この方法は伝統的な最尤法の拡張であって，欠測に対応する尤度を欠測確率で置き換えることによって尤度を定義する．寿命データが独立で密度関数 $f(x|\theta)$ をもつ連続分布に従うとき，尤度は

$$L(\theta)=\prod_{k=1}^{m} f(X_k|\theta)\times\prod_{k=m+1}^{n} P(X_k>c|\theta)$$
$$=\prod_{k=1}^{m} f(X_k|\theta)\times\left(\int_c^{\infty} f(x|\theta)dx\right)^{n-m} \quad (1)$$

と書くことができ，$L(\theta)$ を最大にする $\hat{\theta}$ が，完全情報最尤法による推定量である．実際，指数分布 $f(x|\theta)=\frac{1}{\theta}e^{-x/\theta}(x\geq 0)$ のときは，$E[X]=\theta$ の推定量として $\widehat{E[X]}_2$ が導かれる．なお，FIML の数理的基礎を付録にて補足する．

3　多変量の場合

多変量の場合は記号や場合分けが複雑になるので，二変量の観測ベクトル (X, Y) を例にとり説明する．表1に示すように，二変量の場合，欠測のパターン数は4であり[3]，ここではそれらを $I_{st}\,(s,t=0,1)$ で表している．ここで1は観測を0は欠測を示す．M^X と M^Y はそれぞれ X と Y の欠測指標（missing indicator）とよばれており，先ほどと同様，1は観測，0は欠測を示す．我々がもっている情報は，欠測指標のすべてと (X, Y) については「欠」の記号以外の部分であり，前者を $\boldsymbol{M}$，後者を $\boldsymbol{Y}_{obs}$ と書く．「欠」に本来あるべきデータを $\boldsymbol{Y}_{mis}$ と表す．なお，$\boldsymbol{Y}=[\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}]$ は観測予定のすべてのデータである．

表1：欠測のパターン

	i	X	Y	M^X	M^Y
I_{11}	1	X_1	Y_1	1	1
	⋮	⋮	⋮	⋮	⋮
	n_1	X_{n_1}	Y_{n_1}	1	1
I_{10}	n_1+1	X_{n_1+1}	欠	1	0
	⋮	⋮	⋮	⋮	⋮
	n_2	X_{n_2}	欠	1	0
I_{01}	n_2+1	欠	Y_{n_2+1}	0	1
	⋮	⋮	⋮	⋮	⋮
	n_3	欠	Y_{n_3}	0	1
I_{00}	n_3+1	欠	欠	0	0
	⋮	⋮	⋮	⋮	⋮
	n	欠	欠	0	0

[2] FIML は経済学の分野でよく用いられる用語である．統計学では単に最尤法とよぶことも多い．

[3] 観測ベクトルが p 変量の場合は欠測のパターン数は 2^p である．

完全情報最尤法（FIML）は我々がもつ情報のすべて $[\boldsymbol{M}, \boldsymbol{Y}_{obs}]$ に基づく推測である．(X, Y) の同時分布，X と Y の周辺分布を，それぞれ $f_{11}(x, y|\tau)$，$f_{10}(x|\tau)$，$f_{01}(y|\tau)$ と書く．尤度を具体的に書き下すと以下のようになる．

$$L(\tau|\boldsymbol{M},\boldsymbol{Y}_{obs})=P(\boldsymbol{M}|\boldsymbol{Y}_{obs},\tau)f(\boldsymbol{Y}_{obs}|\tau)$$
$$=\prod_{i\in I_{11}} P(M_i^X=1, M_i^Y=1|X_i,Y_i,\tau)f_{11}(X_i,Y_i|\tau)$$
$$\times\prod_{i\in I_{10}} P(M_i^X=1, M_i^Y=0|X_i,\tau)f_{10}(X_i|\tau)$$
$$\times\prod_{i\in I_{01}} P(M_i^X=0, M_i^Y=1|Y_i,\tau)f_{01}(Y_i|\tau)$$
$$\times\prod_{i\in I_{00}} P(M_i^X=0, M_i^Y=0|\tau)$$

上式に現れる欠測指標に関する（条件付）確率を欠測メカニズムという．ここで，τ は (X, Y) の同時分布を規定する興味あるパラメータと欠測メカニズムに関係するパラメータを合わせたものである．この尤度は前節で議論した一変量の推測の拡張になっていることが容易に確かめられる．

