Approximate multitrait BLUP evaluation to combine functional traits information

Jean-Jacques Colleau, Vincent Ducrocq, Didier Boichard and Hélène Larroque

Station de Génétique Quantitative et Appliquée,
Institut National de la Recherche Agronomique,
78352 Jouy en Josas, France

Abstract

We present an alternative to the classical way of combining breeding values into a total merit index, traditionally based on selection index principles. An approximate multiple trait BLUP sire model is proposed, using deregressed proofs or preferably daughter deviations to build the right hand side of the mixed model equations. The advantages of the proposed approach are: a) a unique definition of the weight for each individual trait in the total merit index (independent from the accuracy of each index); b) new estimated breeding values for each individual trait, optimally combining direct and indirect information; c) the possibility to account for residual covariances between performances of a same animal (in contrast with MACE methodology). An algorithm for an easier solution of the multivariate system of linear equations is proposed as well as a possible extension to the computation of female EBVs. This approach is still to be implemented and tested.

1. Introduction

A total merit objective function for French dairy breeders combining production traits and functional trait information is under development (Colleau *et al*., 1999). It includes genetic values on traits such as milk yield, somatic cells count (SCC), female fertility and functional longevity. These traits are obviously correlated (e.g., see Larroque *et al*., 1999). Most functional traits exhibit rather low heritabilities, leading to genetic evaluations with low reliabilities for young sires. Fortunately, early predictors of, e.g., SCC or functional longevity can be found in the long list of type traits recorded in each breed. An optimal combination of these pieces of information is needed.

The optimal estimation procedure is known to be the multiple trait BLUP evaluation (MT-BLUP): see the reviews by Van der Werf *et al.* (1992) and Ducrocq (1994). It provides an improved accuracy of the evaluation on each trait through an increase of the amount of information, an improved data structure through better connectedness and a correction of biases due to selection on correlated trait. Finally, optimal weighting factors to be used in the total merit index are precisely the economic weights themselves: the multiple trait evaluation automatically accounts for the fact that traits are correlated and that the relative accuracy of the evaluation for each trait varies between animals. This latter property is extremely interesting. Note that MT-BLUP EBVs for predictors are not considered in the total merit index.

However a unique multiple trait BLUP evaluation on all relevant traits together, although conceptually possible, is not routinely feasible. Traits are described by very different models. Some of these models are not linear; others involve repeated measures and/or more than one genetic effect and above all, the amount of data to manipulate in national evaluations

is tremendous. Despite huge and fast improvements of computing power, computational considerations are still a limiting factor. Furthermore, a large set of dispersion parameters should be known or estimated accurately before being included in such an evaluation.

Still, most countries are providing total merit indices, approximately combining indices with various accuracies on correlated traits. For example, correlations between traits or differences in reliabilities between bulls are ignored in the current total merit index in France (ISU), which combines linearly the aggregate production index (INEL) with the overall type and milking speed indices. Low correlations and rather homogeneous reliabilities for all traits justified this. This is obviously no longer the case with the new set of functional traits considered.

A possibility is to rely on index selection theory to account for genetic and phenotypic correlations between traits and unequal information on different traits. But this approach quickly becomes complex if one wants to consider all possible cases (variable family structures). However, some information is lost when the contributions from related animals are not systematically considered. Finally, the nested use of selection index theory (e.g., to compute a more precise functional longevity index and then to compute the total merit index) may be questionable.

Another recent approximation that has been suggested and used (e.g., Druet *et al*., 1999; Larroque and Ducrocq, 1999) is the extension of MACE (Multiple trait Across Country Evaluation) methodology (Schaeffer, 1994). This is a very attractive alternative in order to "recycle" the results of (deregressed) univariate proofs, characterised by

1. an exhaustive use of the pedigree file,
2. an elegant way to include proofs for traits described by non-linear models,
3. the possibility to account for genetic correlations between traits and heterogeneous reliabilities.

However, it relies on a zero residual correlation between any pair of traits, a natural assumption when different traits are recorded in different countries, i.e., on different animals. However, the traits considered both for inclusion in the selection objective and for prediction are observed on the same animals. Then the residual correlations can differ substantially from 0 for some pairs of traits (Larroque and Ducrocq, 1999).

This paper proposes an improved version of MACE, where residual correlations are considered. In section 2, the general, "correct" case will be presented as a starting point for several successive approximations, detailed in section 3. This will illustrate where and how information is lost, for the sake of computational feasibility and simplicity. Section 4 proposes an algorithm for an easier solution of the approximate multiple trait evaluation. The extension of the approach to a cow evaluation will be discussed in section 5.

