Estimation of across-breed metafounder parameters in all-breed US dairy cattle.

A. Legarra[1,2] and M. Bermann[2]

[1] *Council on Dairy Cattle Breeding, Bowie, MD 20716, United States*

[2] *Animal and Dairy Science, University of Georgia, Athens, GA 30602, United States*

Corresponding author: andres.legarra@uscdcb.com

Abstract

Metafounders relationships **Γ** are useful to model missing pedigrees and different breeds. They are estimated from markers and pedigree, yet their estimation is difficult. Here we explore the behavior of two algorithms: General Least Squares (GLS) and Pseudo-Expectation Maximization (Pseudo-EM) to estimate **Γ** . We use a reduced set of US genotypes and pedigree, consisting in six breeds: Ayrshire, Brown Swiss, Guernsey, Holstein and Jersey, with number of genotypes (69 200 SNP markers) equal to 350, 681, 237, 13492 y 3541, with 370 307 animals in the overall pedigree. **Γ** estimates were also transformed to Fst differentiation indexes for ease of interpretation. Algorithms were fast. GLS resulted in biased estimates, with too high diagonal values of **Γ** for small breeds, and too high differentiation indexes across breeds. Algorithm Pseudo-EM resulted in more sensible estimates. Further, estimates of allele frequencies by Pseudo-EM agreed well with upstream estimates of base allele frequencies based on imputation. We recommend use of Pseudo-EM.

Key words: Metafounders, relationships, missing pedigree, breeds

Introduction

Disclaimer: a more complete manuscript has been submitted for peer-review publication. We provide this version as a short reference for the Interbull community. We also present some different results.

Metafounders (MF) are a conceptual framework to model relationships of unknown ancestors, both across breeds, and within breeds across time and selection pathway. They are particularly useful in single-step implementations, in particular for all-breed evaluations like the prospective US one. These relationships are condensed in a matrix **Γ**. Here we describe experiences on the estimation of **Γ** using two algorithms: Maximum Likelihood by Pseudo-Expectation-Maximization (Pseudo-EM) and General Least Squares (GLS). We used the newly released software gammaf90, a part of the blupf90 suite of programs (https://nce.ads.uga.edu/wiki/doku.php?id=start). We characterize a simple case where MF are essentially defined by breed and pathway (males or females ancestors). For this work we focus on results from the females ancestors (i.e. "dam" MF), as these are more representative of the general population.

Materials and Methods

Both algorithms need pedigree and markers. We used pedigree and markers extracted for the April 2025 evaluation of gestation length, which uses a sire model. This included 18 301 genotyped animals with a pedigree including ancestors up to 370 307 animals. All genotyped animals are purebred. Breeds included and genotype counts were Ayrshire (AY, 350), Brown Swiss (BS, 681), Guernsey (GU, 237), Holstein (HO, 13942) and Jersey (JE, 3541). We defined 11 metafounders: dam and sire MF for AY, BS and JE; a single one for GU; dam and sire for HO pre-1994 and dam and sire for

HO post-1994. The latter two were essentially for imported animals and are essentially ignored in the following results. SNP genotypes were the official 69 200 SNP CDCB list, after imputation.

Because we do not split MF across years, we are using a somehow fuzzy definition of base population. The mode of the year of birth of animals whose parents are missing is essentially 1950 to 1960 for all breeds, so our **Γ** estimates refer to that point.

Algorithms

Algorithms were as detailed in Legarra et al. (2024) although that article has some typos. Therefore, here we detail them again for completion. The definition of **Γ** if we knew the allele frequencies of two populations *i* and *j* is $\Gamma_{(i,j)} = \left(\frac{2}{k}\right)(2\boldsymbol{p}_i - 1)\left(2\boldsymbol{p}_j - 1\right)'$. Note that the same expression works for a population with itself.

In GLS, we estimate base allele frequencies using BLUE for $\widehat{\boldsymbol{p}}$ or equivalently, its GLS estimator (Gengler et al., 2007; Garcia-Baccino et al., 2017). Then we plug the estimates to obtain $\widehat{\boldsymbol{\Gamma}}_{GLS} = \frac{2}{k}\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)\left(2\widehat{\mathbf{P}} - \mathbf{11}'\right)'$. There is no iteration involved.

