Approximation of Reliability for Multiple-Trait Animal Models with Missing Data by Canonical Transformation

N. GENGLER* **and I. MISZTAL†**

*Fonds National Belge de la Recherche Scientifique, Unité d'Enseignement et de Recherche de Zootechnie, Faculté Universitaire des Sciences Agronomiques, B-5030 Gembloux, Belgium
†Department of Animal Sciences, University of Illinois, Urbana 61801

ABSTRACT

An algorithm for approximation of reliability for multiple traits by multiple diagonalization was modified to support missing data by weighting transformed contributions of records based on the pattern of missing data. The accuracy of approximation was assessed with simulated and field data by comparing approximate reliabilities with those from direct inversion. Simulated data had several levels of missing data and covariances between traits; correlations were close to those for linear type traits of dairy cattle. Field data were 1) dairy records for milk, fat, and protein yields with 26% of the observations for fat and protein removed and 2) beef records for birth weight, weaning weight, and mean gain after weaning with 43% of observations missing. These files also contained empty fixed effect classes. The algorithm worked best for simulated data, and, when covariances between traits decreased, proportion of missing traits decreased and the number of empty fixed classes decreased. For dairy data, improvement over single-trait reliability occurred only for traits with missing data; for beef data, little or no improvement occurred. The method is useful with multiple diagonalization if the proportion of missing records or number of empty fixed effect classes or covariances between traits is moderate.

(**Key words**: reliability, multiple trait, animal model, missing data)

Abbreviation key: **PEV** = prediction error variance.

INTRODUCTION

The solution for mixed models (7) associated with single-trait animal models and large data files now is possible and practical (11) using techniques that iterate on the data (15). For multiple-trait models, the same principles can be used, but computations take much longer. If an identical model is used for each trait, a canonical transformation can save computing time and resources for multiple-trait models (11). Multiple diagonalization, which is an extended form of canonical transformation, accommodates more than one random effect (8, 11, 12). Recent developments show that missing data for certain traits can be analyzed, as can different models per trait after combining single- and multiple-trait algorithms (4). Thus, multiple-trait models, canonical transformation, and therefore also multiple diagonalization generally are applicable for prediction of breeding values for data files with missing observations. However, similar algorithms for computing approximate measures of accuracy of prediction are not available. For dairy cattle, the concept of reliability is mostly used to describe accuracy, which is defined as a linear function of the prediction error variance (**PEV**), expressed between 0 and 1.

A method to approximate reliabilities for single-trait models has been established (10, 13, 14). For multiple-trait models, methods have been developed to approximate reliabilities for a sire model (6) and for models using multiple diagonalization for data files with no missing observations (11). The objectives of this study were 1) to extend the latter method to accommodate missing data and 2) to demonstrate that different models per trait can be supported with multiple diagonalization without extra programming.

MATERIALS AND METHODS

Statistical Model

A multiple-trait linear model with t traits and records ordered within traits, where no observations are missing, is

$$\mathbf{y} = (\mathbf{I}_t \otimes \mathbf{X})\mathbf{b} + (\mathbf{I}_t \otimes \mathbf{Z})\mathbf{p} + (\mathbf{I}_t \otimes \mathbf{Z})\mathbf{a} + \mathbf{e} \quad [1]$$

where **y** is the vector of observations, **b** is the vector of fixed effects, **p** is the vector of random permanent environmental effects, **a** is the vector of random

Received March 6, 1995.
Accepted October 18, 1995.

animal effects, and **e** is the vector of random residual effects:

$$\mathbf{y} = (\mathbf{y}_1' \quad \mathbf{y}_2' \quad \ldots \quad \mathbf{y}_t')',$$
$$\mathbf{b} = (\mathbf{b}_1' \quad \mathbf{b}_2' \quad \ldots \quad \mathbf{b}_t')',$$
$$\mathbf{p} = (\mathbf{p}_1' \quad \mathbf{p}_2' \quad \ldots \quad \mathbf{p}_t')',$$
$$\mathbf{a} = (\mathbf{a}_1' \quad \mathbf{a}_2' \quad \ldots \quad \mathbf{a}_t')', \text{ and}$$
$$\mathbf{e} = (\mathbf{e}_1' \quad \mathbf{e}_2' \quad \ldots \quad \mathbf{e}_t')';$$

where $\otimes$ denotes a Kronecker product, and **X** and **Z** contain incidence matrices. The (co)variance matrices of the random effects are defined as $\text{Var}(\mathbf{p}) = \mathbf{P} = \mathbf{P}_0 \otimes \mathbf{I}_n$, $\text{Var}(\mathbf{a}) = \mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$, and $\text{Var}(\mathbf{e}) = \mathbf{R} = \mathbf{R}_0 \otimes \mathbf{I}$ where $\mathbf{P}_0$, $\mathbf{G}_0$, and $\mathbf{R}_0$ are the (co)variance matrices between traits, n is the number of animals, and **A** is the numerator relationship matrix. The mixed model equations can be expressed as

$$\begin{bmatrix} \mathbf{R}_0^{-1} \otimes \mathbf{X}'\mathbf{X} & \mathbf{R}_0^{-1} \otimes \mathbf{X}'\mathbf{Z} & \mathbf{R}_0^{-1} \otimes \mathbf{X}'\mathbf{Z} \\ \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{X} & \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{Z} + \mathbf{P}_0^{-1} \otimes \mathbf{I}_n & \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{Z} \\ \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{X} & \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{Z} & \mathbf{R}_0^{-1} \otimes \mathbf{Z}'\mathbf{Z} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{p}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} (\mathbf{R}_0^{-1} \otimes \mathbf{X}')\mathbf{y} \\ (\mathbf{R}_0^{-1} \otimes \mathbf{Z}')\mathbf{y} \\ (\mathbf{R}_0^{-1} \otimes \mathbf{Z}')\mathbf{y} \end{bmatrix}. \quad [2]$$

Multiple Random Effects and Missing Data

Let **L** be the Cholesky factor of $\mathbf{R}_0$ (i.e., $\mathbf{R}_0 = \mathbf{LL}'$). Multiple diagonalization of the (co)variance matrices $\mathbf{P}_0$, $\mathbf{G}_0$, and $\mathbf{R}_0$ is possible if a matrix **B** exists that satisfies the following three equations

$$\mathbf{L}^{-1}\mathbf{R}_0(\mathbf{L}^{-1})' = \mathbf{BIB}',$$
$$\mathbf{L}^{-1}\mathbf{G}_0(\mathbf{L}^{-1})' = \mathbf{BDB}', \text{ and}$$
$$\mathbf{L}^{-1}\mathbf{P}_0(\mathbf{L}^{-1})' = \mathbf{B}\Delta\mathbf{B}'; \quad [3]$$

where **D** and Δ are diagonal matrices, and **I** is an identity matrix. Such a matrix **B** exists if one of the three matrices can be expressed as a linear function of the other two. In other cases, a good approximation often exists (5). If [3] is true, the transformation matrix can be defined as $\mathbf{Q} = (\mathbf{LB})^{-1}$, with $\mathbf{QP}_0\mathbf{Q}' = \Delta$, $\mathbf{QG}_0\mathbf{Q}' = \mathbf{D}$, and $\mathbf{QR}_0\mathbf{Q}' = \mathbf{I}$. Then **y** is transformed to $\mathbf{y}_Q$ by $\mathbf{y}_Q = (\mathbf{Q} \otimes \mathbf{I}_n)\mathbf{y}$.