2. Multiple trait evaluation of functional traits (traits in the objective function and early predictors)

2.1. Notations

Consider the general situation encountered in multiple trait evaluation. For each trait i, i=1,..t, assume that the data vector $\mathbf{y_i}$ can be analysed using a linear model with only one random effect other than the residual $\mathbf{e_i}$:

$$\mathbf{y_i} = \mathbf{X_i}\,\mathbf{b_i} + \mathbf{Z_i}\,\mathbf{u_i} + \mathbf{e_i} \quad [1]$$

where, as usual, $\mathbf{b_i}$ and $\mathbf{u_i}$ represent the vectors of fixed and random effects, $\mathbf{X_i}$ and $\mathbf{Z_i}$ are the corresponding incidence matrices.

Let the (co)variance matrices be written as:

$$\mathbf{G} = \{\mathbf{G_{ij}}\}_{i,j=1,..t} = \{\mathrm{Cov}(\mathbf{u_i}, \mathbf{u_j}')\}_{i,j=1,..t}$$

$$\mathbf{R} = \{\mathbf{R_{ij}}\}_{i,j=1,..t} = \{\mathrm{Cov}(\mathbf{e_i}, \mathbf{e_j}')\}_{i,j=1,..t}$$

and $\mathrm{Cov}(\mathbf{u_i}, \mathbf{e_j}') = 0$ for all i, j.

Define a similar partition for the inverse matrices of $\mathbf{G}$ and $\mathbf{R}$: $\mathbf{G}^{-1} = \{\mathbf{G^{ij}}\}$ and $\mathbf{R}^{-1} = \{\mathbf{R^{ij}}\}$.

The submatrices $\mathbf{G_{ij}}$ and $\mathbf{R_{ij}}$ depend on the pedigree and data structures and on $\mathbf{G_0}$ and $\mathbf{R_0}$, which are functions of the genetic and residual (co)variance parameters. The notation used here is general: the vector $\mathbf{u_i}$ may represent either a sire effect (later referred to as $\mathbf{s_i}$) or an animal additive genetic effect ($\mathbf{a_i}$). In the former case, elements g_{ij} and r_{ij} of $\mathbf{G_0}$ and $\mathbf{R_0}$ are $g_{ij} = 1/4\ \sigma_{g,ij}$ and $r_{ij} = 3/4\ \sigma_{g,ij} + \sigma_{e,ij}$.

2.2. Correct multiple trait BLUP evaluation

Let's assume that, whatever the trait considered, all elementary records that would be used in the univariate evaluation are available and that the linear model [1] can be used to describe them. The general form of the MT-BLUP mixed model equations is:

$$\begin{bmatrix} \vdots & & \vdots & \\ \cdots & \mathbf{X_i'R^{ij}X_j} & \cdots & \mathbf{X_i'R^{ij}Z_j} & \cdots \\ \vdots & & \vdots & \\ \cdots & \mathbf{Z_i'R^{ij}X_i} & \cdots & \mathbf{Z_i'R^{ij}Z_j + G^{ij}} & \cdots \\ \vdots & & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{b_j} \\ \vdots \\ \mathbf{u_j} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathbf{X_i'}\left(\cdots + \mathbf{R^{ii}y_i} + \cdots \mathbf{R^{ij}y_j} + \cdots\right) \\ \vdots \\ \mathbf{Z_i'}\left(\cdots + \mathbf{R^{ii}y_i} + \cdots \mathbf{R^{ij}y_j} + \cdots\right) \\ \vdots \end{bmatrix} \quad [2]$$

3. Various approximations to MT-BLUP

3.1. Approximation 1: multiple trait evaluation using preadjusted records

As previously indicated, the models used for each trait are so different and the amount of information is so large that one may want to get rid of the step consisting in estimating fixed effects. Instead of manipulating raw data, the evaluation system is then based on preadjusted records:

$$\begin{bmatrix} \cdots & \vdots & \vdots & \cdots \\ & \vdots & \vdots & \\ \cdots & \mathbf{Z_i'R^{ii}Z_i + G^{ii}} & \mathbf{Z_i'R^{ij}Z_j + G^{ij}} & \cdots \\ \cdots & \mathbf{Z_j'R^{ji}Z_i + G^{ji}} & \mathbf{Z_j'R^{jj}Z_j + G^{jj}} & \cdots \\ & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} \mu \\ \vdots \\ \mathbf{u_i} \\ \mathbf{u_j} \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{1'}\left(\cdots + \mathbf{R^{ii}y_i^*} + \cdots \mathbf{R^{ij}y_j^*} + \cdots\right) \\ \vdots \\ \mathbf{Z_i'}\left(\cdots + \mathbf{R^{ii}y_i^*} + \cdots \mathbf{R^{ij}y_j^*} + \cdots\right) \\ \mathbf{Z_j'}\left(\cdots + \mathbf{R^{ii}y_i^*} + \cdots \mathbf{R^{ij}y_j^*} + \cdots\right) \\ \vdots \end{bmatrix} \quad [3]$$

where $\mathbf{y_i^* = y_i - X_i \hat{b}_i}$ and the $\mathbf{\hat{b}_i}$'s are obtained from the univariate evaluations. In practice, as the $\mathbf{u_i}$'s are not necessarily centered on the same basis, a general mean vector $\boldsymbol{\mu}=\{\mu_i\}$ must be included in the mixed model equations and estimated together with the vectors $\mathbf{u_i}$.

