Computational techniques in animal breeding

by
Ignacy Misztal
University of Georgia
Athens, USA
## Contents

Preface .................................................................................................................. 5

Software engineering .......................................................................................... 7

Fortran .................................................................................................................. 9

**Fortran 90 - Basics** ....................................................................................... 10
  Compiling ........................................................................................................ 24
  Advanced features in Fortran 95 ..................................................................... 26
    General philosophy of Fortran 90 ............................................................... 37
  Rewriting programs in F77 to F95 ................................................................. 46

**Setting-up mixed model equations** ................................................................. 49
  Fixed 2-way model .......................................................................................... 49
    Extension to more than two effects ........................................................... 52
    Computer program ..................................................................................... 54
  Model with covariables .................................................................................. 56
    Nested covariable ....................................................................................... 58
    Computer program ..................................................................................... 59
  Multiple trait least squares .......................................................................... 61
    Missing traits .............................................................................................. 65
    Different models per trait .......................................................................... 66
    Example of ordering by models and by traits .......................................... 66
    Computer program ..................................................................................... 68

Mixed models ....................................................................................................... 72
  Popular contributions to random effects - single trait .................................. 72
  Numerator relationship matrix $A$ .................................................................. 73
  Rules to create $A^{-1}$ directly ...................................................................... 74
  Unknown parent groups .................................................................................. 75
  Numerator relationship matrix in the sire model ......................................... 76
  Variance ratios .............................................................................................. 78
  Covariances between effects ........................................................................ 80
  Random effects in multiple traits ................................................................... 80
    Multiple traits and correlated effects ......................................................... 81
    Computer pseudo-code .............................................................................. 83

**Storing and Solving Mixed Model Equations** ............................................... 88
  Storage of sparse matrices .......................................................................... 89
    Ordered or unordered triples ...................................................................... 89
    IJA .............................................................................................................. 89
    Linked lists ................................................................................................ 90
    Trees .......................................................................................................... 91
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summing MME</td>
<td>92</td>
</tr>
<tr>
<td>Searching algorithms</td>
<td>93</td>
</tr>
<tr>
<td>Numerical methods to solve a linear system of equations</td>
<td>95</td>
</tr>
<tr>
<td>Gaussian elimination</td>
<td>95</td>
</tr>
<tr>
<td>LU decomposition (Cholesky factorization)</td>
<td>96</td>
</tr>
<tr>
<td>Storage</td>
<td>98</td>
</tr>
<tr>
<td>Triangular</td>
<td>98</td>
</tr>
<tr>
<td>Sparse</td>
<td>98</td>
</tr>
<tr>
<td>Sparse matrix inversion</td>
<td>100</td>
</tr>
<tr>
<td>Numerical accuracy</td>
<td>102</td>
</tr>
<tr>
<td>Iterative methods</td>
<td>103</td>
</tr>
<tr>
<td>Jacobi</td>
<td>104</td>
</tr>
<tr>
<td>Gauss-Seidel and SOR</td>
<td>104</td>
</tr>
<tr>
<td>Preconditioned conjugate gradient</td>
<td>106</td>
</tr>
<tr>
<td>Other methods</td>
<td>107</td>
</tr>
<tr>
<td>Convergence properties of iterative methods</td>
<td>107</td>
</tr>
<tr>
<td>Properties of the Jacobi iteration</td>
<td>112</td>
</tr>
<tr>
<td>Properties of the Gauss-Seidel and SOR iterations</td>
<td>112</td>
</tr>
<tr>
<td>Practical convergence issues</td>
<td>112</td>
</tr>
<tr>
<td>Determination of best iteration parameters</td>
<td>113</td>
</tr>
<tr>
<td>Strategies for iterative solutions</td>
<td>113</td>
</tr>
<tr>
<td>Solving by effects</td>
<td>113</td>
</tr>
<tr>
<td>Iteration on data (matrix-free iteration)</td>
<td>114</td>
</tr>
<tr>
<td>Indirect representation of mixed-model equations</td>
<td>117</td>
</tr>
<tr>
<td>Summary of iteration strategies</td>
<td>122</td>
</tr>
<tr>
<td>Variance components</td>
<td>123</td>
</tr>
<tr>
<td>REML</td>
<td>123</td>
</tr>
<tr>
<td>General Properties of Maximization Algorithms</td>
<td>129</td>
</tr>
<tr>
<td>Some popular maximization algorithms</td>
<td>132</td>
</tr>
<tr>
<td>Acceleration to EM (fixed point) estimates</td>
<td>135</td>
</tr>
<tr>
<td>Convergence properties of maximization algorithms in REML</td>
<td>138</td>
</tr>
<tr>
<td>Accuracy of the D and DF algorithms</td>
<td>140</td>
</tr>
<tr>
<td>Method R</td>
<td>143</td>
</tr>
<tr>
<td>Bayesian methods and Gibbs sampling</td>
<td>149</td>
</tr>
<tr>
<td>Using the Gibbs sampler</td>
<td>155</td>
</tr>
<tr>
<td>Multiple-trait</td>
<td>157</td>
</tr>
<tr>
<td>Convergence properties</td>
<td>158</td>
</tr>
<tr>
<td>Ways to reduce the cost of multiple traits</td>
<td>161</td>
</tr>
<tr>
<td>Canonical transformation</td>
<td>161</td>
</tr>
<tr>
<td>Canonical transformation and REML</td>
<td>163</td>
</tr>
<tr>
<td>Numerical properties of canonical transformation</td>
<td>164</td>
</tr>
</tbody>
</table>
Extensions in canonical transformation ........................................ 165
  Multiple random effects .................................................. 165
  Missing traits ............................................................. 166
  Different models per trait ............................................. 167
Implicit representation of general multiple-trait equations ............ 169

Data manipulation .................................................................................. 173
  Renumbering .............................................................................. 173
  Memory consideration .............................................................. 173
  Look-up tables ........................................................................... 174
  Sort and merge ........................................................................... 175
  Data manipulation options .......................................................... 182
    Unix utilities ............................................................................ 183

Sources and References ......................................................................... 185
Preface

These notes originated as class notes for a graduate-level course taught at the University of Georgia in 1997. The goal of that course was to train students to program new ideas in animal breeding. Programming was in Fortran because Fortran programs are relatively easy for numerical problems, and many existing packages in animal breeding that are still useful were written in Fortran. Since the computer science department at UGA does not teach courses in Fortran any longer, the first few lectures were devoted to Fortran 77 and then to Fortran 90. All programming examples and assignments were in Fortran 90. Because the students had only limited experience in programming, the emphasis was on generality and simplicity even at some loss of efficiency.

Programming examples from the course led to an idea of a general yet simple BLUP program that can be easily modified to test new methodologies in animal breeding. An integral part of that program was a sparse matrix module that allowed for easy but efficient programming of sparse matrix computations.

The resulting program was called BLUPF90 and supported general multiple trait models with missing traits, different models per trait, random regressions, separate pedigree files and dominance. An addition of a subroutine plus a minor change in the main program converted BLUPF90 to REMLF90 - an accelerated EM variance-component program, and BLUP90THR - a bivariate threshold-linear model. An addition of subroutines created a Gibbs sampling program GIBBSF90. I greatly appreciate collaborators who contributed to the project. Tomasz Strabel added a dense-matrix library. Shogo Tsuruta created extensions of in-memory program to iteration-on-data. Benoit Auvray added estimation of thresholds to the threshold model programs. Monchau Duanginda, Olga Ravagnolo, and Deuhwan Lee worked on adding threshold and threshold-linear model capability to GIBBSF90. Shogo Tsuruta and Tom Druet worked on extension to AI-REML. Tom Druet wrote an extension to Method R, which can work with large data sets.... Indirect contributions were made by Andres Legarra and Miguel Perez-Enciso...

Program development of BLUPF90 and the other programs was not painless. Almost every f90 compiler had bugs, and there were many small bugs in the programs. As Edison wrote, “the success is 1% inspiration and 99 perspiration.” Now the programs are already useful but there is room for improvements.

