Maximum likelihood estimation of metafounder parameters for single and multiple breeds

O.F. Christensen[1*] and A. Legarra[2]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark, [2] GenPhySE (Genetique, Physiologie et Systemes d'Elevage), INRAE, 31326 Castanet Tolosan, France, *olef.christensen@qgg.au.dk

Abstract

For single-step genomic best linear unbiased prediction, the use of so-called metafounders provides a coherent approach for combining pedigree and genomic relationships. Here, maximum likelihood estimation of metafounder parameters is presented. For one breed, the maximum likelihood estimate of metafounder parameter can be determined by solving a cubic equation. For multiple breeds, we provide an expression for the likelihood function, and for numerical maximisation of this function using gradient methods, we express the metafounder relationship matrix as a function of partial relationship matrices and metafounder parameters.

Introduction

Single-step genomic best linear unbiased prediction (ssGBLUP) is a methodology for genomic evaluation with partially genotyped populations, i.e. some individuals are genotyped and others are not, and phenotypes may be recorded in both subsets (Legarra *et al* 2014). The core idea in ssGBLUP is the construction of a relationship matrix across all individuals that combines genomic and pedigree relationship matrices. For the combined relationship matrix, the concept of metafounders (Legarra *et al.* 2015, Garcia-Baccino *et al.* 2017) provides a coherent approach to make genomic and pedigree relationships relative to the same base population and therefore compatible. With metafounders, the combined relationship matrix is

$$\mathbf{H}_{\boldsymbol{\Gamma}}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - (\mathbf{A}_{\boldsymbol{\Gamma}})^{-1} \end{bmatrix} + (\mathbf{A}_{\boldsymbol{\Gamma}}^{full})^{-1},$$

where $\mathbf{G} = \mathbf{M}\mathbf{M}^T / (k/2)$, with matrix $\mathbf{M}$ containing genotypes coded as -1, 0, 1, and k being the number of markers, $\mathbf{A}_{\boldsymbol{\Gamma}}^{full}$ is the pedigree relationship matrix for all individuals with relationships among base individuals defined by metafounder parameters in matrix $\boldsymbol{\Gamma}$, and $\mathbf{A}_{\boldsymbol{\Gamma}}$ is the submatrix of $\mathbf{A}_{\boldsymbol{\Gamma}}^{full}$ for genotyped individuals. Further details on metafounders and the computation of these matrices can be found in Legarra *et al.* (2015).

The matrix $\boldsymbol{\Gamma}$ contains parameters determining the relationships among base populations, within breeds (diagonal elements) and between breeds (off-diagonal element). The objective of this paper is to present methods for maximum likelihood estimation of these parameters, first for a single breed, and second for multiple breeds.

Maximum likelihood estimation of metafounder parameter for a single breed

For a single breed, the pedigree relationship matrix for genotyped individuals $\mathbf{A}_\gamma = \mathbf{A}(1-\gamma/2) + \gamma\mathbf{1}\mathbf{1}'$ where $\gamma \in [0;2[$ is the metafounder parameter, and $\mathbf{A}$ is the usual pedigree relationship matrix for genotyped individuals. Inference about γ is from the observed marker genotypes $\mathbf{M}$.

A model for observed marker genotypes is that $\mathbf{m}_j \sim N(2p_j - 1, 2p_j(1-p_j)\mathbf{A})$, $j = 1,\ldots,k$ are independent, where $\mathbf{m}_j$ are columns in $\mathbf{M}$, and allele frequencies $p_j \sim N(0.5, \sigma_p^2)$, $j = 1,\ldots,k$ are independent. The metafounder parameter $\gamma = 8\sigma_p^2$. An expectation-maximisation (EM) algorithm can be derived for this model, but the resulting algorithm involves numerical maximisation for the M-step and numerical integration for the E-step, and the algorithm as a whole becomes computationally troublesome to use in practice, and is not presented here. Computationally more feasible algorithms that approximate this EM-algorithm can be constructed, for example the one presented by Garcia-Baccino et al. (2017), but the statistical properties of these are not known.

Another approach, suggested by Christensen (2012), consists of studying the marginal distribution of $\mathbf{m}_j$ (integrating allele frequency p_j) and approximating this by a multivariate normal distribution with the correct marginal mean and marginal variance-covariance matrix, i.e. $\mathbf{m}_j \sim N(\mathbf{0}, 0.5\mathbf{A}_\gamma)$. Maximum likelihood estimation for this approach is presented here.
The joint density of $\mathbf{M} = [\mathbf{m}_1,\ldots,\mathbf{m}_h]$ is

$$f(\mathbf{M}) \propto \prod_{j=1}^{k} \det(0.5\mathbf{A}_\gamma)^{-1/2} \exp(-\mathbf{m}_j'(0.5\mathbf{A}_\gamma)^{-1}\mathbf{m}_j / 2),$$

where a proportionality constant is ignored. Using that the product of exponential terms is the exponential of the sum, and $\sum_j \mathbf{m}_j'(0.5\mathbf{A}_\gamma)^{-1}\mathbf{m}_j = \mathrm{Tr}((0.5\mathbf{A}_\gamma)^{-1}\mathbf{M}\mathbf{M}')$, likelihood function is

$$L(\gamma) \propto \det(0.5\mathbf{A}_\gamma)^{-k/2} \exp(-(1/2)\mathrm{Tr}((0.5\mathbf{A}_\gamma)^{-1}\mathbf{M}\mathbf{M}')).$$

