Estimating Genetic Parameters in a Dominance Model that Includes Inbreeding

S.P. Smith and A. Mäki-Tanila[1]

Abstract. A dominance model is described coming with inbreeding, and five genetic parameter plus the environmental variance. The linear model is specified by the phenotypic equations and the mid-parent equations. They are placed in an indefinite system that is highly sparse, not the mixed model equations. From this system the augmented matrix **K** is built that is symmetric and indefinite, leading to restricted maximum likelihood. Likelihood evaluation follows from the factorization of **K**, and various sparse matrix tools are described for maximizing the likelihood. The method is used to estimate the six parameters for 2706 egg-laying hens that were part of a selection experiment.

1. Introduction

When there is no inbreeding, much progress has been made with methods for estimating additive, dominance and environmental variances (e.g., Misztal 1997). With inbreeding the situation becomes very challenging.

Gillois (1964) and Harris (1964) worked out the mathematics to define the genetic covariances among inbred relatives. Harris's details are very complicated and may involve 28 genetic parameters[2] to represent all first and second-order gene interactions (i.e., involving two loci) and lesser numbers depending on simplifying stipulations. In the case of no epistasis at all, there are five genetic parameters to modulate additive genetic and dominance effects. In this case, the environmental variance brings the parameter count to six, and these must be estimated from data.

Hoeschele and Vollema (1993) were among the first to actually estimate the five genetic parameters in dairy cows using a method of moments, and they then found restricted maximum likelihood (REML) unfeasible for a large-scale animal model. Estimating variance components by restricted maximum likelihood was proposed by Patterson and

[1] Department of Agricultural Sciences, University of Helsinki, Finland.

[2] Gallais (2003) counts 21 parameters.

Thompson (1971). To define the log-likelihood function that is suitable for REML, consider the following linear model[3].

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

where **y** is a column vector of observations, **b** a vector of fixed effect, and in this case **e** are the residuals that include all the random genetic effects. The variance matrix of the residuals is given below.

$$\text{Var}(\mathbf{e}) = \mathbf{V}$$

One likelihood function that is suitable for REML is the following,

$$\text{Log-likelihood} = -\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| - \tfrac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y}$$

where $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. There are other formulations of the likelihood function that involve matrix components taken from the mixed model equations (e.g., see Smith 1995), which has been recommended particularly if sparse matrix methods are to be employed. However, its not obvious that for the dominance and inbreeding model the methods developed for sparse matrices are most suitable. Moreover, sparse matrix methods may not be used even if the alternative formulation involving the mixed model equations is used. Fernández, Legarra, Martinez, Sánchez and Baselga (2017) note that necessary inverse-variances matrices tend to be non-sparse, making it difficult to use the mixed model equations when the number of observations or the order of matrices becomes large. If the sparse matrix approach is to be followed stickily, that possibly implies that the order of matrices are to be expanded drastically (beyond the number of observations) while complex expressions are utilized to provide the inverse dominance relationship matrix (e.g., Smith and Mäki-Tanila, 1990), and these adjustments may also be prohibitive.

Therefore, it is not surprising that Abney, Peek and Ober (2000) described REML estimation for a dominance and inbreeding model, while making no special reference to sparse matrices that are beyond the matrix **V** that has an order limited to the number of observations. Likewise, Shaw and Woolliams (1999) describe REML by limiting their treatment to dense matrix operations involving **V**. If the order of **V** is not too great, these approaches are feasible. For large matrices, however, vector processing can be employed before having to consider using an alterative formulation that involves sparse-matrix manipulation.

Sparse matrix methods will be evaluated again for the dominance and inbreeding model that is presented in Section (2), with renewed interests. This will again involve a large

[3] This formulation is different to model (6) that occurs later because of how the random genetic effects are placed in **e** rather than **b**.

expansion of genetic effects in the model, well beyond the order of **V**. Restricted maximum likelihood is described in Section (3), but rather than following the mixed model equations, Siegel's (1965) equations are employed because they have superior sparse-matrix handling properties. A sample data set involving egg-laying hens is described in Section (3), and the resulting estimates of the genetic parameters are presented in Section (4) after they are successfully estimated by the new method.

2. Dominance Model

The dominance model with inbreeding[4], but with no epistasis, is indicated for the k-th phenotype:

$$y_k=\alpha+\beta\times F_k + a_{i(k)} + a_{j(k)} + d_{i(k),j(k)} + e_k \qquad (1)$$

where y_k is the k-th phenotype, α is the overall mean or intercept, F_k is the inbreeding coefficient for the k-th zygote, β is a regression indicating inbreeding depression (when negative), i(k) and j(k) are index functions that assign addresses for two gametes that unite to form the k-th zygote, a_m is the additive genetic effect summed over loci that is associated with the m-th gamete, d_{mn} is the dominance effect summed over loci that is derived with the union of the m-th and n-th gametes, and e_k is the environmental error associated with the k-th phenotype.

The variances of e_k are uncorrelated, and independent of k, and collectively denoted by:

$$Var(e_k) = \sigma_e^2$$

To describe the remaining five parameters, and to introduce gametic recursions, it is useful to introduce vectors, $\mathbf{a}_m$ and $\mathbf{d}_{mn}$, representing the genetic effects parsed into loci. If there are L loci, and **1** is an L×1 vector of 1s, then $a_m=\mathbf{1}^T\mathbf{a}_m$ and $d_{nm}=\mathbf{1}^T\mathbf{d}_{mn}$. Therefore, in populations in Hardy-Weinberg and linkage equilibrium, and without inbreeding, the additive genetic and dominance variance are defined below.

$$Var(a_m) = \sigma_a^2 = E\left(\mathbf{a}_m{}^T\mathbf{a}_m\right)$$

$$Var(d_{nm}) = \sigma_d^2 = E\left(\mathbf{d}_{nm}{}^T\mathbf{d}_{nm}\right), \text{ where } n \neq m$$

[4] Those readers that correctly see an anomaly in redundantly having an effect for inbreeding in a model that already includes a dominance term, please suspend judgment for now. This situation will be explained shortly, and (1) is to be used as the traditional starting model.

