相加的血縁行列の逆行列を用いた傾斜勾配法による 混合モデル方程式の解法の比較

佐 藤 正 寛

農業生物資源研究所, つくば市 305-8602

(2003 年 6 月 2 日受付, 2003 年 7 月 23 日受理)

要 約 混合モデル方程式を構築せずあらかじめ A^{-1} のみを算出した後に解を求める Semi-indirect 法と方程式を構築して解を求める Direct 法を, スケーリング付き傾斜勾配法により比較した。データはモンテ・カルロ法により発生させた 8 セットで, データセット 1~4 は単一形質で個体数約 2,000,000 頭, データセット 5~8 は 5 形質で個体数約 400,000 万頭である。混合モデル方程式は母数効果に続き変量効果を配置した。Direct 法によるスケーリング付き傾斜勾配法プログラムは単精度計算では収束しないことがあった。演算時間において, 両者に大きな違いはみられなかった。Semi-indirect 法は Direct 法に比べメモリを必要としなかった。以上の結果より, Semi-indirect 法は必要メモリが少ないことから, より複雑なモデルが用いられたときに有効性が高まるものと考えられた。

緒 言

大規模な情報から最良線形不偏予測(BLUP)法により育種価を推定する場合, 方程式を構築して解を得る方法(Direct法)は多くのメモリを必要とする。この問題に対し, SCHAEFFER と KENNEDY^{1,2)} は混合モデル方程式を構築せず, 最小限のメモリを使って解を求める方法 (Indirect 法) を考案した。その後, 逐次過剰緩和法 (SOR 法) や線形加速法 (SOJ 法) によるアルゴリズムを用いた Indirect 法が広く普及するようになった³⁻⁵⁾。一方, 1990 年代前半を中心に近交係数を算出するアルゴリズムが次々に考案され⁶⁻⁸⁾, パソコンレベルでも大規模な血統情報に基づく近交係数の算出ソフトが開発された⁹⁾。このように, 近交係数の演算法が発達したことから, QUAAAS の方法¹⁰⁾ により相

加的血縁行列の逆行列 (A^{-1}) を導くことが容易になった。また, コンピュータが高速化するとともに記憶容量が激増するなど, コンピュータのハードウェアの発達は, より大きなメモリを必要とするもののプログラミングが容易な Direct 法による演算を可能にするものである。

最近, 混合モデル方程式の解法として, 方程式にその係数行列の対角要素を用いて前処理を施し, 傾斜勾配法により解を得る方法 (スケーリング付き傾斜勾配法; SCG 法) が注目されている。そこで本研究では, SCG 法に基づく方程式の異なるいくつかの解法を提案し, 方程式を格納するために必要な記憶容量とそれを解くための演算時間を比較した。これにより, あらかじめ A^{-1} のみを算出した後に混合モデル方程式を解く演算方法の有効性を明らかにすることを目的とした。

Comparison of Algorithms with Setting up Inverse of Additive Relationship Matrix for Solving Mixed Model Equations Using Scaled Conjugate Gradient Iteration

M. SATOH

National Institute of Agrobiological Sciences, Tsukuba-shi, 305-8602, Japan

方 法

1. データの発生

世代重複のない 1 形質または 5 形質の継代記録をモンテ・カルロ法によるコンピュータシミュレーションにより発生させた。基礎集団（第 0 世代）の各個体は、非近交で、個体間の血縁はないものとした。基礎集団を除く世代数は 5 または 10 の 2 通りとした。1 形質の場合は表型値、5 形質の場合は基礎集団の表型標準偏差の逆数を経済重みとした指数により、各世代雄 1 頭に交配する雌の頭数が 1 または 10 となるように選抜した。産子の雌雄頭数は、それぞれ平均 5、分散 4 の正規分布に従うものとし、小数点以下第 1 位を四捨五入した値とした。一腹から雄 2 頭雌 2 頭を育成し、これに満たない場合、里子などの補充は行わなかった。基礎集団を除く全個体数は、1 形質の場合約 200 万頭、5 形質の場合約 40 万頭となるように世代あたりの繁殖雌の頭数を定めた。基礎集団は記録を持たず、それ以外の個体はすべての形質において 1 記録を持つものとした。

