Genet. Sel. Evol. 39 (2007) 481–494
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:20070016

Available online at:
www.gse-journal.org

Original article

Factor analysis models for structuring covariance matrices of additive genetic effects: a Bayesian implementation

Gustavo de los CAMPOS[a*], Daniel GIANOLA[a,b,c]

[a] Department of Animal Sciences, University of Wisconsin-Madison, WI 53706, USA
[b] Department of Dairy Science and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI 53706, USA
[c] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, 1432 Ås, Norway

(Received 5 January 2006; accepted 28 March 2007)

Abstract – Multivariate linear models are increasingly important in quantitative genetics. In high dimensional specifications, factor analysis (FA) may provide an avenue for structuring (co)variance matrices, thus reducing the number of parameters needed for describing (co)dispersion. We describe how FA can be used to model genetic effects in the context of a multivariate linear mixed model. An orthogonal common factor structure is used to model genetic effects under Gaussian assumption, so that the marginal likelihood is multivariate normal with a structured genetic (co)variance matrix. Under standard prior assumptions, all fully conditional distributions have closed form, and samples from the joint posterior distribution can be obtained via Gibbs sampling. The model and the algorithm developed for its Bayesian implementation were used to describe five repeated records of milk yield in dairy cattle, and a one common FA model was compared with a standard multiple trait model. The Bayesian Information Criterion favored the FA model.

factor analysis / mixed model / (co)variance structures

1. INTRODUCTION

Multivariate mixed models are used in quantitative genetics to describe, for example, several traits measured on an individual [6–8], or a longitudinal series of measurements of a trait, *e.g.*, [23], or observations on the same trait in different environments [19]. A natural question is whether multivariate observations should be regarded as different traits or as repeated measures of the same response variable. The answer is provided by a formal model comparison. However, it is common to model each measure as a different trait,

* Corresponding author: gdeloscampos@wisc.edu

leading to a fairly large number of estimates of genetic correlations [7, 8, 19]. A justification for this is that the multiple-trait model is a more general specification, with the repeated measures (repeatability) model being a special case. However, individual genetic correlations differing from unity is not a sufficient condition for considering each measure as a different trait. While none of the genetic correlations may be equal to one, the vector of additive genetic values may be approximated reasonably well by a linear combination of a smaller number of random variables, or common factors.

Another approach to multiple-trait analysis is to redefine the original records, so as to reduce dimension. For example, [25] suggested collapsing records on several diseases into simpler binary responses (*e.g.*, "metabolic diseases", "reproductive diseases", "diseases in early lactation"). Likewise, for continuous characters, one may construct composite functions that are linear combinations of original traits. However, when records are collapsed into composites, some of the information provided by the data is lost. For instance, consider traits X and Y. If $X + Y$ is analyzed as a single trait, information on the (co)variance between X and Y is lost.

Somewhere in between, is the procedure of using a multivariate technique such as principal components or factor analysis (PCA and FA, respectively), for either reducing the dimension of the vector of genetic effects (PCA) or for obtaining a more parsimonious model without reducing dimension (FA). Early uses of FA described multivariate phenotypes, *e.g.*, [21, 24]. PCA and FA have been used in quantitative genetics [1, 3, 5, 11], and most applications consist of two steps. One approach, *e.g.*, [3], consists of reducing the number of traits first, followed by fitting a quantitative genetic model to some common factors or principal components. In the first step, a transformation matrix (matrix of loadings) is obtained either by fitting a FA model to phenotypic records or by decomposing an estimate of the phenotypic (co)variance matrix into principal components. These loadings are used to transform the original records to a lower dimension. In the second step, a quantitative genetic model is fitted to the transformed data. Another approach fits a multiple trait model in the first step [1, 11], leading to an estimate of the genetic (co)variance matrix, with each measure treated as a different trait. In the second step, PCA or FA is performed on the estimated genetic (co)variance matrix. However, as discussed by Kirkpatrick and Meyer [10] and Meyer and Kirkpatrick [15], two-step approaches have weaknesses, and it is theoretically more appealing to fit the model to the data in a single step.

This article discusses the use of FA as a way of modeling genetic effects. The paper is organized as follows: first, a multivariate mixed model with an

embedded FA structure is presented, and all fully conditional distributions required for a Bayesian implementation via Gibbs sampling are derived. Subsequently, an application involving a data set on cows with five repeated records of milk yield each is presented, to illustrate the concept. Finally, a discussion of possible extensions of the model is given in the concluding section.