It is understood that the mean vectors of $\mathbf{a}_m$ and $\mathbf{d}_{mn}$ are both **0**, the zero vector. That is,

$$E(\mathbf{a}_m) = E(\mathbf{d}_{dm}) = \mathbf{0}$$

These definitions follow from the fact that the variance of a sum of independent random variables is the sum of the variances (as is also needed below), which is a statement that the loci are segregating independently.

When there is inbreeding, three additional parameters are found:

$$Cov(a_m, d_{mm}) = \sigma_{ad} = E\left(\mathbf{a}_m{}^T \mathbf{d}_{mm}\right)$$

$$u_d^2 = E(\mathbf{d}_{mm})^T E(\mathbf{d}_{mm})$$

$$Var(d_{mm}) = \sigma_d^2 = E\left(\mathbf{d}_{mm}{}^T \mathbf{d}_{mm}\right) - E(\mathbf{d}_{mm})^T E(\mathbf{d}_{mm})$$

The five genetic parameters can be used to evaluate the covariance between any two related zygotes or the variance of each, using the path coefficient method (Jacquard 1966) or a tabular method (Smith and Mäki-Tanila, 1990), but the parameters are used at the end of the calculation of identity coefficients (or probabilities) that do not change with the parameters (de Boer and Hoeschele,1993). The tabular method does not actually require a table for storing all like coefficients for gametes and gamete pairs that are encountered during the gametic recursion. If storing such a table is prohibited, identity coefficients that serve as intermediates can be computed on the fly (rather than saved for reuse) by a recursive function that progresses through a list of zygotes and zygote pairs. Provided the pedigree is not too deep, this shortcut is not too demanding on computing time. While it is very useful to conceptualize such a table, particular if a sparse-matrix method is to be used, actual identify coefficients can be restricted to sub-matrices, or blocks, when gametic recursion is applied (e.g., Smith and Mäki-Tanila, 1989). The present paper will take up where Smith and Mäki-Tanila (1989) left off with: with a large and sparse matrix the same dimension as needed for the conceptualized tabular method, and with the same recursive functions that were then developed[5], but by-passing Sections F and G because the inverse table is no longer needed.

To define the conceptualized table, or matrix, where gametic recursions are applied, first number all the gametes in the pedigree and in the zygotes that have records,

[5] Before work started again 28 years later, the prior software was debugged by comparing the previous calculations with newly programmed recursive functions. This uncovered three small errors in the prior computer code, which were corrected. If the prior software is to be used again by others, kindly understand that the prior software needs correcting in three places.

consecutively from 1 to G by a partial ordering where j<i if the i-th gametes descended from the j-th gamete, or if the j-th gamete is from the base population and the i-th gamete is a non-base gamete. Next define the L×(i+1) matrix $\mathbf{H}_i$, and the L×½G(G+3) matrix **H**, as follows.

$$\mathbf{H}_i = \{\mathbf{a}_i, \mathbf{d}_{1i}, \mathbf{d}_{2i}, \ldots, \mathbf{d}_{ii}\}$$

$$\mathbf{H} = \{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_G\}$$

The matrix of second moments, summed over loci, is neatly provided by $E(\mathbf{H}^T\mathbf{H})$ which is of order G(G+3)×½G(G+3) and way too large if closure under gametic recursion[6] is the only requirement that must be met. Let $\hat{\mathbf{H}}$ be a subsequence of the columns of **H** so constructed to be closed under gametic recursion, and let $\hat{\mathbf{H}}_i$ be the corresponding column subsequence of $\mathbf{H}_i$. Appendix A of Smith and Mäki-Tanila (1990) provides a depth-first search algorithm that computes the subsequence, starting with a complete list of zygotes and tracing back through the genetic paths using the available pedigree information.

Something must be done to connect the columns of $\hat{\mathbf{H}}$ with the collection of genetic effects, the additive genetic and dominance effects. This is easy because the row vector $\mathbf{g}=\mathbf{1}^T\hat{\mathbf{H}}$ is such a correspondence where the various effects in the phenotypic equations (1), for all k, are all located in **g**.

Define the row vectors $\mathbf{g}_i=\mathbf{1}^T\hat{\mathbf{H}}_i$, for i=1 to G. Additional model specifications are needed that describe the relationships among the various $\mathbf{g}_i$ while incorporating the known pedigree information.

Regarding the first moments, $E(\mathbf{g}_i)=\mathbf{0}$ (a row vector of 0's) when the i-th gamete belongs to the base population[7], even when the particular expectation involves $\mathbf{1}^T\mathbf{d}_{ii}$. Because the phenotypic equation (1) already incorporates the inbreeding depression, coherence with the equivalent model (2) demands a subtraction of β off the homozygotic dominance effects in (2) to permit a translation from (2) to (1).

$$y_k=\mu+ a_{i(k)} + a_{j(k)} + \underline{d}_{i(k),j(k)} + e_k \qquad (2)$$

---

[6] If $\mathbf{a}_i$, $\mathbf{d}_{ji}$ where j<i, and $\mathbf{d}_{ii}$ are columns in $\hat{\mathbf{H}}$ and the parents of the i-th gamete are x and y, then closure implies that $\mathbf{a}_x$, $\mathbf{d}_{jx}$ or $\mathbf{d}_{xj}$, $\mathbf{d}_{xx}$, $\mathbf{a}_y$, $\mathbf{d}_{jy}$ or $\mathbf{d}_{yj}$, and $\mathbf{d}_{yy}$ are columns in $\hat{\mathbf{H}}$.

[7] Actually, the zero expectation are found for all gametes where model (1) applies, but further declarations of that fact are not needed because of the introduction of mid-parent equations.