Note that this first simplification could be used to estimate in an approximate way (because it is based on preadjusted records) both the genetic and residual correlations between traits at least on a reduced file and for traits described by a linear model. However, this procedure is not adapted to non-linear traits (for which the assumed linear model is not valid) and needs some adaptation for traits for which repeated measurements are available (production traits, SCC, female fertility). Also, the necessary knowledge of all

individual preadjusted records is quite demanding.

3.2. Approximation 2: use of deregressed proofs

An evaluation of males only - through a sire model – advocates the use of a more concise information at the sire level. Let $s_{i,m}$ represent the effect of a particular sire m for trait i. Let $n_{ij,m}$ be the number of daughters of sire m, with performances actually recorded both on trait i and on trait j (for simplicity, we will note $n_{i,m} = n_{ii,m}$; also, whenever possible, an *equivalent* number of daughters should be preferred, to account for loss of information due to the estimation of fixed effects).

Define $\mathbf{N_{ij}} = \text{diag}\{n_{ij,m}\}$; $\mathbf{G_0^{-1}} = \{g^{ij}\}_{i=,j=1..t}$ and similarly $\mathbf{R_0^{-1}} = \{r^{ij}\}$. Then, if we expand the vectors of sire effects so that they have all the same dimension, the typical expression in the coefficient matrix of the mixed model equations in [3] can be simplified into:

$$\mathbf{Z_i' R^{ij} Z_j} + \mathbf{G^{ij}} = r^{ij}\,\mathbf{N_{ij}} + g^{ij}\,\mathbf{A^{-1}} \qquad [4]$$

where $\mathbf{A}$ is the relationship matrix between sires. At the same time, if $\tilde{\mathbf{y}}_\mathbf{i}$ represents the vector of deregressed proofs of all sires for trait i, the right hand side in [3] can be approximated as:

$$\mathbf{Z_i'}\left(\cdots + \mathbf{R^{ii}}\mathbf{y_i^*} + \cdots \mathbf{R^{ij}}\mathbf{y_j^*} + \cdots\right) \approx \left(\cdots + r^{ii}\mathbf{N}_{ii}\,\tilde{\mathbf{y}}_\mathbf{i} + \cdots + r^{ij}\mathbf{N_{ij}}\tilde{\mathbf{y}}_\mathbf{j} + \cdots\right) \qquad [5]$$

and the mixed model equations become:

$$\begin{bmatrix} \vdots & \vdots & \\ \cdots\ r^{ii}\,\mathbf{N_{ii}} + g^{ii}\,\mathbf{A^{-1}} & r^{ij}\,\mathbf{N_{ij}} + g^{ij}\,\mathbf{A^{-1}} & \cdots \\ \cdots\ r^{ji}\,\mathbf{N_{ji}} + g^{ji}\,\mathbf{A^{-1}} & r^{jj}\,\mathbf{N_{jj}} + g^{jj}\,\mathbf{A^{-1}} & \cdots \\ \vdots & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{s_i} \\ \mathbf{s_j} \\ \end{bmatrix} = \begin{bmatrix} \vdots \\ \cdots + r^{ii}\mathbf{N_{ii}}\,\tilde{\mathbf{y}}_\mathbf{i} + \cdots + r^{ij}\mathbf{N_{ij}}\tilde{\mathbf{y}}_\mathbf{j} + \cdots \\ \cdots + r^{ii}\mathbf{N_{ji}}\,\tilde{\mathbf{y}}_\mathbf{i} + \cdots + r^{ij}\mathbf{N_{jj}}\tilde{\mathbf{y}}_\mathbf{j} + \cdots \\ \vdots \end{bmatrix} \qquad [6]$$

The important advantage of this approximation is that there is no longer the need to work at the individual performance level. All the information about sire m is summarised in its deregressed proofs and the $n_{ij,m}$'s. A potential limitation is that, for each bull and for all combinations of traits, the figures $n_{ij,m}$ must be known. This may require the handling of large files summarising information on all traits for all cows. This drawback can be circumvented via two further approximations that we will present now.