The programming work took time away from the notes, so the notes are far from complete. Not all ideas are illustrated by numerical examples or sample programs. Many ideas are too sketchy, and references are incomplete. Hopefully the notes will become better with time..... I am grateful to all who have reported errors in the notes, and particularly to Dr. Suzuki from Obihiro University and his group.

Developing good programs in animal breeding is a work for more than one person. There are many ideas around how to make them more comprehensive and easier to use. I hope that the
ideas presented in the notes and expressed in the programs will be helpful to many of you. In
return, please share your suggestions for improvement as well as programming developments
with me and the animal breeding community.

Ignacy Misztal
July 1998 - June 2007
Software engineering

In programming, we are interested in producing quality code in a short period of time. If that code is going to be used over and over again, and if improvements will be necessary, also clarity and ease of modification are important.

The following steps are identified in software engineering:

<table>
<thead>
<tr>
<th>Step</th>
<th>Purpose</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>Description of the problem to be solved</td>
<td>Variance component estimation for threshold animal models</td>
</tr>
<tr>
<td>Specifications</td>
<td>Specific description with detailed limits on what the program should do</td>
<td>Number of equations up to 200,000; number of effects up to 10; the following models are supported; line-mode code (no graphics); computer independent. Generation of test examples.</td>
</tr>
<tr>
<td>Design</td>
<td>Conceptual program (e.g., in pseudo-code) where all algorithms of the program have been identified</td>
<td>Estimates by accelerated EM-REML, mixed models set up by linked list, sparse matrix factorization. Pseudo-code (or flow diagram) follows. Possibly a program in very high-level, e.g., SAS IML or Matlab (prototyping)</td>
</tr>
<tr>
<td>Coding</td>
<td>Writing code in a programming language</td>
<td>Implementation in Fortran 90</td>
</tr>
<tr>
<td>Testing</td>
<td>Assurance that the program meets the specifications</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>Assurance that the program meets the requirements</td>
<td></td>
</tr>
</tbody>
</table>

In commercial projects, each of these step is documented so that many people can work on the same project, and that departure of any single person won’t destroy the project.

For small programming projects, it appears that only the coding step is done. In fact, the other steps are done also conceptually, just not documented.
It is important that all of the above steps are understood and done well. Changes to earlier steps done late are very expensive. Jumping right into programming without deliberation on specifications and design is a recipe for disaster!

In research programs where many details are not known in advance, programming may take a long time, and final programs may contain many mistakes that are not discovered until after the paper has been accepted or a degree has been awarded.

One possible step towards simpler programming is by reduction of program complexity by using very high-level language. In a technique called prototyping, one would first write a program in a very high-level but possibly inefficient language (SAS, Matlab, R, AWK...). This assures that the problem has been understood and delivers executed examples, and in the end can cut drastically on coding and testing!

Another way of reducing the cost of programming is using “object-oriented” programming, where complicated operations are defined as easily as possible, and complicated (but efficient) code is “hidden” in application modules or classes. This way, one can approach the simplicity of a matrix language with an efficiency of Fortran or C. This approach, using Fortran 90, is followed in the notes and in project BLUPF90.

Literature

Computational Cost

Let \( n \) be a dimension of a problem to solve, for example a size of the matrix. If for large \( n \) cost \( \sim f(n) \) then computational complexity is \( O[f(n)] \)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O(n) )</td>
<td>linear</td>
<td>Reading data</td>
</tr>
<tr>
<td>( O[n\log(n)] )</td>
<td>log-linear</td>
<td>Sorting</td>
</tr>
<tr>
<td>( O(n^2) )</td>
<td>quadratic</td>
<td>Summing a matrix</td>
</tr>
<tr>
<td>( O(n^3) )</td>
<td>cubic</td>
<td>Matrix multiplication</td>
</tr>
<tr>
<td>( O(2^n) )</td>
<td>exponential</td>
<td>exhaustive search</td>
</tr>
</tbody>
</table>

Algorithms usually have separate complexity for memory and computations. Often a memory-efficient algorithm is inefficient in computations and vice versa. Desirable algorithms are linear or log-linear. A fast computer is generally not a substitute for a poor algorithm for serious computations.

Observation
Advances in computing speed are evenly due to progress in hardware and in algorithms.
For efficient implementation of mathematical algorithms, Fortran is still the language of choice. A few standards are coexisting:

- Fortran 77 (F77) with common extensions,
- Fortran 90 (F90),
- Fortran 95 (F95)

The standard for F90 includes all F77 but not necessarily all the extensions. While F90 adds many new features to F77, changes in F95 as compared to F90 are small although quite useful. Plenty of high-quality but often hard to use software is available for F77. On the other hand, programs written in F90/F95 are simpler to write and easier to debug. The best use of old F77 subroutines is to “encapsulate” them into much-easier to use but just efficient F90 subroutines.

The next chapter provides a very basic overview of F90. The subsequent chapter is partly a conversion guide from F77 to F90 and partly an introduction to more advanced features of F90 and F95. There is also a chapter on converting old programs in F77 to F95.

Only the most useful and “safe” features of F90 and F95 are covered. For a more detailed description of Fortran, read regular books on F90/95 programming.
Example program

! This is a simple example

program test
! program for testing some basic operations
implicit none
integer:: a,b,c
!
read*,a,b
c=a**b !exponentiation
print*,’a=’,a,’b=’,b,’a**b=’,c
! Case is not important
PRINT*,’a**b=’,C
end

Rules:
- program starts with optional statement “program”, is followed by declarations and ends with “end”
- case is not important except in character strings
- text after ! is ignored and treated as comment; comments are useful for documentation

Variables
Variables start with letters and can include letters, numbers and character “_”.
Examples: a, x, data1, first_month

Data types
integer 4 byte long, range approx. -2,000,000,000 - 2,000,000,000
Guaranteed range is 9 decimal digits

real 4 byte floating point with accuracy of about 7 digits and range of about
\(10^{-34} - 10^{34}\). Examples: 12.4563, .56e-10

real (r8) (r8 to be defined)
or
double precision
8 byte floating point with accuracy of 15 digits and range of about
\(10^{300} - 10^{100}\). Examples: 12.4563, .56d-10
logical - 4 bytes, values of .true. or .false.
character (n) - a character string n characters long. One byte = one character.
Examples: ‘abcdef’, “qef”, “we’ll”. Strings are either single ‘’ or double quotes “”. If a string is enclosed in double quotes, single quotes inside are treated as normal characters, and vice versa.

complex not used much in animal breeding; see textbooks

Initialization
Variables can optionally be initialized during declaration; without declarations, they hold undetermined values (some compilers initialize some variables to equivalents of integer 0; this should not be relied on).

Examples

```fortran
implicit none
integer::number_effect=5, data_pos,i=0,j
real::value,x,y,pi=3.1415
character (10)::filename='data.run2000'
logical::end_of_file, file_exists=.false.
```

If a line is too long, it can be broken with the & symbol:

```fortran
integer::one, two, three,&
        four, five
```

Warning:
An undeclared variable in F90 is treated as integer if its name starts with “i-n” and as real otherwise. A statement: ‘implicit none’ before declarations disables automatic declarations. Always use ‘implicit none’ before declaring your variables. Declaration of all variables (strong typing) is an easy error-detection measure that requires a few extra minutes but can save hours or days of debugging.

```fortran
implicit none
integer::a,b,c....
.....
```

Arrays
integer::x(2,100)  integer array with 2 rows and 100 columns
character(5)::x(10)  character array of 10 elements, each 5-character long
real:: c(3,4,5)  3-D array of 60 elements total

By default, indexing of each element of an array starts with 1. There are ways to make it start from any other number.