Note that in (2) the underscore distinguishes the dominance effects that are potentially impacted by the reformulation. In particular, $\underline{d}_{ij}=d_{ij} + \beta \times F_{ij}$ where F_{ij} is the inbreeding coefficient for the zygote[8] formed by the union of the i-th and j-th gametes. With $E(\underline{\mathbf{d}}_{ii})=\underline{\mathbf{u}}$, then $\mathbf{1}^T\underline{\mathbf{u}}=\beta$ and this specification can be made with a linear model outside of (2), i.e., in equation (3) below. However, returning to the model that follows from (1) where $E(\mathbf{d}_{ii})=\mathbf{u}$, then $\mathbf{1}^T\mathbf{u}=\mathbf{1}^T\underline{\mathbf{u}} - \beta=0$. To build (1) from (2) its not necessary to distribute the subtraction over the various loci, and this action is technically invalid[9] because it generates a subtle change to u_δ^2. Nevertheless, to maintain expediency with notation thereby letting us stipulate that $E(\mathbf{g}_i)=\mathbf{0}$, it is handy to apply the subtraction to loci while ignoring the impact on u_δ^2. The sensible way to make the distribution is to define $\mathbf{u}=\underline{\mathbf{u}} - (\beta/L)\mathbf{1}$, i.e., to subtract a constant off the effect from each locus in the homozygotic condition.

When $\mathbf{g}_i$ belongs to the base population, then the variance-covariance matrix for $\mathbf{g}_i$ is diagonal (with one σ_a^2, perhaps several σ_d^2 and no more than one σ_δ^2), except for the possibility of one element ($\sigma_{a\delta}$) that is off diagonal. When $\mathbf{g}_j$ also belongs to the base population, then the covariance matrix between $\mathbf{g}_i$ and $\mathbf{g}_j$ has every element set to zero. Nowhere does the parameter u_δ^2 enter into the covariance structure for base population gametes.

Partition the entire row vector $\mathbf{g}$ as $\mathbf{g}=[\mathbf{g}_B, \mathbf{g}_N]$, where $\mathbf{g}_B$ corresponds to all the base population gametes and $\mathbf{g}_N$ the non-base population gametes. To summarize, additional information for the base population gametes is needed and is specified as the following,

$$E(\mathbf{g}_B)=\mathbf{r} \quad \text{and} \quad Var(\mathbf{g}_B)=\mathbf{G}_B \tag{3}$$

where $\mathbf{G}_B$ is an almost diagonal variance-covariance matrix with the occasional off diagonal element that is set to $\sigma_{a\delta}$, as previously noted. When model (1) applies then $\mathbf{r}=\mathbf{0}$, but when (2) applies the entries of $\mathbf{r}$ are set to zero except for an occasional β corresponding to an expectation of the form $\mathbf{1}^TE(\underline{\mathbf{d}}_{ii})$. Because further specifications for $\mathbf{g}_N$ are by mid-parent equations (Smith and Mäki-Tanila, 1990), there is no further need for treating β, other than calculating F_k as required only for model (1). The mid-parent equations are identical for models (1) and (2), and come with residual terms representing random segregation effects that are uncorrelated with $\mathbf{g}_B$. Once the mid-parent model is presented the model specifications are complete.

Consider the column vector $\mathbf{g}_i=\mathbf{1}^T\hat{H}_i$ again where $\mathbf{g}_i$ is a sub-vector of $\mathbf{g}_N$. Denote the parents of the i-th gamete by x and y, which are by definition part of the pedigree information where index numbers are consistent with the partial ordering previously

---

[8] This could be a real life zygote, or one that is merely theoretical.

[9] Because model (2) is the coherent starting place despite the traditional use of a regression on F to depict inbreeding depression as in (1), the parameter u_δ^2 is understood to be $\underline{\mathbf{u}}^T\underline{\mathbf{u}}$, and alternative interpretations cannot be supported.

determined. From the larger matrix $\hat{H}$ define the two matrices, $\hat{H}_{i=x}$ and $\hat{H}_{i=y}$, to be a symbolic transformations of $\hat{H}_i$ where the index-x or the index-y, respectively, substitute for the index-i. If $\mathbf{a}_i$, $\mathbf{d}_{ij}$ and $\mathbf{d}_{ii}$ are columns of $\hat{H}_i$, then $\mathbf{a}_x$, $\mathbf{d}_{xj}$ or $\mathbf{d}_{jx}$, and $\mathbf{d}_{xx}$ are columns of $\hat{H}_{i=x}$. Note specifically that $\hat{H}_{i=x}$ and $\hat{H}_{i=y}$ are not the same as $\hat{H}_x$ and $\hat{H}_y$.

The general mid-parent equation is the following.

$$\hat{H}_i = ½\, \hat{H}_{i=x} + ½\, \hat{H}_{i=y} + \mathbf{S}_i \qquad (4)$$

Where $\mathbf{S}_i$ represents the segregation residual that is realized when the i-th gamete was spawn from parent gametes x and y during meiosis. Post multiplying (4) by $\mathbf{1}^T$ gives equation (5), a linear model (as collection of column vectors) representing the mid-parent equation in terms of effects that have been enumerated for modeling.

$$\mathbf{g}_i = ½\, \mathbf{g}_{i=x} + ½\, \mathbf{g}_{i=y} + \mathbf{s}_i \qquad (5)$$

where $\mathbf{g}_{i=x} = \mathbf{1}^T \hat{H}_{i=x}$, $\mathbf{g}_{i=y} = \mathbf{1}^T \hat{H}_{i=y}$, and $\mathbf{s}_i = \mathbf{1}^T\mathbf{S}_i$.

The first moments of $\mathbf{S}_i$ and $\mathbf{s}_i$ are a matrix and column vector of zeros. This is derived below.

$$\mathbf{S}_i = \hat{H}_i - ½\, \hat{H}_{i=x} - ½\, \hat{H}_{i=y}$$

$$\begin{aligned} E(\mathbf{S}_i) &= E(\hat{H}_i - ½\, \hat{H}_{i=x} - ½\, \hat{H}_{i=y}) \\ &= E(\hat{H}_i) - ½\, E(\hat{H}_{i=x}) - ½\, E(\hat{H}_{i=y}) \\ &= ½\, E(\hat{H}_i \mid i=x) + ½\, E(\hat{H}_i \mid i=y) - ½\, E(\hat{H}_{i=x}) - ½\, E(\hat{H}_{i=y}) \\ &= ½\, E(\hat{H}_{i=x}) + ½\, E(\hat{H}_{i=y}) - ½\, E(\hat{H}_{i=x}) - ½\, E(\hat{H}_{i=y}) \\ &= \mathbf{0} \end{aligned}$$

