EMB アルゴリズムを用いた Multiple Imputation

○矢田 真城[1]　浜田 知久馬[2]
[1]エイツーヘルスケア株式会社 開発戦略本部 生物統計部
[2]東京理科大学 工学部 情報工学科

Multiple Imputation with the EMB algorithm

Shinjo Yada[1] and Chikuma Hamada[2]
[1]*Biostatistics Department, Development Strategy Division, A2 Healthcare Corporation*
[2] *Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science*

要旨

多重補完法のひとつとして，MCMC 法に基づく方法が知られている．Honaker and King (2010)は，新しいアプローチとして，EMB (Expectation-Maximization with Bootstrapping)アルゴリズムによる多重補完法を示した．これは，MCMC 法に基づく方法において，事後分布からのモンテカルロ標本により欠測値を補完するプロセスを，ブートストラップ法を用いた EM アルゴリズムによる補完に置き換える方法である．本稿では，解析対象となるデータが全て連続変数である場合に対し，EMB アルゴリズムによる多重補完法を適用し，SAS を用いて実行するための具体的な方法を紹介する．

キーワード：Multiple Imputation，Expectation-Maximization with Bootstrapping，MI プロシジャ，%BOOT

1. はじめに

多重補完法 (multiple imputation) は，欠測値を，真の値に対する不確実性を反映させた値で置き換える方法である(Rubin, 1987)．多重補完法は，1)欠測値に対してなんらかの方法で生成した値で補完したデータを M 組用意する($M>1$)，2)M 組のデータセットに対し標準的な方法で解析する，3)得られた M 組の解析結果を 1 つに統合する，という，3 つの手順で構成される．

MCMC (Markov Chain Monte Carlo)法は，欠測データの事後予測分布から生成した標本を用いて欠測値を補完する方法である．観測データを $\mathbf{Y}_{\mathbf{obs}}$，欠測データを $\mathbf{Y}_{\mathbf{mis}}$，興味のあるパラメータを $\boldsymbol{\theta}$ とし，t 番目の反復過程における $\boldsymbol{\theta}$ の推定値を $\boldsymbol{\theta}^{(t)}$と表すとき，MCMC 法の具体的な手順は以下のようになる．Imputation step（以下 I step と表記）として，$\mathbf{Y}_{\mathbf{obs}}$ と $\boldsymbol{\theta}^{(t)}$ 所与のもとでの $\mathbf{Y}_{\mathbf{mis}}$ の条件付分布 $p(\mathbf{y}_{\mathbf{mis}}|\mathbf{y}_{\mathbf{obs}}, \boldsymbol{\theta}^{(t)})$から欠測値 $\mathbf{Y}_{\mathbf{mis}}^{(t)}$ を生成する．Posterior step（以下 P step と表記）として，$\mathbf{Y}_{\mathbf{obs}}$ と I step において生成された $\mathbf{Y}_{\mathbf{mis}}^{(t)}$ から一時的に構成される完全データ $\mathbf{Y} = (\mathbf{Y}_{\mathbf{obs}}, \mathbf{Y}_{\mathbf{mis}}^{(t)})$に基づき，事後分布 $p(\boldsymbol{\theta}|\mathbf{y}_{\mathbf{obs}}, \mathbf{y}_{\mathbf{mis}}^{(t)})$から $\boldsymbol{\theta}^{(t+1)}$を生成する．このようにして形成

されるマルコフ連鎖($\mathbf{Y}_{\mathbf{mis}}^{(1)}$,$\boldsymbol{\theta}^{(1)}$), ($\mathbf{Y}_{\mathbf{mis}}^{(2)}$,$\boldsymbol{\theta}^{(2)}$),…が収束するまで I step と P step を繰り返し，反復過程の最後に得られた標本を用いて欠測値を補完する．

解析対象となるデータが全て連続変数の場合，それらが多変量正規分布に従うと想定し，興味のあるパラメータである平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ に対して適当な事前分布を仮定したもとで，I step と P step をマルコフ連鎖が収束するまで繰り返すことになる．ただし，実務上の問題点として，取り扱う変数の数が増えるほど $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の事後分布からのサンプリングが困難になることがあげられる．p 個変数がある場合，$\boldsymbol{\mu}$ で p 個，$\boldsymbol{\Sigma}$ で $p(p+1)/2$ 個の合計 $p(p+3)/2$ のパラメータが存在するため, 変数の個数が増えればそれだけ計算負荷が急速に増大する．