### Parameter (read-only) variables

integer,parameter:: n=5, m=20
real,parameter::pi=3.14159265
! parameter variables can be used to indicate dimensions in declarations
real::x(n,m)

### Operations on variables

#### Expressions

Numerical expressions are intuitive except that ** stands for exponentiation. Fortran does automatic conversions between integers and reals. This sometimes causes unexpected side effects

```fortran
real::x
integer::i=2,j=4
x=i*j/10
print*,i*j/10,x !This prints 0, why?
!
x=i*j/10.0
print*,i*j/10.0=x !This print 0.8
end
```

With matrices, operation involving a scalar is executed separately for each matrix element:

```fortran
real::x
x=1   !all elements of x set to 1
x(1)=1; x(2)=4; x(3)=7
x=x*5   !all elements of x multiplied by 5
```

#### Character manipulations

```fortran
character (2)::a='xy',b
c=’pqrs'
```
Basic control structures

IF statement

if (condition) statement ! statement is executed only if condition is true

if (condition) then ! as above but can contain many statements
.....
endif

If statements can contain many conditions using elseif. If no condition is fulfilled, an else keyword can be used.

if (condition1) then !
.....! executed if condition1 is true
elseif (condition2)
.....! executed if condition2 is true
elseif(condition3)
.....
else
.....!optional; executed when none of previous conditions are true
endif

Conditions in if

< <= == (equal) >= > /= (not equal)
.and. .not. .or.

! IF example
real::x
read*,x
if (x>5 .and. x<20) then
print*,’number greater than 5 and smaller than 20’
elseif (x<0) then
print*,’x smaller than 0’,x
Select case statement

These statements are optimized for conditions based on integer values

```fortran
select case ( 2*i) !expression in () must be integer or character
    case( 10)
        print*, 10
    case (12:20,30)
        print*, '12 to 20 or 30'
    case default !everything else, optional
        print*, 'none of 10:20 or 30'
end select
```

Loops

Statements inside the loop are executed repeatedly, until some termination condition occurs. In the simplest loop below, statements inside the loop below are executed for i=1,2,...,n

```fortran
do i=1,n !basic do loop
    ....
enddo
```

The example below shows creation and summation of all elements of a matrix in a do loop

```fortran
integer::,i,j
integer,parameter::n=10
real::x(n,n),sumx
!c
! initialize the matrix to some values, e.g., x(i,j)=i*j/10
do i=1,n
    do j=1,n
        x(i,j)=i*j/10.0
    enddo
enddo
! now total
```
sumx=0  
do i=1,n  
do j=1,n  
    sumx=sumx+x(i,j)  
  enddo  
enddo  
print*,sumx  
!  
!can do the same using fortran90 function sum()  
print*,sum(x)

Indices in loops can be modified to count with a stride other than one and even negative

do j=5,100,5  !j =5,10,15,...,100  
...  
do k=5,1,-1 !k=5,4,..,1  
...  

Another loop operates without any counter. An exit statement ends the loop.

do  
.....  
  if (x==0) exit  
.....  
end do  

x=1  
do  ! no condition mandatory  
   x=x*2  
   if (x==8) cycle ! skip just one iteration  
   if (x>=32) exit ! quit the loop  
   print*,x  
end do

Functions and subroutines

!Program that calculates trace of a few matrices  
integer,parameter::n=100  
real::x(n,n),y(n,n),z(n,n),tr  
......  
! trace calculated in a loop
tr=0
do i=1,n
    tr=tr+x(i,i)
enddo
!
! trace calculated in a subroutine; see body of subroutine below
! call trace_sub(tr,x,n)
!
! trace calculated in a function; see body of function below
tr=trace_f(x,n)
end

subroutine trace(t,x,n)
integer::n,i
real x(n,n),t
!
t=0
do i=1,n
    t=t+x(i,i)
enddo
end

function trace(x,n) result(t)
integer::n,i
real x(n,n),t
!
t=0
do i=1,n
    t=t+x(i,i)
enddo
end

In a subroutine, each argument can be use as input, output, or both. In a function, all arguments are read only, and the “result: argument is write only. Functions can simplify programs because they can be part of an arithmetic expression

! example with subroutine
call trace(t,x,n)
y=20*t**2+5

!example with a function
Subroutines and functions are useful if similar operations need to be simplified, or a large program decomposed into a number of smaller units.

**Argument passing**
In a subroutine, each variable can be read or written. In a program:

```fortran
    call trace(a,b,m)
    ...
    subroutine trace(t,x,n)
    a and t are pointing to the same storage in memory. Same with b and x, and m and n. In some cases, no writing is possible:
    call trace(a,b,5) !writing to p either does not work or changes constant 5
    call trace(x, y+z, 2*n+1) ! expressions cannot be changed
```

**Save feature**
Ordinarily, subroutines and functions do not “remember” variables inside them from one call to another. To cause them to remember, use the “save” statement.

```fortran
    function abcd(x) result(y)
    ...
    integer,save::count=0 !keeps count; initialized to 0 first time only
    real,save::old_x !keeps previous value of x
    ....
    count=count+1
    old_x=x
    end
```

Without save, the variables are initialized the first time only, and on subsequent calls have undetermined values.

“return” from the middle of a subroutine or function:

```fortran
    subroutine ....
    ....
    if (i > j) return
    ...
    end
```
Intrinsic (built-in) functions

Fortran contains large number of predefined functions. More popular are:

- int - truncating to integer
- nint - rounding to integer
- exp, log, log10
- abs - absolute value
- mod - modulo
- max - maximum of possibly many variables
- min
- sqrt - square root
- sin, cos, tan
- len - length of character variable

INPUT/OUTPUT

General statements to read or write are read/write/print statements. The command:

```
print*,a,b,c
```

prints a, b and c on the terminal using a default format. Variables in the print statement can include variables, expressions and constants. These are valid print statements:

- `print*, 'the value of pi is ', 3.14159265`
- `print*, 'file name =', name, ' number of records =', nrec`

Formatted print uses specifications for each printed item. Formatted print would be:

```
print ff,a,b,c
```

where ff contains format specifications. For example:

```
print '(5x, "value of i,x,y = ",i4,2x,f6.2,f6.2)',i,x,y
```

Assuming i=234, x=123.4567 and y=0.9876, this prints:

```
value of i,x,y = 234 123.46 0.99
```

Selected format descriptions

<table>
<thead>
<tr>
<th>i5</th>
<th>integer five characters wide</th>
</tr>
</thead>
<tbody>
<tr>
<td>a5</td>
<td>character 5 characters wide</td>
</tr>
<tr>
<td>a</td>
<td>character as wide as the variable (character a*8 would read as a8)</td>
</tr>
<tr>
<td>Format</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| `f3.1` | 3 characters wide  
On write, 1 digit after comma  
On read, if no comma read as if last 1 digit after comma (123 read as 12.3, 1 as .1)  
if real with comma, read as is (1.1 read as 1.1, .256 read as .256) |
| 5x | On write, output five spaces  
On read, skip five characters |
| `'text'` | On write only, output string 'text' |
| `''text''` | use double quotes when format is a character variable or string |
| `/` | Skip to next line |

**Shortcuts**

(i5,i5,i5) = ‘(3i5)’  
‘(i2,2f5.1, i2,2f5.1)’ = ‘(2(i2,2f5.1))’

**Example**

```plaintext
print '(i3,“squared=”,i5)',5,25
```

This prints: 5 squared= 25

Character variables can be used for formatting:

```plaintext
character(20)::ff
...
ff='(i3,“ squared=”,i5)'
print ff,5,25
```

Reading from a keyboard is similar in structure to writing:

```plaintext
read*,a,b,c !default, sometimes does not work with alphanumeric variables  
read ff,a,b,c !ff contains format
```

**Reading from files and writing to files**

File operations involve units, which are integer numbers. To associate unit 2 with a file:

```plaintext
open(2,file='/home/abc/data_bb')
```

Read and write from/to file has a syntax:

```plaintext
read(2,ff)a,b
```
write(2,ff)a,b

Formats can be replaces by * for default formatting. Units can be replaced by * for reading from keyboard or writing to keyboard. Thus:

read(*,*)a,b !read from keyboard using defaults
read*,a,b ! identical to above simplified

and

write(*,*)a,b,c !write to keyboard using defaults
print*,a,b,c !identical to above simplified

are equivalent.