各形質の記録 y_{ijklm} は次の無限遺伝子座を仮定したモデルにより発生させた。

$$y_{ijklm} = S_{ij} + HGS_{ik} + g_{im} + e_{ijklm}$$

ここで、 S_{ij} は i 番目の形質の j 番目の性の効果、 HGS_{ik} は i 番目の形質の k 番目の群-世代-季節の効果、 g_{im} は i 番目の形質の m 番目の個体の育種値、 e_{ijklm} は無作為誤差である。基礎集団における g_{im} は正規乱数と遺伝分散に基づき、またそれ以外の世代では個体 m の両親平均および両親の近交係数をメンデルアンサンプリングに考慮して発生させた。形質数が 1 のとき、表型分散は 100、遺伝率は 0.4 とした。また、形質数が 5 のとき、表型分散はそれぞれ 1, 10, 100, 1,000, 1,0000、これに対応する遺伝率は 0.2, 0.3, 0.4, 0.5, 0.6、形質間の遺伝相関および環境相関はすべて 0.1 とした。各形質の表型平均は変動係数を 0.1 として定めた。群の数は 1 群の平均が 100 頭となるように定め、また季節の効果は 4 水準とした。群-世代-季節の各水準に属する個体は同一世代内で無作為に配置した。性の効果（雌に対する雄の効果）の大きさは、すべての形質において 1 表型標準偏差 ($1.0\sigma_p$) と

した。また、群-世代-季節の効果の大きさは、 $-1.0\sigma_p \sim 1.0\sigma_p$ の範囲で一様分布する値を各水準に無作為に割り当てた。

2. 演算方法

データの発生と同じモデルおよび同じ遺伝的パラメーターを用いて、以下の混合モデル方程式 (1) を解くことにより育種値を推定した。すなわち、混合モデルを、

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

とする。ここで、 $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$, $\mathbf{e}$ はそれぞれ観測値、母数効果、育種値、誤差のベクトル、 $\mathbf{X}$ および $\mathbf{Z}$ はそれぞれ $\mathbf{y}$ と $\mathbf{b}$, $\mathbf{y}$ と $\mathbf{u}$ を関連づける生起行列である。また、 $E(\mathbf{u}) = E(\mathbf{e}) = \mathbf{0}$, $\text{var}(\mathbf{u}) = \mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$, $\text{var}(\mathbf{e}) = \mathbf{R}$, $\text{cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$ とする。ただし、 $\mathbf{G}_0$ は個体ベースの遺伝分散共分散行列、 $\mathbf{A}$ は相加的血縁行列、 $\otimes$ はクロネッカー積を表す。このとき、 $\mathbf{u}$ の BLUP である $\mathbf{u}^\circ$ は、

$$\begin{aligned} & \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b}^\circ \\ \mathbf{u}^\circ \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad \dots(1) \end{aligned}$$

を解くことによって得られる。ただし、 $\mathbf{b}^\circ$ は母数効果 $\mathbf{b}$ に関する混合モデル方程式 (1) の解である。いま、(1) を、

$$\mathbf{M}\mathbf{x} = \mathbf{v} \quad \dots(2)$$

とおく。さらに、 i 番目の対角要素が $\text{diag}\{d_i\} = 1/\text{diag}\{m_i\}^{1/2}$ である対角行列を $\mathbf{D}$ とする。ただし、 $\text{diag}\{m_i\}$ は係数行列 $\mathbf{M}$ の i 番目の対角要素である。本研究では、傾斜勾配法 (CG 法) の前処理に方程式のスケールリングを行った。すなわち、 $\mathbf{x} = \mathbf{D}\mathbf{z}$ とおき、(2) 式を、

$$\mathbf{DMDz} = \mathbf{Dv}$$

と変形する。次に、初期ベクトルを $\mathbf{z}_0 = \mathbf{0}$, $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{Dv} - \mathbf{DMDz}_0$ とおき、

$$\left. \begin{aligned} \mathbf{z}_{k+1} &= \mathbf{z}_k + \mathbf{p}_k(\mathbf{p}_k, \mathbf{r}_k) / (\mathbf{p}_k, \mathbf{DMDp}_k) \\ \mathbf{r}_{k+1} &= \mathbf{Dv} - \mathbf{DMDz}_{k+1} \\ \mathbf{p}_{k+1} &= \mathbf{r}_{k+1} - \mathbf{p}_k(\mathbf{r}_{k+1}, \mathbf{DMDp}_k) / (\mathbf{p}_k, \mathbf{DMDp}_k) \end{aligned} \right\} \dots(3)$$