Honaker and King (2010)は，この問題を解決するための新しいアプローチとして，興味あるパラメータ $\boldsymbol{\theta}$ の事後分布からのサンプリングを，ブートストラップ法を用いた EM アルゴリズムで置き換えるという方法を提案した．以下，Honaker et al. (2011) に従い，これを EMB (expectation-maximization with bootstrapping) アルゴリズムと表記する．EMB アルゴリズムでは，欠測データを含むデータ $(\mathbf{Y}_{\mathbf{obs}}, \mathbf{Y}_{\mathbf{mis}})$からブートストラップ法を用いて M 組のデータを生成し，生成された M 組のデータごとに EM アルゴリズムにより $\boldsymbol{\theta}$ を推定し，推定された $\boldsymbol{\theta}$ と $\mathbf{Y}_{\mathbf{obs}}$ から欠測値を補完する．Honaker and King (2010)は，この方法の利点として，扱えるデータの大きさは解析を行う PC のメモリーサイズのみに依存していることを挙げ，時系列横断データのような巨大なデータを扱う社会科学の分野にて適用することを推奨した．

本稿では，解析対象となるデータが全て連続変数である場合をとりあげ，EMB アルゴリズムによる多重補完法について示す．そして，SAS を用いて行うための方法を，具体的に説明する．

2. EMB アルゴリズムによる多重補完法とは

EMB アルゴリズムによる多重補完法は, 補完回数を M (>1)として, 以下の 4 つのステップから構成される．

Step1：欠測値を含んだ不完全データセット D^{orign} から，M 組のブートストラップ標本 D_i^{resap} (i = 1,2,…,M)を生成する．

Step2：D_i^{resap} (i = 1,2,…,M)に対し，EM アルゴリズムを用いて組ごとに $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定する．

Step3：組ごとに推定された $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を用いて，D_i^{resap} の欠測値を補完したデータセット D_i^{imput} を生成する．

Step4：D_i^{imput} (i = 1,2,…,M)に対し解析を実行し，これらの解析結果を統合する．

EMB アルゴリズムを用いた多重補完法では，まず，欠測値を含んだ不完全データセットから，ブートストラップ法を用いて，M 組のブートストラップ標本を生成する．ブートストラップ法とは，観測されたデータからリサンプリングすることにより生成された擬似的なデータを用いて，パラメータ推定量の統計的誤差や統計量の分布を推定する方法のことである．n 個のデータが観測されたとき，大きさ n のブートストラップ標本とは，(0,1]上で発生させた n 個の一様乱数 $u_1,u_2,\ldots,u_n$ を用いて，観測されたデータから，$[nu_1]$+1 番目，$[nu_2]$+1 番目,…, $[nu_n]$+1 番目にあたるデータを抽出することで構成される（小西, 2008）．ここに[ ]はガウス記号である．生成されたブートストラップ標本を用いて，観測されたデータが従う確率分布に関するパラメータを推定し，ブートストラップ推定値を算出する．ブートストラップ推定値は，リサンプリング回数（ブートストラップ反復回数）を M とするとき，大きさ n のブートストラップ標本を M 組生成させ，各ブートストラップ標本に基づき，モンテカルロ法によって数値的に近似することができる．

リサンプリング回数 M が有限である以上，ブートストラップ推定値には数値近似による誤差が含まれる．釣り合い型ブートストラップは，数値近似による誤差を抑えるための方法のひとつである．一様リサンプリングを用いたブートストラップ法に比べ，同一のリサンプリング回数で推定量の分散を減少させることができる（汪・桜井, 2011）．一様リサンプリングでは，n 個の観測データ $(\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n})$ から大きさ n のブートストラップ標本を生成する作業を，独立に M 回繰り返す．このため，生成された $M \times n$ 個のブートストラップ標本において，観測データ $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}$ が全て M 回ずつ登場するとは限らない．これに対して，釣り合い型ブートストラップ法では，$M \times n$ 個のデータにおいて $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}$ がいずれも M 回ずつ登場するように，ブートストラップ標本が生成される．Step1 では，欠測値を含んだ n 個のデータから，リサンプリング回数を M として大きさ n のブートストラップ標本を生成するのみであり，ブートストラップ推定値の算出は不要である．