3.3. Approximation 3: approximate computation of the number of animals recorded both on trait i and trait j

To avoid the actual computation of $n_{ij,m}$, one may assume that all the daughters of a particular sire m have the same probability to be recorded on any trait. Let $n^*_m = \max_i(n_{i,m})$, i.e., the largest number of daughters of sire m recorded on a particular trait. This trait will usually be milk production. Then, the probability that a given daughter of sire m is recorded on trait i is simply $p_{i,m} = n_{i,m} / n^*_m$. The probability that a cow is recorded on both traits i and j is $p_{i,m}\, p_{j,m}$ and the number of daughters recorded on both traits is:

$$\begin{aligned} n_{ij,m} &= (p_{i,m}\, p_{j,m})\ n^*_m \\ &= n_{i,m}\, n_{j,m} / n^*_m \end{aligned} \qquad [7]$$

Of course, there are combinations of traits when this expression would be known to be invalid, but then the approximate value can easily be replaced by the correct one. For example, if i and j are type traits, one expects $n_{ij,m} = n_{i,m} = n_{j,m}$.

3.4. Approximation 4: no residual correlations

If one can assume that residual correlations between any combination of traits i and j are small enough so they can

be ignored, or equivalently, that traits i and j are recorded on two distinct batches of daughters ($n_{ij,m}$=0), then the MT-BLUP mixed model equations are further simplified into:

$$\begin{bmatrix} \vdots & \vdots & \\ \cdots r^{ii}\mathbf{N_{ii}}+g^{ii}\mathbf{A^{-1}} & g^{ij}\mathbf{A^{-1}} & \cdots \\ \cdots\ g^{ji}\mathbf{A^{-1}} & r^{jj}\mathbf{N_{jj}}+g^{jj}\mathbf{A^{-1}} & \cdots \\ \vdots & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{s_i} \\ \mathbf{s_j} \\ \end{bmatrix} = \begin{bmatrix} \vdots \\ r^{ii}\mathbf{N_{ii}}\tilde{\mathbf{y}}_\mathbf{i} \\ r^{ij}\mathbf{N_{jj}}\tilde{\mathbf{y}}_\mathbf{j} \\ \vdots \end{bmatrix} \quad [8]$$

with:

$$\mathbf{N_{ii}}\ \tilde{\mathbf{y}}_\mathbf{i} = \sum_m n_{i,m}\ \tilde{y}_{i,m} \quad [9]$$

This is exactly the model, which is considered in the MACE approach. Such an approximation is indeed perfectly valid when traits i and j are recorded in different countries, but may result in biases in the framework considered here (Larroque and Ducrocq, 1999).

It is worthwhile noting that intermediate options exist. In particular, the assumption of zero residual correlation between two type traits recorded simultaneously on the same animal is often not acceptable. For such combinations of traits, accurate estimates of residual correlations usually exist and could be used, while the choice of r_{ij}=0 is kept for other combinations (e.g., somatic cell counts and fertility?).

2.6. Where to stop?

Which of these approximations should be promoted? Obviously, there is a balance to be found between ease of computation and accuracy. If a good data base exists and/or in the long run, **approximation 1 is probably the one which should be envisioned**. On a shorter horizon, our conjecture is that approximation 2 is "better" (less assumptions => less bias) than the MACE approach (approximation 4). This needs to be tested: are the differences between the two large enough to justify the extra computational effort ?

4. An EM-type algorithm for the MT-BLUP evaluation

4.1. MT-BLUP on preadjusted records

An attractive feature of system [3] is that it implicitly describes a model for which all traits are analysed using the same linear model (same fixed effect = $\boldsymbol{\mu}_i$; only one random effect other than the residual). Animals are not necessarily recorded on all traits but the extension of the canonical transformation (Thompson, 1976; Quaas, 1984) to the missing value case proposed by Ducrocq and Besbes (1993) and Ducrocq and Chapuis (1997) can be implemented. The underlying idea is that the missing values are iteratively replaced by their expectation given the current values of all parameters. Then, the resulting system is solved as if they were not missing, i.e., applying the canonical transformation. This technique leads to the same solutions as the original multiple trait system and has a formal justification based on the Expectation-Maximisation (EM) algorithm of Dempster *et al*. (1977). Substantial benefits are expected: such a transformation leads to decreased computing requirements, faster convergence and simplified (univariate) programming.

Let's first extend the notations. In the case of no missing values, let $\mathbf{y}^*_{m,q}$ be the vector of all preadjusted records that represents the typical contribution to the right-hand side of [3] of a daughter q of sire m. In practice, some traits may be missing. Denote as δ the list of indices describing the combination of traits actually recorded on cow q and and δ^- the list of the missing ones, i.e., $\delta + \delta^- = \{1,2\ldots t\}$.