を反復する。ここで、下付き文字 k は k 回目の反復を表す。収束条件は、

$$\|\mathbf{v} - \mathbf{M}\mathbf{x}\| / \|\mathbf{x}\| < 10^{-8}$$

とした。ここで、 $\|\mathbf{x}\|$ は $\mathbf{x}$ のノルムを表す。

演算の手順は、まず前処理として、各個体の両

親, 当該個体が属する各母数効果の水準のリナ
ンバリング, 欠測記録のある形質のコード化, $\mathbf{A}^{-1}$ の
算出等を行う。次に, (3)において, ①(1)の左辺を
構築せず, コード化した情報と $\mathbf{A}^{-1}$ を用いて反復
毎に $\mathbf{DMDz}_{k+1}$ や $\mathbf{DMDp}_k$ を計算する方法 (M1),
②(1)の左辺の部分行列 ($\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$, $\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}$ または
 $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}$) を構築・格納し, それ以外はコード
化した情報と $\mathbf{A}^{-1}$ を用いて反復毎に $\mathbf{DMDz}_{k+1}$
や $\mathbf{DMDp}_k$ を計算する方法 (それぞれ, 構築した
部分行列に対応して, M2, M3, M4), ③左辺のす
べてを構築・格納する方法 (M5: Direct 法), の
それぞれについて, 方程式を格納するために必要
な記憶容量とそれを解くための演算時間を調べ
た。なお, 本研究では M1~M4 を Semi-indirect
法と称することとする。混合モデル方程式は係数
行列の上三角部分におけるノンゼロ要素とそれら
の行 (列) 番号を格納した。また, SCG 法のアル
ゴリズムは単精度実数 (4 バイト) と倍精度実数
(8 バイト) の両者により実行した。計算には,
Windows Xp Workstation (3.06 GHz, Pentium
IV, キャッシュ 512 KB, メインメモリ 1.5 GB) を
用いた。

結 果

コンピュータシミュレーションにより発生させ
た 8 種類のデータセットを表 1 に示した。方程式

の数はいずれのデータセットも 210~230 万の範
囲であった。しかし, ノンゼロ要素数は 1 形質の
とき約 1,000 万個であるのに対し, 5 形質のとき
約 4,400 万個で, 5 形質の混合モデル方程式の係
数行列は 1 形質のそれに比べ密であった。

各データセットにおける前処理時間および収
束までの反復計算時間, 反復回数およびメモリ使用
量を表 2 に示した。なお, 反復計算時間は方程式
の構築に要する時間を含んでいる。いずれのデー
タセットにおいても, M2 および M3 の計算時間
は M1, M4, M5 のいずれよりも長くなったため,
これらの結果は省略した。また, M1 および M4
ではすべてのデータセットにおいて解は収束した
が, M5 による単精度計算では 8 種類のうち 5 種
類のデータセットにおいて解が収束しなかった。
そのため表 2 にはすべての方法が収束した倍精度
計算による結果を示した。前処理時間は個体数
が多いほど, また世代数が多いほど長くなり, これ
らの演算時間の多くは $\mathbf{A}^{-1}$ の算出, 中でも近交係
数の算出に費やされた。反復計算時間は 1 形質よ
りも 5 形質で長く, 雄 1 対雌 1 交配よりも雄 1 対
雌 10 交配のほうが長くなった。1 形質の場合には
M1 の反復計算時間が最も短い傾向にあり, 次い
で M4, M5 の順となった。一方, 5 形質の場合に
は M4 の反復計算時間が最も短く, 次いで M5, M
1 の順となった。メモリ使用量は M1 が他に比べ

表 1. 各データセットの要約
Table 1. Summary for each data set

Data set	Number						
	sires	dams	generations	traits	Animals	Equations	Non-zero elements ¹⁾
1	100,000	100,000	5	1	2,185,944	2,265,945	10,056,801
2	10,000	100,000	5	1	2,095,928	2,175,929	9,786,679
3	50,000	50,000	10	1	2,085,917	2,165,918	9,806,559
4	5,000	50,000	10	1	2,041,013	2,121,014	9,672,163
5	20,000	20,000	5	5	437,032	2,265,165	44,435,744
6	2,000	20,000	5	5	419,223	2,176,120	44,186,459
7	10,000	10,000	10	5	417,317	2,166,590	44,164,244
8	1,000	10,000	10	5	408,126	2,120,635	44,010,479

¹⁾ Upper triangular matrix of BLUP equations.

表 2. 各データセットにおける前処理および収斂までの反復計算時間 (秒), 反復回数およびメモリ使用量 (MB)

Table 2. Central processing unit time in seconds used for preprocessing and iterating until convergence, number of iterations, and random memory access requirements in megabytes for each data set

Data set	CPU time (second)				Number of iterations	Memory requirements (MB)		
	Pre-processing	Iterating				M1	M4	M5
		M1 ¹⁾	M4 ¹⁾	M5 ¹⁾				
1	32.6	62.0	79.3	80.5	50	101.6	203.2	255.0
2	29.5	94.6	107.7	105.1	79	100.2	200.4	249.5
3	767.1	59.6	70.0	113.9	59	100.1	200.1	249.7
4	525.7	132.8	130.3	138.6	115	99.4	198.8	247.0
5	9.2	555.3	434.0	471.8	129	20.3	461.4	698.3
6	8.0	1,011.6	812.1	819.0	240	20.0	457.2	694.3
7	96.7	669.5	503.4	521.7	162	20.0	456.8	693.9
8	55.6	1,374.2	1,019.6	1,052.1	341	19.9	454.4	691.4

¹⁾ See the text.