The mixed model Equations [2] associated with the original Model [1] can be simplified and split into t independent single-trait mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \tau_i\mathbf{I} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \alpha_i\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}_{Qi} \\ \hat{\mathbf{p}}_{Qi} \\ \hat{\mathbf{a}}_{Qi} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y}_{Qi} \\ \mathbf{Z}'\mathbf{y}_{Qi} \\ \mathbf{Z}'\mathbf{y}_{Qi} \end{bmatrix}$$

$$\Leftrightarrow \mathbf{C}_i\hat{\mathbf{s}}_i = \mathbf{r}_i \quad [4]$$

where $\tau_i = 1/\delta_i$, $\alpha_i = 1/d_i$, and δ_i and d_i = diagonal elements i of Δ and **D**, respectively. Original solutions can be obtained by backtransformation:

$$\hat{\mathbf{b}} = (\mathbf{Q}^{-1} \otimes \mathbf{I})\hat{\mathbf{b}}_Q,$$
$$\hat{\mathbf{p}} = (\mathbf{Q}^{-1} \otimes \mathbf{I}_n)\hat{\mathbf{p}}_Q, \text{ and}$$
$$\hat{\mathbf{a}} = (\mathbf{Q}^{-1} \otimes \mathbf{I}_n)\hat{\mathbf{a}}_Q.$$

A condition for the use of the multiple diagonalization is that no observations for certain traits are missing. Ducrocq and Besbes (4) described a method that permits canonical transformation with missing data and the same incidence matrices for all traits. Their method is based on the replacement of a missing observation with its expectation (3) using an expectation-maximization algorithm (9). Assume that a record of animal j contains two groups of traits, traits that are observed and traits that are missing. If $\mathbf{y}_j$ represents this record, then $\mathbf{y}_{j\alpha}$ is the part of the record containing the observed traits, and $\mathbf{y}_{j\beta}$ is the part of the record with the missing traits. For iteration k, $\hat{\mathbf{y}}_{j\alpha}^{[k]} = \mathbf{y}_{j\alpha}$, and

$$\hat{\mathbf{y}}_{j\beta}^{[k]} = \text{E}[\mathbf{y}_{j\beta} | \mathbf{y}_{j\alpha}, \mathbf{b} = \mathbf{b}^{[k]}, \mathbf{p} = \mathbf{p}^{[k]}, \mathbf{a} = \mathbf{a}^{[k]}],$$

which can be rewritten as

$$\hat{\mathbf{y}}_{j\beta}^{[k]} = \mathbf{X}_{j\beta}\hat{\mathbf{b}}_{j\beta}^{[k]} + \hat{\mathbf{p}}_{j\beta}^{[k]} + \hat{\mathbf{a}}_{j\beta}^{[k]} + \hat{\mathbf{e}}_{j\beta}^{[k]} \quad [5]$$

where $\mathbf{X}_{j\beta}$ is the submatrix of **X** that associates $\mathbf{y}_{j\beta}$ and $\hat{\mathbf{b}}_{j\beta}$. All the terms on the right side of the equation are obtained as solutions from iteration k, except for the residuals for missing observations, which are estimated to be the regression of those residuals on the current estimates of the residuals for observed traits. These calculations represent the expectation step of the expectation-maximization algorithm.

At iteration k + 1, a missing observation is replaced by its expectation, and new solutions for **b**, **p**, and **a** are obtained (the maximization step). Ducrocq and Besbes (4) proposed a method to avoid backtransformation that simplified computations.

Different Models per Trait

Ducrocq and Besbes (4) also showed that canonical transformation can support different sets of fixed

effects per model by computing these effects as in a regular multiple-trait procedure and then transforming data adjusted for the fixed effects. Although this method extended canonical transformation to general models, computer programming had to include a procedure to solve multiple-trait models and, therefore, was complicated. The same result can be accomplished for multiple diagonalization and, therefore, also for canonical transformation without additional programming by 1) declaring all fixed effects for each trait, 2) splitting each record into multiple records such that each new record contains the same combination of fixed effects, and 3) assigning values of unneeded fixed effects to a "dummy" level for each new record. If every trait has a different model, each new observation contains one known trait with all remaining traits unknown. This approach results in increased storage requirements for data files unless splitting the records is incorporated into the iteration program.

Numerical example. Consider a joint analysis for production traits and final score, in which milk, fat, and protein records are distributed in fixed herd-year-season classes, and final score records are grouped according to fixed herd-year-month-classification classes and are also affected by the classifier effect. Consider the following records for three cows A, B, and C:

Cow	Herd-year-season	Herd-year-month-classification	Classifier	Milk	Fat	Protein	Final score
A	1	1	1	6000	200	175	78
B	2	2	1	7000	225	200	80
C	2	3	2	8000	250	225	82

They can be rewritten using the algorithm just described:

Cow	Main fixed effect	Classifier	Milk	Fat	Protein	Final score
A	1	D	6000	200	175	M
B	2	D	7000	225	200	M
C	2	D	8000	250	225	M
A	1	1	M	M	M	78
B	2	1	M	M	M	80
C	3	2	M	M	M	82

where the main fixed effect represents herd-year-season for yield traits or herd-year-month-classification for final score, M is the code for missing values, and D is the code for a dummy level, which in this case is 1.

Approximation of Reliability

Let $\mathbf{W}_{Qj}$ be the diagonal matrix of PEV of the t transformed traits for animal j. Following the method of Misztal et al. (11), if $\mathbf{W}_j$ is the matrix of PEV for the original traits, then

$$\mathbf{W}_{Qj} = \mathbf{Q}\mathbf{W}_j\mathbf{Q}' \Leftrightarrow \mathbf{W}_j = \mathbf{Q}^{-1}\mathbf{W}_{Qj}(\mathbf{Q}^{-1})'. \quad [6]$$

The PEV for the transformed traits can be obtained using the method proposed by Misztal and Wiggans (13): $w_{Qij} = 1/(\alpha_i + b_{ij})$, where w_{Qij} is PEV of transformed trait i of animal j, $\alpha_i = 1/d_i$, and b_{ij} is the information on animal j expressed as effective records (11, 13). This information is assumed to be a sum of contributions from own records (f_{ij}) and from immediate relatives (parent or progeny) of animal k (g_{ijk}):

$$b_{ij} = f_{ij} + \sum_k g_{ijk},$$

where f_{ij} and g_{ijk} were derived as in Misztal and Wiggans (13). For a repeatability model with one important fixed effect, reduction of information because of fitting the permanent environmental effect is reflected by

$$f_{ij} = \tau_i z_{ij}/(\tau_i + z_{ij}), \quad [7]$$

where τ_i is a variance ratio for permanent environment for trait i, and z_{ij} is the numbers of records adjusted to reflect the reduction of information because of fitting the fixed effect:

$$z_{ij} = \sum_l (1 - 1/n_l), \quad [8]$$

where n_l is the number of records in fixed effects subclass l when cow j has a record.