To read numerical data from file bbu_data using default formatting

integer::a
character (2)::b
real::x

! open file “bbu-data” as unit 1
open (1,file=’bbu_data’)

! read variables from that file in free format and then print them
read(1,*)a,b,x
print*,a,b,x
! asterisk above denotes free format
!
! read same variables formatted
rewind 1
read(1,’(i5,a2,f5.2)’)a,b,x

In most fortran implementations, unit 5 is preassigned to keyboard and unit 6 to console. “Print” statement prints to console. Thus, the following are equivalent:

write(6,’(i3,’’ squared=’’,i5)’)5,25

and

print ’(i3,’’ squared=’’,i5)’, 5,25

Also these are equivalent:

read(5,*)x

and

read*,x

Free format for reading the character variables does not always work. One can use the following
from the keyboard:

    read (5,'(a)')string

or

    read '(a)',string

Attention: One read statement generally reads one line whether all characters have been read
or not! Thus, the next read statement cannot read characters that the previous one
has skipped.

Assume file “abc” with contents:

1 2 3
4 5 6
7 8 9

By default, one read statement reads complete records. If there is not enough data in one record,
reading continues to subsequent records.

    integer::x, y(4)
    open(1, file=’abc’)
    read(1,*)x
    print*,x
    read(1,*)y
    print*,y
    end

produces the following

1
4 5 6 7

**Implied loop**

The following are equivalent:

    integer::x(3),i
    ! statements below are equivalent
    read*,x(1),x(2),x(3)
    read*,x
    read*,(x(i),i=1,3)

Implied loops can be nested

    integer::y(n,m),i,j
read*,((y(i,j),j=1,m),i=1,n)  

! statement below reads by columns  
read*,x  
! and is equivalent to  
read*.,((y(i,j),i=1,n),j=1,m)  !note reversal of indices;  
or  
read*.,((y(:,j),j=1,m)

Implied loop is useful in assignment of values to arrays

real::x(n,n),d(n)  
integer::cont(100)
!
d=/(x(i,i),i=1,n)/  !d contains diagonal of x  
cont=/(i,i=1,100)/  !cont contains 1 2 3 4 ... 100

Detection of end of files or errors

integer::status
read(...,iostat=status)  
if (status == 0) then  
    print*,’read correct’
else  
    print*,’end of data’
endif

Unformatted input/output

Every time a formatted I/O is done, the programs performs a conversion from character format to computer internal format (binary). This takes time. Also, an integer or real variable can use up to 10 characters (bytes) in formatted form but only 4 bytes in binary, realizing space savings. Since binary formats in different computers and compilers may be different, data files in binary usually cannot be read on other systems.

Unformatted I/O statements are similar to formatted ones except that the format field is omitted.

real x(1000,1000)  
open(1, file=’scratch’, form=’unformatted’)  
.......  
write(1)x
rewind 1
read(1)x
rewind 1
write(1)m,n,x(1:m,1:n) !writing and reading section of a matrix
rewind 1
read(1)m,n,x(1:m,1:n)
close(1,status=’delete’)
end

Statement CLOSE is needed only if the file needs to be deleted, or if another file needs to be opened with the same unit number.

**Reading from/to variables**

One can read from a variable or write to a variable. This may be useful in creating names of formats. For example:

```fortran
character (20)::ff
write(ff, ‘(i2)’)i
ff=’( // ff/ ’’f10.2)’ !if i=5, ff=’( 5f10.2)’
```

The same can be done directly:

```fortran
write(ff, ‘( “(, i2, ”f10.2)” )’ )i ! spaces only for clarity
```

Reading/writing with variables always require a format, i.e., * formatting does not work.
Compiling

Programs in Fortran 90 usually use the suffix f90. To compile program pr.f90 on a Unix system, type:
f90 pr.f90
where f90 is the name of the fortran 90 compiler. Some systems use a different name.

To execute, type:
a.out

To name the executable pr rather than default a.out:
f90 -o pr pr.f90

To use optimization
f90 -O -o pr pr.f

Common options:
-g add debugging code
-static or -s compile with static memory allocation for subroutines that includes
initializing memory to 0; use for old programs that don’t work otherwise.

Hints

If a compile statement is long, e.g.,
f90 -C -g -o mm mm.f90
to repeat it in Unix under typical conditions (shell bash or csh), type
!f

If debugging a code is laborious, and the same parameters are typed over and over again, put them in a file, e.g., abc, and type:
a.out <abc

To repeat, type
!a

Programs composed of several units are best compiled under Unix by a make command. Assume that a program is composed of three units: a.f90, b.f90 and c.f90. Create a file called Makefile as below:
a.out: a.o b.o c.o
    f90 a.o b.o c.o

a.o: a.f90
f90 -c a.f90
b.o: b.f90
f90 -c b.f90
c.o: c.f90
f90 -c c.f90

After typing
make
all the programs are compiled and linked into an executable. Later, if only one program changes, typing
make
causes only that program to be recompiled and all the programs linked. The make command has a large number of options that allows for high degree of automation in programming.

Attention: Makefile requires tabs after : and in front of commands; spaces do not work!
Advanced features in Fortran 95

This section is a continuation of the previous chapter that introduces more advanced features of F95. This section also has some repetitive information from the former section to also be an upgrade guide from F77.

Free and fixed formats
Fortran 77 programs in traditional fixed format are supported, usually as files with suffix “.f”. Fixed format programs can be modified to include features on the f90 fortran. Free format programs usually have suffix “.f90”.

Free format (if suffix .f90)
! Comments start after the exclamation mark
! One can start writing at column 1 and continue as long as convenient
!
integer:: i1, i2, i3, j1, j2, j3,

& integer variables

k1, k2, m, n  !& is line continuation symbol

real :: xx !XX is name of the

! coefficient matrix

&,yy !if continuation is broken by a blank or comment, use & as here

a=0; b=0 !multiple statements are separated by semicolon

Long names
Up to 32 characters

New names for operators
== (.eq.), /= (.ne.), <=, =>, <, >

Matrix operations
real :: x(100,100),y(100,100),z(100),v(20),w(20),p(20,20),a
...

x(:,1)=y(:,1)
p=x(21:40,61:80)   ! matrix segments
v=z(1:100:5)  ! stride of 5:  v=(*z(1), z(6),z(11),...,z(96)/)

v=x(10:100:10,1:10)**5  ! operation with a constant
y=matmul(z,y)  ! matrix multiplication
p=x(:20,81:)  ! same as x(1:20,81:100)
print*,sum(z(1,:)

x=transpose(y)  !x=y’
print*,dot_product(v,w)  !v’w
a=maxval(x)  ! maximum value in x
print*,count(x,x>100)  ! number of elements in x that satisfy the condition
a=sum(y)  ! sum of all elements of y
print*,sum(y,dim=1)  ! vectors that contains sums by row
print*,sum(y,y>2.5)  ! sum of all elements > 0.5

See other functions like maxloc(), any(), all(), size....

Vector and matrix operations involve conformable matrices

**Constructors**

real :: x(2),y(1000),g(3,3)
...
x=(/13.5,2.7/)  !x(1)=3.5, x(2)=2.7
y=(/ (i/10.0, i=l,l000) /)  !y=[1 2 3 ... 999 1000]
...
i=2; j=5
x=(/i**2,j**2/)  ! constructors may include expressions; x=[4  25]
y(1:6)=(/x,x,x/)  ! constructors may include arrays

g(:,1)=(/1 2 3/); g(:,2)=(/4,5,6/); g(:,3)=(/7,8,9/)  ! matrix initialized by columns
g=reshape(  (/1,2,3,4,5,6,7,8,9/), (/3,3/) )  ! matrix initialized by reshape()

Constructors can be used in indices!

integer::i(2), p(10)
...
i=(/2,8/)
...
p(i)=p(i)+7  !p(2)=p(2)+7, p(8)=p(8)+7

**New declarations with attributes**

integer,parameter::n=10,m=20  ! parameters cannot be changed in the program
real,dimension(n,m)::x,y,z  ! x, y and z have the same dimension of n x m
real::x(n,m),y(n,m),z(n,m)  ! equivalent to above

real (r4),save::z=2.5  ! Real variable of precision r4 with value saved (in a subroutine)
character (20):: name  ! character 20 bytes long
New control structures

```fortran
select case ( 2*i) !expression in () must be integer or character
  case( 10)
    print*, 10
  case (12:20,30)
    print*,'12 to 20 or 30'
  case default
    print*, 'none of 10:20 or 30'
end select

x=1
do ! no condition mandatory
  x=x*2
  if (x==8) cycle ! skip just one iteration
  if (x>=32) exit ! quit the loop
  print*,x
end do
```

```fortran
real:: x(100,1000)
where (x >500)
  x=0 !done on all x(i)>500
elsewhere
  x=x/2
end where
```