て最も少なく, 特に形質数が多いときその傾向は顕著であった。

考 察

近年, 混合モデル方程式の反復解法として SCG 法の有効性が示されている¹¹⁻¹³⁾。SCG 法は式 (2) から式 (3) に示した反復過程における行列 $\mathbf{M} = \{m_{ij}\}$ とベクトル $\mathbf{v} = \{v_j\}$ の積の演算 $c_j = \sum m_{ij}v_j$ において, $m_{ij}v_j$ を加える順序が任意であるため, 飛び離れたブロックの積を同時に処理することが可能である。また SCG 法は, これまで混合モデル方程式の解法に用いられてきた SOR 法や SOJ 法に比べて, 収束性の高いことが知られている。一方, $\mathbf{A}^{-1}$ にはいくつかのアルゴリズムが提唱されており, またその上三角行列のノンゼロ要素数も個体数の 6 倍を超えることはない。このことは, あらかじめ $\mathbf{A}^{-1}$ を算出することにより, 混合モデル方程式の解を効率的に得る可能性を示唆するものである。

本研究は, あらかじめ $\mathbf{A}^{-1}$ のみを算出し, SCG 法により混合モデル方程式を解くための, 方程式の構築・格納方法の異なる 2 つの解法 (Semi-indirect 法と Direct 法) において, 必要な記憶容量

と演算時間を比較したものである。Semi-indirect 法は大きく前処理計算と反復計算に分けることができる。前者の計算時間はその大半が $\mathbf{A}^{-1}$ の算出に費やされる。その中でも, 近交係数の算出時間が最も長く, たとえばデータセット 4 の場合, 前処理計算のうち近交係数の算出に占める割合は約 98% であった。一方, 方程式の構築や反復計算に要する時間は係数行列の階数が同じであれば, その疎度に影響を受け, 疎行列のほうが計算時間は短くなると考えられる。しかし, 本研究における方程式の構築時間はほとんどが 20 秒前後であり, 係数行列のノンゼロ要素数とは無関係であった。また, 雄 1 対雌 1 交配よりも雄 1 対雌 10 交配のデータセットにおける収斂までの反復回数は多くなったが, これは $\mathbf{A}^{-1}$ におけるノンゼロ要素の位置に影響すると考えられる。すなわち, 扱う全個体数が同じであれば交配比にかかわらず $\mathbf{A}^{-1}$ のノンゼロ要素数はほぼ同じになる。しかし, 雄 1 対雌 10 の交配は雄 1 対雌 1 の交配に比べ特定の雄の子が多く存在することにより, 当該雄の行 (列) に相当する $\mathbf{A}^{-1}$ のノンゼロ要素数が多くなる。この偏りが SCG 法による収束性に影響していると思われる。

本研究では、方程式の格納法として M1 から M5 までの 5 通りの方法を比較した。M5 は方程式のすべてを格納する、いわゆる Direct 法である。Direct 法はプログラミングが容易である反面、形質数をこれ以上増やしたり、母性効果などの変量効果を含むモデルに対応させるためにはより多くのメモリを必要とする。一方、M1 は反復回数が多い場合、計算時間は長くなることが予想されるものの、メモリの使用量は他の方法に比べて非常に少ない。M4 は M5 よりも計算時間は短く、特に形質数が多い場合にその傾向が顕著であった。また、M4 のメモリ使用量は M1 と M5 の中間であった。

