Article

Approximate Information and Accelerating for High-throughput Heterogeneous Data Analysis with Linear Mixed Models

Shengxin Zhu [1,2,†]

[1] Department of Mathematics, Xi'an Jiaotong-Liverpool University

[2] Research Institute of Big Data Analytics, Xi'an Jiaotong-Liverpool University; Shengxin.Zhu@xjtlu.edu.cn; Tel.: +86-512-8188-4753.

[†] This paper is based on the presentation *Fast Calculation of Restricted Maximum Likelihood Methods for Unbalanced High-throughput Data Analysis* on 2017 IEEE 2nd International Conference on Big Data Analysis, paper ID (ICBDA2017-289)

Abstract: Linear mixed models are frequently used for analysing heterogeneous data in a broad range of applications. The restricted maximum likelihood method is often preferred to estimate co-variance parameters in such models due to its unbiased estimation of the underlying variance parameters. The restricted log-likelihood function involves log determinants of a complicated co-variance matrix. An efficient statistical estimate of the underlying model parameters and quantifying the accuracy of the estimation requires the first derivatives and the second derivatives of the restricted log-likelihood function, i.e., the observed information. Standard approaches to compute the observed information and its expectation, the Fisher information, is computationally prohibitive for linear mixed models with thousands random and fixed effects. Customized algorithms are of highly demand to keep mixed models analysis scalable for increasing high-throughput heterogeneous data sets. In this paper, we explore how to leverage an averaged information splitting technique and dedicate matrix transform to significantly reduce computations and to accelerate computing. Together with a fill-in reducing multi-frontal sparse direct solver, the averaged information splitting approach improves the performance of the computation process.

Keywords: observed information; fisher information; averaged information splitting; approximate information

1. Introduction

Real-world data like these in animal/plant breeding, clinic trials, ecology and evolution, genome-wide association and many other fields are often heterogeneous. For example, one can not control how many child animals can one animal sire each time. In the clinic trials, individuals in every group are usually not equaled. For many repeated measurements, missing data are very common, which results in heterogeneous data. These cases are different with those in controlled experiments (in a laboratory) where the data usually enjoy a nice controlled block structure. For such controlled structured block data, the classical analysis of variance(ANOVA) approach can give a statistical efficient estimate. On contrast, the *restricted maximum likelihood method*(REML) introduced in [1] is an attractive method for heterogeneous data analysis. This approach is conceptually simple and is widely used in in animal/plant breeding [2], clinic trials, ecology and evolution [3], genome-wide association [4–7]. And it is receiving increasing attention. In this approach, the underlying observation $y \in \mathbb{R}^{n\times 1}$ is modeled by the mixed model

$$y = X\tau + Zu + \epsilon, \tag{1}$$

where $\tau \in \mathbb{R}^{p\times 1}$ and $u \in \mathbb{R}^{b\times 1}$ are fixed effects and random effects respectively. ϵ is noise. The random effects and noise are independent normal distribution such that $E(u) = 0$, $E(\epsilon) = 0$, $u \sim N(0, \sigma^2 G)$, $\epsilon \sim N(0, \sigma^2 R)$ and

$$\text{var}\begin{bmatrix} u \\ \epsilon \end{bmatrix} = \sigma^2 \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}, \tag{2}$$

where $G = G(\gamma)$ and $R = R(\phi)$ are parametric co-variance matrices.

Efficient statistical estimates $\hat{\tau}$ and $\tilde{u}$ for the fixed and random effects and uncertainty quantifications for such estimates require estimates of the unknown co-variance parameter $\theta = (\sigma^2, \gamma, \phi)$. Where $\hat{\tau}$ and $\tilde{u}$ satisfy the mixed model equations [8]

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}. \tag{3}$$

The uncertainty of the estimates are quantified by

$$\text{var}\begin{pmatrix} \tau - \hat{\tau} \\ u - \tilde{u} \end{pmatrix} = \sigma^2 \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1} := \sigma^2 C^{-1}.$$