Step2 では，Step1 で生成された M 組のデータセットに対し，EM アルゴリズムを用いて興味のあるパラメータ $\boldsymbol{\theta}$ の推定値を算出する．EM アルゴリズムは，Expectation Step（以下 E Step と表記）と Maximization Step（以下 M Step と表記）の 2 つのステップを繰り返すことで，$\boldsymbol{\theta}$ の最尤推定量を導くためのアルゴリズムである．具体的には，k 回目の反復過程で得られた $\boldsymbol{\theta}$ の推定値を $\boldsymbol{\theta}^{(k)}$ と表すとき，E step では，観測データ $\mathbf{Y_{obs}}$ と $\boldsymbol{\theta}^{(k)}$ が所与のもとで，完全データ $\mathbf{Y} = (\mathbf{Y_{obs}},\mathbf{Y_{mis}})$ の対数尤度の条件付期待値 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k+1)})$ を計算する．続く M step では，E step にて求めた $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k+1)})$ を最大化する $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k+1)}$ を求める．$\boldsymbol{\theta}^{(k+1)}$ を更新後のパラメータとして，$(k+1)$ 回目の E step を行う．このようにして，E step と M step を反復過程の収束条件が満たされるまで繰り返し，最後に得られたパラメータ値を，$\boldsymbol{\theta}$ の最尤推定値とする．

解析対象となるデータが全て連続変数の場合，これらが従う確率密度関数として多変量正規分布を想定すると，興味のあるパラメータ $\boldsymbol{\theta}$ は平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ である．多変量正規分布は指数分布族であるため，E step は，観測データが与えられたもとでの十分統計量 $\mathbf{t(Y)}$ の条件付き期待値 $E_{\boldsymbol{\theta}^{(k)}}[t(\mathbf{Y}) \mid \mathbf{Y_{obs}}]$ を求めるステップと捉えることができる．また，M step は，方程式 $E_{\boldsymbol{\theta}}[t(\mathbf{Y})] = E_{\boldsymbol{\theta}^{(k)}}[t(\mathbf{Y}) \mid \mathbf{Y_{obs}}]$ を最大化する $\boldsymbol{\theta}$ を求めるステップとみなすことができる（小西, 2008）．欠測値を含む多変量正規確率変数に対する EM アルゴリズムの適用例は，渡辺（2008）に簡潔にまとめられている．

Step3 では，Step2 にて組ごとに推定された平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ の推定値を用いて欠測値を補完する．p 個の変数からなる n 個のデータに対し，i 番目のデータ $x_{i1}, x_{i2}, \ldots, x_{ip}$ を成分とする列ベクトル $\mathbf{x_i}$ $(p \times 1)$ が欠測に応じて以下のように分解されたとする．

$$\mathbf{x_i} = \begin{bmatrix} \mathbf{x_i^0} \\ \mathbf{x_i^1} \end{bmatrix} = \begin{bmatrix} \mathbf{x_{i,mis}} \\ \mathbf{x_{i,obs}} \end{bmatrix} \tag{1}$$

(1)の欠測パターンに応じて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{bmatrix} \tag{2}$$

と分解する（渡辺, 2008）．Step2 にて求められた $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値 $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_0 \\ \tilde{\boldsymbol{\mu}}_1 \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{00} & \tilde{\boldsymbol{\Sigma}}_{01} \\ \tilde{\boldsymbol{\Sigma}}_{10} & \tilde{\boldsymbol{\Sigma}}_{11} \end{bmatrix} \tag{3}$$

と表すとき，$\mathbf{x_i}$ にて欠測となっている $\mathbf{x_{i,mis}}$ の補完値 $\tilde{\mathbf{x}}_{\mathbf{i,mis}}$ を

$$\tilde{\mathbf{x}}_{\mathbf{i,mis}} = \tilde{\boldsymbol{\mu}}_{\mathbf{0}} + \tilde{\boldsymbol{\Sigma}}_{\mathbf{01}}\tilde{\boldsymbol{\Sigma}}_{\mathbf{11}}^{-1}(\mathbf{x}_{\mathbf{i,obs}} - \tilde{\boldsymbol{\mu}}_{\mathbf{1}}) + \tilde{\boldsymbol{\varepsilon}}_{\mathbf{i}} \quad (4)$$

により算出する(i=1,2,…,n). p 個の変数に多変量正規分布を想定し，欠測である変数を目的変数，非欠測の変数を説明変数として線型回帰モデルをあてはめ，「予測値」として欠測値を補完する．線型回帰モデルにおいて，誤差項が互いに独立で平均 0，均一分散をもつ正規分布に従うとき，回帰係数の最尤推定値は，解析対象となるデータの平均値及び分散・共分散を用いて求めることができる（竹内, 2011）．この最尤解に Step2 にて求められた $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$ を与えることで，(4)が導かれる．