The contribution from pedigree is obtained by an iterative procedure (13). If no observations are missing for the original traits, the contributions f_{ij} differ only between transformed traits because of different permanent environmental variances. If some observations are missing, b_{ij} is overestimated, and, consequently, w_{Qij} is underestimated. Contributions g_{ijk} are affected indirectly because they are functions of b_{ij}.

To examine the reduction of b_{ij} because of missing data, assume that after multiple diagonalization, the left side of the coefficient matrix of Equation [4] can be approximated:

$$\mathbf{C}_i = \begin{bmatrix} \mathbf{X}'\mathbf{H}_{id}\mathbf{X} & \mathbf{X}'\mathbf{H}_{id}\mathbf{Z} & \mathbf{X}'\mathbf{H}_{id}\mathbf{Z} \\ \mathbf{Z}'\mathbf{H}_{id}\mathbf{X} & \mathbf{Z}'\mathbf{H}_{id}\mathbf{Z} + \tau_i\mathbf{I} & \mathbf{Z}'\mathbf{H}_{id}\mathbf{Z} \\ \mathbf{Z}'\mathbf{H}_{id}\mathbf{X} & \mathbf{Z}'\mathbf{H}_{id}\mathbf{Z} & \mathbf{Z}'\mathbf{H}_{id}\mathbf{Z} + \alpha_i\mathbf{A}^{-1} \end{bmatrix}, \quad [9]$$

where $\mathbf{H}_{id}$ is a diagonal matrix of weights that reflect the contributions of observations. New formulas for the contribution of records are derived by analyzing submatrices of Equation [9].

The fixed effect and animal equations for trait i of animal j with a record in fixed effects subclass l are

$$\begin{bmatrix} n_l' & \cdots & \gamma_{ijl} & \cdots & \gamma_{ijl} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{ijl} & \cdots & \sum_l \gamma_{ijl} + \tau_i & \cdots & \sum_l \gamma_{ijl} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{ijl} & \cdots & \sum_l \gamma_{ijl} & \cdots & \sum_l \gamma_{ijl} + \alpha_i(\ldots) \end{bmatrix},$$

where γ_{ijl} is the part of transformed trait i with observations present, and n_l' is the sum of γ_{ijl} in fixed effect subclass l. After the sequential absorption of the equations for fixed effect subclasses and permanent environment into the animal equation, Equation [7] does not change, and Equation [8] becomes

$$z_{ij} = \sum_l \{\gamma_{ijl}(1 - \gamma_{ijl}/n_l')\}, \quad [10]$$

where γ_{ijl} represents a reduction in reliability because of missing data. Once γ_{ijl} is approximated, approximate reliabilities can be calculated.

If a coefficient matrix for only one animal and one observation is considered and all other fixed and permanent environmental effects are ignored, then PEV on the transformed scale is assumed to be

$$\mathbf{W}_{Qj}^* \approx (\mathbf{\Gamma} + \mathbf{D}^{-1})^{-1}, \quad [11]$$

where $\mathbf{\Gamma} =$

$$\begin{bmatrix} \gamma_{1j} & & & \\ & \cdot & & \mathbf{0} \\ & & \gamma_{ij} & \\ & \mathbf{0} & & \cdot \\ & & & \gamma_{tj} \end{bmatrix},$$

and γ_{ij} are contributions from one record, reduced to reflect lack of contributions from missing data. Exact PEV on the original scale ($\mathbf{W}_j^*$) is

$$\mathbf{W}_j^* = [\mathbf{R}_j^* + \mathbf{G}_0^{-1}]^{-1}, \quad [12]$$

where $\mathbf{R}_j^*$ is in the case of ordered traits (present before missing):

$$\mathbf{R}_j^* = \begin{bmatrix} \mathbf{R}_\alpha^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and $\mathbf{R}_\alpha^{-1}$ is the inverse of the part of $\mathbf{R}_0$ that is associated with nonmissing traits. $\mathbf{R}_j^*$ is also a particular generalized inverse of $\mathbf{R}_0$, the residual (co)variance matrix of original traits with zeros in rows and columns corresponding to missing data. On the transformed scale, we can compute

$$\mathbf{W}_{Qj}^* = \mathbf{Q}\mathbf{W}_j^*\mathbf{Q}'. \quad [13]$$

By isolating $\mathbf{W}_j^*$,

$$\mathbf{W}_j^* = \mathbf{Q}^{-1}\mathbf{W}_{Qj}^*(\mathbf{Q}^{-1})'. \quad [14]$$

For computing PEV of individuals, different possibilities exist to approximate $\mathbf{G}$. They can be based on equating complete formulas or only diagonals on either the original or transformed scale. The three most obvious possibilities are based on an approximate diagonalization of $\mathbf{R}_j^*$, on equating diagonals of Equations [12] and [14], or on equating diagonals of Equations [11] and [13]. Preliminary tests using these and other possibilities were done and showed the superiority of the last approach, which can be rewritten as

$$\mathbf{\Gamma} \approx [\text{diag}(\mathbf{Q}\mathbf{W}_j^*\mathbf{Q}')]^{-1} - \mathbf{D}^{-1}. \quad [15]$$

In Equation [15], an approximate diagonalization of $\mathbf{W}_j^*$ is performed, and the resulting off-diagonals are discarded. With all traits recorded, the equation is exact because $[\text{diag}(\mathbf{Q}\mathbf{W}_j\mathbf{Q}')]^{-1} - \mathbf{D}^{-1} = \mathbf{I}_t$, or $\gamma_{ij} = 1$. If some traits are missing, $\mathbf{Q}\mathbf{W}_j^*\mathbf{Q}'$ has off-diagonal elements, and Equation [15] is an approximation only. Equation [15] also gives exact results ($\gamma_{ij} = 0$ if trait i is missing, and $\gamma_{ij} = 1$ if i is not missing) if (co)variances between traits are 0, which is equivalent to the single-trait model. As a matter of fact, given Equation [15], the approximation improves if off-diagonals of $\mathbf{Q}\mathbf{W}_j^*\mathbf{Q}'$ are small.