Allocatable arrays

```fortran
real,allocatable::a(:,;) ! a declared as 2-dimensional array
allocate (a(n,m)) ! a is now n x m array
!
if (allocated(a)) then ! the status of allocatable array can be verified
  deallocate(a) ! memory freed
endif
```

Pointer arrays
real, pointer:: a(:, :), b(:, :, c(:, :), d(:)) ! a-c declared as 2-dimensional arrays
allocate(a(n, m)) ! a is now n x m array

b=>a ! b points to a, no copy involved
c=>a ! a-c point to the same storage location

allocate(a(m, m)) ! a points to new memory;
nullify(b) ! association of b eliminated
b=null() ! same as above

allocate(b(m, m))
b=a ! b copied to a

deallocate(c) ! deallocate memory initially allocated to a
c=>b(1:5, 1:10:2) ! pointer can point to section of a matrix
d=>b(n,:) ! vector pointer pointing to a row of a matrix

if (associated(b)) then ! test of association
deallocate(b) ! memory freed and b nullified
endif

real, target:: x(2, 3) ! pointer can point to non-pointer variables with target attribute
real, allocatable, target:: y(:, :)
real:: pointer:: p1(:, :), p2(:, :)

... p1=>x
p2=>y

Read more on pointers later.

**Data structures**

type animal
    character (len= 13) :: id, parent(2)
    integer :: year_of_birth
    real :: birth_weight, weaning_weight
end type animal

type (animal) :: cow

... read( 1 , form) cow%id, cow%sire, cow%dam, cow%yob, cow%bw, cow%ww

To avoid typing the type definitions separately in each subroutine, put all structure definitions in a
module, e.g., named definitions

module aa
    type animal
        ...
    end type
    type ...
        ...
    end type
end module

Program main
use aa !use all definitions in the module
type(animal): :cow
...
end

subroutine pedigree
use aa
type(animal): :cow
...
end

To include modules in the executable file:
1) compile files containing those modules,
2) if modules are not in the same directory as programs using them, use appropriate compile switch,
3) if modules contains subroutines, the object code of the modules needs to be linked.

Elements of data structures can be pointers (but not allocatable arrays) and can be initialized.

type sparse_matrix
    integer::n=0
    integer,pointer::row(:)=null(),col(:)=null()
    real,pointer::val(:)=null()
end type

**Automatic arrays in subroutines**
real:: x(5,5)
...
call sub1(x,5)
...
subroutine sub1(x,n)
real:: x(n,n),work(n,n) !work is allocated automatically and deallocated on exit
.. end

New program organization - internal functions/subroutines and modules
Functions and subroutines as described in the previous section are separate program units. They can be compiled separate from main programs and put into libraries. However, there are prone to mistakes as there is no argument checking, e.g., subroutines can be called with wrong number or type of arguments without any warning during compilation.

Fortran 90 allows for two new possibilities for placement of procedures: internal within a program and internal within a module.

! This is a regular program  with separate procedures
program abcd
...
call sub1(x,y)
...
z=func1(h,k)
end program

subroutine sub1(a,b)
.... end subroutine

function func1(i,j)
.... end function

! This is the same program with internal subroutines
program abcd
...
call sub1(x,y)
...
z=func1(h,k)

contains ! end program is moved to the end and “contains” is inserted

subroutine sub1(a,b)
....
end subroutine
function func1(i,j)
.....
end function
end program

Internal procedures have access to all variables of the main program. Procedures internal to the program cannot be part of a library.

! This program uses modules
module mm1
  contains
  subroutine sub1(a,b)
  ....
  end subroutine
  function func1(i,j)
  .....  
  end function
end module

program abcd
use mm1 !use module mm1
...
call sub1(x,y)
...
z=func1(h,k)
end program

An example of internal functions or subroutines
program abc
integer :: n=25
real :: x(n)
...
call abc(i)
contains
  subroutine abc(i)
  integer :: i
  integer :: i
if (i<=n) x(i)=0 !variables declared in the main program need not
    ! be declared in internal subroutines
end subroutine

end program

Procedures that are internal or contained within modules can have special features that make them
simpler or more convenient to use.

**Deferred size arrays in subroutines or functions**
real:: x(5,5)
..
call sub1(x)
...

subroutine sub1(x)
real:: x(:,:), !size is not specified; passed automatically
real:: work(size(x,dim=1),size(x,dim=2)) !new variable created with dimension of x()
.. end subroutine

This works only if the program using the subroutine sub1 is informed of its arguments, either
by (1) preceding that program by an interface statement, (2) by putting sub1 in a module and using
the module in the program, or 3) by making the subroutine an internal one.

(1)
interface
    subroutine sub1(a)
        real::a(:,:)
    end subroutine
end interface
real::x(5,5)
....

(2)
module abc
    interface
....
subroutine sub1(x)
  real:: x(:,:)
  real::work(size(x,dim=1),size(x,dim=2))
  ..
end subroutine
end module

program main
use abc
real::x(5,5)
.....
end program

Use of interface statements results in interface checking, i.e., whether arguments in subroutine calls and actual subroutines agree.

The module can be compiled separately from the main program. The module should be compiled before it is used by a program.

**Function returning arrays**

A function can return an array. This array can be either deferred size if its dimensions can be deduced from the function arguments, or it can be a pointer array.

function kronecker1(y,z) result(x)
  !this function returns y “kronecker product” z
  real::y(:,:),z(:,:), z(size(y,dim=1)*size(z,dim=1), size(y,dim=2)*size(z,dim=2))
  ...
end function
function kronecker2(y,z) result(x)
!this function returns y “kronecker product” z
real::y(:,:),z(:,:)
real,pointer:: z(:,:)
.. allocate(z(size(y,dim=1)*size(z,dim=1), size(y,dim=2)*size(z,dim=2))
... end function

function consec(n) result (x)
! returns a vector of [1 2 3 ... n]
integer::n,x(n),i
!
x=(/(i,i=1,n)/)
end function

**Functions with optional and default parameters**

subroutine test(one,two,three)
integer, optional::one,two,three
integer :: one1,two1,three1       !local variables
!
if (present(one)) then
   one1=one
else
   one1=1
endif
if (present(two)) then
   two1=two
else
   two1=2
endif
if (present(three)) then
   three1=three
else
   three1=3
!
endif
print*,one1,two1,three1
end

.....
call test(15,6,7) call test(10) ! equivalent to call test(10,2,3)
call test(5, three=67) ! parameter three out of order

**Subroutine/Function overloading**
It is possible to define one function that will work with several types of arguments (e.g., real and integer). This is best done using modules and internal subroutines.

module prob
  !function rand(x) is a random number generator from a uniform distribution.
  ! if x is real - returns number in interval <0-x)
  ! if x is integer - returns integer between 1 and x.

  interface rand
    module procedure rand_real, rand_integer
  end interface

  contains

  function rand_real(x) result (y)
  real::x,y
  !
  call random_number(y) ! system random number generator in interval <0,1)
  y=y*x
  end function

  function rand_integer(x) result (y)
  integer::x,y
  real::z
  !
  call random_number(y) ! system random number generator in interval <0,1)
  y=int(y*z)+1
  end function

end module

program overload
! example of use of function rand in module prob
use prob  ! name of module where overloaded functions/subroutines are located
integer::n
real::x

! generate an integer random number from 1 to 100
print*,rand(100)

! generate a real random number from 0.0 to 50.0
print*,rand(50.0)
end

Operator overloading

Arithmetic operators (+ - / *) as well as assignment(=) can be overloaded, i.e., given different meaning for different structures.

For example,

Type (bulls)::
  bull_breed_A,&
  bull_breed_B,&
  bull_combined,&
  bull_different,&
  bull_common

...  