(4)において，$\tilde{\boldsymbol{\varepsilon}}_{\mathbf{i}}$ は根本的な不確実性（fundamental uncertainty）である（Honaker and King,2010;高橋・伊藤, 2013）．例えば，$\mathbf{x_1}$ では x_{11} のみが欠測，$\mathbf{x_2}$ では x_{21} のみが欠測であり，$\mathbf{x_1}$ と $\mathbf{x_2}$ では欠測以外全て同値である場合，(4)にて $\tilde{\boldsymbol{\varepsilon}}_{\mathbf{i}}$ がなければ補完された値 $\tilde{x}_{11}$ と $\tilde{x}_{21}$ は確実に同値となる．このように，欠測以外の変数が全て同一であっても欠測である変数の補完値は不確実であることを考慮する必要がある．本稿では，Step2 にて求めた分散の推定値と標準正規分布に従う乱数をかけあわせた値を用いることとした．

以下，表 1 に示した不完全データを用いて，これら 4 つのステップについて説明する．表 1 のデータは，SAS/STAT(R) 14.1 User's Guide, The MI Procedure にて用いられているデータセット Fitness1 から作成したものである．データセット Fitness1 は，SAS/STAT(R) 14.1 User's Guide, The REG Procedure の Example 97.2 にあるデータセット Fitness をもとに作成されており，格納されている 3 つの変数 Oxygen，RunTime，RunPulse にはいずれも欠測値がある．Fitness1 は 31 個のデータがあったが，ここでは説明を簡単にするため，10 個のデータに絞った．表 1 で「NA」が欠測値を意味する．

表 1：Aerobic Fitness のデータ

ID No.	1	2	3	4	5	6	7	8	9	10
Oxygen	44.609	45.313	54.297	NA	39.442	60.055	50.541	37.388	44.754	51.855
RunTime	11.37	10.07	8.65	11.95	13.08	8.63	NA	8.63	11.12	10.33
RunPulse	178	185	156	176	174	170	NA	170	176	166

Step1 では，欠測値を含んだ 10 個の不完全データに対し，大きさ 10 のブートストラップ標本を M 組生成する．生成された M 組のデータセットには，欠測データが含まれず補完する必要のない組も考えらえるが，欠測データが含まれている組が存在する可能性が高い．

Step2 では, Step1 で生成された M 組のデータセットに対し, EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定する．表 1 の Aerobic Fitness のデータにある Oxygen，RunTime，RunPulse の 3 変数に対する欠測値の補完を考えており，M 組のデータセットごとに $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定するため，Step2 終了時には，3×1 の列ベクトル $\boldsymbol{\mu}$ と 3×3 の行列 $\boldsymbol{\Sigma}$ の推定値が M 組算出される．

Step3 では，組ごとに Step2 で算出された $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を用い，組ごとに，欠測値を補完した 10 個のデータセットを生成することになる．例えば，Step2 で生成された M 組のデータセットのうち，1 組目では，表 2 のように, 10 個のデータのうち欠測データが, 2 番目の Oxygen (X_1), 5 番目の RunTime (X_2)及び RunPulse (X_3) であったとする．

表 2：リサンプリングされた Aerobic Fitness のデータ

ID No.	1	2	3	4	5	6	7	8	9	10
Oxygen	45.313	NA	37.388	45.313	50.541	44.754	45.313	44.609	45.313	54.297
RunTime	10.07	11.95	8.63	10.07	NA	11.12	10.07	11.37	10.07	8.65
RunPulse	185	176	170	185	NA	176	185	178	185	156

1 組目のデータに対して EM アルゴリズムを用いて得られた $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{pmatrix},\ \tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \tilde{\sigma}_{11} & \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \\ \tilde{\sigma}_{21} & \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{31} & \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{pmatrix}$$

と表す．2 番目のデータでは欠測は Oxygen(X_1)のみであるため，(3)に従い欠測パターンに応じて $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を分解すると

$$\tilde{\boldsymbol{\mu}}_{\mathbf{0}} = \tilde{\mu}_1,\ \tilde{\boldsymbol{\mu}}_{\mathbf{1}} = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{00}} = \tilde{\sigma}_{11},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{01}} = \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{10}} = \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{11}} = \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}$$