For heterogeneous and unbalanced data sets, the restricted maximum likelihood (REML) is preferred for an unbiased or less biased estimate of the covariance parameter[1]. On contrast to the standard maximum likelihood estimate which requires maximizing a log-likelihood function directly, the REML estimate requires maximizing a marginal log-likelihood function. Alternatively, we can view that the REML maximizing a log-likelihood function in a restricted subspace which can remove redundant or correlated information used in estimating the fixed effects. The benefit is that it results in an unbiased estimate or less biased estimate which has been articulated in [9]. The restricted maximum likelihood method results in a restricted log-likelihood function of the form [10]

$$\ell_R = -\frac{1}{2}\left\{ const + \log|V| + \log|X^T V^{-1} X| + y^T P y \right\}, \tag{4}$$

where $V(\theta) = \sigma^2(R + ZGZ^T)$ and

$$P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}. \tag{5}$$

2. Objective

For many realistic problems, the variance-variance matrices of the random effects u and the noise ϵ are unknown. In these cases to quantify estimates of the fixed effects and random effects, one has to estimate the variance parameters θ first. A problem of great interest is to obtain a statistical efficient estimate of the variance parameter θ for high-throughput biological data sets according to the maximum (restricted) likelihood principle:

$$\hat{\theta} = \arg_{\theta > 0} \max \ell_R. \tag{6}$$

The first derivative of the log-likelihood is called a *score function*, which is given by [10, p.252]

$$\text{Score for } \theta_i : \quad S(\theta_i) = -\frac{1}{2}\{\text{tr}(P\frac{\partial V}{\partial \theta_i}) - y^T P \frac{\partial V}{\partial \theta_i} P y\}. \tag{7}$$

This formulae is standard, for a detailed self-contained derivation, the reader is directed to [9]. Therefore finding the REML estimate requires the stationary point of log-likelihood function ℓ_R, ie. the solution to the nonlinear score function $S(\hat{\theta}) = 0$. It usually requires the negative Jacobian matrix of the score, which is usually referred to as the *observed information*. The expected value of the observed information is called the *Fisher information*.

3. Main Challenges

The objective function, ℓ_R, involves log determinant terms and is computationally prohibitive for high-throughput biological data sets. Challenges exist for the derivative Newton approach[11], derivative-free approach and the Fisher-scoring approach. For a derivative (Newton-Ramphson) approach, the element of the observed information is [10][9],

$$2\mathcal{I}_O(\theta_i,\theta_j) = \text{tr}(P\frac{\partial^2 V}{\partial\theta_i\partial\theta_j}) - \text{tr}(P\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}) + 2y^TP\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}Py - y^TP\frac{\partial^2 V}{\partial\theta_i\partial\theta_j}Py. \tag{8}$$

It is too complicated for practical use. For a derivative-free approach, it often requires more total time even though less time per iterate because the method converges very slow and even doesn't converge for some difficult problems [12]. The Fisher-scoring approach employs the Fisher information matrix instead of the observed information matrix. Elements of the Fisher information matrix are given by [13,14]

$$\mathcal{I}(\theta_i,\theta_j) = \frac{1}{2}\text{tr}(P\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}). \tag{9}$$

The Fisher-scoring approach is a quasi-Newton method. It enjoys a simper formulae than the exact Newton-Ramphson approach does. But still, it is not scalable for high-throughput biological data sets due to the four matrix-matrix products in the Fisher information, $\mathcal{I}$.

4. Averaged Information splitting

When V depends linearly on the variance parameter θ, i.e. $\frac{\partial^2 V}{\partial\theta_i\partial\theta_j} = 0$, it was observed that the average of the observed and Fisher information enjoys a simper form [15]

$$\mathcal{I}_A = \frac{\mathcal{I}_O + \mathcal{I}}{2} = \frac{1}{2}y^TP\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}Py, \tag{10}$$

which can be efficiently evaluated by four matrix-vector multiplication and one inner product. And this formulae was used in [16] for general case without any proof. For more general cases, we prove that

[Averaged Information Splitting Theorem[17]] Let $\mathcal{I}_O$ and $\mathcal{I}$ be the observed information and the Fisher information for the residual log-likelihood of the linear mixed model respectively, then the average of the observed and the Fisher information can be split as $\frac{\mathcal{I}_O+\mathcal{I}}{2} = \mathcal{I}_A + \mathcal{I}_Z$, such that the expectation of $\mathcal{I}_A$ is the Fisher information matrix and $E(\mathcal{I}_Z) = 0$.