! The following operations could be programmed

bull_combined = bull_breed_A + bull_breed_B
bull_different = (bull_breed_A - bull_breed_B) + (bull_breed_B - bull_breed_A)
bull_common = bull_breed_A * bull_breed_B

General philosophy of Fortran 90

A fortran 90 program can be written in the same way as fortran 77, with the main program and many independent subroutines. During compilation it is usually not verified that calls to subroutines and the actual subroutines match, i.e., the number of arguments and their types match. With a mismatch, the program can crash or produce wrong results. Fortran 90 allows to organize the program in a series of modules with a structure as follows:

MODULE m1
  interfaces
  declarations of data structures
  data declarations
CONTAINS
  subroutines and functions
END MODULE m1

MODULE m2
  interfaces
  declarations of data structures
  data declarations
    CONTAINS
      subroutines and functions
    END MODULE m2

MODULE m3
  ...

PROGRAM title
USE MODULE m1,m2,..
  declarations
  body of main program
    CONTAINS
      internal subroutines and functions
  END PROGRAM title

Each set of programs is in a module. If there is a common data to all programs together with corresponding subroutines and functions, it can be put together in one module. Once a program (or a module) accesses a module, all the variables become known to that program and matching of arguments is verified automatically.

It is a good practice to have each module in a separate file. Modules can be compiled separately and then linked. In this case, the compile line needs to include a directory where modules are compiled.

Other selected features

New specifications of precision

integer,parameter::r4=selected_real_kind(6,30) ! precision with at least 6 decimal digits
  ! and range of 10^{30}
real(r4)::x ! x has accuracy of at least 6 digits (same as default)
integer, parameter :: r8 = selected_real_kind(15) ! as above but 15 decimal digits
real(r8):: y ! y has accuracy of at least 15 decimal digits (same as double precision)

integer, parameter :: i2 = selected_int_kind(4) ! integer precision of at least 4 digits
integer, parameter :: i4 = selected_int_kind(9) ! integer precision of at least 9 digits

To avoid defining the precision in each program unit, all definitions can be put into a module.

module kinds
  integer, parameter :: i2 = SELECTED_INT_KIND(4)
  integer, parameter :: i4 = SELECTED_INT_KIND(9)
  integer, parameter :: r4 = SELECTED_REAL_KIND(6, 37)
  integer, parameter :: r8 = SELECTED_REAL_KIND(15, 307)
  integer, parameter :: r16 = SELECTED_REAL_KIND(18, 4931)
  integer, parameter :: rh = r8
end module kinds

Variable i2 denotes precision of an integer with a range of at least 4 decimal digits or -10000 to 10000. It usually can be a 2-byte integer. The variable i4 denotes precision of an integer with a range of at least 9 decimal digits. It can be a 4-byte integer. The variable r4 denotes precision of a real with the precision of at least 6 significant digit and a range of $10^{37}$. It can be a 4-byte real. The variable r8 denotes precision of a real with the precision of at least 15 significant digit and a range of $10^{307}$. It can be a 8-byte real. Finally rh is set to r8.

In a program

program test
  use kinds
  real (r4):: x
  real (rh):: y
  integer (i2):: a
... end

Variables x, y, and a have appropriate precision. If variable y and other variables of precision rh need to be changed to precision r4 to save memory, all what is needed is a change in the module and recompilation of the remaining programs.

Lookup of dimensions

real x(100,50)
integer p(2)

.. print*,size(a) ! total number of elements(5000)
print*,size(x,dim= 1) ! contains size of the first dimension (100)
print*,size(x,dim=2)  ! contains size of the second dimension (50)
p=shape(x) ! p=(/100,50/)

Functions with results and intents

subroutine add (x,y,z)
real, intent(in):: x ! x can be read but not overwritten
real, intent(inout):: y ! y can be read and overwritten, this is default
real, intent(out):: z ! z cannot be read before it is assigned a value
....
end subroutine

Pointers II

The following example shows how a pointer variable can be enlarged while keeping the original contents and with a minimal amount of copying

real,pointer::p1(:),p2(:)

allocate(p1(100)) !p1 contains 100 elements
.....
! now p1 needs to contain extra 100 elements but with initial 100 elements intact
allocate(p2(200))
p2(1:100)=p1
deallocate(p1)
p1=>p2 "! now p1 and p2 point to the same memory location
nullify(p2) "! p2 no longer points to any memory location
...
end

If deallocate is not executed, memory initially assigned to p1 would become inaccessible. Careless use of pointers can result in memory leaks, where the amount of memory used is steadily increasing.

real,pointer::x(:)
do i=1,n
    x=>diagonal(y)
    ...
    deallocate(x) !memory leak if this line omitted
dendo

Pointers can be used for creation of linked lists, where each extra element is added dynamically.

type ll_element
   real::x=0
   integer::column=0
   type (ll_element),pointer::next=>null()
end type

type ll_matrix
   type(ll_element),pointer::rows(:)=>null() !pointer to each row of linked list
   integer::n=0 ! number of rows
end type

program main
   type (ll_matrix)::xx !declaration
   ...
   Subroutine add_ll(a,i,j,mat)
   ! mat(i,j)=mat(i,j)+a
   type (ll_matrix)::mat
   real::a
   integer::i,j
   type (ll_element)::current,last
   !
   current=>mat%rows(i) !points to first element in row i
   do
      if (.not. associated(current) exit !exit loop to create new element
      if (current%column == j) then
         current%x=current%x+a !found element; add to it
         return !return from subroutine
      else
         current=>current%next !switch to next element and loop
      enddo
   !element not found
   allocate(last)
current%next=>last !set link from the previous element
last%column=j; last%x=a !set values
end subroutine

Interfaces

operators (+, *, user defined)

One can create interfaces so that operations could look as
a+b
where a and b are structures of, say, type matrix.

The following would make it possible

interface operator (+)
    Module procedure add
end interface
    contains

function add(x,y) result (z)
    type(matrix)::z
    type(matrix), intent(in)::x,y
    ...
end function

In this case
a+b
have the same effect as
add(a,b)

Please note that matrices being arguments of this function have “intent(in)”, i.e., the arguments of the function are defined as non-changeable. User defined operators need to be characters within points, e.g. .add. or .combine.

equal sign

To create an assignment that would be
a=b
create the following interface and function

interface assignment(=)
    Module procedure equalxy
end interface
contains

subroutine equalxy(x,y)
type(matrix),intent(out)::x
type(matrix), intent(in)::y
...
end function

Please note locations of intent(in) and intent(out). In this case,

a=b
and
call equalab(a,b)
will result in identical execution.

**Differences between Fortran 90 and 95**

Fortran 95 includes a number of “fixes + improvements to F90. Some of them are shown here.

**Timing function**

The CPU timing function is now available as :

call cpu_time(x)

**Automatic deallocation of allocatable arrays**

Arrays allocated in procedures are automatically deallocated on exit.

**Pointer and structure initialization**

Pointers can be nullified when declared; most compilers but not all do it automatically.

real,pointer::x(:,;)=>null()

**Elements of the structure can be initialized**
type IJA
  integer::n=0
  integer,pointer::,ia(:)=null(),ja(:)=null
  real,pointer::a(:)=null()
end type

Elemental subroutines

A subroutine/function defined as elemental for a scalar argument, can be automatically used with arrays. In this case, the same operation is performed on every element of the array.

real::x,y(10),z(5,5)

call pdf(x)
call pdf(y)
call pdf(z)
..
contains

elemental subroutine pdf(x)
real::x
  x=1/(2*sqrt(3.14159265))*exp(-x**2/2)
end subroutine

end program

In the program above, one can precalculate 1/(2*sqrt(3.14159265)) to avoid repeated calculations, but optimizing compilers do it automatically.

Many functions in Fortran 90 are elemental or seem to be elemental. This may include a uniform pseudo-random number generator.

real::x,y(12)
!
call random_number(x) ! x ~ un[0,1)
call random_number (y) ! y(i)~un[0,1), i=1,12
x=sum(y)-6 ! x ~ approx N(0,1)

Including the random_number subroutine as part of another elemental routine works well with some compilers and generates errors (not pure subroutine) with others. In this case, random_number may be implemented as overloaded.
Rewriting programs in F77 to F95

The purpose of the rewrite is primarily simplification of old programs and not computing speed. In F77 it is very easy to make simple but hard-to-find programming mistakes while in F95 it is much more difficult. One can consider a rewrite or a “touch-up” when upgrade or fixes to an old program are too time consuming or seem impossible.