となり，2 番目のデータでの欠測 $\mathbf{x}_{\mathbf{2,mis}}$ の補完値 $\tilde{\mathbf{x}}_{\mathbf{2,mis}}$ を

$$\tilde{\mathbf{x}}_{\mathbf{2,mis}} = \tilde{x}_{2,1} = \tilde{\mu}_1 + \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix} \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}^{-1} \begin{bmatrix} x_{2,2} - \tilde{\mu}_2 \\ x_{2,3} - \tilde{\mu}_3 \end{bmatrix} + \tilde{\varepsilon}_{9,1}$$

$$= \tilde{\mu}_1 + \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix} \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}^{-1} \begin{bmatrix} x_{9,2} - \tilde{\mu}_2 \\ x_{9,3} - \tilde{\mu}_3 \end{bmatrix} + \sqrt{\tilde{\sigma}_{11}} z_1$$

により算出する．ここに，z_1 は標準正規乱数である．5 番目のデータでは，欠測は RunTime (X_2)と RunPulse (X_3)であるから，欠測パターンに応じて $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を分解すると

$$\tilde{\boldsymbol{\mu}}_{\mathbf{0}} = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix},\ \tilde{\boldsymbol{\mu}}_{\mathbf{1}} = \tilde{\mu}_1,\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{00}} = \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{01}} = \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{10}} = \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix},\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{11}} = \tilde{\sigma}_{11}$$

となる．よって，5 番目のデータでの欠測 $\mathbf{x}_{\mathbf{5,mis}}$ の補完値 $\tilde{\mathbf{x}}_{\mathbf{5,mis}}$ を

$$\tilde{\mathbf{x}}_{\mathbf{5,mis}} = \begin{bmatrix} \tilde{x}_{5,2} \\ \tilde{x}_{5,3} \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix} + \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix} \sigma_{11}^{-1} (x_{5,1} - \tilde{\mu}_1) + \begin{bmatrix} \tilde{\varepsilon}_{5,2} \\ \tilde{\varepsilon}_{5,3} \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix} + \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix} \sigma_{11}^{-1} (x_{5,1} - \tilde{\mu}_1) + \begin{bmatrix} \sqrt{\tilde{\sigma}_{22}} z_2 \\ \sqrt{\tilde{\sigma}_{33}} z_3 \end{bmatrix}$$

により算出する．ここに，z_2, z_3 は独立な標準正規乱数である．

このようにして欠測値が補完された M 組のデータセットに対し，個別に解析を実行する．パラメータ推定値と標準誤差を算出し，最終的な解析結果としてそれらをひとつに統合する．

3. SAS による実装例

EMB アルゴリズムによる多重補完法は，2.に示したとおり，1)欠測値を含む不完全データからブートストラップ標本を M 組生成する，2) M 組のブートストラップ標本に対し EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を求める，3) $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を用いて欠測値に対する補完を行う，4)補完されたデータセットごとに解析し得られた推定量を 1 つに統合する，という 4 つのステップから構成される．

ブートストラップ法に対応しているプロシジャは存在しないが，ブートストラップ法に対応するサンプルマクロ%BOOT が公開されており，SAS Institute Inc.の Technical Support よりダウンロードが可能である（http://support.sas.com/kb/22/220.html）．%BOOT を使用するにあたっては，まず，サンプルマクロダウンロード先にある%analyze を読みこんでおく必要がある．その上でブートストラップ法を行いたいデータセット名をマクロ引数 data =により指定すれば，BOOTDATA という名前でリサンプリングされたデータセットができあがる．デフォルトではリサンプリングの回数が 200 であるため，例えば，リサンプリングを 20 回行ないたい場合には samples = 20 と指定する．%BOOT では，特に指定しない限り，釣り合い型ブートストラップ法を行う（表 1）．

表 1: %BOOT のマクロ引数

data = 入力データセット名を指定.
samples = リサンプリング回数（ブートストラップ反復回数）を指定．デフォルトは200.
random = 擬似乱数を生成するときのシードを指定.
size= ブートストラップ標本の大きさを指定. デフォルトはdata =で指定される入力データセットのオブザベーション数として処理される.
balanced = デフォルトではsize =を指定しない限り1: 釣り合い型ブートストラップが適用される. size =を指定すると0: 一様リサンプリングが適用される.
biascorr = デフォルトは1: ブートストラップ偏り修正済み推定量
alpha = ブートストラップ信頼区間の信頼係数を指定．デフォルトは0.05（95%信頼区間）.
print = デフォルトは1: ブートストラップ推定値を出力，出力しない場合には0と指定.
chart = デフォルトは1: ブートストラップ分布を出力，出力しない場合には0と指定.