Decompose the vector $\mathbf{y}^*_{m,q}$ into the sum of two vectors, distinguishing between observed and missing preadjusted records:

$$\mathbf{y}^*_{m,q} = \mathbf{y}^*_{\delta,m,q} + \mathbf{y}^*_{\delta^-,m,q} \qquad [10]$$

For example, if t=3 and δ ={1,3}, then δ^-={2} and

$$\mathbf{y}^*_{m,q} = \begin{bmatrix} y^*_{1,m,q} \\ 0 \\ y^*_{3,m,q} \end{bmatrix} + \begin{bmatrix} 0 \\ y^*_{2,m,q} \\ 0 \end{bmatrix} = \mathbf{y}^*_{\delta,m,q} + \mathbf{y}^*_{\delta^-,m,q} \qquad [11]$$

i.e., records $y^*_{1,m,q}$ and $y^*_{3,m,q}$ are observed whereas $y^*_{2,m,q}$ is missing.

To implement the EM algorithm, consider that $\mathbf{y}^* = \{\mathbf{y}^*_{m,q}\}$ is the complete (augmented) data vector. Given $\mathbf{s}$, the vector of all sire effects on all traits and $\boldsymbol{\mu}$, the vector of grand means, $\mathbf{y}^*$ follows a multivariate normal distribution with mean:

$$(\mathbf{I_t} \otimes \mathbf{I})\boldsymbol{\mu} + (\mathbf{I_t} \otimes \mathbf{N}^*)\mathbf{s} \qquad [12]$$

Here, $\mathbf{N}^* = \text{diag}\{n^*_m\}$ implies that in the complete data vector, records of all daughters of sire m on all traits are available. The right hand side of system [3] is a vector of sufficient statistics for the estimation of $\mathbf{s}$ and $\boldsymbol{\mu}$. At each EM iteration [k], the vectors $\mathbf{y}^*_{\delta^-,m,q}$ (for all m and q) being unknown, they are replaced by their expectation given the current parameter estimates, when constructing the right hand side of [3]. The vector $\mathbf{y}^*_{\delta^-,m,q}$ is replaced at iteration [k] by:

$$\hat{\mathbf{y}}^{*[k]}_{\delta^-,m,q} = \hat{\boldsymbol{\mu}}^{[k]}_{\delta^-} + \hat{\mathbf{s}}^{[k]}_{\delta^-,m} + \mathbf{R}_{\mathbf{0},\delta^-\delta}\,\mathbf{R}^{-1}_{\mathbf{0},\delta\delta}\,(\hat{\mathbf{y}}^*_{\delta,m,q} - \hat{\boldsymbol{\mu}}^{[k]}_{\delta} - \hat{\mathbf{s}}^{[k]}_{\delta,m}) \qquad [13]$$

where the notation $\hat{\boldsymbol{\mu}}_\delta$ refers to vector $\hat{\boldsymbol{\mu}}$ where rows that are not in combination δ were set to 0 and $\mathbf{R}_{\mathbf{0},\delta^-\delta}$ represents matrix $\mathbf{R_0}$ where rows not in δ^- and columns not in δ were set to 0 (similarly for $\hat{\boldsymbol{\mu}}_{\delta^-}$, $\hat{\mathbf{s}}_{\delta,m}$, $\hat{\mathbf{s}}_{\delta^-,m}$ and $\mathbf{R}^{-1}_{\mathbf{0},\delta\delta}$). The last term of [13] is the regression of the residuals for the missing values of cow q onto the residual estimates for the observed records. he resulting system is the-BLUP mixed model equations in the situation of no missing values: if $\mathbf{n}^*$ =$\{n^*_m\}$ and N is the total number of records ($N = \Sigma_m\, n^*_m$), [3] becomes:

$$\begin{bmatrix} \mathbf{R_0^{-1}} \otimes \mathrm{N} & \mathbf{R_0^{-1}} \otimes \mathbf{n}^{*\prime} \\ \mathbf{R_0^{-1}} \otimes \mathbf{n}^* & \mathbf{R_0^{-1}} \otimes \mathbf{N}^* + \mathbf{G_0^{-1}} \otimes \mathbf{A^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{s} \end{bmatrix} = \mathbf{R_0^{-1}} \otimes \begin{bmatrix} \mathbf{n}^{*\prime} \\ \mathbf{N}^{*\prime} \end{bmatrix} \hat{\mathbf{y}}^{*[k]} \qquad [14]$$