演算方法の選択は、計算誤差、計算時間およびメモリ使用量の優先度に依存する。表には示していないが、M5 に対する平均平方誤差は M2 が $10^{-7} \sim 10^{-4}$, M1, M3 および M4 はいずれも $10^{-1} \sim 10^1$ となった。しかし、Direct 法と Indirect 法では、方法の違いが誤差の大きさに影響する。たとえば、(3)式の $\mathbf{DMDz}_{k+1}$ の計算において、あらかじめ算出した $\mathbf{DMD}$ を $\mathbf{z}_{k+1}$ に乗じる場合と、 $\mathbf{Dz}_{k+1}$ を計算し、次に $\mathbf{M}$, $\mathbf{D}$ を左から順に乗じる場合とでは計算結果に違いの生じることがある。この場合、大規模な演算では計算誤差の大小が判別できない。このように、一般に真の解は不明であるため、計算誤差を論じることは難しい。一方、SCG 法では倍精度計算で解が収束しても、単精度計算では収束しない場合がある¹⁰⁾。本研究では Semi-indirect 法による単精度計算はすべてのデータセットにおいて解が収束したものの、Direct 法による単精度計算では半数以上のデータセットにおいて解が収束しなかった。また、Direct 法と Semi-indirect 法を比較した場合、計算時間の差よりもメモリ使用量の差が大きかった。そのため、より複雑なモデルに対して汎用性の高いプログラムを構築する場合には、Direct 法よりも Semi-indirect 法 (M1 や M4) が好ましいと考えられる。

文 献

- 1) SCHAEFFER, L.R. and B.W. KENNEDY : Computing strategies for solving mixed model equations, *J. Dairy Sci.*, **69**, 575-579, 1986.
- 2) SCHAEFFER, L.R. and B.W. KENNEDY : Computing solutions to mixed model equations, 3rd World Congr. Appl. Livest. Prod., **12**, 382-393, 1986.
- 3) MISZTAL, I. and D. GIANOLA : Indirect solution of mixed model equations, *J. Dairy Sci.*, **70**, 716-723, 1987.
- 4) BERGER, P.J., G.R. LUECKE and J.A. HOEKSTRA : Iterative algorithms for solving mixed model equations, *J. Dairy Sci.*, **72**, 514-522, 1989.
- 5) STRANDÉN, I. and E.A. MIÄNTYSAARI : Animal model evaluation in Finland : experience with two algorithms, *J. Dairy Sci.*, **75**, 2017-2022, 1992.
- 6) TIER, B. : Computing inbreeding coefficients quickly, *Genet. Sel. Evol.*, **22**, 419-430, 1990.
- 7) GOLDEN, B.L., J.S. BRINKS and R.M. BOURDON : A performance programmed method for computing inbreeding coefficients from large data sets for use in mixed-model analyses, *J. Anim. Sci.*, **69**, 3564-3573, 1991.
- 8) MEUWISSEN, T.H.E. and Z. LUO : Computing inbreeding coefficients in large populations, *Genet. Sel. Evol.*, **24**, 305-313, 1992.
- 9) 佐藤正寛 : 大規模血統情報から近交係数を算出するプログラムの開発, *日豚会誌*, **37**, 122-126, 2000.
- 10) QUAAS, R.L. : Computing the diagonal elements and inverse of a large numerator relationship matrix, *Biometrics*, **32**, 949-953, 1976.
- 11) TSURUTA, S., I. MISZTAL and I. STRANDÉN : Use of the preconditioned conjugate gradient algorithm as a genetic solver for mixed-model equations in animal breeding applications, *J. Anim. Sci.*, **79**, 1166-1172, 2001.
- 12) STRANDÉN, I. and M. LIDAUER : Solving large mixed linear models using preconditioned conjugate gradient iteration, *J. Dairy Sci.*, **82**, 2779-2787, 1999.
- 13) STRANDÉN, I., S. TSURUTA and I. MISZTAL :

Simple preconditionedrs for the conjugate gradient method : experience with

test day models, J. Anim. Breed. Genet., 119, 166-174, 2002.

Comparison of Algorithms with Setting up Inverse of Additive Relationship Matrix for Solving Mixed Model Equations Using Scaled Conjugate Gradient Iteration

Masahiro SATOH

National Institute of Agrobiological Sciences, Tsukuba-shi, 305-8602, Japan

A semi-indirect approach, where not the whole equations but inverse of additive relationship matrix was set up, was proposed in order to compute the solutions to mixed model equations. The semi-indirect and direct algorithms for solving mixed model equations using scaled conjugate gradient iteration were compared for 8 data sets generated by the Monte Carlo simulation. Data sets 1 through 4 contained a single dependent variable with about 2,000,000 animals. Data sets 5 through 8 consisted of five traits with about 400,000 animals. The ordering of mixed model equations was fixed effects followed by additive genetic random effects. In direct algorithm implemented with single precision, the scaled conjugate gradient program did not converge for several data sets. The CPU time using the new technique was not different from that by the direct approach. The semi-indirect approach required smaller size of memory than the direct approach. Advantage of semi-indirect approach depends on the model ; more complicated models will produce clearer differences.

Jpn. J. Swine Science, **40**, 3 : 149-154

Key words : Relationship matrix, Conjugate gradient, Mixed model equations