2. A COMMON FACTOR MODEL FOR CORRELATED GENETIC EFFECTS

In a standard FA model, a vector of random variables ($\mathbf{u}$) is described as a linear combination of fewer unobservable random variables called common factors ($\mathbf{f}$), *e.g.*, [12,13,16]. The model equation for the i^{th} subject when q common factors are considered for modeling the p observed variables can be written as,

$$\begin{pmatrix} u_{1i} \\ . \\ . \\ . \\ u_{pi} \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1q} \\ . & . & . \\ . & . & . \\ . & . & . \\ \lambda_{p1} & \dots & \lambda_{pq} \end{pmatrix} \begin{pmatrix} f_{1i} \\ . \\ . \\ . \\ f_{qi} \end{pmatrix} + \begin{pmatrix} \delta_{1i} \\ . \\ . \\ . \\ \delta_{pi} \end{pmatrix},$$

or, in compact notation,

$$\mathbf{u}_i = \mathbf{\Lambda}\mathbf{f}_i + \boldsymbol{\delta}_i. \tag{1}$$

Above, $\mathbf{u}_i = \left(u_{1i}, ..., u_{pi}\right)'$; $\mathbf{\Lambda} = \left\{\lambda_{jk}\right\}$ is the $p \times q$ matrix of factor loadings; $\mathbf{f}_i = \left(f_{1i}, ..., f_{qi}\right)'$ is the $q \times 1$ vector of common factors peculiar to individual i, and $\boldsymbol{\delta}_i = \left(\delta_{1i}, ..., \delta_{pi}\right)'$ is a vector of trait-specific factors peculiar to i. From (1) the equation for the entire data can be written as,

$$\mathbf{u} = (\mathbf{I}_n \otimes \mathbf{\Lambda})\,\mathbf{f} + \boldsymbol{\delta}, \tag{2}$$

where $\mathbf{u} = (\mathbf{u}_1', ..., \mathbf{u}_n')$, $\mathbf{f} = (\mathbf{f}_1', ..., \mathbf{f}_n')$, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1', ..., \boldsymbol{\delta}_n')$.

Equation (1) can be seen as a multivariate multiple regression model where both the random factor scores and the incidence matrix ($\mathbf{\Lambda}$) are unobservable. Because of this, the standard assumption required for identification in the linear model, *i.e.*, $\boldsymbol{\delta}_i \perp \mathbf{f}_i$, is not enough. To see that, following [16], let $\mathbf{H}$ be any non-singular matrix of appropriate order, and form the expression $\mathbf{\Lambda f} = \mathbf{\Lambda A H H^{-1} f} = \mathbf{\Lambda}^*\mathbf{f}^*$, where $\mathbf{\Lambda}^* = \mathbf{\Lambda H}$ and $\mathbf{f}^* = \mathbf{H^{-1}f}$. This implies that (1) can also be written as $\mathbf{u}_i = \mathbf{\Lambda}^*\mathbf{f}_i^* + \boldsymbol{\delta}_i$ so that neither $\mathbf{\Lambda}^*$ nor $\mathbf{f}^*$ are unique. In the orthogonal factor model this identification problem is solved by assuming that common factors are mutually uncorrelated. However, even with this assumption, factors are determined up to an orthonormal transformation only. To

verify this, following [16], let $\mathbf{T}$ be an orthonormal matrix such that $\mathbf{T}'\mathbf{T} = \mathbf{I}$. Then, from (1), $\text{Cov}(\mathbf{u}_i) = \boldsymbol{\Sigma}_u = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\mathbf{T}'\mathbf{T}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi} = \text{Cov}(\boldsymbol{\delta}_i)$ and $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}'$. This means that, to attain identification, factor loadings need to be rotated in an arbitrary q-dimensional direction. The restrictions discussed above are arbitrary and not based on substantive knowledge; because of this, the method is particularly useful for exploratory analysis [9, 12, 13].

In addition to the restrictions described above, maximum likelihood or Bayesian inference necessitate distributional assumptions. The standard probability assumption for a Gaussian model with orthogonal factors is

$$\begin{pmatrix} \mathbf{f}_i \\ \boldsymbol{\delta}_i \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{pmatrix}\right], \tag{3}$$

where "*iid*" stands for "independent and identically distributed", and $\boldsymbol{\Psi}$, of order $p \times p$, is assumed to be a diagonal matrix. Combining (1) and (3), the marginal distribution of $\mathbf{u}_i$ is,

$$\mathbf{u}_i \overset{iid}{\sim} N\left[\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}\right]. \tag{4}$$

Consider now a standard multivariate additive genetic model for p traits measured on each of n subjects

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{y}_i = \left(y_{i1}, ..., y_{ip}\right)'$, is a $p \times 1$ vector of phenotypic measures taken on subject i ($i = 1, ..., n$); $\boldsymbol{\beta}$ and $\mathbf{u}_f$ are unknown vectors of regression coefficients and of additive genetic effects, respectively; $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known incidence matrices of appropriate order, and $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ vector of model residuals. Stacking the records of the n subjects, the equation for the entire data set is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\mathbf{y} = (\mathbf{y}_1', ..., \mathbf{y}_n')'$, $\mathbf{X} = (\mathbf{X}_1', ..., \mathbf{X}_n')'$, $\mathbf{Z} = \text{Diag}\{\mathbf{Z}_i\}$, $\mathbf{u} = (\mathbf{u}'_1, ..., \mathbf{u}'_n)'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', ..., \boldsymbol{\varepsilon}_n')'$. A standard probability assumption in quantitative genetics is,

$$\begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{u} \end{pmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix} \mathbf{I}_n \otimes \mathbf{R}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_n \otimes \mathbf{G}_0 \end{pmatrix}\right], \tag{6}$$

where $\mathbf{R}_0$ and $\mathbf{G}_0$ are each $p \times p$ (co)variance matrices of model residuals and of additive genetic effects, respectively, and $\mathbf{A}$ is the $n \times n$ additive relationship matrix.