The computing time in F95 may be longer due to management of memory allocation. However, if a simpler program allows easy upgrade to a more efficient algorithm, the program in F95 may end up being faster.

The complexity of the program is roughly the number of subroutines in a program times the number of arguments per subroutine, plus the number of declared variables times the size of code (excluding comments). A program can be simplified by:

1. Decreasing the number of variables
2. Decreasing the number of subroutines
3. Decreasing the number of arguments per subroutine/function.
4. Decreasing the length of the code (without using tricks)

The rewrite may be done at a few levels.

Simplest

Eliminate some loops by a matrix statement or a built-in functions

```
! p=0
! do i=1,n
! p=p+x(i)
! enddo

p=sum(x(:n))
```

Compound functions

```
! x~MVN(y,inv(z*p))
! .... ! multiply
! .... ! invert
! .... ! sample MVN(0,..)
! .... ! add constant

x=normal(y,inv(mult(z,p))
```

Replace all work storage passed to subroutines with automatic arrays
real::work(n)
... call mult(x,y,z,work,n)
... subroutine mult(a,b,c,n)
real::work(n) !work is used as scratch
... call mult(x,y,n)
... subroutine(x,y,n)
real::work(n)

Possibly initialize variables in the declaration statement; beware of consequences
real::a,b(20) real::a=0,b(20)=0
integer::j integer:j=0
a=0 !values initialized once
do i=1,20
doi=0
end do
j=0

Use memory allocation to have matrices the right size and then eliminate separate bounds for declared and used indices
real::x(10000) real,allocatable::x(:)
m=2*(n+7) ! used dimension of x allocate (x(2*(n+7))
call comp(x,m,n) call comp(x)

Level II

Simplify interfacing with old subroutines either by rewriting (if simple) or by reusing (writing a F95 interface); the rewritten subroutine as below must be either internal or in a module.
call fact3dx(x,,n,m,w) call factor(x)
... subroutine fact3dx(mat,n,m,work) subroutine factor(mat)
! mat=fact(mat) 
real::mat(:,,:),w(size(mat,dim=1))
integer::n,m interface
integer::n,m
real::mat(n,n),work(n)[end interface
... call fcat3dx(mat,size(w),size(w),w)
end subroutine

Replace all “common” variables by a module
program
integer p(n)
real x(m,m), z(n,n)
common //p,x,z
....
subroutine aa...
integer p(n)
real x(m,m), z(n,n)
common //p,x,z
....
module stor
integer p(n)
real x(m,m), z(n,n)
end module
program ..
use stor
subroutine ..
use stor

Change subroutines into internal subroutines (if specific to a program) or put them into module (if subroutines useful in other programs).

**Level III**

Replace old libraries by new, simpler to use

Overload whatever makes sense

Organize variables into data structures

**Level IV**

Rewrite from scratch. The old program is utilized as a source of general algorithms and examples.
### Setting-up mixed model equations

Good references on setting up models of various complexity are in Groeneveld and Kovacs (1990), who provide programming examples, and in Mrode (1996), who provides many numerical examples and a large variety of models. The approach below provides both numerical examples and programming. Almost all programming is in fortran 90, emphasizing simplicity with programming efficiency.

**Fixed 2-way model**

Mixed model: \( y = X\beta + e \). Assuming \( E(y) = X\beta \), \( V(y) = V(e) = I\sigma^2 \), the estimate of \( \beta \) is a solution to normal equations: \( X'X \beta = X'y \)

Because \( X \) contains few nonzero elements and matrix multiplication is expensive (the cost is cubic), matrices \( X'X \) and \( X'y \) are usually created directly.

### Data

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>( y_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Write the model as \( y_{ij} = a_i + b_j + e_{ij} \)

The mixed model for the given data:

\[
\begin{bmatrix}
5 \\
8 \\
7 \\
9 \\
6
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
b_1 \\
b_2 \\
b_3
\end{bmatrix}
+ e
\]

or
\[
\begin{bmatrix}
5 \\
8 \\
7 \\
9 \\
6
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
b_1 \\
b_2 \\
b_3
\end{bmatrix} + \varepsilon
\]

\[y = (X_1 + X_2 + X_3 + \ldots) \beta + \epsilon\]

Observation: first second third

One can write: \[\mathbf{x} = \sum_{k=1}^{\text{all observations}} \mathbf{x}_k \quad \mathbf{x}'\mathbf{x} = \sum_{k=1}^{\text{all observations}} \mathbf{x}_k'\mathbf{x}_k \quad \mathbf{x}'\mathbf{y} = \sum_{k=1}^{\text{all observations}} \mathbf{x}_k'\mathbf{y}\]

For the first two observations:

\[
\mathbf{x}_1'\mathbf{x}_1 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}_1'\mathbf{x}_1 \mathbf{y} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

\[
\mathbf{x}_2'\mathbf{x}_2 = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}_2'\mathbf{x}_2 \mathbf{y} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

For a two-factor model, the contribution to \(\mathbf{X}'\mathbf{X}\) from one observation is always 4 elements of ones, in a square.
The contribution to \( \mathbf{X}'\mathbf{y} \) from one observation is always two elements with the values of the current observation.

\[
\mathbf{x}_{ij}' \mathbf{y} = \begin{bmatrix}
1 & 0 & 0 & 0 & 5 \\
0 & 0 & 0 & 0 & 8 \\
1 & 0 & 0 & 0 & 7 \\
0 & 0 & 0 & 0 & 6 \\
0 & 0 & 0 & 0 & 9 \\
\end{bmatrix} \begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 \\
y_5 \\
\end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}
\]

Rules

Define a function \texttt{address}(effect) that calculates the solution number corresponding to the i-th effect and the j-th level.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Address(effect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>1</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>2</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>3</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>4</td>
</tr>
<tr>
<td>( b_3 )</td>
<td>5</td>
</tr>
</tbody>
</table>

In the model \( y_{ij} = a_i + b_j + e_{ij} \), the contribution to \( \mathbf{X}'\mathbf{X} \) (also called the Left Hand Side or \textbf{LHS}) from observation \( y_{ij} \) are 4 ones on positions:

\[
[\text{address}(a_i), \text{address}(a_i)] [\text{address}(a_i), \text{address}(b_j)] [\text{address}(b_j), \text{address}(a_i)] [\text{address}(b_j), \text{address}(b_j)]
\]

The contributions to \( \mathbf{X}'\mathbf{y} \) (also called the right-hand side or \textbf{RHS}) are 2 values of \( y_{ij} \) on positions:

\text{address}(a_i) \text{address}(b_j)
If the number of levels were large, say in the thousands, the explicit calculation of $X'X$ would involve billions of arithmetic operations, mostly on zeros. By calculating only nonzero entries of LHS and RHS, the number of operations is in the thousands.

Using the rules, calculate contributions to LHS and RHS from all five observations

<table>
<thead>
<tr>
<th>$i$</th>
<th>$j$</th>
<th>Contributions to LHS</th>
<th>Contributions to LHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>(1,1),(1,3),(3,1),(3,3)</td>
<td>add 5 to element 1 and 3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>(1,1),(1,4),(4,1),(4,4)</td>
<td>add 8 to element 1 and 4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>(2,2),(2,3),(3,2),(3,3)</td>
<td>add 7 to element 2 and 3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(2,2),(2,4),(4,2),(4,4)</td>
<td>add 6 to element 2 and 4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>(2,2),(2,5),(5,2),(5,5)</td>
<td>add 9 to element 2 and 5</td>
</tr>
</tbody>
</table>

Summing all the contributions,

$$
X'X = \begin{bmatrix}
2 & 0 & 1 & 1 & 0 \\
0 & 3 & 1 & 1 & 1 \\
1 & 1 & 2 & 0 & 0 \\
1 & 1 & 0 & 2 & 0 \\
0 & 1 & 0 & 0 & 1 \\
\end{bmatrix}
\quad \text{and} \quad
X'y = \begin{bmatrix} 13 \\ 22 \\ 12 \\ 14 \\ 9 \end{bmatrix}
$$

Please note the diagonal and symmetrical structure for $X'X$.