Theorem 1 indicates that the $\mathcal{I}_Z$ part which involves intensive computations is negligible. This builds up the foundation of using $\mathcal{I}_A$ as a good approximation to the negative Jacobian matrix in general cases. As a byproduct, it shows that $\mathcal{I}_A$ can be used as an alternative of the Fisher information matrix. In a natural language, the Theorem reads as

[1′] The average of the Fisher information and the observed information of the restrict log-likelihood for the linear mixed models can be split as a simper Fisher information matrix plus an random zero matrix.

5. Matrix transforms

The next problem to be handled is that the formulae for the matrix P in (5)

$$P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$$

doi:10.20944/preprints201704.0044.v1

is also too complicated and the matrix vector multiplication Py still involves considerable computations. The next results shows that Py is equivalent to a simper formulae, for a detail derivation, the reader is directed to [9, Theorem 6]. $Py = R^{-1}e$, where $e = y - X\hat{\tau} - Z\tilde{u}$, $\hat{\tau}$ and $\tilde{u}$ is the solution to the mixed model equation (3).

Theorem 3 states the equivalence between the projection of the observations Py and the weighted fitted residual $R^{-1}e$. The variance-variance matrix R, often enjoys a very simple structure, for example a diagonal matrix. The residual e can be obtained by solving the mixed model equation (3) of order $(p+b) \times (p+b)$. The freedom of unknowns in the mixed model equation is often much smaller than the order of P, or the number of observations.

According to Theorem 1 and Theorem 3, the approximate information matrix can be computed according to Algorithm 1.

Algorithm 1 Main steps to compute $\mathcal{I}_A$

```
Let $C$ be the coefficient matrix in (3), $W = [X, Z]$,
Factorize $C = LDL^T$
Solve (3) with the help of $LDL^T$ factorization
Calculate $\eta = Py = R^{-1}e = R^{-1}(y - X\tau - Zu)$
Calculate $\dot{V}_i = \frac{\partial V}{\partial \theta_i}$, let $Y = [\dot{V}_1\eta, \dot{V}_2\eta, \cdots, \dot{V}_m\eta]$
Solve the following mixed model equations with multiple right-hand side.
    $\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} T \\ U \end{pmatrix} = \begin{pmatrix} X^TR^{-1}Y \\ Z^TR^{-1}Y \end{pmatrix}. \quad (11)$
Compute $\Xi = PY = R^{-1}(Y - XT - ZU)$
Compute $\mathcal{I}_A = \frac{1}{2}Y^T\Xi$
```

The choice of information or (negative Jacobian matrix) distinguishes the exact Newton-Raphson method, the Fisher-scoring algorithm and the average information splitting approach. Difference between them are illustrated in Algorithm 2.

The approximate information $\mathcal{I}_A$ is much simper than information used in [5,6], [7, eq.6, eq.7]. And such an approximation significantly reduces computations and speeds up the linear mixed model together with the sparse techniques for sparse linear systems[18].

6. Accelerating techniques: sparse factorization and fill-in reducing algorithms

One of the main computational kernel is to compute the negative Jacobian matrix. In the averaged information splitting(AIS) approach, compute the approximate information matrix $\mathcal{I}_A$ depends heavily on the efficient factorization of the coefficient matrix in the mixed model equations (3). Since the coefficient matrix C is often very sparse, and the factorization is reused to solve the linear system with multiple right hand-sides (11). The computation amount of the factorization depends on the sparsity of the L factor. Therefore, it is of great importance to keep L as sparse as possible to reduce all the computations. Such an algorithm is often referred to as a *fill-in* reducing algorithm. It is a preprocess of the factorization by exchanging rows and columns of the coefficient matrix C.

Another functionality of an fill-in reducing algorithm is to make the factorization process more scalable. The LDL^T factorization is based on the Gaussian elimination process which was often believed essentially essentially sequential. Fill-in reducing is also a key technique which can increase the parallelism as many as possible. Figure 1 compares the sparsity of the LDL^T factor L of a matrix C and that of a reorder matrix. Here we employ the approximated minimum degree ordering [19] and the parallel multi-frontal LDL^T factorization algorithm[20].