Assume now that (2) holds for the vector of additive genetic effects in (5) so that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\left(\mathbf{I}_n \otimes \boldsymbol{\Lambda}\right)\mathbf{f} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{7}$$

where $\boldsymbol{\Lambda}$ is as before, and $\mathbf{f}$ and $\boldsymbol{\delta}$ are interpreted as vectors of common and specific additive genetic effects, respectively. Combining the assumptions of the orthogonal FA model described above with those of the additive genetic model leads to the joint distribution

$$\begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{f} \\ \boldsymbol{\delta} \end{pmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix} \mathbf{I}_n \otimes \mathbf{R}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_n \otimes \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_n \otimes \boldsymbol{\Psi} \end{pmatrix}\right], \tag{8}$$

where $\boldsymbol{\Psi}$ $(p \times p)$ is the (co)variance matrix of specific additive genetic effects, assumed to be diagonal, a stated earlier. Note that in (8), unlike in the standard FA model, *i.e.*, (3), different levels of common and specific factors are correlated due to genetic relationships. With these assumptions, the conditional distribution of the data, given $\boldsymbol{\beta}$, $\mathbf{u}$ and $\mathbf{R}_0$ is

$$\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}_0 \sim N\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I} \otimes \mathbf{R}_0\right]. \tag{9a}$$

Alternatively, using (2), one can write

$$\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}_0 = \mathbf{y}|\mathbf{f}, \boldsymbol{\delta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{R}_0 \sim N\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\left(\mathbf{I}_n \otimes \boldsymbol{\Lambda}\right)\mathbf{f} + \mathbf{Z}\boldsymbol{\delta}, \mathbf{I} \otimes \mathbf{R}_0\right]. \tag{9b}$$

2.1. Bayesian analysis and implementation

In a multivariate linear mixed model, a Bayesian implementation can be entirely based on Gibbs sampling because, under standard prior assumptions, the fully conditional posterior distributions of all unknowns have closed form, *e.g.*, [20]. It turns out that in the model defined by (7) and (8), and under prior assumptions to be described below, all fully conditional distributions have closed form, and a Bayesian analysis can be based on a Gibbs sampler as well. Next, the prior assumptions are described, and the fully conditional distributions required for a Bayesian implementation of our FA model via Gibbs sampling are presented.

2.1.1. Prior distribution

Let $\boldsymbol{\lambda} = \mathrm{Vec}\left(\boldsymbol{\Lambda}\right)$, and consider the following specification of the joint prior distribution (omitting the dependence on hyper-parameters, for ease of notation)

$$p\left(\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{R}_0, \boldsymbol{\Psi}\right) = p\left(\mathbf{u}|\boldsymbol{\lambda}, \boldsymbol{\Psi}\right) p\left(\boldsymbol{\beta}\right) p\left(\boldsymbol{\lambda}\right) p\left(\mathbf{R}_0\right) p\left(\boldsymbol{\Psi}\right). \tag{10}$$

The prior distribution of the genetic effects implied by (7) and (8) is $N[\mathbf{u}|\mathbf{0}, \mathbf{A} \otimes (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})]$, where the randomness of $\mathbf{u}$ is made explicit to the left of the conditioning bar. Next, assume bounded flat priors for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$; an inverted Wishart distribution for $\mathbf{R}_0$, with scale matrix $\mathbf{S}_{R0}$ and υ_R prior degrees of freedom, denoted as $IW_p(\mathbf{R}_0|\mathbf{S}_{R0}, \upsilon_R)$, and independent scale inverted chi-square distributions for each of the diagonal elements of $\boldsymbol{\Psi}$, denoted as $\chi^{-2}\left(\Psi_{jj}\middle|\upsilon_j, S_j\right)$, $j = 1,\ldots, p$. With these prior-assumptions, and using (9a) as sampling model, the joint posterior distribution is

$$\begin{aligned} p(\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{R}_0, \boldsymbol{\Psi}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \mathbf{R}_0)\, p(\mathbf{u}|\boldsymbol{\lambda}, \boldsymbol{\Psi})\, p(\boldsymbol{\beta})\, p(\boldsymbol{\lambda})\, p(\mathbf{R}_0)\, p(\boldsymbol{\Psi}) \\ &\propto N[\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I} \otimes \mathbf{R}_0]\, N[\mathbf{u}|\mathbf{0}, \mathbf{A} \otimes (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})]\, IW(\mathbf{R}_0|\mathbf{S}_{R0}, \upsilon_{R0}) \\ &\quad \times \prod_{j=1}^{p} \chi^{-2}\left(\Psi_{jj}|S_j, \upsilon_j\right). \end{aligned} \quad (11)$$

2.1.2. Fully conditional posterior distributions

In what follows, when deriving fully conditional distributions, use is made of many well-known results for the Bayesian multivariate linear mixed model; a detailed description of these results is in [20].