欠測がある場合は，(M^X, M^Y, X, Y) の同時分布，特に欠測メカニズム $P(M^X, M^Y|X, Y)$ の規定が重要であることが理解されよう．また，この尤度は4個の母集団からそれぞれ標本サイズ $\#I_{st}$ のサンプルを採取したときの尤度と一致することから，複数個の母集団の同時分析とみなすこともできる[4]．通常の同時分析と異なる点は i）観測ベクトルが (M^X, M^Y, X, Y) の部分集合であり母集団ごとに異なること，ii）各母集団単独で推定を行うと推定にバイアスが生じるか，もしくはパラメータ τ が識別できないこと，である．

[4] 集合 A に対して $\#A$ は A の濃度(要素の数) を表す．

例を挙げる．入学試験（X）と入学後の成績（Y）との関係（相関係数）を調べたいとする（図1）．受験者全員について入学試験の成績は存在するが，不合格者には入学後の成績が存在しない．したがって，Yにのみ欠測が生じ得ることから，表1による分類ではI_{11}とI_{10}のみを考えればよい．

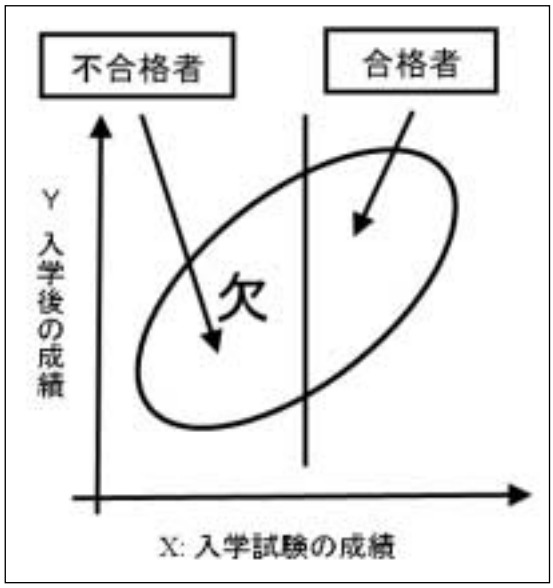

図1：入学試験と入学後の成績

欠測メカニズムは

$$P(M^X=r, M^Y=s|X, Y, \phi)=P(M^Y=s|X, \phi)$$

$$\begin{cases} P(M^Y=1|X, \phi)=1, \text{ if } X \geq \phi \\ P(M^Y=0|X, \phi)=1, \text{ if } X < \phi \end{cases}$$

となる．ここでϕは合格最低点である．尤度は

$$L(\theta|\boldsymbol{M}, \boldsymbol{Y}_{obs}) = \prod_{i \in I_{11}} f_{11}(X_i, Y_i|\theta) \times \prod_{i \in I_{10}} f_{10}(X_i|\theta)$$

で与えられる．(X, Y)に二変量正規分布を仮定すると，パラメータは$\theta=[\mu_x, \mu_y, \sigma_{xx}, \sigma_{yy}, \sigma_{xy}]^T$であり，上記の尤度を最大化することによって推定することができる．最尤推定量$\hat{\theta}$は反復法を必要とせず陽に解くことができる[e.g., 岩崎(2001)]．$\hat{\theta}$を用いて相関係数の推定量rを得ることができる．それは，合格者のみを用いた（偏りのある）相関係数r'の単純な関数となっており，具体的には

$$r=\frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_{xx}\,\hat{\sigma}_{yy}}}=\frac{r'}{\sqrt{(1-k^2)(r')^2+k^2}} \quad (2)$$

で与えられる．ここで$k^2=\frac{s_{xx}}{\sigma_{xx}}$であり，これは合格者の$X$の分散と受験者の$X$の分散の比を表している．$k^2$は合格率（倍率と同等）と直接的な関係がある．合格者の相関係数を$r'=0.3$として，いくつかの合格率に対してrがどのように変化するかを表2に示した．たとえば，合格率が10％の場合，本来の相関係数が$r=0.59$であるにもかかわらず，欠測を無視し合格者だけで相関係数を計算すると$r'=0.30$となり，本来の相関係数の推定値rを大きく過小評価してしまうことがわかる．