In Pseudo-EM we start with a guess of $\boldsymbol{\Gamma} = 0.1\mathbf{I}$ and iterate on (at iteration t):

1. Build $\mathbf{A}_{\Gamma(t)}$
2. For each marker i
   - $\widehat{2\boldsymbol{p}}_{i(t)} - \mathbf{1} = \mathbf{A}_{\Gamma(2,mf)(t)}\mathbf{A}^{-1}_{\Gamma 22(t)}\mathbf{z}_i$
3. $\widehat{\boldsymbol{\Gamma}}_{(t)} = \widehat{\boldsymbol{\Gamma}}_{(t-1)} - \mathbf{A}_{\Gamma(t)(mf,2)}\mathbf{A}^{-1}_{\Gamma(t)(2,2)}\mathbf{A}_{\Gamma(t)(2,mf)} + \frac{2}{k}\left(2\widehat{\mathbf{P}}_t - \mathbf{11}'\right)\left(2\widehat{\mathbf{P}}_t - \mathbf{11}'\right)'$

Where **z** contains genotypes coded as {-1,0,1} The likelihood is monitored. At convergence of the algorithm, we have estimates for **P** and **Γ**.

Results & Discussion

Computation took reasonable time – 13 hours for Pseudo-EM, and 3 minutes for GLS. Pseudo-EM took 950 iterations to reach a convergence criterion (squared relative difference of successive estimates) of 10^{-11}. Convergence was slow for elements with little information, e.g. Ayrshire (the least represented breed) with itself, but fast for well represented elements, e.g. Jersey with Holstein. Likelihood increased monotonically for Pseudo-EM (Figure 1), with the exception of the very first rounds, where there's a decrease – we believe that this is due to numerical round-off.

We report values only for the dam pathway (and in HO, for the dam pathway pre-1994). Table 1 and 2 contain estimates of Gamma coefficients. Table 3 contains Fst (Wright's fixation index) of differentiation across breeds, which is obtained as a direct transformation of **Γ** estimates using (Legarra et al., 2024): $F_{ST(i,j)} = \frac{\Gamma_{i,i} + \Gamma_{j,j} - 2\Gamma_{i,j}}{2+2-2\Gamma_{i,j}}$.

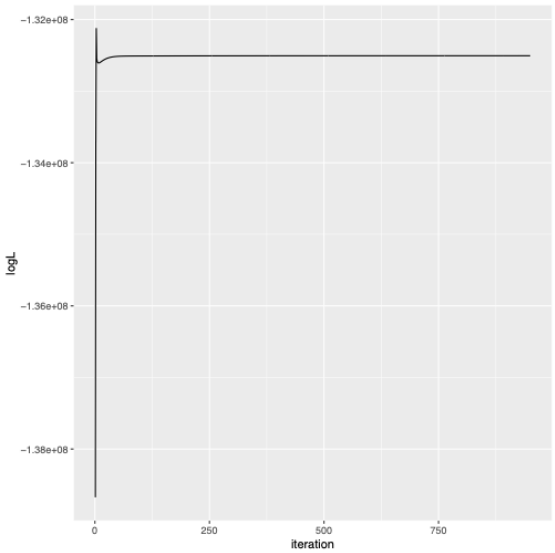


Figure 1. Evolution of the log-Likelihood during iterations of Pseudo-EM.

Table 1: Estimated Gamma values across breeds, General Least Squares method.

	AY	BS	GU	HO	JE
AY	1.20	0.17	0.21	0.18	0.18
BS	0.17	1.05	0.26	0.21	0.23
GU	0.21	0.26	0.60	0.26	0.31
HO	0.18	0.21	0.26	0.60	0.24
JE	0.18	0.23	0.31	0.24	0.85

Table 2: Estimated Gamma values across breeds, Pseudo-Expectation Maximization

	AY	BS	GU	HO	JE
AY	0.57	0.29	0.30	0.27	0.26
BS	0.29	0.59	0.31	0.25	0.29
GU	0.30	0.31	0.53	0.26	0.33
HO	0.27	0.25	0.26	0.39	0.26
JE	0.26	0.29	0.33	0.26	0.58