SAS プログラム 1 は，%BOOT を用いてブートストラップ法を行うプログラムの一例である．データセット Fitness1 に格納されている 3 つの変数 Oxygen，RunTime，RunPulse に対しブートストラップ法を行いたい場合，%analyze の MEANS プロシジャにてこれらの変数を VAR ステートメントで指定し，OUTPUT ステートメントで指定する変数もそれにあわせて更新しておく必要がある．Fitness1 には欠測データを含めて 31 個のデータがある．SAS プログラム 1 では，%BOOT において sample = 20 と指定し，size = は特に指定していないため，データセット BOOTDATA には，釣り合い型ブートストラップ法によりリサンプリングされた 31 個のデータが 20 組セットされる．変数_OBS_は，リサンプリングされたデータが Fitness1 のどのデータから得られたものかを表す. 例えば_OBS_ = 24 ならそれは Fitness1 の 24 番目のデータに該当することを意味する．_OBS_の頻度集計を行うことで，リサンプリングされた 31×20 = 620 個のブートストラップ標本において，Fitness1 の 31 個のデータがそれぞれ 20 回ずつ抽出されていることが確認できる．

SAS プログラム 1: %BOOT によるブートストラップ法の実行

```
%macro analyze(data = , out = ) ;
    proc means noprint data = &data vardef = n ;
        output out = &out(drop = _freq_ _type_) var = Oxygen RunTime RunPulse ;
        var Oxygen RunTime RunPulse ;
        %bystmt ;
    run ;
%mend ;
%boot(data = Fitness1, samples = 20, random = 170723, print = 0, chart = 0);
```

EM アルゴリズムによる $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値の算出には，MI プロシジャを用いる．

MI プロシジャは，欠測値を補完した完全データを複数組作成するためのプロシジャである．どのような方法で欠測値を補完するかは，欠測パターンと補完したいデータのタイプに依存する．補完したいデータが連続量で欠測パターンが単調であれば，多変量正規性を仮定するパラメトリック回帰，あるいは傾向スコアを用いたノンパラメトリック法のいずれかが適用される．補完したいデータが 2 値で欠測パターンが単調であればロジスティック回帰が適用される．補完したいデータが連続量で任意の欠測パターンのときには，多変量正規性を仮定する MCMC 法が適用される．また Ver.9.3 から，Fully Conditional Specification による多重補完も実行可能となった（SAS Institute Inc., 2011；多田, 2013）．これは，欠測パターンや補完したいデータのタイプによらず適用できる．

このように，MI プロシジャでは，様々なアルゴリズムを用いて欠測値を補完することができるが，EM ステートメントを用いることにより，多重補完は行わず EM アルゴリズムにより $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定を行うことも可能である．そこでこの機能を用いて，EMB アルゴリズムによる多重補完法を行う際の 2 つ目のステップ（*M* 組のブートストラップ標本に対し EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を求める）を実装する．

SAS プログラム 2 はその一例である. MI プロシジャにて, EM ステートメントにてオプション OUTEM = を指定することにより，EM アルゴリズムを用いた平均値，分散，共分散の推定値を保存することが可能となる．前述のとおり，多重補完なしで EM アルゴリズムを行いたいため，オプション nimpute = 0 と指定した．SAS プログラム 3.1 で作成されたデータセット BOOTDATA では，リサンプリングされた 31 個のデータが 20 組存在している．SAS プログラム 2 では，20 組のデータがデータセット BOOTDATA の変数_sample_により識別できることを利用して，20 組のデータそれぞれに対し $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を算出させている．MI プロシジャでは， $\boldsymbol{\mu}$ の初期値として入力データセットの変数ごとに算出された平均値を用い， $\boldsymbol{\Sigma}$ の初期値は，非対角成分を全て 0，対角成分を入力データセットの変数ごとに算出された分散として，E-step と M-step の反復計算が行われる．