Algorithm 2 Newton-Raphson (NR)/Fisher Scoring (FS)/Averaged Information Splitting (AIS) method to solve $S(\theta) = 0$.

```
Give an initial guess of $\theta_0$
for $k = 0, 1, 2, \cdots$ until convergence do
    Solve
        $\begin{cases} \mathcal{I}_O(\theta_k)\delta_k = S(\theta_k) & \text{for NR} \\ \mathcal{I}(\theta_k)\delta_k = S(\theta_k) & \text{for FS} \\ \mathcal{I}_A(\theta_k)\delta_k = S(\theta_k) & \text{for AIS} \end{cases}$
    $\theta_{k+1} = \theta_k + \delta_k$
end for
```

Table 1. Date sets for the benchmark problem. The column titles are the number of years (y), the number of centres (c), the number of varieties(v), thes number of levels of cross terms(y.c , y.v, v.c), the average varieties per year (v/year), and the averages year per variety (y/v), and the number of controlled varieties all year (control varieties all year).

DataSet	y	c	v	y.c	y.v	v.c	units	v/y	y/v	c.v
prob_1	12	22	130	132	673	2518	6667	56.1	5.2	10
prob_2	15	25	160	180	888	3527	9595	59.2	5.6	10
prob_3	22	25	188	264	1177	4215	12718	53.5	6.3	12
prob_4	25	25	262	300	1612	5907	17420	64.5	6.2	12
prob_5	25	25	390	300	2345	8625	25334	93.8	6.0	15
prob_6	25	35	390	425	2345	12249	35887	93.8	6.0	15
prob_7	30	35	470	510	3013	15087	46113	100.4	6.4	20
prob_8	30	35	620	510	3835	19737	58685	127.8	6.2	20
prob_9	35	40	720	700	4522	26432	81396	129.2	6.3	20
prob_10	40	50	820	1000	5262	37701	118403	131.6	6.4	20