表2：欠測が相関係数へ及ぼす影響

	合格率		
	30%	20%	10%
k^2	0.27	0.22	0.17
r'	0.30	0.30	0.30
r	0.51	0.54	0.59

この例のように，欠測する変数Yの欠測メカニズムが（他の変数Xに依存し得るが）Y自身には依存しないとき，欠測メカニズムはMAR [Missing At Random; Little and Rubin (2002)] であるといい，統計的推測が簡略化されることが多い．MARの定義をシンボリックに表記すると

$$P(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})=P(\boldsymbol{M}|\boldsymbol{Y}_{obs})$$

となる．

(2) 式の公式は教育（心理）学の分野ではずいぶん昔から知られており[5]，新規性はない [e.g., Lord and Novick (1968)]．しかし，この公式が不完全データの解析という統一的な観点から解釈できることは興味深い．

4　カテゴリカルデータの場合

2×2分割表はカテゴリカルデータの中で最も基本的である．二つのカテゴリカル変数（Y_1, Y_2）がそれぞれ二つのカテゴリー$Y_1=1, 2$, $Y_2=1, 2$をもつとする．得られるデータ（不完全分割表データ）

[5] 選抜効果という．

表3：2×2分割表データ

補助的周辺度数をもつ分割表データ

		$M_2=1$		$M_2=0$
		$Y_2=1$	$Y_2=2$	$Y_2=?$
$M_1=1$	$Y_1=1$	$n_{11,11}$	$n_{11,12}$	$n_{10,1+}$
	$Y_1=2$	$n_{11,21}$	$n_{11,22}$	$n_{10,2+}$
$M_1=0$	$Y_1=?$	$n_{01,+1}$	$n_{01,+2}$	$n_{00,++}$

対応する生起確率

		$M_2=1$		$M_2=0$
		$Y_2=1$	$Y_2=2$	$Y_2=?$
$M_1=1$	$Y_1=1$	$\pi_{11,11}$	$\pi_{11,12}$	$\pi_{10,1+}$
	$Y_1=2$	$\pi_{11,21}$	$n_{11,22}$	$\pi_{10,2+}$
$M_1=0$	$Y_1=?$	$\pi_{01,+1}$	$\pi_{01,+2}$	$\pi_{00,++}$

と対応する生起確率は表3のようになる．2×2分割表の周辺には，一方または両方の変数において欠測がある個体（観測値）の数（またはその確率）が示されている．たとえば，$n_{10,1+}$ は $Y_1=1$ であるが Y_2 の情報がない（欠測）個体の数を表す．また，$+$ の記号をもつものは，たとえば

$$\begin{aligned}\pi_{10,1+} &= \pi_{10,11} + \pi_{10,12} \\ &= P(M_1=1, M_2=0, Y_1=1, Y_2=1) \\ &\quad + P(M_1=1, M_2=0, Y_1=1, Y_2=2)\end{aligned}$$

である．ただし，個々のパラメータ $\pi_{10,11}$ と $\pi_{10,12}$ は直ちに推定できないことに注意する．

欠測を含む分割表データの分析についても歴史があり多くの統計学者が議論を積み重ねてきたが，現在はFIMLによる分析に統一されている[e.g., Molenberghs et al.(1999)]．前節での議論と同様に考えると，FIMLは

$$\begin{aligned}&L(\pi|\boldsymbol{M}, \boldsymbol{Y}_{obs}) \\ &= \prod_{y_1,\, y_2=1,2} P(M_1=1, M_2=1, Y_1=y_1, Y_2=y_2)^{n_{11,\, y_1 y_2}} \\ &\quad \times \prod_{y_1=1,2} P(M_1=1, M_2=0, Y_1=y_1)^{n_{10,\, y_1+}} \\ &\quad \times \prod_{y_2=1,2} P(M_1=0, M_2=1, Y_2=y_2)^{n_{01,+y_2}} \\ &\quad \times P(M_1=0, M_2=0)^{n_{00,++}} \\ &= \prod_{y_1,\, y_2=1,2} \pi_{11,\, y_1 y_2}^{n_{11,\, y_1 y_2}} \times \prod_{y_1=1,2} \pi_{10,\, y_1+}^{n_{10,\, y_1+}} \\ &\quad \times \prod_{y_2=1,2} \pi_{01,\, +y_2}^{n_{01,+y_2}} \times \pi_{00,++}^{n_{00,++}}\end{aligned}$$