And because $E(\mathbf{s}_i)=\mathbf{1}^T E(\mathbf{S}_i)$, then $E(\mathbf{s}_i)=\mathbf{0}$

Now because the first moments are zero, the collection of variances and covariances are just the corresponding second moments. Moreover, because the variance of a sum of independent random variables is the sum of the individual variances, the variance-covariance matrix reduces to the following.

$$\begin{aligned} Var(\mathbf{s}_i) &= E(\mathbf{S}_i^T\mathbf{S}_i) \\ &= ½E(\mathbf{S}_i^T\mathbf{S}_i \mid i=x) + ½E(\mathbf{S}_i^T\mathbf{S}_i \mid i=y) \\ &= ½E(\hat{S}_i^T\hat{S}_i \mid i=x) + ½E([-\hat{S}_i]^T[-\hat{S}_i] \mid i=y) \text{ , where } \hat{S}_i = ½\, E(\hat{H}_{i=x}) - ½\, E(\hat{H}_{i=y}) \\ &= E[\hat{S}_i^T\hat{S}_i] \end{aligned}$$

The very last specification is to commute the variance blocks, $\mathbf{B}_i=E[\hat{S}_i^T\hat{S}_i]$ for all gametes

that are in $\mathbf{g}_N$, and Smith and Mäki-Tanila (1989) developed software for doing just that.[10] These segregation effects for the i-th gamete are uncorrelated with $\mathbf{g}_B$ and other segregation effects for the j-th gamete when $i \neq j$, and therefore, the entire variance matrix for $\mathbf{g}_B$, and all segregation effects together, is block diagonal, with the blocks of $\mathbf{G}_B$ and $\mathbf{B}_i$. Moreover, these blocks are of order of modest size, if not mostly small, and tend to be sparse.

Even though $\text{Var}(\mathbf{s}_i) = E(\mathbf{S}_i^T\mathbf{S}_i) = E[\hat{\mathbf{S}}_i^T\hat{\mathbf{S}}_i]$, its important to note that $\mathbf{S}_i \neq \hat{\mathbf{S}}_i$, and in general, $E(\hat{\mathbf{S}}_i) \neq \mathbf{0}$ even if $E(\mathbf{S}_i) = \mathbf{0}$. Because of this we find that the parameter u_δ^2 can enter into the calculation of $\mathbf{B}_i$ even if its never used in $\mathbf{G}_B$.

3. Restricted Maximum Likelihood

A general linear model is described as follows:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \epsilon \qquad (6)$$

where $\mathbf{y}$ is a column vector of observations, or imputed observations[11], $\mathbf{b}$ is a column vector of effects in the model that have historically been called fixed but in this version of (6) some elements can now be called random, and ϵ is a column vector of residuals.

Lastly, define the variance-covariance matrix $\mathbf{V}$, where $\text{var}(\epsilon) = \mathbf{V}$. In this system, $\mathbf{V}$ is permitted to be singular or non-negative definite.

To make the correspondence between the dominance models of Section 2, and with (6), make the following assignments shown in Table 1.

---

[10] It is not necessary to pre-treat the model, and the subsequence extraction, to remove singularities that are known to exist within the genetic variance matrix as a whole. That prior treatment described in Smith and Mäki-Tanila (1989, 1990) is not followed.

[11] In what historically has been called a collection of "random effects" comes now with a second set of equations: the mid-parent equations. This takes the historical random effects and places them in $\mathbf{b}$, puts the residues (or the segregation effects from the mid-parent equations) in ϵ, and lastly imputes values for the corresponding elements of $\mathbf{y}$. The imputed values are generally a collection of zeros (or known quantities that come from model specifications). When fully specified, (6) represents both the phenotypic and mid-parent equations.

Table 1. Assignments of column vectors.					
Model (6)	**y**	notes	**b**	ϵ	notes
Models (1) & (5) or Models (2) & (5)	$\begin{bmatrix} \{y_k\} \\ \mathbf{0} \end{bmatrix}$	✓ ✗	$\begin{bmatrix} \alpha \\ \beta \\ \mathbf{g}_B^T \\ \mathbf{g}_N^T \end{bmatrix}$	$\begin{bmatrix} \{e_k\} \\ \mathbf{g}_B^T \\ \{\mathbf{s}_i^T\} \end{bmatrix}$	✓ ✓

✓ The function $\{\mathbf{v}_w\}$ returns a column vector, where $\mathbf{v}_w$ is a column vector, or scalar, and is stacked one on bottom of the other to build $\{\mathbf{v}_w\}$, as the index w varies from smallest to largest over its intended range.
✗ The column vector **0** contains zeros and has the same dimensions as $\mathbf{g}^T$.

The variance matrix, **V**, is block-diagonal (i.e., with zeros off the diagonal blocks) and has the following form,

$$\mathbf{V} = \begin{bmatrix} \sigma_e^2 \mathbf{I}_{N\times N} & & \\ & \mathbf{G}_B & \\ & & \text{Diag}\{\mathbf{B}_i\} \end{bmatrix}$$

where k varies between 1 and N observations, and the function Diag{$\mathbf{B}_i$} returns a block diagonal matrix with diagonals $\mathbf{B}_i$, as the index i varies from smallest to largest over its intended range.

The matrix **X** has the following form,

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{N\times 1} & \mathbf{x}_1 & \mathbf{Z}_1 & \mathbf{Z}_2 \\ \mathbf{0} & \mathbf{x}_2 & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{P}_1 & \mathbf{I}-\mathbf{P}_2 \end{bmatrix}$$

where for Model (1), $\mathbf{x}_1$ is a column vector of inbreeding coefficients in an order that matches the phenotypic equations, and $\mathbf{x}_2$=**0**; or for Model (2), $\mathbf{x}_1$=**0** and $\mathbf{x}_2$ contains mostly zeros except for occasional entries that equal 1 corresponding the homozygotic dominance effect that are matched with $\mathbf{g}_B^T$. The matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ contain ones and zeros and collect the genetic effects from $\mathbf{g}^T$ to match the phenotypic equations. The matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ contain zeros and ½'s, and collect all the parental effects that contribute to the mid-parent equations.