The M step of the EM algorithm consists in solving [14]. Now the canonical transformation (Thompson, 1976; Quaas, 1984) can be applied. This leads to t univariate systems to solve, each one of the form:

$$\begin{bmatrix} \mathrm{N} & \mathbf{n}^{*\prime} \\ \mathbf{n}^* & \mathbf{N}^* + \lambda_{ic} \otimes \mathbf{A^{-1}} \end{bmatrix} \begin{bmatrix} \mu_{ic} \\ \mathbf{s}_{ic} \end{bmatrix} = \begin{bmatrix} \mathbf{n}^{*\prime} \\ \mathbf{N}^{*\prime} \end{bmatrix} \mathbf{y}^{*[k]}_{ic} \qquad [15]$$

where ic refers to the i^{th} canonical trait with variances ratio λ_{ic} on the canonical scale. Once these systems are solved, back-solutions on the original scale are obtained and a new EM iteration starts, and so on until convergence. If an iterative algorithm is used to solve [15], iterations for the solutions of the linear system and for the EM algorithm can be interlaced (Ducrocq and Besbes, 1993). Again, this repeated solution of univariate systems is expected to lead to faster convergence, lower computational requirements and easier coding.

4.2. MT-BLUP on deregressed proofs

As already indicated, the approach in section 3.1 may be unattractive because it implies to work on individual (preadjusted) records rather than on deregressed proofs.

To avoid the manipulation of individual records, an approximation of the EM-type algorithm described in the previous section is necessary. Most of the notations remain identical, except that preadjusted records are replaced by deregressed proofs (e.g., any $\mathbf{y}^*_{m,q}$ is replaced by $\tilde{\mathbf{y}}_m$, etc…).

Let $n_{\delta,m}$ be the number of daughters of sire m that have only combination δ of traits known. At each EM iteration [k], the vector $\{\tilde{\mathbf{y}}_{\delta^-,m}\}$ (for all m and δ^-) being unknown, it is replaced by its expectation given the current parameter estimates in the computation of the right hand side of [6]. In particular, one needs $\left(n^*_m \tilde{\mathbf{y}}_m\right)^{[k]}$ which is equal to:

$$\sum_{\text{all}\,\delta} n_{\delta,m}\left(\tilde{\mathbf{y}}_{\delta,m} + \hat{\tilde{\mathbf{y}}}^{[k]}_{\delta^-,m}\right) \qquad [16]$$

However, when working on deregressed proofs, there are no individual data to estimate the residual vectors for observed traits $(\hat{\mathbf{y}}^*_{\delta,m,q} - \hat{\boldsymbol{\mu}}^{[k]}_{\delta} - \hat{\mathbf{s}}^{[k]}_{\delta,m})$, that are needed in [13]! An extra hypothesis must be added in order to approximate $n_{\delta,m}\hat{\tilde{\mathbf{y}}}^{[k]}_{\delta^-,m}$ in [16]. If we assume that there is no systematic bias in the distribution of the daughters of sire m across the different combinations of recorded traits, i.e., for all combination δ:

$$E\left[\hat{\mathbf{y}}^*_{\delta,m,q} - \hat{\boldsymbol{\mu}}^{[k]}_{\delta} - \hat{\mathbf{s}}^{[k]}_{\delta,m}\right] \approx 0 \qquad [17]$$

it may be approximately considered that:

$$\sum_{q\subset\delta}\left[\mathbf{R}_{\mathbf{0},\delta^-\delta}\,\mathbf{R}^{-1}_{\mathbf{0},\delta\delta}(\hat{\mathbf{y}}^*_{\delta,m,q} - \hat{\boldsymbol{\mu}}^{[k]}_{\delta} - \hat{\mathbf{s}}^{[k]}_{\delta,m})\right] \approx 0 \qquad [18]$$

and therefore:

$$n_{\delta,m}\hat{\tilde{\mathbf{y}}}^{[k]}_{\delta^-,m} \approx n_{\delta,m}\,(\hat{\boldsymbol{\mu}}^{[k]}_{\delta^-} + \hat{\mathbf{s}}^{[k]}_{\delta^-,m}) \qquad [19]$$

Note that the quality of this assumption will depend not only on the true absence of samplingbias but also on the number $n_{\delta,m}$ of observations in each combination: the larger this number, the smaller the *average* estimated residual for a given sire. If the approximation is not a too strong one, the missing value algorithm described in section 4.1 can be applied… except that now all $n_{\delta,m}$ records are processed at the same time. This may be extremely advantageous for some bulls with thousands of daughters, since the number of possible combinations of recorded trait is usually very small.

5. An EM-type approximation for female MT-BLUP evaluation

Until now, only the computation of male breeding values in a multivariate context was discussed. Obviously, an extension of the procedure to obtain female EBVs is also needed. One demanding approach would be to apply either one of the above approximations ([3] or [4]) simultaneously on males and females assuming an animal model.