を最大化する．通常，興味のあるパラメータは

$$P(Y_1=y_1, Y_2=y_2) = \sum_{m_1,\, m_2=0,1} \pi_{m_1 m_2,\, y_1 y_2}^{n_{m_1 m_2,\, y_1 y_2}}$$

であるが，先に指摘したように $\pi_{10,11}$ や $\pi_{10,12}$ などは簡単には得られない．

$P(Y_1=y_1, Y_2=y_2)$ を推定する有力な方法の一つは欠測メカニズムにMARを仮定することである．ここではMARが仮定できないときの推測を考える．表3のデータを4変数 M_1, M_2, Y_1, Y_2 のカテゴリカルデータと考え，4変数間の関係をグラフィカルモデルで記述する．図2には2種類のグラフィカルモデルが示されている．これらは無向独立グラフとよばれ，線によって結ばれた変数間には直接的な関係があることを示す[e.g., 宮川(1997)]．図2の左のモデルには M_1 がなく，これは Y_1 に欠測が生じないことを示す．Y_1 と M_2 を結ぶ線は Y_2 の欠測確率が Y_1 と関係すること，M_2 と Y_2 は線で結ばれていないことは両者には直接的な関係がないことを示す．より正確には，同モデルは

$$P(M_2|Y_1, Y_2) = P(M_2|Y_1)$$

を満たす．すなわち，欠測メカニズムはMARであることを示している．この構造は Y_i が第 i 回目の測定という経時測定データによく現れる．前節の入試選抜の例はこのモデルに対応する．

右のグラフでは M_1 と Y_1，M_2 と Y_2 に直接的な関係がある．したがって，欠測するかどうかが欠測変数と直接的に関係しておりMAR ではない．二変数の両者に欠測が生じ，M_1, M_2 と Y_1, Y_2 とが何らかの線で結ばれている場合は基本的にMAR とはならない[高井(2008)][6]．

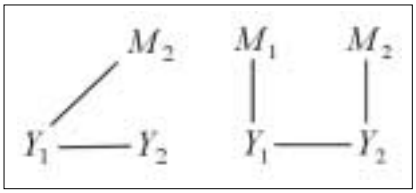

図2：グラフィカルモデル

[6] (M_1, M_2) と (Y_1, Y_2) とが線で結ばれていないことはそれらが独立であることを示し，このとき，欠測は完全にランダムである(MCAR) という．MCAR はMAR の特殊な場合であり，MCAR のときは欠測が生じたケースを削除して（通常の）分析を行っても推定にバイアスは生じない．

図２の右のグラフの下でパラメータが推定できるためには，Y_1 と Y_2 が線で結ばれていることが必要である[Ma et al. (2003)]．それは，$M_1 - Y_1$ なるモデルが（単独では）推定できないことからも明らかであろう．この仮定はパラメータ推定を行うときには概ね満たされていると考えてよいが，２×２分割表における基本的な解析である２変数間の独立性の検定を行うときには決定的になる．帰無仮説の下で Y_1 と Y_2 が独立であるからである．Takai と Kano (2008) は独立性の検定を可能とするような適当な仮定を導入し，FIML と既存の検定統計量のパフォーマンスを数値実験によって比較している．

５　おわりに

実証研究とはデータによって理論を検証することである．実験研究であれ調査研究であれ予定していたデータが採取できないことがあり，それが実証研究を歪めることがある．本稿では，欠測に対するモデリングと不完全データの適切な分析方法について最新の研究を交えて紹介した．

参考文献

[1] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd edition). New York: Wiley.

[2] Lord, F. M. and Novick, M. R.(1968). *Statistical Theories of Mental Test Scores: With Contributions by Allan Birnbaum*. Addison-Wesley Educational Publishers Inc.

[3] Ma, W.-Q., Geng, Z. and Li, X.-T. (2003). Identification of nonresponse mechanisms for two-way contingency tables. *Behaviormetrika*, 30, 125-144.

[4] Molenberghs, G., Goetghebeur, E., Lipsitz, S. R. and Kenward, M. G. (1999). Non-random missingness in categorical data: strengths and limitations. *The American Statistician*, 53, 110-118.