Estimates of **b**, denoted by $\hat{\mathbf{b}}$, are computed by solving Siegel's (1965) equations:

$$\begin{bmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}^{\mathsf{T}} & \end{bmatrix} \begin{bmatrix} \lambda \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \end{bmatrix} \qquad (7)$$

The empty space in system (7) is intended to be a collection of zeros, and **V** and **X** are also very sparse. Standard errors for various estimates are obtained as the square-root of corresponding negative elements of the inverse coefficient matrix of (7). At this point, all the novelty that went into the dominance model in Section (2), including the translation that involved (6), gives way to various matrix operations that can be standardized and optimized. The theoretical developments in this paper have been completed, and the paper now turns to computational ideas that have previously been developed; what remains is just plug and play[12].

Smith (2001b) described how to take the matrix elements of (7), and turn them into one coefficient matrix **K** that is presented below, and how to use **K** to calculate the log-likelihood function for restricted maximum likelihood (REML).

$$\mathbf{K}_{M\times M} = \begin{bmatrix} \mathbf{V} & \mathbf{X} & \mathbf{y} \\ \mathbf{X}^{T} & & \\ \mathbf{y}^{T} & & \end{bmatrix}$$

The M×M matrix **K** is symmetric and indefinite. Such matrices are not generally known to have a Cholesky decomposition, unlike the case with positive definite matrices. However, it is feasible to re-order the rows and columns of **K** in such a way to permit Cholesky's factorization (Smith 2001a), leading directly to likelihood evaluation. Let **Q** be a permutation matrix that permutes the rows and columns of **K** such that factorization can proceed as the following,

$\mathbf{LDL}^{\mathsf{T}}=\mathbf{QKQ}^{\mathsf{T}}$

[12] There is no need for **V** to be non-singular, there is no need for a rapid method to compute the inverse dominance matrix, there is no need for the mixed model equations. These historical approaches in the past only hobble the calculations. Today a sharper focus can be directed at sparse matrix methods applied to (7), or the **K** matrix that follows, which are incredibly sparse and simple by comparison with the mixed model equations. Moreover, **V** is a linear combination of the genetic parameters, and come with coefficients that are computed only once and saved for reuse with different genetic parameters.

where **L** is lower triangular, **D** is diagonal with diagonal elements +1 or -1, and **Q** is a permutation matrix that leaves the last row and column of **K** unchanged.

The REML likelihood is provided by the following,

$$\log - likelihood = -\frac{L_{MM}^{\ 2}}{2} - \sum_{i \in \Omega} \log |L_{ii}| \qquad (8)$$

where $\Omega=\{i: L_{ii}>0, i<M\}$.

There is a second preferred way to calculate the likelihood. It is useful to factor σ_e^2 out of part of the likelihood by expressing the remaining parameters as variance ratios. Computationally, **K** is evaluated by substituting the variance ratios for variances, as presented in Table 2.

Table 2. Variance ratios to substitute for variances when calculating **K**.					
$\sigma_a^2 \leftarrow \frac{\sigma_a^2}{\sigma_e^2}$	$\sigma_d^2 \leftarrow \frac{\sigma_d^2}{\sigma_e^2}$	$\sigma_\delta^2 \leftarrow \frac{\sigma_\delta^2}{\sigma_e^2}$	$\sigma_{a\delta} \leftarrow \frac{\sigma_{a\delta}}{\sigma_e^2}$	$u_\delta^2 \leftarrow \frac{u_\delta^2}{\sigma_e^2}$	$\sigma_e^2 \leftarrow 1$

Factorization then follows, while using the prior notation, $\mathbf{LDL}^T=\mathbf{QKQ}^T$. A closed form estimate for σ_e^2 becomes available with the variance ratios fixed, as follows.

$$\hat{\sigma}_e^2 = \frac{L_{MM}^{\ 2}}{df}$$

The degrees of freedom[13] are denoted by df, and in the present example with two fixed effects, df= N - 2. This estimate is substituted back into the likelihood to derive the concentrated likelihood (9) which depends on five variance ratios rather than the six variance components that impacted (8).

$$\log - likelihood = -df \log(L_{MM}) - \sum_{i \in \Omega} \log |L_{ii}| \qquad (9)$$

[13] This number can be counted during factorization following Fact 10 of Smith (2001a).

The first real big calculation involves finding the permutations that permit factorization. This is done in sparse matrix mode, using an expensive double-linked list to keep track of rows or columns that may be pulled forward in the pivot order while the outer-product form of the Cholesky decomposition is followed. Unlike the situation when **K** is positive definite, the initial factorization is not done symbolically. Rather, it is done with real floating point calculations that operate on the **K** that was built for a particular set of parameters. The results are saved symbolically, however, to build the sparse structure of **L**. Factorization of a new matrix, $\mathbf{LDL}^T=\mathbf{QKQ}^T$, can now be repeated using the same sparse structure but with a different set of parameters, and provided the new parameters are not too different from the starting set. The success of factorization is very likely[14] and dependent on the discovered permutations being robust in permitting factorizations that follow.

The concentrated likelihood (9) is maximized following a variety of potential techniques (e.g., Fletcher 1987), including derivative-free, conjugate gradient, or Netwon's methods. The derivative-free approach can be fashioned as a direction set search with evolving directions (Powell 1964), or by the popular simplex algorithm (Nelder and Mead, 1965). The conjugate gradient and Newton's methods require calculation of some or all of the first and second derivatives.

Likelihoods (8), and (9), are functions of **L**, which can be differentiated using forward or backward differentiation of the algorithm used to perform the Cholesky decomposition of an indefinite matrix. Smith (2001b) used the outer-product form, and provides the backward derivatives. However, that account of differentiation has small errors that have been corrected by Smith, Nikolic and Smith (2012). Smith (2017a, b) describes the backward differentiation of the bordering method of the Cholesky decomposition, including for the Cholesky decomposition of an indefinite matrix. Murray (2916) describes the differentiation on the inner-product form of the Cholesky decomposition, but these results must be extended for indefinite matrices which is easy enough as found by Smith (2017b). Algorithms developed from the inner-product form and the bordering algorithm are likely to be superior[15] to those that come from the outer-product form.