Table 3: Estimated Fst values across breeds. Upper triangular: General Least Squares method. Lower triangular: Pseudo-Expectation Maximization

	AY	BS	GU	HO	JE
AY		0.52	0.38	0.39	0.47
BS	0.17		0.32	0.34	0.41
GU	0.15	0.15		0.19	0.24
HO	0.12	0.14	0.11		0.27
JE	0.18	0.17	0.14	0.13	

The values of **Γ** in Tables 1 and 2 are hard to interpret. However, the values in the diagonal indicate homozygosity, and for GLS we see high values for small breeds such as AY or BS. This is unexpected. In Table 2 the values in the diagonal smaller and more similar across breeds, which agrees better with our knowledge of the breeds. Fst coefficients in Table 3 are in a better understood scale, with previously published reports (e.g. Gibbs et al., 2009). Typical Fst values across dairy European breeds fall in the 0.10 – 0.20 interval, which we found in estimates by Pseudo-EM. However, GLS gives very high values, comparable to differentiations between *Bos Indicus* and *Bos Taurus*. This concerns even breeds with abundant information such as HO and JE. We conclude that the GLS algorithm is generally biased.

Finally, Figure 2 presents a scatterplot of allele frequencies in BS as estimated by the imputation process at CDCB (upstream of the files used here) using findhap.f90 (VanRaden et al. 2013), and by Pseudo-EM. The correlation between Pseudo-EM and Imputation-based frequencies is 0.95. Note that Imputation uses a more correct model considering haplotype tracing, recombinations, and linkage disequilibrium patterns (VanRaden et al. 2013), whereas Pseudo-EM uses an approximated linear model which ignores the physical nature of the chromosome (Legarra et al., 2024). It is satisfactory to see that both methods largely agree.

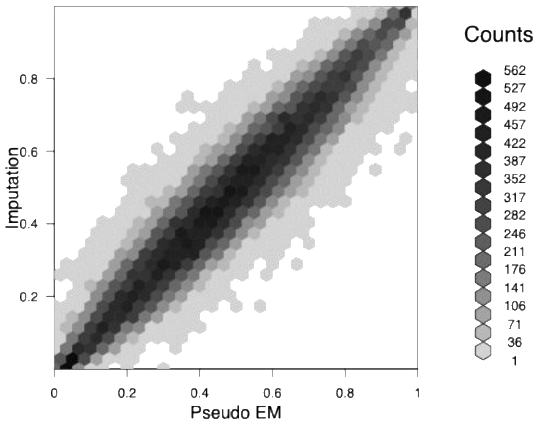


Figure 2. Hexbin scatterplot of base allele frequencies in BS, estimated by Pseudo-EM and Imputation.

Conclusions

For this data set, Pseudo-EM resulted in seemingly unbiased estimates of the MF matrix of relationships **Γ**, unlike GLS which gave biased estimates as observed in Fst values. Results concerning allele frequencies from Pseudo-EM agree well with Imputation results. Computing time was acceptable. We recommend Pseudo-EM.

Acknowledgments

This work uses the National Cooperator Database stewarded by CDCB. We would like to acknowledge the participating dairy producers for supplying data, DHI

organizations and DRPCs for processing and relaying the information to the Council on Dairy Cattle Breeding (CDCB) and the purebred breed associations for providing pedigree data.

References

Garcia-Baccino, C.A., A. Legarra, O.F. Christensen, I. Misztal, I. Pocrnic, Z.G. Vitezica, and R.J. Cantet. 2017. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol*. 49:1,14.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal* 1:21,28.

The Bovine HapMap Consortium, R.A. Gibbs, et al. 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* 324,528–532. https://doi.org/10.1126/science.1167936.

Legarra, A., M. Bermann, Q. Mei, and O.F. Christensen. 2024. Redefining and interpreting genomic relationships of metafounders. *Genet. Sel. Evol.* 56,34. https://doi.org/10.1186/s12711-024-00891-w.

VanRaden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans, M.E. Tooker, J.B. Cole, T.S. Sonstegard, E.E. Connor, M. Winters, J.B.C.H.M. van Kaam, A. Valentini, B.J. Van Doormaal, M.A. Faust, and G.A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* 96:668,678. https://doi.org/10.3168/jds.2012-5702 .