Taking logarithm of the likelihood function, introducing the notation $\mathbf{G} = \mathbf{M}\mathbf{M}' / (k/2)$, and ignoring constants, we obtain the log-likelihood function
$\ell(\gamma) = -(k/2)\log(\det(\mathbf{A}_\gamma)) - (k/2)\mathrm{Tr}((\mathbf{A}_\gamma)^{-1}\mathbf{G})$.

Using formulas from Henderson and Searle (1981), we obtain
$(\mathbf{A}_\gamma)^{-1} = \mathbf{A}^{-1}/(1-\gamma/2) - \mathbf{A}^{-1}\mathbf{1}\mathbf{1}'\mathbf{A}^{-1}\gamma / ((1-\gamma/2)(1-\gamma/2+\gamma\mathbf{1}'\mathbf{A}^{-1}\mathbf{1}))$ and
$\det(\mathbf{A}_\gamma) = \det(\mathbf{A})(1-\gamma/2)^n(1+\gamma\mathbf{1}'\mathbf{A}^{-1}\mathbf{1}/(1-\gamma/2)) = \det(\mathbf{A})(1-\gamma/2)^{n-1}(1-\gamma/2+\gamma\mathbf{1}'A^{-1}\mathbf{1})$.
Inserting those into the log-likelihood function, and using short-hand notation $a = \mathbf{1}'\mathbf{A}^{-1}\mathbf{1}$, $b = \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{G})$ and $c = \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{1}\mathbf{1}'\mathbf{A}^{-1}\mathbf{G})$, we obtain

$$\ell(\gamma) = -\frac{k(n-1)}{2}\log(1-\gamma/2) - \frac{k}{2}\log(1-\gamma/2+\gamma a) - \frac{k}{2}\frac{b}{1-\gamma/2} + \frac{k}{2}\frac{c\gamma}{(1-\gamma/2)(1-\gamma/2+\gamma a)}.$$

The maximum likelihood estimate is obtained by differentiating $\ell(\gamma)$ and setting equal to zero, where we note that the $k/2$ in front of each term in the expression can be ignored,

$$0 = \frac{(n-1)/2}{1-\gamma/2} - \frac{-1/2+a}{1-\gamma/2+\gamma a} - \frac{b/2}{(1-\gamma/2)^2}$$

$$+c\frac{(1-\gamma/2)(1-\gamma/2+\gamma a) - \gamma\left((1-\gamma/2)(-1/2+a) - (1-\gamma/2+\gamma a)/2\right)}{(1-\gamma/2)^2(1-\gamma/2+\gamma a)^2}.$$

Multiplying by $(1-\gamma/2)^2(1-\gamma/2+\gamma a)^2$ on both sides of the equation followed by some algebraic manipulations we obtain a cubic equation

$$e_3\gamma^3 + e_2\gamma^2 + e_1\gamma + e_0 = 0,$$

where $e_3 = -n(-1/2+a)^2/2$, $e_2 = (n(-3/2+a)+a-b(-1/2+a)+c)(-1/2+a)$, $e_1 = (n-1)(-3/2+2a)-(-1+2a)(-3/2+a+b)$ and $e_0 = n-2a-b+2c$.

Solving a cubic equation, first requires the computation of the discriminant $\Delta = 18e_3e_2e_1e_0 - 4e_2^3e_0 + e_2^2e_1^2 - 4e_3e_1^3 + 27e_3^2e_0^2$. If $\Delta > 0$ then there are three distinct real roots, and if $\Delta < 0$ then there is only one real root (and two complex roots). Following the computation of the discriminant, root(s) need to be computed, it needs to checked whether root(s) are within the interval $[0;2[$, and investigated which root (or possibly the boundary value 0) is the maximum.

Maximum likelihood estimation of metafounder parameters for multiple breeds

Here, we present maximum likelihood estimation for multiple breeds or genetic origins.
The columns of the genotype matrix $\mathbf{M}$ are assumed independent and distributed as $\mathbf{m}_j \sim N(\mathbf{0}, 0.5\mathbf{A}_\Gamma)$, where