4. Example Data

These methods were applied to data derived from a selection experiment involving eqq-laying hens at the former MTT Agrifood Research Finland (Luke Natural Resources

[14] Discovered empirically.

[15] Based on computing the Cholesky decomposition of the largest matrices that press the computer's limits in terms of memory and computing time, and this expectation being the case for sparse matrix manipulation (Ng and Peyton, 1993) and possibly vector processing.

Institute Finland), Jokioinen, Finland. Previously, the same data were used to illustrate the inverse calculation for the extended genomic table (Smith and Mäki-Tanila 1989), which is not needed for the present situation.

The selection experiment generated two divergent lines over four generations of selection, and the pedigree information contained as many as five generations back to the base population where inbreeding was taken as non-existent (see Table 3).

Table 3. Breakdown of 2706 Hens by amount of Pedigree Information

# Hens	Ancestor Generations Recorded
488	1
645	2
540	3
704	4
329	5

Of the total 2706 hens, only 686 were inbred coming with inbreeding coefficients that varied between 0.00195 and 0.125. The minimal inbreeding creates limits on information available to estimate some of the genetic parameters, and therefore, this data set serves as a good benchmark for testing the suitability of the methods with limited data sets. Challenges come to bare on numerical stability of the calculations and estimability (or definability) of the parameters.

5. Results

The symmetric M×M matrix **K** (defined in Section 3) that resulted for the hen data has order M=332,937, but it is very sparse starting with 672,690 non-zero elements (half-stored). With a minimum degree ordering[16] of rows and columns to permit the Cholesky factorization of the indefinite matrix ($\mathbf{K}=\mathbf{L}^{T}\mathbf{L}$), the number of non-zero elements in **L** grew to

[16] Among diagonals that are candidate pivots that are commuted dynamically and are found different from the operational zero. Diagonals that start out as zero must experience fill-in before they can be selected as pivots.

the huge number, 27,947,586. The minimum degree ordering required 40 hours of computing on a Windows machine that runs at 3.30 GHz with 8.0 GB of random access memory. The very modest data set, with minimal inbreeding, generates a very challenging example for testing the methods.

By comparison, likelihood evaluation only required 17 minutes of computing following the bordering method, and once the sparse structure of **L** is specified following the very expensive re-ordering[17] of rows and columns. A work array containing 27,947,586 double precision numbers representing **L** is required, and is stored inside random access memory. Equal amounts of memory are required for storing the sparse structure of **L** as two integer arrays that speed the application of the bordering method.

To calculate the gradient required the computation of a second array **F** the same size as **L**, and requires **L** to be available in the computer's memory. Because the memory demand was now beyond the limits of the home computer, the array **F** had to be treaded outside of random access memory using direct access reading and writing to disc. This slowed the gradient calculation way down, and required about 217 minutes per gradient vector.

To calculate five second derivatives from the 5×5 Hessian matrix requires the calculation of two more arrays, **Q** and **S**,[18] the same size as **L**. Both **Q** and **S** can be calculated separately. However, while **Q** is calculated and stored in random access memory the array **F** must be available by reading from disc. The array **S** is calculated in two parts. In the first part, array **S** is calculated and stored in random access memory while arrays **Q** and **F** are available by reading from disc. In the second part, **S** is treaded outside of random access memory using direct access reading and writing to disc, while array **L** is available in random access memory. Direct access reading and writing from disc slows the calculations way down. The calculation of **Q** required 238 minutes of computing. The calculation of **S** for part 1 required 252 minutes of computing, and part 2 used 217 minutes. Beyond the calculation of the likelihood and gradient vector, computing a row or column from the Hessian matrix required a total of 707 minutes, or 11.8 hours. Calculating the entire Hessian matrix required 58.9 hours.

Derivative free searches only require 17 minutes of computing per likelihood evaluation. By comparison, the conjugate-gradient method requires a minimum of 234 minutes per step, and up to 724 minutes per step if the second directional derivative[19] is computed to

---

[17] The re-ordering is none implicitly by defining when diagonals become pivots, rather than actually assigning new numbers that act as labels for rows and columns.

[18] The arrays **F**, **Q** and **S** are all defined in Smith (2017b). The array **S**, in particular, has no relation to the matrix **S** in Section 2.

[19] In the direction of the gradient, and using the same software for computing the arrays **Q** and **S**.

determine step size. One iteration of Newton's method uses 62 hours of computing.

The derivative free method was used to maximize the likelihood because it was much faster per step than the alternatives that require derivatives for calculations performed on the home computer. Days, weeks and months of computing time were used up testing the conjugate-gradient and Newton's methods on a preliminary data set that was only later found to be inaccurate and needed to be corrected. While the methods based on derivatives tended to move estimates in the right direction, i.e., to the maximum, stability issues were also encountered. Newton's method in particular, and to a lesser extent the conjugate-gradient method, can overshoot the maximum, and fail to converge, if the starting iterate is too far from the maximum. The present data set, coming with limited amounts of inbreeding, only exacerbated the known challenges facing both the Conjugate-gradient and Newton's methods. The likelihood was found very flat for two of the parameters. The variance estimate of σ_δ^2 tended to drift to zero, and to prevent a possible failure[20] in the Cholesky decomposition of $\mathbf{QKQ}^T$ it was eventually held fixed at the value 0.048. Moreover, the covariance parameter $\sigma_{a\delta}$ tended to drift to the minimum value permitted, corresponding to a negative correlation of -100%. It was possible to set the implicit correlation to -100% or +100% without creating a singularity that causes the Cholesky decomposition of $\mathbf{QKQ}^T$ to fail because the permutation matrix $\mathbf{Q}$ was found using an implicit correlation that was actually set to +100%.[21] Nevertheless, the implicit correlation was eventually held fixed at -99%. In any regard, with two parameters that tended to the edge of the parameter space, the first derivatives for those parameters will not push themselves to zero as they would if the maximum values were in the interior. This enabled possible overshooting. Furthermore, the negative Hessian matrix was found non-negative definite with rank three, implying that σ_δ^2 and $\sigma_{a\delta}$ are approximated by a linear combination of the remaining three parameters where the likelihood is not so flat. Lastly, the methods based on derivatives work well but only if the radius of convergence is bigger than the machine precision, and its doubtful if this condition was met given the aforementioned challenges.