Algorithm

The algorithm to calculate the approximate reliabilities for a large-scale multitrait animal model with missing data follows:

1. For every observation, compute γ_{ij} as in Equation [15].
2. For every animal, compute adjustments for fixed effects as in Equation [10], and then adjust for permanent environmental effect as in Equation [7].
3. After calculating f_{ij}, the adjustment for effect of permanent environment, compute the single-trait PEV algorithm of Misztal and Wiggans (13) for each of t traits.
4. Backtransform PEV as in Equation [6].
5. Compute reliabilities for every animal and original trait as $1 - (w_{ij}/\sigma^2_{gj})$, where w_{ij} is the PEV for original trait i and animal j and σ^2_{gj} is the genetic variance of original trait i.

Comparison of Reliabilities Approximated with and Without Correction for Missing Data

Reliabilities were calculated for simulated and field data. (Co)variance matrices used were checked to be positive definite.

Simulated data. Data were simulated for 900 cows with records having 400 ancestors. All cows had a minimum of one record with a random number of additional records with up to 10 records for one animal. A total of 3000 records was grouped in 100 classes for fixed effects. The design was unbalanced, and the classes for fixed effects contained from 17 to 46 observations. Records were distributed randomly among fixed effect classes.

For the first simulated data file (F1), random effects included additive genetic, permanent environmental, and residual effects with respective (co)variances:

$$\mathbf{P}_0 = \begin{bmatrix} 1000 & 500 & -900 \\ 500 & 2000 & -100 \\ -900 & -100 & 2000 \end{bmatrix},$$

$$\mathbf{G}_0 = \begin{bmatrix} 4000 & 500 & 700 \\ 500 & 3000 & -1000 \\ 700 & -1000 & 2000 \end{bmatrix},$$

and

$$\mathbf{R}_0 = \begin{bmatrix} 2000 & 1000 & -1800 \\ 1000 & 4000 & -200 \\ -1800 & -200 & 4000 \end{bmatrix}.$$

These genetic parameters correspond to correlations reported by Misztal et al. (12) for type traits. Three other data files (F2, F3, and F4) were simulated with identical variances but with covariances of 0.1, 0.5, and 1.4 times the covariances of F1, respectively. These four data files were examined with various percentages (0, 11, or 33%) of data considered missing to determine whether correlations decreased as covariances increased. Observations to be considered missing were randomly chosen, but three missing observations for one record were avoided.

Dairy field data. Data for dairy cattle were obtained from A. Toussaint, and A. De Bast (Élevage Informatique, Ciney, Belgium) and included 5809 305-d records for milk, fat, and protein from 500 cows chosen randomly and their 3108 contemporaries in 614 herd-year-seasons. The number of animals totaled 6072 with ancestors included. Variance and covariance components, expressed in square kilograms, were obtained from another study with similar but more data (P. Coenraets and N. Gengler, 1994, personal communication):

$$\mathbf{P}_0 = \begin{bmatrix} 126{,}102 & 4508 & 3538 \\ 4508 & 271 & 145 \\ 3538 & 145 & 115 \end{bmatrix},$$

$$\mathbf{G}_0 = \begin{bmatrix} 153{,}767 & 4925 & 3890 \\ 4925 & 347 & 156 \\ 3890 & 156 & 123 \end{bmatrix},$$

and

$$\mathbf{R}_0 = \begin{bmatrix} 484{,}291 & 19{,}429 & 15{,}098 \\ 19{,}429 & 973 & 641 \\ 15{,}098 & 641 & 510 \end{bmatrix}.$$

Data for protein and fat were considered to be missing for 26% of records, with 9% of the missing observations concentrated in the 58 most recent herd-year-seasons and the other 17% dispersed randomly among other records. The purpose of eliminating some data was to provide a realistic situation for Belgium and the US in which information on protein or on fat and protein was missing.

Beef field data. Data for beef cattle were obtained from R. E. Golden (1994, Colorado State University, Fort Collins) and included 7270 records for birth weight (7270 observations), weaning weight (7270 observations), and postweaning gain (4150 observations) with a total of 7864 cattle. The model included fixed management effects, which were assumed to be common to all traits, and random animal effects, but not permanent environmental or maternal effects. (Co)variance, expressed as square kilograms, was based on information from B. Klei (Cornell Univer-

sity, Ithaca, New York, 1994, personal communication):

$$\mathbf{G}_0 = \begin{bmatrix} 7.9 & 21.1 & 11.8 \\ 21.1 & 236.8 & 103.7 \\ 11.8 & 103.7 & 174.6 \end{bmatrix},$$

and

$$\mathbf{R}_0 = \begin{bmatrix} 12.4 & 15.3 & 10.1 \\ 15.3 & 588.6 & -91.9 \\ 10.1 & -91.9 & 496.7 \end{bmatrix}.$$

Missing data (43% of observations) were essentially all grouped in specific classes for fixed effects, unlike the dairy data, for which only 9% of missing observations were concentrated in recent herd-year-seasons.

Method of comparison. Approximate reliabilities were calculated with the algorithm that corrects for missing values and with the normal algorithm that ignores missing data; then those approximate values were compared with exact reliabilities obtained from a direct inversion approach by MTDFREML (1), a package using the SPARSPAK sparse matrix solver (2), which was modified to obtain the reliabilities. For field data files, single-trait reliabilities were calculated using the same algorithm but ignoring (co)variances between traits. Accuracy of the methods was assessed by Pearson's correlation between approximated and exact reliabilities and by means, standard deviations, and maxima of the approximation error, computed separately for all animals, animals with records, and sires of animals with records.

RESULTS AND DISCUSSION

Table 1 shows correlations, mean errors, standard deviations of errors, and maximum absolute errors between approximate and exact reliabilities for simulated data observed for all the animals, with and without the correction for missing traits. With no missing observations (data file F1), approximate

TABLE 1. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for simulated data with and without correction for missing observations, mean errors of approximation (mean error), standard deviation of errors (SD error), and maximum absolute errors (max error) observed for all animals.