From (11), the joint fully conditional distribution of location effects is proportional to

$$p\left[(\boldsymbol{\beta}', \mathbf{u}')'|else\right] \propto N[\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I} \otimes \mathbf{R}_0]\, N[\mathbf{u}|\mathbf{0}, \mathbf{A} \otimes (\boldsymbol{\Lambda}\boldsymbol{\Lambda} + \boldsymbol{\Psi})],$$

where "else" denotes everything in the model that is not specified to the left of the conditioning bar (*i.e.*, data, hyper parameters and all other unknowns). The expression above is recognized as the kernel of the fully conditional distribution of location effects in a standard multivariate mixed model. Therefore, the fully conditional distribution of $(\boldsymbol{\beta}', \mathbf{u}')'$ is as in the standard multivariate mixed model, that is,

$$p\left[(\boldsymbol{\beta}', \mathbf{u}')'|else\right] = N\left[\hat{\mathbf{r}}_1, \mathbf{C}_1^{-1}\right], \quad (12)$$

where $\hat{\mathbf{r}}_1$ and $\mathbf{C}_1$ are the solution vector and coefficient matrix of the following standard mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{X} & \mathbf{X}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{Z} \\ \mathbf{Z}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{X} & \mathbf{Z}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{Z} + \mathbf{A}^{-1} \otimes (\boldsymbol{\Lambda}\boldsymbol{\Lambda} + \boldsymbol{\Psi})^{-1} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{bmatrix} \mathbf{X}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{y} \\ \mathbf{Z}'\left(\mathbf{I} \otimes \mathbf{R}_0^{-1}\right)\mathbf{y} \end{bmatrix}.$$

Similarly, from (11), the fully conditional distribution of the residual (co)variance matrix is proportional to

$$p(\mathbf{R}_0|else) \propto N[\mathbf{y}|\mathbf{X\beta} + \mathbf{Zu}, \mathbf{I} \otimes \mathbf{R}_0] IW(\mathbf{R}_0|\mathbf{S}_{R0}, v_{R0}),$$

which is the kernel of the fully conditional distribution of the residual (co)variance matrix in the standard multivariate mixed model. Thus,

$$p(\mathbf{R}_0|else) = IW(\mathbf{E}'\mathbf{E} + \mathbf{S}_{R0}, n + v_{R0}), \tag{13}$$

and $\mathbf{E} = \left(\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_p\right)$ is an $n \times p$ matrix, in which the column $\boldsymbol{\varepsilon}_j$ is an $n \times 1$ vector of residuals for trait j.

Consider now the fully conditional distribution of the parameters of the FA model. From (7), (8) and (11), the fully conditional distribution of the parameters of the FA model is proportional to

$$p(\mathbf{f}, \boldsymbol{\lambda}, \boldsymbol{\Psi}|else) \propto p(\mathbf{u}|\boldsymbol{\lambda}, \mathbf{f}, \boldsymbol{\Psi})\, p(\mathbf{f})\, p(\boldsymbol{\Psi})$$

$$\propto N[\mathbf{u}|(\mathbf{I}_n \otimes \boldsymbol{\Lambda})\mathbf{f}, \mathbf{A} \otimes \boldsymbol{\Psi}] N\left[\mathbf{f}|\mathbf{0}, \mathbf{A} \otimes \mathbf{I}_q\right] \prod_{j=1}^{p} \chi^{-2}\left(\Psi_{jj}|S_j, v_j\right) \tag{14a}$$

$$\propto N\left[\mathbf{u}|\left(\mathbf{F} \otimes \mathbf{I}_p\right)\boldsymbol{\lambda}, \mathbf{A} \otimes \boldsymbol{\Psi}\right] N\left[\mathbf{f}|\mathbf{0}, \mathbf{A} \otimes \mathbf{I}_q\right] \prod_{j=1}^{p} \chi^{-2}\left(\Psi_{jj}|S_j, v_j\right) \tag{14b}$$

where $\mathbf{F} = \left(\mathbf{f}_1, ..., \mathbf{f}_q\right)$ is a matrix of $n \times q$ common factor values. From (14a) the fully conditional distribution of the vector of common factors is proportional to,

$$p(\mathbf{f}|else) \propto N[\mathbf{u}|(\mathbf{I}_n \otimes \boldsymbol{\Lambda})\mathbf{f}, \mathbf{A} \otimes \boldsymbol{\Psi}] N\left[\mathbf{f}|\mathbf{0}, \mathbf{A} \otimes \mathbf{I}_q\right]$$

$$\propto \exp\left\{-\frac{1}{2}[\mathbf{u} - (\mathbf{I}_n \otimes \boldsymbol{\Lambda})\mathbf{f}]' \left[\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right][\mathbf{u} - (\mathbf{I}_n \otimes \boldsymbol{\Lambda})\mathbf{f}]\right\}$$

$$\times \exp\left\{-\frac{1}{2}\mathbf{f}'\left[\mathbf{A}^{-1} \otimes \mathbf{I}_q\right]\mathbf{f}\right\}.$$