---

[20] The Cholesky decomposition of an indefinite matrix is not known to be numerically stable compared to the Cholesky decomposition of a positive definite matrix. Therefore, the calculations should be monitored for stability, particularly to make sure the same set of non-zero pivots are encountered each time the Cholesky decomposition is calculated. The number of non-zero pivots, and the degrees of freedom that are calculated from counting them by following Fact 10 of Smith (2001a), should be checked for each evaluation. Moreover, the operational zero is tuned for counting pivots, and not set too small to avoid declaring a real zero pivot as non-zero. The primary goal is to calculate the likelihood consistently for different sets of genetic parameters. To improve numerical stability, the observations and inbreeding coefficients in (1) should first be centered before building $\mathbf{K}$.

[21] The evaluation of $\mathbf{Q}$ was not attempted by setting $\sigma_\delta^2=\sigma_{a\delta}=0$, but that might have been useful.

The derivative free approach worked well on the home computer, evaluating the likelihood function hundreds of times, with none of the stability problems noted with the other methods. However, the derivative free method did encounter stability issues caused by parameters being too far from the initial parameter set that was used to compute the permutation matrix **Q** where $\mathbf{QKQ}^T$ is subjected to factorization. Restricting the step size eliminated this problem, and for every successful likelihood evaluation the same non-zero pivots were found from round to round. Despite the present success with using the derivative free method, the other methods might still show better performance on a bigger computer, or with a data set that shows more inbreeding.

The REML estimates of the six parameters are presented in Table 4. The relative sizes of the standard errors are large for all the parameters, except for the estimate of σ_e^2. Therefore, the following observations are very guarded. The point estimate of σ_a^2 was huge relative to σ_e^2, corresponding to a heritability of $h^2=2\sigma_a^2/(2\sigma_a^2+\sigma_d^2+\sigma_e^2)=95.3\%$ in a non-inbred population. While the statistical error is such that its unlikely that heritability is as high as 95.3%, the data does imply that the heritability is high for egg-number. The point estimate of σ_d^2 actually looks very reasonable compared to σ_e^2. The point estimates of σ_δ^2 and $\sigma_{a\delta}$ are close to zero, and probably should be completely discounted due to the statistical error. There are theoretical reasons why σ_δ^2 and $\sigma_{a\delta}$ can tend to zero, or $|\beta|$ approach infinity, when the number of loci become arbitrarily large in models describing directional dominance (e.g., Robertson and Hill, 1983), but there is little to suggest this is occurring in the present analysis.

The big surprise is the size of the estimate of u_δ^2 relative to σ_e^2, being 263.2 times bigger. The impact that u_δ^2 has on the genetic variances is not large, however, because the associated coefficients in **V**, in (7), corresponding to u_δ^2 are found to be an order of magnitude smaller than most of the other coefficients.

Table 4. Estimates of Genetic Parameters for Egg Number Among Egg Laying Hens.		
Parameter [A]	Estimate	Standard Error [B]
$\frac{\sigma_a^2}{\sigma_e^2}$	12.21	26.28
$\frac{\sigma_d^2}{\sigma_e^2}$	0.208	0.5809
$\frac{\sigma_\delta^2}{\sigma_e^2}$	0.048	83.62 [C]
$\frac{\sigma_{a\delta}}{\sigma_e^2}$	-0.756	26.78 [C]
$\frac{u_\delta^2}{\sigma_e^2}$	263.2	532.6
σ_e^2	21.22	0.5730 [D]

A. The parameters in the various numerators (of variance ratios) are defined in Section 2. The environmental, or residual, variance for the phenotypic equation is denoted by σ_e^2.
B. Except where indicated, the standard errors are estimated using the Hessian matrix computed for the concentrated log-likelihood with σ_e^2 removed.
C. With the negative Hessian matrix non-negative definite and rank 3, the standard error was approximated as the root of the negative reciprocal of the respective diagonal element of the Hessian matrix.
D. Approximated from the second derivative of the log-likelihood with respect to σ_e^2 while holding all the variance ratios fixed at their estimated values.

Estimates for two of the fixed effects in Model (1) are presented in Table 5. The estimate of β should be negative to reflect an expected inbreeding depression, but here its estimated as a positive 14.50. However, the estimate comes with a large standard error of 39.37, and therefore its not significant.

Table 5. Estimates of Fixed Effects.		
Parameter	Estimate	Standard Error
α: Intercept	113.02	1.53
β: Slope on F	14.24	41.62

Considering the size of β=14.24 in Table 5, a comparison with the size of $u_\delta^2=\sigma_e^2\times263.2$ in Table 4 is possible, albeit using a very naive method. If there are L loci, then f=β/L is the average effect coming from one locus. If there is no locus to locus variance, then $Lf^2= u_\delta^2$ or $\beta^2/L=u_\delta^2$ or $L= \beta^2/u_\delta^2 =14.24\times14.24/(21.22 \times263.2)=0.0363$. By this simple comparison, the estimate of u_δ^2 was found much larger than β, because it should be that $L\geq1$.

6. Conclusion

Twenty eight years ago the inverse genomic table calculation became theoretically feasible (Smith and Mäki-Tanila, 1989), but the complexity of using the mixed model equations with such a matrix made practical applications difficult at best. It remains too big of a price to pay for the convenience offered by sparse matrices. While the REML calculations remain extensive, new sparse-matrix approaches involving Siegel's (1965) equations, and the corresponding **K** matrix, did lead to calculations that gave practical results on the home computer. Moreover, better methods are available for calculating derivatives of the log-likelihood, even if they were found unhelpful in maximizing the particular likelihood function in the present application. While the example data set is small, coming with large standard errors on some of the genetic parameter, so is the home computer. Larger data sets will improved the standard errors, and bigger and faster computers can replace the home computer. The advances that came as hardware and software have now made it possible to calculate the REML estimates for a dominance model that comes with inbreeding, or at lease an expectation now exists on what it takes to solve these hard problems.