Data file	Observations missing	Measure	Without correction			With correction		
			Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
	(%)							
F1	0	r	1.000	1.000	1.000	1.000	1.000	1.000
		Mean error	0.005	0.005	0.004	0.005	0.005	0.004
		SD Error	0.003	0.003	0.004	0.003	0.003	0.004
		Max error	0.036	0.024	0.026	0.036	0.024	0.026
	11	r	0.979	0.979	0.981	0.998	0.999	0.996
		Mean error	0.016	0.017	0.015	0.004	0.005	0.004
		SD Error	0.037	0.026	0.025	0.010	0.007	0.012
		Max error	0.542	0.342	0.300	0.183	0.099	0.156
	33	r	0.895	0.905	0.900	0.992	0.991	0.970
		Mean error	0.057	0.054	0.057	-0.001	0.010	0.007
		SD Error	0.084	0.057	0.058	0.024	0.018	0.032
		Max error	0.591	0.368	0.374	0.174	0.152	0.220
F2[1]	33	r	0.899	0.887	0.893	1.000	1.000	0.999
		Mean error	0.055	0.059	0.056	0.006	0.006	0.005
		SD Error	0.076	0.061	0.051	0.004	0.004	0.004
		Max error	0.554	0.406	0.345	0.028	0.019	0.019
F3[2]	33	r	0.899	0.891	0.898	0.998	0.996	0.983
		Mean error	0.056	0.058	0.057	0.002	0.008	0.007
		SD Error	0.077	0.060	0.051	0.012	0.012	0.021
		Max error	0.560	0.399	0.348	0.085	0.100	0.138
F4[3]	11	r	0.975	0.983	0.976	0.999	0.999	0.997
		Mean error	0.012	0.015	0.011	0.004	0.004	0.003
		SD Error	0.044	0.024	0.035	0.011	0.007	0.012
		Max error	0.579	0.306	0.400	0.216	0.114	0.204

[1]Covariances are 0.1 times those for data file F1.

[2]Covariances are 0.5 times those for data file F1.

[3]Covariances are 1.4 times those for data file F1.

reliabilities were close to exact reliabilities from direct inversion.

The improvement with the correction was most dramatic for simulated data with extremely low covariances among traits (data file F2). The improvement with the correction was smaller for data with higher covariances between traits. Thus, the algorithm works best with fewer missing observations and low correlations between traits. This last fact is obvious because Equation [15] is only an approximation because of discarded off-diagonals.

Tables 2 and 3 show values for the same measures observed for animals with records and for sires of animals with records. The minimum correlation for animals with records without the correction is only 0.408, but for sires it is never <0.969. After the corrections, these correlations increase to 0.888 and 0.992, which suggests that, although the correction for missing data helped the reliability approximation for animals with records, these approximations were still not as good as for sires with several progeny.

Also, for the field data for dairy cattle, correlations between approximate and exact reliabilities (Table 4) were much higher than for simulated data. For milk and protein yields, they were >0.990, even without the use of the algorithm for the correction for missing data. The correction has increased the correlations for fat from 0.931 to 0.997, marginally decreased the correlation for protein, and reduced the correlation for milk from 0.995 to 0.969. The correction changed bias for milk from 0.013 to -0.010, meaning that reliability for milk changed from generally being overestimated to being underestimated. The single-trait approximation had a bias up to four times higher than the multiple-trait approximation with the correction, but the standard deviation of the single-trait approximation was better for milk, the only trait that was always recorded; the maximum error was smaller for milk and fat. Relatively modest gains from the correction in yield traits were likely caused by high correlations among the traits.

The most accurate method for beef data was the single-trait approach. The correction in the multitrait

TABLE 2. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for simulated data with and without correction for missing observations, mean errors of approximation (mean error), standard deviation of errors (SD error), and maximum absolute errors (max error) observed for all animals with records.

Data file	Observations missing	Measure	Without correction			With correction		
			Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
	(%)							
F1	0	r	0.995	0.995	0.995	0.995	0.995	0.995
		Mean error	0.005	0.005	0.005	0.005	0.005	0.005
		SD Error	0.003	0.003	0.003	0.003	0.003	0.003
		Max error	0.012	0.012	0.011	0.012	0.012	0.011
	11	r	0.656	0.831	0.778	0.978	0.993	0.955
		Mean error	0.019	0.020	0.018	0.004	0.006	0.004
		SD Error	0.043	0.031	0.029	0.012	0.007	0.014
		Max error	0.542	0.342	0.300	0.183	0.099	0.156
	33	r	0.549	0.651	0.619	0.968	0.980	0.888
		Mean error	0.068	0.063	0.067	-0.001	0.011	0.008
		SD Error	0.098	0.065	0.067	0.028	0.021	0.038
		Max error	0.590	0.368	0.374	0.174	0.152	0.220
F2[1]	33	r	0.576	0.625	0.638	1.000	0.999	0.999
		Mean error	0.065	0.068	0.065	0.007	0.006	0.006
		SD Error	0.088	0.070	0.057	0.003	0.003	0.004
		Max error	0.554	0.406	0.345	0.019	0.019	0.019
F3[2]	33	r	0.571	0.631	0.642	0.993	0.993	0.950
		Mean error	0.066	0.067	0.066	0.002	0.010	0.008
		SD Error	0.090	0.069	0.058	0.014	0.013	0.025
		Max error	0.560	0.399	0.348	0.085	0.100	0.138
F4[3]	11	r	0.408	0.844	0.519	0.975	0.989	0.957
		Mean error	0.014	0.017	0.014	0.004	0.005	0.004
		SD Error	0.052	0.028	0.042	0.012	0.008	0.014
		Max error	0.579	0.306	0.400	0.216	0.114	0.204

[1]Covariances are 0.1 times those for data file F1.

[2]Covariances are 0.5 times those for data file F1.

[3]Covariances are 1.4 times those for data file F1.

TABLE 3. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for simulated data with and without correction for missing observations, mean errors of approximation (mean error), standard deviation of errors (SD error), and maximum absolute errors (max error) observed for sires of animals with records.

Data file	Observations missing	Measure	Without correction			With correction		
			Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
	(%)							
F1	0	r	0.999	0.999	0.999	0.999	0.999	0.999
		Mean error	0.005	0.005	0.004	0.005	0.005	0.004
		SD Error	0.003	0.003	0.003	0.003	0.003	0.003
		Max error	0.024	0.024	0.019	0.024	0.024	0.019
	11	r	0.993	0.994	0.996	0.999	0.999	0.998
		Mean error	0.011	0.014	0.012	0.004	0.005	0.004
		SD Error	0.011	0.011	0.008	0.005	0.004	0.006
		Max error	0.141	0.091	0.059	0.047	0.025	0.029
	33	r	0.969	0.980	0.978	0.996	0.998	0.992
		Mean error	0.038	0.042	0.041	0.000	0.009	0.005
		SD Error	0.025	0.019	0.018	0.009	0.007	0.011
		Max error	0.177	0.091	0.097	0.032	0.028	0.039
F2[1]	33	r	0.970	0.974	0.975	1.000	0.999	0.999
		Mean error	0.039	0.046	0.046	0.005	0.005	0.005
		SD Error	0.025	0.022	0.021	0.003	0.004	0.003
		Max error	0.182	0.103	0.118	0.020	0.019	0.015
F3[2]	33	r	0.971	0.976	0.976	0.999	0.999	0.996
		Mean error	0.039	0.045	0.026	0.002	0.007	0.005
		SD Error	0.025	0.021	0.020	0.005	0.005	0.008
		Max error	0.180	0.104	0.111	0.023	0.026	0.033
F4[3]	11	r	0.992	0.995	0.995	0.998	0.999	0.998
		Mean error	0.009	0.013	0.009	0.004	0.004	0.003
		SD Error	0.012	0.010	0.008	0.006	0.004	0.005
		Max error	0.147	0.078	0.054	0.055	0.019	0.022

[1]Covariances are 0.1 times those for data file F1.