A simpler approach is proposed here, which makes use of the previous results. Again, we will suppose -at least initially- that all traits can be described by a linear model. The starting point is the correct MT-BLUP system of equations described in [2], but this time, an animal model is assumed ($\mathbf{u_i} = \mathbf{a_i}$ with $\mathbf{G_0}$ and $\mathbf{R_0}$ appropriately modified). A first level of approximation consists in working on preadjusted records, i.e., after correction of the records using univariate estimates of fixed effects. A second level of approximation relies on the assumption that the mean vector $\boldsymbol{\mu}$ and male EBVs are known: they are the solutions of one of the approximate MT-BLUP evaluations described above ($\hat{\mathbf{a}}_{\mathbf{i},m} = 2\hat{\mathbf{s}}_{i,m}$ for all sires m).

The resulting multivariate mixed model equations (comparable to system [3] with sire solutions known) is huge. Applying once more the missing value algorithm of Ducrocq and Besbes (1993),

the system can be modified to a configuration where the canonical transformation is applicable. In practice, the vector of missing values $\mathbf{y}^{*}_{\delta^-,q}$ for female q at iteration [k] must be computed as:

$$\hat{\mathbf{y}}^{*[k]}_{\delta^-,q} = \hat{\boldsymbol{\mu}}^{[k]}_{\delta^-} + \hat{\mathbf{a}}^{[k]}_{\delta^-,q} + \mathbf{R}_{\mathbf{0},\delta^-\delta}\,\mathbf{R}^{-1}_{\mathbf{0},\delta\delta}\,(\hat{\mathbf{y}}^{*}_{\delta,q} - \hat{\boldsymbol{\mu}}^{[k]}_{\delta} - \hat{\mathbf{a}}^{[k]}_{\delta,q}) \quad [20]$$

Again, the matrix $\mathbf{R_0}$ in [20] is different from $\mathbf{R_0}$ for the sire model. On the canonical scale, t univariate systems of the form:

$$\begin{bmatrix} F & \mathbf{f}^{*\prime} \\ \mathbf{f}^{*} & \mathbf{F}^{*\prime} + \lambda_{ic} \otimes \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mu_{ic} \\ \mathbf{a}_{ic} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{*\prime} \\ \mathbf{F}^{*\prime} \end{bmatrix} \mathbf{y}^{*[k]}_{ic} \quad [21]$$

must be solved, where **F***, **f*** and F are the matrix, vector and scalar analogous to **N***, **n*** and N in [15]. In particular, **N*** is a diagonal matrix with q^{th} diagonal element equal to the maximum number of records of cow q.

In equation [21], the mean vector $\boldsymbol{\mu}$ and the male EBVs are assumed known. Isolating the rows of [21] corresponding to females, we get the system:

$$\left[\mathbf{F}^{*\prime} + \lambda_{ic} \otimes \mathbf{A}^{-1}_{ff}\right] \mathbf{a}_{ic,f} = \mathbf{F}^{*\prime}\,(\mathbf{y}^{*[k]}_{ic} - \hat{\boldsymbol{\mu}}) - 2\,\mathbf{A}^{-1}_{fm}\,\hat{\mathbf{s}}_{ic,m} \quad [22]$$

where f refers to females and m to males, $\mathbf{A}^{-1}_{ff}$ represents the matrix composed of the terms of $\mathbf{A^{-1}}$ relating females (diagonal terms for females + contribution of their dam and their female progeny) and $\mathbf{A}^{-1}_{fm}$ includes the terms relating females to males (sire, mate or male progeny). System [22] is of the form:

$$\mathbf{H}_{ic}\ \mathbf{a}_{ic,f} = \mathbf{h}^{[k]}_{ic} \quad [23]$$

Poivey (1986) showed that, when the rows of $\mathbf{H}_{ic}$ are ordered such that progeny always preceed parents, then the Cholesky factor $\mathbf{L}_{ic}$ of $\mathbf{H}_{ic}$ (such that $\mathbf{H}_{ic} = \mathbf{L}_{ic}\,\mathbf{L}'_{ic}$) has exactly the same structure as the lower diagonal part of $\mathbf{H}_{ic}$: there is at most one nonzero element in each column q of $\mathbf{L}_{ic}$, located in the row corresponding to the dam of animal q. Then the Cholesky decomposition of $\mathbf{H}_{ic}$ is extremely fast, as well as the *exact* solution of system [23] as:

$$\begin{aligned} &\text{Solve } \mathbf{L}_{ic}\ \mathbf{z}_{ic,f} = \mathbf{h}^{[k]}_{ic} \text{ for } \mathbf{z}_{ic,f} \\ &\text{Solve } \mathbf{L}'_{ic}\ \mathbf{a}_{ic,f} = \mathbf{z}_{ic,f} \text{ for } \mathbf{a}_{ic,f} \end{aligned} \quad [24]$$

An application of this particular decomposition for the solution of large univariate problems can be found in Ducrocq *et al.* (1990). Once this system has been solved for all canonical traits, a new prediction of missing values using [20] is possible. The algorithm is iterated until convergence is reached.