This is the kernel of the fully conditional distribution in a Gaussian model of random effects, **f,** with incidence matrix $(\mathbf{I}_n \otimes \boldsymbol{\Lambda})$, **u** as "data", model residual (co)variance matrix $\mathbf{A} \otimes \boldsymbol{\Psi}$ and prior distribution of the random effects $N\left[\mathbf{f}|\mathbf{0}, \mathbf{A} \otimes \mathbf{I}_q\right]$. Therefore, the fully conditional distribution of the common factors is

$$p(\mathbf{f}|else) = N\left[\hat{\mathbf{f}}, \mathbf{C}_2^{-1}\right], \tag{15}$$

where $\hat{\mathbf{f}}$ and $\mathbf{C}_2$ are the solution vector and coefficient matrix, respectively, of the following mixed model equations:

$$\left[(\mathbf{I}_n \otimes \boldsymbol{\Lambda}')\left(\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right)(\mathbf{I}_n \otimes \boldsymbol{\Lambda}) + \mathbf{A}^{-1} \otimes \mathbf{I}_q\right]\hat{\mathbf{f}} = \left[(\mathbf{I}_n \otimes \boldsymbol{\Lambda}')\left(\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right)\mathbf{u}\right],$$

or,

$$\left[\mathbf{A}^{-1} \otimes \left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right) + \mathbf{A}^{-1} \otimes \mathbf{I}_q\right]\hat{\mathbf{f}} = \left(\mathbf{A}^{-1} \otimes \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\right)\mathbf{u}.$$

Similarly, from (14b), the fully conditional distribution of the vector of factor loadings $\boldsymbol{\lambda}$ is proportional to

$$\begin{aligned} p(\boldsymbol{\lambda}|else) &\propto N\left[\mathbf{u}|\left(\mathbf{F} \otimes \mathbf{I}_p\right)\boldsymbol{\lambda}, \mathbf{A} \otimes \boldsymbol{\Psi}\right] \\ &\propto \exp\left\{-\frac{1}{2}\left[\mathbf{u} - \left(\mathbf{F} \otimes \mathbf{I}_p\right)\boldsymbol{\lambda}\right]'\left[\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right]\left[\mathbf{u} - \left(\mathbf{F} \otimes \mathbf{I}_p\right)\boldsymbol{\lambda}\right]\right\}, \end{aligned}$$

which is the kernel of the fully conditional distribution in a Gaussian model of "fixed" effects $\boldsymbol{\lambda}$ with bounded flat priors; incidence matrix $\left(\mathbf{F} \otimes \mathbf{I}_p\right)$, residual (co)variance matrix $\mathbf{A} \otimes \boldsymbol{\Psi}$, and $\mathbf{u}$ as "data". Therefore, the fully conditional posterior distribution of the vector of factor loadings is the truncated multivariate normal process (truncation points are the bounds of the prior distribution of $\boldsymbol{\lambda}$)

$$p(\boldsymbol{\lambda}|else) \propto N\left[\hat{\lambda}, \mathbf{C}_3^{-1}\right], \tag{16}$$

where, $\hat{\boldsymbol{\lambda}}$ and $\mathbf{C}_3$ are the solution and coefficient matrix, respectively, of the linear system

$$\left[\left(\mathbf{F}' \otimes \mathbf{I}_p\right)\left(\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right)\left(\mathbf{F} \otimes \mathbf{I}_p\right)\right]\hat{\boldsymbol{\lambda}} = \left[\left(\mathbf{F}' \otimes \mathbf{I}_p\right)\left(\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right)\mathbf{u}\right],$$

or,

$$\left[\mathbf{F}'\mathbf{A}^{-1}\mathbf{F} \otimes \boldsymbol{\Psi}^{-1}\right]\hat{\boldsymbol{\lambda}} = \left[\mathbf{F}'\mathbf{A}^{-1} \otimes \boldsymbol{\Psi}^{-1}\right]\mathbf{u}.$$

Finally, from (15a), the fully conditional distribution of the variances of the specific factors is

$$\begin{aligned} p(\boldsymbol{\Psi}|else) &\propto N\left[\mathbf{u}|(\mathbf{I}_n \otimes \boldsymbol{\Lambda})\mathbf{f}, \mathbf{A}_n \otimes \boldsymbol{\Psi}\right]\prod_{j=1}^{p}\chi^{-2}\left(\Psi_{jj}|S_j, v_j\right) \\ &= \prod_{j=1}^{p}N\left[\mathbf{u}_j\middle|\mathbf{F}\boldsymbol{\lambda}_j, \mathbf{A}\psi_j\right]\prod_{j=1}^{p}\chi^{-2}\left(\Psi_{jj}|S_j, v_j\right). \end{aligned} \tag{17}$$

Above, $\mathbf{u}_j$ and $\boldsymbol{\lambda}_j$ are the vector of random effects for the j^{th} trait and the j^{th} row of $\boldsymbol{\Lambda}$, respectively. Hence, the fully conditional posterior distributions

of the p diagonal elements of $\mathbf{\Psi}$ are scaled inverse chi-square, with posterior degree of belief $v_i' = n + v_i$, and posterior scale parameter $S'_j = \frac{\boldsymbol{\delta}_j' \mathbf{A}^{-1} \boldsymbol{\delta}_j + v_j S_j}{n + v_j}$. Here, $\boldsymbol{\delta}_j = \mathbf{u}_j - \mathbf{F}\boldsymbol{\lambda}_j$ is a vector of specific effects for the j^{th} trait.