[2]Covariances are 0.5 times those for data file F1.

[3]Covariances are 1.4 times those for data file F1.

TABLE 4. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for field data with and without correction for missing observations and single-trait reliabilities obtained by ignoring covariances between traits; mean errors of approximation (mean error), standard deviation of errors (SD error) and maximum absolute errors (max error), measures observed for all the animals.

		Correlations between multiple-trait reliabilities								
		Without correction			With correction			Single-trait reliabilities with correction for missing data		
Data	Measure	Milk	Fat	Protein	Milk	Fat	Protein	Milk	Fat	Protein
Dairy	r	0.995	0.931	0.997	0.969	0.997	0.994	0.995	0.972	0.870
	Mean error	0.013	0.036	0.009	−0.010	0.008	0.002	−0.042	−0.038	−0.029
	SD Error	0.015	0.052	0.011	0.032	0.013	0.012	0.013	0.031	0.054
	Max error	0.151	0.334	0.167	0.143	0.153	0.124	0.126	0.080	0.286
		Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain
Beef	r	0.994	0.984	0.805	0.971	0.990	0.918	0.994	0.982	0.989
	Mean error	0.013	0.025	0.092	−0.011	0.001	0.067	0.008	−0.018	−0.034
	SD Error	0.011	0.016	0.064	0.024	0.013	0.043	0.011	0.017	0.019
	Max error	0.335	0.319	0.342	0.259	0.251	0.267	0.328	0.344	0.144

TABLE 5. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for field data with and without correction for missing observations and single-trait reliabilities obtained by ignoring covariances between traits; mean errors of approximation (mean error), standard deviation of errors (SD error, and maximum absolute errors (max error), measures observed for animals with records.

		Correlations between multiple-trait reliabilities								
		Without correction			With correction			Single-trait reliabilities with correction for missing data		
Data	Measure	Milk	Fat	Protein	Milk	Fat	Protein	Milk	Fat	Protein
Dairy	r	0.981	0.785	0.990	0.900	0.992	0.982	0.980	0.970	0.738
	Mean error	0.019	0.054	0.013	−0.014	0.013	0.002	−0.007	−0.041	−0.042
	SD Error	0.014	0.059	0.011	0.040	0.012	0.014	0.015	0.033	0.066
	Max error	0.151	0.334	0.167	0.143	0.153	0.124	0.126	0.160	0.286
		Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain
Beef	r	0.986	0.964	0.711	0.933	0.980	0.902	0.986	0.963	0.992
	Mean error	0.014	0.025	0.094	−0.012	0.001	0.068	0.008	−0.019	−0.035
	SD Error	0.009	0.016	0.065	0.024	0.013	0.044	0.009	0.016	0.017
	Max error	0.147	0.172	0.311	0.101	0.104	0.248	0.138	0.108	0.121

model has decreased bias and maximum error for all traits, but the standard errors actually increased from 0.011 to 0.024, and the correlation for postweaning gain was only 0.918. The failure of the correction could be due to an inability to account for missing management classes.

Tables 5 and 6 show the results for animals with records and sires of animals with records. As for simulated data, the correction is much better for sires than for cows. Correlations for sires were all >0.99, which was better than with the single-trait model for which the correlation for protein was 0.974. Correlations for sires with beef data were all >0.97; however, the correlations by the single-trait model were all >0.99. One explanation for the different behavior observed for the three types of data files is the different distribution of missing values for the fixed effects. For the simulated data files, missing values were not grouped in complete empty levels of the fixed effect. For the field data for dairy cattle, 9% of the missing values were concentrated in certain herd-year-season classes, and, for the field data for beef cattle, nearly all missing data were concentrated in certain fixed effect classes. This result suggests that the algorithm is less accurate if complete classes of fixed effects were missing.

TABLE 6. Correlations (r) between exact reliabilities from multiple-trait direct inversion and reliabilities from multiple-trait approximation by multiple diagonalization for field data with and without correction for missing observations and single-trait reliabilities obtained by ignoring covariances between traits; mean errors of approximation (mean error), standard deviation of errors (SD error, and maximum absolute errors (max error), measures observed for sires of animals with records.

		Correlations between multiple-trait reliabilities								
		Without correction			With correction			Single-trait reliabilities with correction for missing data		
Data	Measure	Milk	Fat	Protein	Milk	Fat	Protein	Milk	Fat	Protein
Dairy	r	0.998	0.986	0.998	0.993	0.998	0.998	0.998	0.992	0.974
	Mean error	0.011	0.029	0.011	−0.005	0.009	0.005	0.002	−0.017	−0.020
	SD Error	0.013	0.026	0.013	0.018	0.013	0.012	0.011	0.017	0.028
	Max error	0.079	0.190	0.077	0.085	0.073	0.067	0.062	0.141	0.223
		Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain	Birth weight	Weaning weight	Post-weaning gain
Beef	r	0.991	0.992	0.954	0.992	0.994	0.974	0.991	0.992	0.996
	Mean error	0.024	0.030	0.100	0.007	0.012	0.079	0.022	0.007	−0.010
	SD Error	0.024	0.023	0.056	0.024	0.021	0.042	0.024	0.026	0.028
	Max error	0.335	0.319	0.342	0.259	0.251	0.267	0.328	0.264	0.095

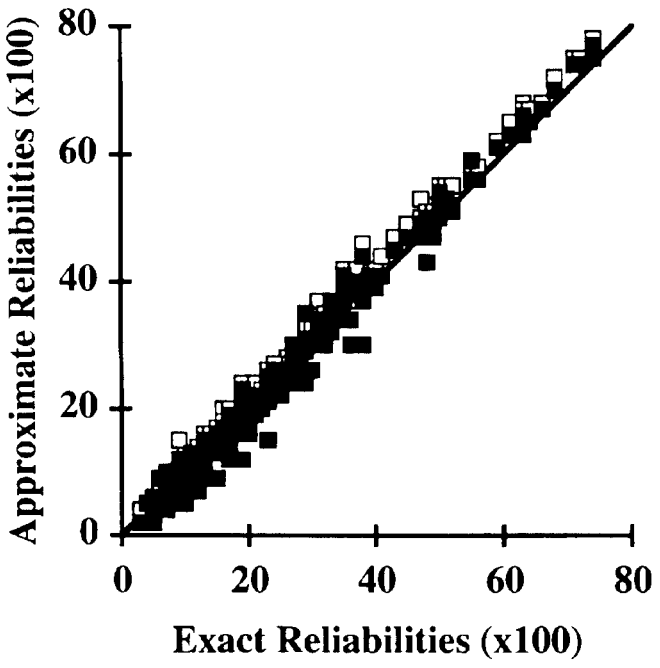

Figure 1. Plot of exact versus approximate reliabilities for milk yield obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of cows with records.