The methods described in this paper involving the indefinite **K** matrix, and the derivative calculations of the log-likelihood, can possible find application with much less ambitions models. Whenever there is a possible depiction of mid-parent equations, even when the inbreeding is treated as non-existent as a simplifying assumption (e.g., Hoeschele and Van Raden, 1991; Van Raden and Hoeschele, 1991), perhaps new applications can be found.

Before ending this paper a few speculative comments about the feasibility of extending the above approach for some examples of epistasis. Even though this method is becoming more feasible when there is inbreeding, but with no epistasis as the present paper is attempting to demonstrate, broader success seems a long way off even if some extensions are theocratically feasible for 2nd order interactions involving locus pairs. A 2nd order system can be specified separate comparted to the 1st order system with L loci. Theoretically, there will be ½L(L-1) pairs with L loci, turning matrices with L rows into

matrices with ½L(L-1) rows. A much bigger complication comes with gametic recursion because here alleles from four different gametes may need to be considered in combination, further inflating the length of the subsequence needed to represent gametic recursion within a large matrix, further leading into over- parameterization, and otherwise making the calculations impossible.[22] In any regard, there are theoretical expectations that imply that epistatic variance is small relative to additive genetic variance (Hill, Goddard and Visscher 2008; Mäki-Tanila and Hill 2014), and therefore, further extensions may be unnecessary.

References

Abney, M., Mc Peek, M.S., and C. Ober, 2000, Estimation of variance components of quantitative traits in inbred populations, *Science Direct*, 66 (2), 629-650.

de Boer, I.J., and I. Hoeschele, 1993, Genetic evaluation methods for populations with dominance and inbreeding, *Theoretical and Applied Genetics*, 86 (2-3), 245-258.

Fernández, E.N., A. Legarra, R. Martinez, J.P. Sánchez and M. Baselga, 2017, Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler equivalent model, *Journal of Animal Breeding and Genetics*, 134, 184-195.

Fletcher, R., 1987, *Practical Methods of Optimization*, Second Edition, John Wiley & Sons, New York.

Gallais, A., 2003, *Quantitative Genetics and Breeding Methods in Autopoplypoids Plants*, INRA Editions, Paris, page 257.

Gillois, M. 1964, La relation d'identité en génétique, Thesis, Faculty of Sciences, Paris, pp 205.

Hill, W.G., M.E. Goddard, P.M. Visscher, 2008, Data and theory point to mainly additive genetic variance for complex traits, PLOS Genetics, 4(2), 1-10.

Hoeschele, I., and P.M. Van Raden, 1991, Rapid inversion of dominance relationship matrices for non-inbred populations by including sire by dam subclass effects, *Journal of Dairy Science*, 74, 557-569.

Hoeschele, I., and A.R. Vollema, 1993, Estimation of variance components with dominance and inbreeding in dairy cattle, Journal of *Animal Breeding and Genetics*, 110, 93-104.

[22] This is not to say that higher order identity coefficients cannot be evaluated efficiently through gametic recursion, they can.

Harris, D.L., 1964, Genotypic covariances between inbred relatives, *Genetics* 50, 1319-1348.

Jacquard, A., 1966, Logique du calcul des coefficients d'intdetité entre deux individus, Population, 21, 751-776.

Mäki-Tanila, A, and W.G. Hill, 2014, Influence of gene interaction of complex trait variation with multilocus models, *Genetics*, 198, 355-367.

Misztal, I.,1997, Estimation of variance components with large-scale dominance models, *Journal of Dairy Science*, 80 (5), 965-974.

Murray, I. (2016), Differentiation of the Cholesky decomposition, arXiv archived.

Nelder, J.A., and R. Mead, 1965, A simplex method for function minimization, *The Computer Journal*, 7, 308-313.

Ng, E.G. and B.W. Peyton, 1993, Block spare Cholesky algorithms on advanced uniprocessor computers, *SIAM Journal of Scientific Computing*, 14, 1034-1055.

Patterson, H.D., and R. Thompson, 1971, Recover of inter-block information when block sizes are unequal, *Biometrika*. 58, 545-554,

Robertson, A., and W.G. Hill, 1983, Population and quantitative genetics of many linked loci in finite populations, Proceedings of the Royal Society of London B, 219, 253-264.

Powell, M.J.D., 1964, An efficient methods for finding the minimum without calculating derivatives, *The Computer Journal*, 7, 155-162.

Siegel, I.H., 1965, Deferment of computation in the method of least squares, *Mathematics of Computation*, 19 (90): 329-331.

Smith, J.R., M. Nikolic and S.P. Smith, 2012, Hunting the Higgs boson using the Cholesky decomposition of an indefinite matrix, memo, vixRa archived.

Smith, S.P. (1995), Differentiation of the Cholesky algorithm, *Journal of Computational and Graphical Statistics*, 4, 134-147.

Smith, S.P., 2001a, Factorability of symmetric matrices, *Linear Algebra and Its Application*, 335: 63-80.

Smith, S.P., 2001b, Likelihood-based analysis of linear state-space models using the Cholesky decomposition, *Journal of Computational and Graphical Statistics*, 10 (2): 350-369.

Smith, S.P., 2017a, The bordering method of the Cholesky decomposition and its backward differentiation, memo, vixRa archived.

Smith, S.P., 2017b, The backward differentiation of the bordering algorithm for an indefinite Cholesky factorization, memo, vixRa archived

Smith, S.P, and A. Mäki-Tanila, 1989, Inverting the extended genomic table: a case study, memo, Agricultural Research Centre, Jokioinen, Finland, viXra archived in 2018.

Smith, S.P, and A. Mäki-Tanila, 1990, Genotypic covariances matrices and their inverses for models allowing dominance and inbreeding, *Genetics Selection Evolution*, 22, 65-91.

Van Raden, P.M., and I. Hoeschele, 1991, Rapid inversion of additive by additive relationship matrices by including sire-dam combination effects, *Journal of Dairy Science*, 74, 570-579.