The preceding developments imply that one can sample location parameters ($\boldsymbol{\beta}$ and $\mathbf{u}$) and the residual (co)variance matrix with a Gibbs sampler for the standard multivariate linear mixed model, with $\mathbf{G}_0 = \mathbf{\Lambda}\mathbf{\Lambda} + \mathbf{\Psi}$. Once $\mathbf{u}$ has been sampled, the parameters of the common factor model can be sampled using (15), (16) and (17). In practice, the Gibbs sampler can be implemented by sampling iteratively along the cycle:

- location parameters ($\mathbf{u}, \boldsymbol{\beta}$) using distribution (12),
- residual (co)variance matrix using distribution (13),
- vector of common factors using (15),
- vector of factor loadings using (16); if desired, rotate loadings, and,
- variances of the specific factors using (17).

3. FA OF GENETICS EFFECTS: APPLICATION TO REPEATED RECORDS OF MILK YIELD IN PRIMIPAROUS DAIRY COWS

The concepts are illustrated by fitting an FA model to data consisting of five repeated records of milk yield on each of a set of first lactation dairy cows. In particular, a one common factor structure is used to model the random effect of the sire on each of the five traits, and this model is compared with a multiple trait (MT) model. In a one common factor model for five traits, the (co)variance matrix of the sire effects is modeled using 10 parameters (5 loadings and 5 variances of the specific factors), that is, 9 more dispersion parameters that in a repeatability model, but 5 less parameters than in the standard MT model, *i.e.*, unstructured $\mathbf{G}_0$.

3.1. Data and methods

Data consisted of five repeated records of MY on 3827 first lactation daughters of 100 Norwegian red (NRF) sires having their first progeny test in 1991 and 1992. Only complete records (*i.e.*, five test day records) of cows with a first calving in 1990 through 1992, and from herds with at least five daughters of any of these bulls were included. Data was pre-adjusted with predictions of herd effects as described in [4]. First lactation was divided into five 60-day periods starting at calving. For each cow, a test-day record (the one closest to the mid-point of the period) was assigned to each period.

A standard multiple trait sire model for this data set is $MY_{ijk} = \mu_k + s_{ik} + \varepsilon_{ijk}$, where μ_k ($k = 1, ..., 5$) is a test-day-specific mean, s_{ik} is the effect of sire i on trait k, ($i = 1, ..., 100$), and ε_{ijk} is a residual specific to the k^{th} record of the j^{th} daughter ($j = 1, ..., n_i$) of sire i. The probability assumption was standard, as in (6), with **A** now being the additive relationship matrix due to sires and maternal grand sires.

A single common genetic factor model for this data specifies $s_{ik} = \lambda_k f_i + \delta_{ik}$, so that the equation for the k^{th} record on the j^{th} daughter of sire i is, $MY_{ijk} = \beta_k + \lambda_k f_i + \delta_{ik} + \ell_{ijk}$, with probability assumption as in (8), with $p = 5$ (number of traits), $q = 1$ (number of common factors), and $n = 100$ (number of sires).

The MT model was compared with the FA model using the Bayesian Information Criterion (BIC), computed as $BIC_{FA,MT} = -2\left(\bar{l}_{FA} - \bar{l}_{MT}\right) - 5\log(N)$, where $\bar{l}_{FA} - \bar{l}_{MT}$ is the difference between the average (across iterations of the Gibbs sampler) log-likelihoods of the FA and the MT model, respectively, 5 is the difference in number of parameters between the two models and $N = 3827$. A negative $BIC_{FA,MT}$ provides evidence in favor of the FA model.

Both models were fitted using a collection of R-functions [18] written by the senior author[1] that can be used for fitting multivariate linear mixed, and some R functions that were created to sample the unknowns of the FA structure. R-packages used by these function are: MASS [22], MCMCpack [14] and Matrix [2]. Post Gibbs analysis was performed using the coda package of R [17].

3.2. Results

Posterior means of the log-likelihoods were −19 706.57 and −19 696.85 for the FA and MT models, respectively, indicating that both models had similar "fit". The $BIC_{FA,MT}$ was −21.81, indicating that the data favored the FA model over the MT model.

Table I shows posterior summaries for test-day means. Posterior means and posterior standard deviations were similar for both models, and this is expected because the FA model imposes no restriction on the mean vector. Table II shows posterior summaries for the vector of loadings and the variances of the specific factors in the FA model. The posterior mean of loadings increased from the first lactation period (0.751) to the second lactation period (0.984) and decreased thereafter. The sire variances of the specific factors were all small; those for test-days 1 and 5 were the largest. The relative importance of specific and common factors can be assessed by evaluating the proportion

[1] These functions are available by request.