Figures 1, 2, and 3 show plots of approximate versus exact reliabilities for sires of cows with records for milk yield traits. For most sires, the multitrait approximation overestimated the reliabilities, and the correction has reduced that overestimation. The underestimation occurred only for lower repeatability sires (<55%), suggesting that the approximation does not affect sires with more offspring. Except for milk yield, for which certain values were slightly overcorrected, those corrected values were close to the exact reliabilities. Figures 4, 5, and 6 show plots of approximates versus exact reliabilities for sires of animals with beef records. For birth weight and weaning weight, the multitrait approximation overestimated slightly the reliabilities for most animals, and the correction for missing values reduced that overestimation. The approximations were worse for two sires that had progeny in fewer herds than other sires and whose relationships were less complete. For certain animals, values were slightly overcorrected. For postweaning gain, multitrait approximation showed big estimation errors. The correction reduced these errors, but not as much as for the two other traits. A systematic bias existed for animals with high exact reliabilities, and approximate reliabilities were overestimated.

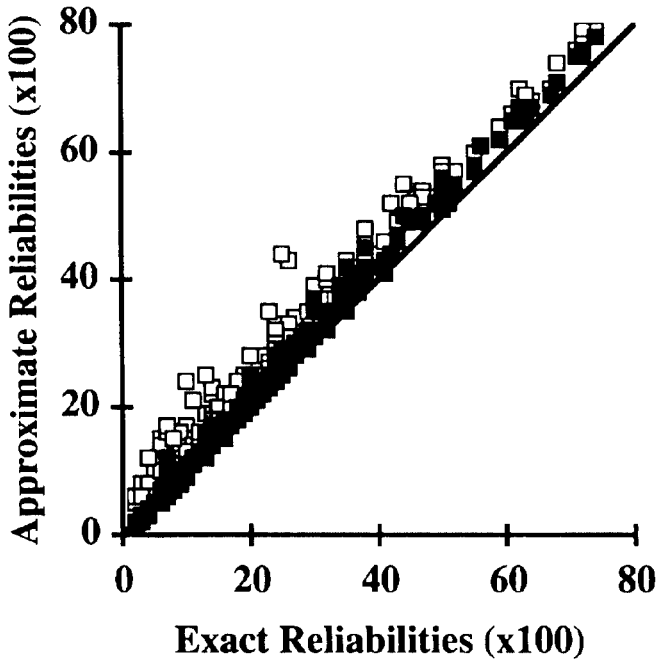

Figure 2. Plot of exact versus approximate reliabilities for fat yield obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of cows with records.

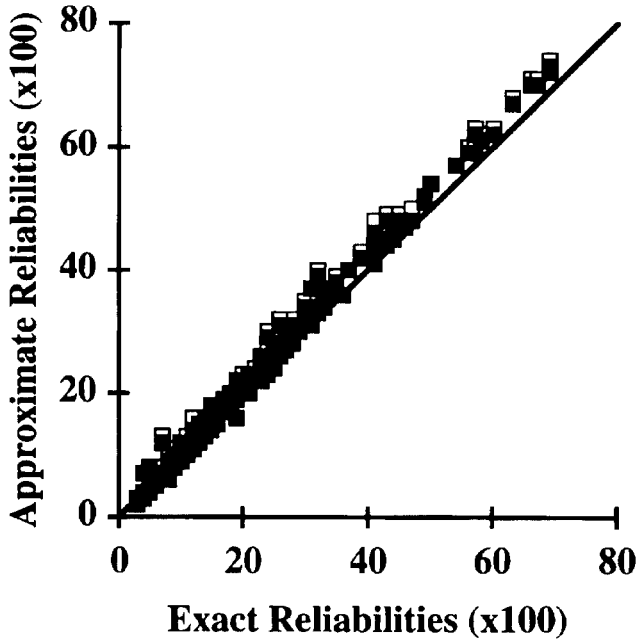

Figure 3. Plot of exact versus approximate reliabilities for protein yield obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of cows with records.

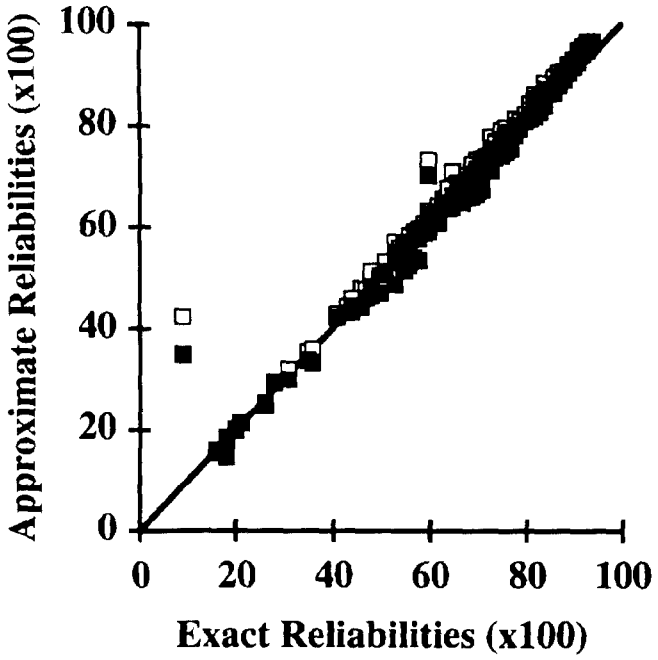

Figure 4. Plot of exact versus approximate reliabilities for birth weight obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of animals with records.

CONCLUSIONS

The accuracy of the algorithm for approximating PEV or reliabilities by multiple diagonalization with a multiple-trait repeatability model with missing data decreased as the percentage of observations missing and the correlation between traits increased. For simulated data with 33% of records missing and correlations ≤50% between traits, correlations of approximate reliabilities with those obtained by direct inversion were >0.97, means of estimation errors were <0.01, and standard deviations of estimation errors were <0.032. The algorithm worked best for sires of animals with records.