SAS プログラム 2: EM アルゴリズムによる推定

```
%macro m_miem(inds = , samples = , seed = , outem = , outmi= ) ;
    data m_wkst1 ; run ; data m_wkst2 ; run ;
    %do irep = 1 %to &samples. ;
        data m_wrk ; set &inds. ;if _sample_ = &irep. ;run ;
```

```
        proc mi data = m_wrk seed = &seed. nimpute = 0 noprint ;
          em outem = m_outem out = m_outmi initial = ac ;
          var Oxygen RunTime RunPulse ;
        run ;
        data m_outem ;set m_outem ; _sample_ = &irep. ; run ;
        data m_outmi ;set m_outmi ; _sample_ = &irep. ; run ;
        data m_wkst1 ;set m_wkst1 m_outem ;if _sample_ ^= . ;run ;
        data m_wkst2 ;set m_wkst2 m_outmi ;if _sample_ ^= . ;run ;
    %end ;
    data &outem. ;set m_wkst1 ;run ; data &outmi. ;set m_wkst2 ;run ;
    proc datasets library = work nolist nowarn ;
      delete m_outem m_outmi m_wrk m_wkst1 m_wkst2 ;
    run ;quit ;
%mend m_miem ;
%m_miem(inds = BOOTDATA, samples = 20, seed = 170723, outds = outem, outmi = outmi) ;
```

MI プロシジャの EM ステートメントにてオプション OUT =を指定すると, 欠測値が補完されたデータセットが生成され, EM アルゴリズムに基づき算出された期待値が出力される（SAS Institute Inc., 2015）. ただし, これは $\mathbf{x}_{i,mis}$ の補完値 $\tilde{\mathbf{x}}_{i,mis}$ を

$$\tilde{\mathbf{x}}_{i,mis} = \tilde{\boldsymbol{\mu}}_0 + \tilde{\boldsymbol{\Sigma}}_{01}\tilde{\boldsymbol{\Sigma}}_{11}^{-1}(\mathbf{x}_{i,obs} - \tilde{\boldsymbol{\mu}}_1) \tag{5}$$

により算出した場合に相当する. 従って, (4)により欠測値を補完するためには, $\tilde{\boldsymbol{\varepsilon}}_i$ を加えるための SAS プログラムが別途必要となる. SAS プログラム 3 としてその一例を示す.

SAS プログラム 3: EMB アルゴリズムによる補完

```
data m_outmi ;set OUTMI ;keep _sample_ Oxygen RunTime RunPulse ;run ;
data m_outmi ;set m_outmi ;by _sample_ ;if first._sample_ = 1 then rep = 0 ;rep +1 ;run ;
data m_ind ;set BOOTDATA ;
  keep _sample_ M1-M3 ;
  array aa M1-M3 ; array bb Oxygen RunTime RunPulse ;
  do over aa ;if bb = . then aa = 0 ;else if bb^= . then aa = 1 ;end ;
run ;
data m_ind ;set m_ind ;by _sample_ ;if first._sample_ = 1 then rep = 0 ;rep +1 ;run ;
data m_var ;set OUTEM ;if _TYPE_ ="COV" ;run ;
data m_var ;set m_var ;
  keep _sample_ varno var ;
  if _NAME_ = "Oxygen" then do ;varno = 1 ;var = Oxygen ;end ;
  else if _NAME_ = "RunTime" then do ;varno = 2 ;var = RunTime ;end ;
```

```
   else if _NAME_ = "RunPulse" then do ;varno = 3 ;var = RunPulse ;end ;
run ;
proc transpose data = m_var out = m_var prefix = var ;var var ;by _sample_ ;run ;
data m_ind ; merge m_ind m_var(drop = _NAME_) ; by _sample_ ;run ;
data m_outmi ;
   merge m_outmi m_ind ;
   by _sample_ rep ;
   array aa Oxygen RunTime RunPulse ;
   array bb I_Oxygen I_RunTime I_RunPulse ;
   array cc M1-M3 ;
   array dd var1-var3 ;
   do over aa ;if cc = 1 then bb = aa ;else if cc = 0 then bb = aa + sqrt(dd)*rand('normal') ; end ;
run ;
data outemb ;
   set m_outmi ;
   keep _sample_ I_Oxygen I_RunTime I_RunPulse ;
run ;
```

SAS プログラム 2 まで実行した時点で，既にデータセット OUTMI に欠測値に対し(5)により補完された値が代入されている．このため，あとはデータセット OUTEM に保存されている $\tilde{\boldsymbol{\Sigma}}$ と標準正規乱数から $\tilde{\boldsymbol{\varepsilon}}_i$ を算出し，データセット OUTMI に出力されている補完値に加えればよい．SAS プログラム 3 では，データセット BOOTDATA から，欠測なら 0，非欠測なら 1 をとる欠測識別変数をもたせたデータセット M_IND を用意し，この処理を行っている．

補完されたデータセットごとに解析し得られた解析結果を 1 つに統合するステップは，SAS プログラム 4 のとおり, UNIVARIATE プロシジャと MIANALYZE プロシジャを使って簡単に実行できる（SAS Institute Inc., 2015; 石田・斉藤,2014）.