Table I. Summaries of the posterior distributions of test-day means for each of the models fitted.

Parameter[1]	Multiple trait model		Common genetic factor model	
	Mean[2]	SD	Mean[2]	SD
μ_1	21.46	0.1621	21.47	0.1609
μ_2	21.40	0.1879	21.39	0.1945
μ_3	19.60	0.1767	19.58	0.1825
μ_4	17.45	0.1775	17.44	0.1824
μ_5	14.14	0.1704	14.13	0.1750

[1] μ_k is the mean of the k^{th} trait.

[2] Time-series Monte Carlo standard errors were all < 0.0001.

Table II. Summaries of the posterior distributions of the loadings and of the variances of the specific factors in the common factor model.

Lactation period	Loadings on the common factor		Variance of the specific factor	
	Mean[1]	SD	Mean[1]	SD
1	0.751	0.0720	0.0797	0.0375
2	0.984	0.0749	0.0367	0.0149
3	0.921	0.0705	0.0318	0.0123
4	0.918	0.0714	0.0380	0.0155
5	0.815	0.0762	0.0997	0.0411

[1] Time-series Monte Carlo standard errors were all < 0.002.

of the sire variance due to common factors (called communality). In this case, the contribution of common factors to the sire variance on a trait is obtained by squaring the factor loading on the trait. Communality evaluated at the posterior means was high for all traits, ranging from around 0.88 (first and fifth lactation periods) to around 0.96 (lactation periods 2, 3 and 4).

Table III shows posterior summaries of the dispersion parameters. Estimates of the residual (co)variance components were similar between the two specifications; again, this is expected because the FA model imposes no structure on such parameters. However, for sire (co)variance components, some differences between estimates from the two specifications were observed. These differences arise because of the restrictions imposed by the FA specification. Another consequence of those restrictions is that posterior standard deviations were smaller in the FA specification.

Table III. Summaries of the posterior distributions of residual and sire (co)variance components.

Entry[1]	Residual (co)variance matrix				Sire (co)variance matrix			
	Multiple trait model		Common genetic factor model		Multiple trait model		Common genetic factor model	
	Mean[2]	SD	Mean[2]	SD	Mean[2]	SD	Mean[2]	SD
(1,1)	11.67	0.2705	11.69	0.2710	0.652	0.1533	0.649	0.1112
(1,2)	5.77	0.2082	5.78	0.2083	0.678	0.1551	0.743	0.1172
(1,3)	3.88	0.1878	3.86	0.1879	0.599	0.1391	0.695	0.1060
(1,4)	2.60	0.1817	2.58	0.1818	0.594	0.1372	0.692	0.1040
(1,5)	1.30	0.2068	1.27	0.2067	0.508	0.1278	0.614	0.0954
(2,2)	10.97	0.2543	10.96	0.2537	0.932	0.1951	1.011	0.1494
(2,3)	6.86	0.2050	6.85	0.2045	0.816	0.1718	0.911	0.1330
(2,4)	5.05	0.1903	5.04	0.1897	0.801	0.1680	0.907	0.1301
(2,5)	3.37	0.2069	3.36	0.2063	0.705	0.1569	0.805	0.1210
(3,3)	10.01	0.2313	10.01	0.2313	0.821	0.1716	0.886	0.1317
(3,4)	6.78	0.1978	6.78	0.1974	0.761	0.1623	0.850	0.1255
(3,5)	4.82	0.2061	4.82	0.2058	0.685	0.1525	0.754	0.1170
(4,4)	9.96	0.2305	9.96	0.2312	0.829	0.1744	0.885	0.1326
(4,5)	6.60	0.2187	6.60	0.2189	0.684	0.1564	0.752	0.1196
(5,5)	13.53	0.3120	13.54	0.3124	0.719	0.1699	0.770	0.1302

[1] Numbers between parentheses give the row and column of the element of the (co)variance matrix for which posterior summaries are provided.

[2] All time-series Monte Carlo standard errors < 0.001.

4. CONCLUDING REMARKS

Multivariate linear mixed models are increasingly important in animal breeding, because the number of traits included in genetic evaluation programs has increased steadily over time. When the number of traits is large, FA can provide a useful way of structuring (co)variance matrices without reducing dimensionality. A more parsimonious specification is expected to lead to smaller posterior standard deviations, but may show lack of fit. Since the FA model imposes restrictions on the parameterization of the standard multiple trait model, a natural benchmark for evaluating the goodness/lack of fit of the model is the multiple trait model. In the example presented here, the "lack of fit" of the FA model (mainly due to differences in estimates of the sire (co)variance components) was more than overcome by the parsimony of the specification.

Although we focused on the orthogonal common factor model, only minor modifications are needed for confirmatory factor analysis schemes, which may be of interest in some applications. The model presented here did not consider

the possibility of missing values. However, it can be shown that the fully conditional distribution of missing values of the model presented here is as in the standard multivariate linear mixed model (*e.g.* [20]). Similarly, the model can be easily extended to include generalized-linear models (*e.g.*, probit, multi-threshold, censored log-normal).