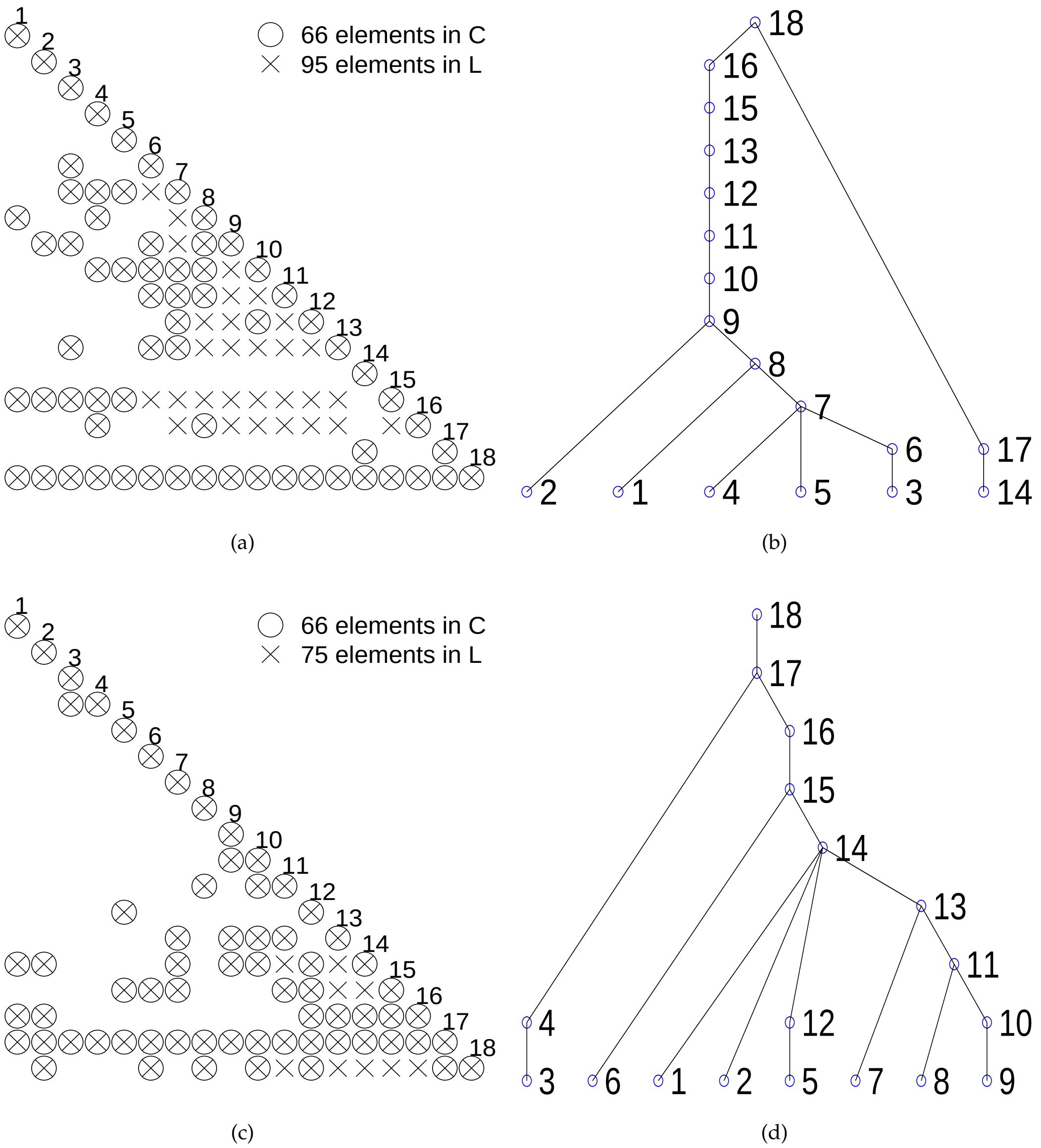
7. Numerical examples

A serial of plant breeding benchmark problems are used to verify performance of algorithms implemented here. These examples are based on a second-stage analysis of a set of variety trials with linear mixed models, i.e. based on variety predicted values from each trial. Trials are conducted in a number of years across a number of locations (centres). The data in Table 1 are generated by a program which allows you to specify the number of years, total number of centres and proportion of centres used per year, the number of control varieties (used every year), the number of test varieties entering the system per year and the average persistence of the test varieties, and the proportion of missing varieties per trial, where proportions of things are selected. They are sampled at random, and the life of each variety is generated from a Poisson distribution. This gives a three-way crossed structure (year*variety*site) with some imbalance. In the current model, all terms except a grand mean are fitted as random. The random terms are generated as independent and identically distributed normal distribution with variance components generated from a test program with similar structure used for the original SAS REML program, so it is just a variance components model.

The main part of the computing is the reordering and the factorization. Table 2 compares the algorithm implemented here and the counterpart in an existing commercial software, in which the second column is the number of effects, third to fifth columns are speedup of algorithms implemented here over the counterpart of an existing software package.

8. Discussion and future work

The aim of averaged information splitting is to remove computationally expensive and negligible terms so that a simper approximate information matrix is obtained. Such a splitting keeps the essential information and can be used as a good approximation to the observed information matrix



(a) (b)

(c) (d)

Figure 1. Illustration of the sparse structure of matrix C and its factor L in (a) and the sparse structure of the reordered matrix and its Cholesky factor L in (c). (b) and (d) are the elimination trees which indicate the elimination process correspond to the matrix C and its reordered matrix. The Cholesky elimination starts from a leaf node and ends in the root node. The height of he elimination tree stands for the sequential steps in the elimination, and the width of the elimination tree stands for the parallelism.

which is required for a derivative Newton method. The resulted formula is significantly simper than that used in the software package used in *Nature Genetics* [7, p.825, eq.8]. Together with the fill-in reducing and multi-frontal factorization sparse matrix techniques, the splitting can significantly improve the performance of a quasi-Newton approach to estimate the co-variance parameters in the linear mixed models. Part of the result has been implemented in a leading commercial breeding software package by VSN international(VSNi) Ltd. On a suit of test examples the method described here ran 10 times faster than counterpart of VSNi's existing software [18]. The splitting approach builds up a framework to analysis unbalanced data sets modeled by liner mixed models. Mathematically, theoretical proof on the convergence order of the quasi-Newton method based on the