SAS プログラム 4: パラメータ推定値の出力

```
proc univariate data = OUTEMB noprint ;
   var I_Oxygen I_RunTime I_RunPulse ;
   output out = out mean = Oxygen RunTime RunPulse stderr = SOxygen SRunTime SRunPulse ;
   by _sample_ ;
run ;
proc mianalyze data = out edf = 30 ;
   modeleffects Oxygen RunTime RunPulse ;
   stderr SOxygen SRunTime SRunPulse ;
run ;
```

4. おわりに

EMB アルゴリズムによる多重補完法を把握し，4 つのステップに分割した．分割したステップに対応するプロシジャあるいは SAS マクロを探し，SAS プログラム上でどのような処理をしているかを確認した．その上で，対応するプロシジャあるいは SAS マクロと分割したステップとを比較し，対応するプロシジャで必要なオプションを指定，足りない部分を実行するためのプログラムを用意した．以上により，%BOOT，MI プロシジャ，DATA ステップによる加工プログラムを連動させることで，EMB アルゴリズムによる欠測値補完の実装が可能となった．

代替案として SAS と R とを連動させる方法がある．R は，フリーで配布されている統計分析ソフトウェアであり，AmeliaII パッケージを用いて，EMB アルゴリズムによる多重補完法を行うことができる．そこで，SAS で用意した欠測値を含むデータセットを R にわたし，R に EMB アルゴリズムによる欠測値補完を行わせ，欠測値が補完されたデータセットを SAS に返し，SAS を用いて解析結果を統合することが考えられる．SAS 上で R を動かして解析結果を得る方法については，船尾(2009)が詳しく紹介している．

本稿では，多重補完法のひとつとして，EMB アルゴリズムによる多重補完法をとりあげた．この方法は，既存のプロシジャでは実行できない．しかし，対応するプロシジャがなくとも，SAS Institute Inc.より公開されている SAS マクロや関連するプロシジャを連動させることにより，簡単に実装することができた．

参考文献

[1] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[2] Honaker, J., King, G. (2010). What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science*. **54**: 561–581.

[3] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*. **45**:1–47.

[4]小西貞則, 越智義道, 大森裕浩 (2008). 計算統計学の方法 ―ブートストラップ･EM アルゴリズム･MCMC. 朝倉書店.

[5]汪金芳, 桜井裕仁 (2011). ブートストラップ入門. 共立出版.

[6]渡辺美智子 (2008). EM アルゴリズム. 21 世紀の統計科学. 日本統計学会 HP 版. http://park.itc.u-tokyo.ac.jp/atstat/jss75shunen/Vol3.pdf (閲覧日：2017 年 4 月 21 日).

[7]竹内啓（監修），市川伸一, 大橋靖雄, 岸本淳司, 浜田知久馬,下川元継,田中佐智子（著）(2011). SAS によるデータ解析入門 第 3 版. 東京大学出版会.

[8]高橋将宜, 伊藤孝之 (2013). 経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価.～統計研究彙報 **70**(2):19–86.

[9]高橋将宜, 伊藤孝之 (2014). 様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～. 統計研究彙報 **71**(3):39–82.

[10]SAS Institute Inc. (2015). *SAS/STAT(R) 14.1 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[11]SAS Institute Inc. (2011). *SAS/STAT(R) 9.3 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[12]多田圭佑 (2013). MI Procedure による多重代入 SAS ver 9.3 における新機能の紹介. SAS ユーザー総会 論文集, 23-44.

[13]石田和也,斉藤和弘 (2014). PROC MIANALYZE を用いた、多重代入法による結果の統合. SAS ユーザー総会 論文集, 627-640.

[14]舟尾暢男 (2009). SAS でベイズ推定を行う方法 -proc MCMC vs R&WinBUGS-. SAS ユーザー総会論文集, 95-114.

連絡先

E-mail: *yada-s@a2healthcare.com*