[5] Takai, K. and Kano, Y. (2008). Test of independence in a 2 × 2 contingency table with nonignorable nonresponse via constrained EM algorithm. *Computational Statistics & Data Analysis*, 52, 5229-5241.

[6] 岩崎学（2001）. 不完全データの統計解析. エコノミスト社.

[7] 宮川雅巳（1997）．グラフィカルモデリング．朝倉書店.

[8] 高井啓二（2008）．グラフィカルモデルによる欠測のモデリングとその周辺．科学研究費シンポジウム「多変量解析における最近の話題」報告集．pp.94-103.

付録　次の定理が成立する．

定理 $(M, X) \sim P(M = m|x, \boldsymbol{\theta}_0) f(x|\boldsymbol{\theta}_0); m = 0, 1; x \in \boldsymbol{\mathcal{X}} (\subset \mathbb{R}^1); \boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \boldsymbol{\Theta} (\subset \mathbb{R}^q)$. $KL(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ を次式で定義する．

$$KL(\theta|\theta_0) = E\Big[M \log \{P(M=1|X,\theta) f(X|\theta)\} + (1-M) \log P(M=0|\theta) \,\Big|\, \theta_0\Big]$$

ただし，この期待値は $P(M = m|x, \boldsymbol{\theta}_0) f(x|\boldsymbol{\theta}_0)$ について取るものとする．このとき，$\boldsymbol{\theta} = \boldsymbol{\theta}_0$ は最大化問題 $\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} KL(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ の解である．

証明

$$\begin{aligned}
&KL(\theta|\theta_0) \\
&= E\Big[M \log \frac{P(M=1|X,\theta) f(X|\theta)}{P(M=1|\theta)} \\
&\quad + M \log P(M=1|\theta) + (1-M) \log P(M=0|\theta) \,\Big|\, \theta_0\Big] \\
&= E\Big[\log \frac{P(M=1|X,\theta) f(X|\theta)}{P(M=1|\theta)} \,\Big|\, M=1, \theta_0\Big] P(M=1|\theta_0) \\
&\quad + P(M=1|\theta_0) \log P(M=1|\theta) \\
&\quad + P(M=0|\theta_0) \log P(M=0|\theta) \qquad (3)
\end{aligned}$$

情報量不等式を適用すると，$\boldsymbol{\theta} = \boldsymbol{\theta}_0$ のとき (3) が最大になることが示される．　Q.E.D.

この定理は多次元のモデルへ容易に拡張できる．確率変数 M は欠測指標である必要はない．不完全データの分析においては，この一般的な結果を，M を欠測指標として適用しているのである．一般に，最大化問題の解 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ の一意性は保証されない．各個別問題においてパラメータの識別性を調べる必要がある．

さて，上記定理で扱った確率分布に従う母集団から採取した独立同一分布をもつ標本 $(M_1, X_1), \cdots, (M_n, X_n)$ を得たとし，（必要ならば）順序を入れ替えて $M_1 = \cdots = M_m = 1,\ M_m + 1 = \cdots = M_n = 0$ とする．次式は $KL(\theta|\theta_0)$ の標本版であり不偏一致推定量である．

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\Big[M_k \log\Big\{P(M_k=1|X_k,\theta)f(X_k|\theta)\Big\} \\
&\qquad + (1-M_k)\log P(M_k=0|\theta)\Big] \\
&= \frac{1}{n}\Bigg[\sum_{k=1}^{m}\log\Big\{P(M_k=1|X_k,\theta)f(X_k|\theta)\Big\} \\
&\qquad + (n-m)\log P(M=0|\theta)\Bigg] \\
&= \frac{1}{n}\log\Bigg[\prod_{k=1}^{m}\Big\{P(M_k=1|X_k,\theta)f(X_k|\theta)\Big\} \\
&\qquad \times P(M=0|\theta)^{n-m}\Bigg] \qquad (4)
\end{aligned}$$

$KL(\theta|\theta_0)$ を最大にする解が $\theta=\theta_0$ であるので，適当な条件の下で，$KL(\theta|\theta_0)$ の不偏一致推定量である (4) を最大にする推定量 $\hat{\theta}$ は真値 θ_0 に収束することが期待される．

なお，(4) 式（の対数の真数）は (1) 式に対応することに注意する．