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{b_1} & \cdots & \gamma_{b_1,b_r} \\ \vdots & \vdots & \cdots \\ \gamma_{b_1,b_r} & \cdots & \gamma_{b_r} \end{bmatrix},$$

is a symmetric positive definite matrix with all $\gamma_b < 2$, and $\mathbf{A}_\Gamma$ is the metafounder relationship matrix. Then the joint density of $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_p]$ is

$$f(\mathbf{M}) \propto \prod_{j=1}^{k} \det(0.5\mathbf{A}_\Gamma)^{-1/2} \exp(-\mathbf{m}_j'(0.5\mathbf{A}_\Gamma)^{-1}\mathbf{m}_j/2),$$

where proportionality constants have been ignored. Using that the product of exponential terms is the exponential of the sum, and that $\sum_j \mathbf{m}_j'(0.5\mathbf{A}_\Gamma)^{-1}\mathbf{m}_j = \mathrm{Tr}((0.5\mathbf{A}_\Gamma)^{-1}\mathbf{M}\mathbf{M}')$, the likelihood function is

$$L(\mathbf{\Gamma}) \propto \det(0.5\mathbf{A}_\Gamma)^{-k/2} \exp(-(1/2)\mathrm{Tr}((0.5\mathbf{A}_\Gamma)^{-1}\mathbf{M}\mathbf{M}')).$$

Taking logarithm of the likelihood function, noting that $\mathbf{G} = \mathbf{M}\mathbf{M}'/(k/2)$, and ignoring constants, we obtain the log-likelihood function

$$\ell(\mathbf{\Gamma}) = -(k/2)\log(\det(\mathbf{A}_\Gamma)) - (k/2)\mathrm{Tr}((\mathbf{A}_\Gamma)^{-1}\mathbf{G}).$$

Maximum likelihood estimates of metafounder parameters can then be obtained by numerical maximization of this function.

For numerical maximization using gradient methods, derivatives of the log-likelihood function are needed. From the definition of $\mathbf{A}_\Gamma$ it can be shown using proof by induction that

$$\mathbf{A}_\Gamma = \sum_b \mathbf{A}^b (1-\gamma_b/2) + \sum_{b,b':b'>b} \mathbf{A}^{b,b'} ((\gamma_{b'}+\gamma_b)/2) - \gamma_{b,b'})/4 + \sum_b \mathbf{C}^b \gamma_b + \sum_{b,b':b'>b} \mathbf{C}^{b,b'} \gamma_{b,b'},$$

with $\mathbf{A}^b$ being the breed b specific partial relationship matrix, $\mathbf{A}^{b,b'}$ being the breeds b, b' segregation partial relationship matrix, matrix $\mathbf{C}^b$ having entries $C^b_{i,i'} = f^b_i f^b_{i'}$, and matrix $\mathbf{C}^{b,b'}$ having entries $C^{b,b'}_{i,i'} = f^b_i f^{b'}_{i'} + f^{b'}_i f^b_{i'}$. Further details about partial relationship matrices can be found in Garcia-Cortes and Toro (2006). From this expression for matrix $\mathbf{A}_\Gamma$ derivatives with respect to metafounder parameters can be obtained.

Discussion

We have here presented maximum likelihood estimation of metafounder parameters. With genotypes on one breed, the parameter can be estimated by solving a cubic equation, whereas with genotypes on several breeds, numerical maximisation of the likelihood function is required for estimating parameters. Therefore, we have a theoretically based procedure for estimating metafounder parameters in simple cases with one or few breeds/ populations.

More challenging cases with large number of populations or genetic origins exist. In particular, for dairy cattle genetic evaluation, pedigree usually is incomplete with unknown parents of certain groups of individuals, e.g. due to imported bulls or incomplete pedigrees, and can be handled using so-called unknown parent groups. For metafounder approach this provides a challenge because number of unknown parent groups is large, making the number of parameters in $\mathbf{\Gamma}$ matrix prohibitively large, and either a reduction in number of unknown parent groups, as in Kudinov et al (2020), or a parameterisation of $\mathbf{\Gamma}$ using fewer parameters, as in Koivula et al (2021), needs to done.

References

Christensen O.F. (2012) Genet. Sel. Evol. 44:37. http://doi.org/10.1186/1297-9686-44-37

Garcia-Baccino C.A., Legarra A., Christensen O.F., Misztal I., Pocrnir I. *et al.* (2017) Genet. Sel. Evol. 49:34. http://doi.org/10.1186/s12711-017-0309-2

García-Cortés L.A. and Toro M.A. (2006) Genet. Sel. Evol. 38:601-615. http://doi.org/10.1051/gse:2006024

Henderson and Searle (1981) SIAM Review. 23(1):53-60. https://doi.org/10.1137/1023004

Koivula M., Stranden I., Aamand G.P. and Mantysaari E.A. (2021) Interbull Bulletin 56, Leeuwarden, The Netherlands, April 26 – 30, 2021

Kudinov A.A., Mantysaari E.A., Aamand G.P., Uimari P., and Stranden. I. (2020) J. Dairy Sci. 103(7):6299:6310. https://doi.org/10.3168/jds.2019-17483

Legarra A., Christensen O.F., Aguilar I., and Misztal I. (2014) Livest. Sci. 166:54-65. http://dx.doi.org/10.1016/j.livsci.2014.04.029

Legarra A., Christensen O.F., Vitezica Z., Aguilar I. and Misztal I. (2015) Genetics 200(2):455-468. http://doi.org/10.1534/genetics.115.177014