If large percentages of data were missing and concentrated in certain levels of fixed effects, and correlations between traits were high, reliabilities obtained with the multiple-trait model and multiple diagonalization were inaccurate. For the field data for dairy cattle, half-sib families were small, and parents were not known for all animals. For simulated data with the same (co)variances and similar patterns of missing observations, correlations (not shown) between approximate and exact reliabilities were higher than for the field data. Accuracy of the algorithm may be

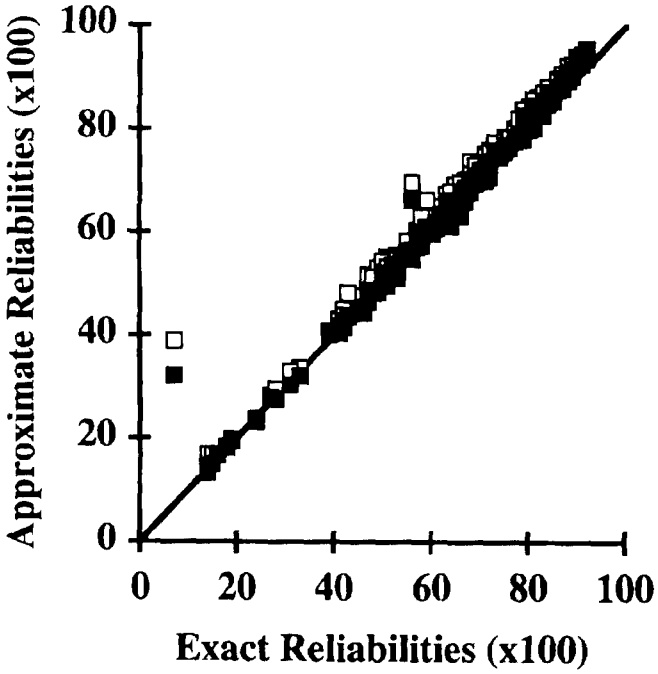

Figure 5. Plot of exact versus approximate reliabilities for weaning weight obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of animals with records.

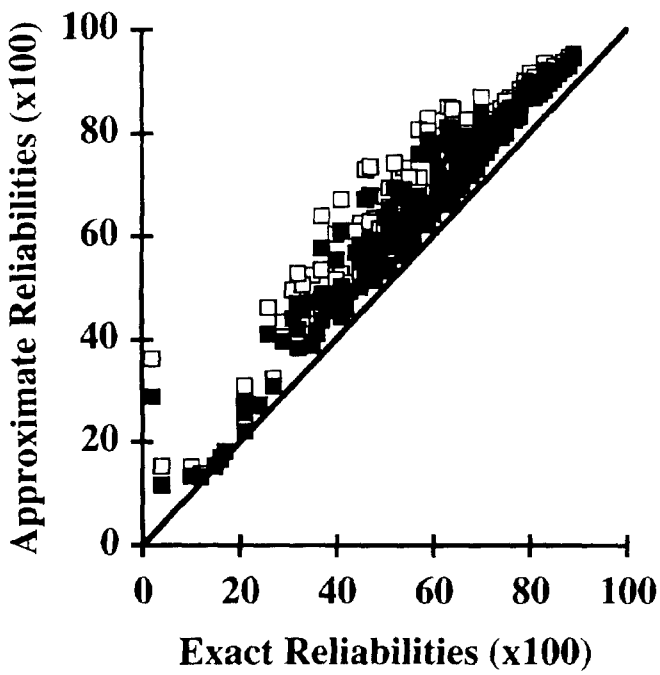

Figure 6. Plot of exact versus approximate reliabilities for post-weaning gain obtained by multiple diagonalization with (■) and without (□) correction for missing data for sires of animals with records.

greater for larger data files with more complete pedigree information.

This algorithm is a first approach to the problem of calculating reliabilities using multiple diagonalization with large-scale, multiple-trait animal models with missing data. Possible extensions would be to calculate the effective number of observations per trait as described and then 1) to develop multiple-trait algorithms that are equivalent to the single-trait algorithm proposed by Misztal and Wiggans (13) or 2) to calculate single-trait PEV that are combined to approximate multiple-trait PEV.

ACKNOWLEDGMENTS

Suggestions of L. D. Van Vleck, USDA-ARS, Clay Center, Nebraska and P. Dagnelie, Faculté Universitaire des Sciences Agronomiques (retired), Gembloux, Belgium are appreciated. Partial funding was provided by the Fonds National Belge de la Recherche Scientifique, Brussels, Belgium and Holstein Association USA, Inc., Brattleboro, Vermont. N. Gengler is Aspirant of the Fonds National Belge de la Recherche Scientifique. The authors thank Suzanne Hubbard, USDA-AIPL, for her help in preparing the manuscript.

REFERENCES

1 Boldman, K. G., L. A. Kriese, L. D. Van Vleck, and S. D. Kachman. 1993. A Manual for the Use of MTDFREML: A Set of Programs to Obtain Estimates of Variances and Covariances. USDA, Clay Center, NE.

2 Chu, E., A. George, J. Liu, and E. Ng. 1984. SPARSPAK: Waterloo Sparse Matrix Package User's Guide for SPARSPAK-A (CS-84-36). Dep. Comput. Sci., Univ. Waterloo, Waterloo, ON, Canada.

3 Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. 39:1.

4 Ducrocq, V., and B. Besbes. 1993. Solution of multiple trait animal models with mising data on some traits. J. Anim. Breed. Genet. 110:81.

5 Flury, B. N., and G. Constantine. 1985. Algorithm AF211: the F-G diagonalization algorithm. Appl. Stat. 34:177.

6 Greenhalgh, S. A., R. L. Quaas, and L. D. Van Vleck. 1986. Approximating prediction error variances for multiple trait sire evaluations. J. Dairy Sci. 69:2877.

7 Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. Univ. Guelph, Guelph, ON, Canada.

8 Lin, C. Y., and S. P. Smith. 1990. Transformation of multitrait to unitrait mixed model analysis of data with multiple random effects. J. Dairy Sci. 73:2494.

9 Little, R.J.A., and D. B. Rubin. 1987. Statistical Analysis with Missing Data. John Wiley & Sons, New York, NY.

10 Meyer, K. 1989. Approximate accuracy of genetic evaluation under an animal model. Livest. Prod. Sci. 21:87.

11 Misztal, I., T. J. Lawlor, and T. H. Short. 1993. Implementation of single- and multiple-trait animal models for genetic evaluation of Holstein type traits. J. Dairy Sci. 76:1421.

12 Misztal, I., K. Weigel, and T. J. Lawlor. 1995. Approximation of estimates of (co)variance components with multiple-trait restricted maximum likelihood by multiple diagonalization for more than one random effect. J. Dairy Sci. 78:1862.

13 Misztal, I., and G. R. Wiggans. 1988. Approximation of prediction error variance in large-scale animal models. J. Dairy Sci. 71(Suppl. 2):27.

14 Misztal, I., G. R. Wiggans, T. J. Lawlor, and T. H. Short. 1991. Continuous genetic evaluation of dairy cattle. J. Dairy Sci. 74: 2001.

15 Schaeffer, L. R., and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575.