The advantage of this approach is that all females get EBVs on the same traits as males, even if they do not get a cow EBV in the univariate evaluation on some traits. For example, so far, only sire EBVs for direct functional longevity are available. It seems difficult and even questionable (for animals still alive) to compute female EBVs (Ducrocq, 1999), based on direct information only. The procedure described here offers an appealing framework to get female EBVs on functional longevity: Pedigree information on the male side for functional longevity is optimally combined with information from female predictors such as type traits or somatic cell scores. The same total merit index (same economic weight) would apply to males and females.

6. Conclusion

This paper presents a number of approximations to the optimal MT-BLUP evaluation, for the computation of both male and female EBVs. It is shown that, at least theoretically, there are better alternatives than the often-recommended

MACE approach. The implementation of these alternatives will tell us whether the extra computing effort is worthwhile.

There are still *many* unsolved problems to face (estimation of dispersion parameters, sensitivity of the evaluation to the values of these parameters, adaptation when univariate analyses are based on repeated records, inclusion of foreign information and groups of unknown parents, etc…). Without any doubt, the most difficult ones relate to the inclusion of the information on traits described via nonlinear models, and in particular, functional longevity: preadjusted records are *not* available then, especially for censored records. Subsequently, the use of deregressed proofs for the functional longevity part is almost compulsory.

In conclusion, it is strongly argued that the (approximate) MT-BLUP evaluation of males and females offers an appealing framework for the computation of total merit indices, with, as by-products, improved EBVs on low reliability traits (such as functional longevity) that optimally combine direct and indirect information.

References

Dempster, A.P., Laird, N.M., Rubin, D.R., 1977. Maximum likelihood estimation for incomplete data via the EM algorithm (with discussion). *J. Royal Stat. Soc., B.*, 39: 1-38.

Druet, T., Solkner, J., Groen, A.F., Gengler, N., 1999. Improved genetic evaluation of survival using MACE to combine direct and correlated information from yield and functional traits. *Proceedings International Workshop on Genetic Improvement of Functional Traits in cattle (GIFT) - Longevity,* Jouy-en-Josas. INTERBULL Bulletin, 21, 122-127.

Ducrocq, V., 1994. Multiple trait prediction: principles and problems. in *Proc. 5th World Congr. Genet. App. Livest. Prod.,* Guelph, Ontario, Canada. Vol. 18, 455-462.

Ducrocq, V., 1999. Topics that may deserve further attention in survival analysis applied to dairy cattle breeding – some suggestions. *Proceedings International Workshop on Genetic Improvement of Functional Traits in cattle (GIFT) - Longevity,* Jouy-en-Josas. INTERBULL Bulletin, 21, 181-189.

Ducrocq, V., Boichard, D., Bonaiti, B., Barbat, A., Briend, M., 1990. A pseudo-absorption strategy for dolving animal model equations for large data files. *J. Dairy Sci.,* 73: 1945-1955.

Ducrocq, V., Besbes, B., 1993. Solution of multiple trait animal models with missing data on some traits. *J. Anim. Breed. Genet.*, 110:81-92.

Ducrocq, V., Chapuis, H., 1997. Generalizing the use of the canonical transformation for the solution of multivariate mixed model equations. *Genet. Sel. Evol*., 29: 205-224.

Larroque, H., Ducrocq, V., 1999. An indirect approach for the estimation of genetic correlations between longevity and other traits. *Proceedings International Workshop on Genetic Improvement of Functional Traits in cattle (GIFT) - Longevity,* Jouy-en-Josas. INTERBULL Bulletin, 21, 128-135.

Poivey, J.P. , 1986. Méthode simplifiée de calcul des valeurs génétiques des femellles tenant compte de toutes les parentés. *Génét. Sél. Evol.,* 18: 321-332.

Quaas, R.L., 1984. Linear prediction. in*: BLUP school handbook*, *Use of mixed models for prediction and for estimation of (co)variance components,* AGBU, Univ. New England, Armidale, New South Wales, Australia.

Schaeffer L. R., 1994. Multiple-country comparison of dairy sires. *Journal of Dairy Science,* 77 :2671-2678.

Thompson, R, 1976. Estimation of quantitative genetic parameters. In *Proc. Int. Conf; on Quant. Genet.,* E. Pollak, O. Kempthorne, T.B. Bailey (eds). 639-657. Ames, Iowa, USA.

Van der Werf, J., Van Arendonk, J.A.M., de Vries, A.G., 1992. Improving selection of pigs using correlated characters. 43rd Ann. EAAP meeting, Madrid, Spain.