Finally, although we addressed the use of FA for modeling genetic effects only, one may consider using the same strategy for modeling permanent environmental random effects, or model residuals. Extension to such models does not pose special difficulties. Similar ideas may also be used in the context of random regression models, with random regression coefficients modeled as functions of common and specific factors.

ACKNOWLEDGEMENTS

The authors thank Professors Kent Weigel and Robert Hauser, and Drs. Bjørg Heringstad and Yu-Mei Chang for comments. Access to the data was given by the Norwegian Dairy Herd Recording System in agreement number 004.2005. Financial support from the Babcock Institute for International Dairy Research and Development, University of Wisconsin, Madison and by grants NRICGP/USDA 2003-35205-12833, NSF DEB-0089742, and NSF DMS-044371 is greatly appreciated. Constructive comments by two anonymous reviewers are greatly appreciated.

REFERENCES

[1] Atchley W., Rutledge J.J., Genetic components of size and shape, I. Dynamics of components of phenotypic variability and covariability during ontogeny in the laboratory rat, Evolution 34 (1980) 1161–1173.

[2] Bates D., Maechler M., Matrix: A Matrix package for R. R-project (2006), http://rh-mirror.linux.iastate.edu/CRAN/ [consulted: 6 October 2006].

[3] Chase K., Carrier D.R., Alder F.R., Jarvik T., Ostrander E.A., Lorentzen T.D., Lark K.G., Genetic basis for systems of skeletal quantitative traits: Principal component analysis of the canid skeleton, Proc. Natl. Acad. Sci. USA 99 (2002) 9930–9935.

[4] de los Campos G., Gianola D., Heringstad B., A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows, J. Dairy Sci. 89 (2006) 4445–4455.

[5] Hashiguchi S., Morishima H., Estimation of genetic contribution of principal components to individual variates concerned, Biometrics 25 (1969) 9–15.

[6] Hazel L.N., The genetic basis for constructing selection indexes, Genetics 28 (1943) 476–490.

[7] Heringstad B., Chang Y.M., Gianola D., Klemetsdal G., Multivariate threshold model analysis of clinical mastitis in multiparous Norwegian dairy cattle, J. Dairy Sci. 87 (2004) 3038–3046.
[8] Heringstad B., Chang Y.M., Gianola D., Klemetsdal G., Genetic analysis of clinical mastitis, milk fever, ketosis, and retained placenta in three lactations of Norwegian red cows, J. Dairy Sci. 88 (2005) 3273–3281.
[9] Johnson R.A., Wichern D.W., Applied multivariate statistical analysis, 5th edn., Prentice Hall, 2002.
[10] Kirkpatrick M., Meyer K., Direct estimation of genetic principal components: simplified analysis of complex phenotypes, Genetics 168 (2004) 2295–2306.
[11] Leclerc A., Fikse W.F., Ducrocq V., Principal components and factorial approaches for estimating genetic correlations in international sire evaluation, J. Dairy Sci. 88 (2005) 3306–3315.
[12] Manly Bryan F.J., Multivariate Statistical Methods. A primer, Chapman & Hall/CRC, 2005.
[13] Mardia K.V., Kent J.T., Bibby J.M., Multivariate analysis, 7th reprinting, Academic Press, 1979.
[14] Martin A.D., Quinn K.M., MCMCpack: Markov chain Monte Carlo (MCMC) package. R-project (2006), http://rh-mirror.linux.iastate.edu/CRAN/ [consulted: 6 October 2006].
[15] Meyer K., Kirkpatrick M., Restricted maximum likelihood estimation of genetic principal components and smoothed (co)variance matrices, Genet. Sel. Evol. 37 (2005) 1–30.
[16] Peña D., Análisis de datos multivariantes, Mc Graw Hill, 2002.
[17] Plummer M., Best N., Cowles K., Vines K., Coda: output analysis and diagnostics for MCMC. R-project (2006), http://rh-mirror.linux.iastate.edu/CRAN/ [consulted: 6 October 2006].
[18] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2006 [consulted: 6 October 2006].
[19] Schaeffer L.R., Multiple-country comparison of dairy sires, J. Dairy Sci. 77 (1994) 2671–2678.
[20] Sorensen D., Gianola D., Likelihood, Bayesian, and MCMC methods in quantitative genetics, Springer-Verlag, New York, 2002.
[21] Spearman C., General intelligence, objectively determined and measured, Amer. J. Psychol. 15 (1904) 201–293.
[22] Venables W.N., Ripley B.D., Modern applied statistics with S., 4th edn., Springer, New York, 2002.
[23] Wood P.D.P., Algebraic model of the lactation curve in cattle, Nature 216 (1967) 164–169.
[24] Wright S., On the nature of size factors, Genetics 3 (1918) 367–374.
[25] Zwald N.R., Weigel K.A., Chang Y.M., Welper R.D., Clay J.S., Genetic selection for health traits using producer-recorded data. II. Genetic correlations, disease probabilities, and relationships with existing traits, J. Dairy Sci. 87 (2004) 4295–4302.