Table 2. Speedup of of algorithms implemented here and the counterpart in an existing commercial software

prob	No. of Effects	Reordering	Factorization	all
Prob2	4796	*	25.98	6.46
Prob3	5892	15.4	*	4.95
Prob4	8132	14.78	*	7.55
Prob5	11711	12.32	*	6.65
Prob6	15470	15.15	8.91	7.16
Prob7	17.25	12.67	8.17	
Prob8	24768	19.01	11.01	9.22
Prob9	32450	26.29	14.92	10.64
Prob10	44874	36.78	5.93	9.31

averaged information splitting deserves to be further investigated. It is of great interest to leverage this method to actuarial science and other econometrics where the linear mixed models are used intensively.

Acknowledgments: The project is supported by the Natural Science Fund of China(No.11501044) and partially supported by NSFC(No.11571002, 11571047, 11671049, 11671051, 61672003). There is no funds for covering the costs to publish in open access.

The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Bibliography

1. Patterson, H.D.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554.
2. Masuda, Y.; Baba, T.; Suzuki, M. Application of supernodal sparse factorization and inversion to the estimation of (co) variance components by residual maximum likelihood. *Journal of Animal Breeding and Genetics* **2013**.
3. Bolker, B.M.e. Generalized linear mixed models:a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **2008**, *24*.
4. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.; Davidson, R.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nature Methods* **2011**, *8*, 833–835.
5. Listgarten, J.; Lippert, C.; Kadie, C.M.; Davidson, R.; Eskin, E.; Heckerman, D. Improved linear mixed models for genome-wide association studies. *Nature methods* **2012**, *9*, 525–526.
6. Zhang, Z.; Ersoz, E.; Lai, C.Q.; Todhunter, R.; Tiwari, H.K.; Gore, M.; Bradbury, P.; Yu, J.; Arnett, D.; Ordovas, J.; others. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* **2010**, *42*, 355–360.
7. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **2012**, *44*, 821–824.
8. Henderson,C.R.; Kempthorne, O.; Searle, S.R.; Krosigk,C. M. von. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **1959**, *15(2)*, 192–218.
9. Zhu,S.; Gu, T.; Xu, X; and Mo, Z Information splitting for big data analytics *International Conference on Cyber-enabled Distributed Computing and Knowledge Discovery* bf 2016 294-302
10. Searle, S.R.; Casella, G; McCulloch, C.E. Variance components Wiley Series in Probability and Statisitcs. John Wiley & Sons **2006**
11. Efron, B.; Hinkley, D. Assessing the accurancy of the maximum likelihood estimator:Observed versus expected Fisher information. *Biometrika* **1978**, *65*, 457–487.
12. Misztal, I. Comparison of computing properties of derivative and derivative-free algorithms in variance-component estimation by REML. *Journal of Animal Breeding and Genetics* **1994**, *111*, 346–355.
13. Jennrich, R.; Sampson, P. Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics* **1976**, *18*, 11–17.

14. Longford, N. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **1987**, *74*, 817–827.
15. Johnson, D.L.; Thompson, R. Restricted maximum likelihood estimation of variance componnets for univariate animal models using sparse matrix techniques and average information *Journal of Dairy Science* **1995**,*78(2)*, 449-456
16. Gilmour,A.R.; Thompson, R.; Cullis, B.R. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **1995**,*51(4)*, 1440-1450
17. Zhu, S.; Gu, T.; Liu, X. Information matrix splitting. *arXiv* **2016**. arXiv:1605.07646v1.
18. Sue Welham, S.; Zhu, S.; Wathen, A.J. Big data, fast models: faster calculation of models from high-throughput biological data sets. Knowledge Transfer Project Reprot IP12-009, Smith Industry Mathematics Institute, The University of Oxford, Oxford, 2013.
19. Amestoy, P.R.; Davis, T.A.; Duff, I.S. An approximate minimum degree ordering algorithms *SIAM Journal on Matrix Analysis and Applications***1996**,*17(4)*, 886-905
20. Davis, T.A. Algorithm 849: A concise sparse cholesky factroization package *ACM Transactions on Mathematical Software***2005**,*31(4)*, 587-591