**Extension to more than two effects**

For $n$ effects, each observation will contribute $n^2$ ones to the LHS and $n$ values to the RHS. Proceed as follows:
1. For each observation, calculate addresses: address$_i$, $i=1,n$ for each of the effect
2. Add 1 to LHS elements (address$_i$, address$_j$), $i=1,n$, $j=1,n$
3. Add observation value to RHS, elements (address$_i$), $i=1,n$
4. Continue for all observations.
Example

Assume a four-effect model with the following number of levels:

<table>
<thead>
<tr>
<th>effect</th>
<th>number of levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (a)</td>
<td>50</td>
</tr>
<tr>
<td>2 (b)</td>
<td>100</td>
</tr>
<tr>
<td>3 (c)</td>
<td>40</td>
</tr>
<tr>
<td>4 (d)</td>
<td>10</td>
</tr>
</tbody>
</table>

Assume the following observation:
\( a_{11}, b_{67}, c_{24}, d_{5}, y=200 \)

The addresses will be calculated as:

<table>
<thead>
<tr>
<th>effect</th>
<th>starting address for this effect</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0+11=11</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>50+67=117</td>
</tr>
<tr>
<td>3</td>
<td>100+50=150</td>
<td>150+24=174</td>
</tr>
<tr>
<td>4</td>
<td>50+100+40</td>
<td>190+5=195</td>
</tr>
</tbody>
</table>

1 would be added to the following LHS locations: (11,11), (11,117), (11,174), (11,195), (117,11), (117,117), (117,174), (117,195), (174,11), (174,117), (174,174), (174,195), (195,11), (195,117), (195,174), (195,195)

200 will be added to the following RHS locations: 11, 117, 174 and 195.
Computer program

program lsq
implicit none
!
!This program calculates least square solutions for a model with 2 effects.
!All the storage is dense, and solutions are obtained iteratively by
!Gauss-Seidel.
!This model is easily upgradable to any number of effects
!
real, allocatable:: xx(:,,:),xy(:,),sol(:,)  !storage for the equations
integer, allocatable:: indata(,:)           !storage for one line of data
integer,parameter:: neff=2,nlev(2)=(/2,3/)  !number of effects and levels
real :: y                                     !observation value
integer :: neq,io,i,j                       !number of equations and io-status
integer,allocatable::address(:)
!
neq=sum(nlev)
allocate (xx(neq,neq), xy(neq), sol(neq),indata(neff),address(neff))
xx=0; xy=0
!
open(1,file='data_pr1')
!
do
read(1,*),io,indata,y
if (io.ne.0) exit
!
call find_addresses
!
do i=1,neff
   do j=1,neff
      xx(address(i),address(j))=xx(address(i),address(j))+1
   enddo
   xy(address(i))=xy(address(i))+y
enddo
!
print*,"left hand side"
do i=1,neq
   print '(100f5.1)',xx(i,:)
enddo
!
print '( '' right hand side:' ,100f6.1)',xy
!
call solve_dense_gs(neq,xx,xy,sol)        !solution by Gauss-Seidel
!
contains
!
subroutine find_addresses
integer :: i
!
do i=1,neff
   address(i)=sum(nlev(1:i-1))+indata(i)
enddo
!
end subroutine
end program lsq

subroutine solve_dense_gs(n,lhs,rhs,sol)
! finds sol in the system of linear equations: lhs*sol=rhs
! the solution is iterative by Gauss-Seidel
integer :: n
real :: lhs(n,n),rhs(n),sol(n),eps
integer :: round
!
round=0
do
    eps=0; round=round+1
    do i=1,n
        solnew=sol(i)+(rhs(i)-sum(lhs(i,:)*sol))/lhs(i,i)
        eps=eps+ (sol(i) -solnew)**2
        sol(i)=solnew
    end do
    if (eps.lt. 1e-10) exit
end do
print*, 'solutions computed in ',round,' rounds of iteration'
end subroutine
Model with covariables

Data

<table>
<thead>
<tr>
<th>i</th>
<th>x_i</th>
<th>y_{ij}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Write the model as \( y_{ij} = a_i + \alpha x_j + e_{ij} \), where \( \alpha \) is coefficient of regression.

The mixed model for the given data:

\[
\begin{bmatrix}
5 \\
8 \\
7 \\
6 \\
9
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 & 1 \\
1 & 0 & 2 \\
0 & 1 & 1 \\
0 & 1 & 2 \\
0 & 1 & 3
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_3
\end{bmatrix}
+ \epsilon
\]

For the first two observations:

\[
X_i^T X_i = 
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 & 1 \\
0 & 0 & 0 \\
1 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\]
For a two-factor model with covariates, the contribution to $X'X$ from one observation is always 4 elements in a square, with values being one, the value of the covariate or its square.

$$X_j'X_j = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

The contribution to $X'y$ from one observation is always two elements with the values of the current observation possibly multiplied by the value of the covariate.

$$X_j'y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 8 \\ 1 & 0 & 0 & 0 & 0 & 7 \\ 0 & 2 & 0 & 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 5 \\ 8 \\ 7 \\ 9 \end{bmatrix}$$

$$X_j'y = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 8 \\ 0 & 2 & 0 & 0 & 0 & 7 \\ 0 & 2 & 0 & 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 8 \\ 0 \\ 16 \end{bmatrix}$$

The contribution to $X'y$ from one observation is always two elements with the values of the current observation possibly multiplied by the value of the covariate.

**Rules**

Define a function `address(effect)` that defines the solution number corresponding to i-th effect and j-th level

<table>
<thead>
<tr>
<th>Effect</th>
<th>Address(effect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>1</td>
</tr>
<tr>
<td>$a_2$</td>
<td>2</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
</tr>
</tbody>
</table>
In the model $y_{ij} = a_i + b_j + e_{ij}$, the contribution to $X'X$ (also called the Left Hand Side or LHS) from observation $y_{ij}$ are 4 ones on positions:

1 to  $[\text{address}(a_i), \text{address}(a_i)]$
$x_{ij}$ to  $[\text{address}(a_i), \text{address}(\alpha)]$ and $[\text{address}(\alpha), \text{address}(a_i)]$
$x_{ij}^2$ to  $[\text{address}(\alpha), \text{address}(\alpha)]$

The contributions to $X'y$ (also called the right-hand side or RHS) are 2 contributions:

$y_{ij}$ to  $\text{address}(a_i)$
$y_{ij}x_{ij}$ to  $\text{address}(b_j)$

Summing all the contributions,

\[
X'X = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 3 & 6 \\ 3 & 6 & 19 \end{bmatrix}, \quad X'y = \begin{bmatrix} 13 \\ 22 \\ 67 \end{bmatrix}
\]

Nested covariable

If the covariable are nested in effect p, one estimates p regressions, one for each level of effect n. The rules remain the same as above except that for each level of p, contributions are for the regression specific to that covariable. If the regression were nested within the effect 1 in the example data set, LHS and RHS would be:

\[
X_1'X = \begin{bmatrix} 2 & 0 & 3 & 0 \\ 0 & 3 & 0 & 6 \\ 3 & 0 & 5 & 0 \\ 0 & 6 & 0 & 14 \end{bmatrix}, \quad X_1'y = \begin{bmatrix} 13 \\ 22 \\ 21 \\ 46 \end{bmatrix}
\]
Computer program
Changes relative to the previous program are highlighted.

program lsqr
implicit none
!
! As lsq but with support for regular and nested regressions
!
integer, parameter:: effcross=0, & !effects can be cross-classified
effcov=1 & !or covariables
real, allocatable:: xx(:,:), xy(:,), sol(:) !storage for the equations
integer, allocatable:: indata(:,:)!storage for one line of effects
integer, parameter:: neff=2, nlev(2)=(/2,3/) !number of effects and levels
integer :: effecttype(neff)=(/effcross, effcov/)
integer :: nestedcov(neff)=(/0,1/)
real :: weight_cov(neff)
!
real :: y !observation value
integer :: neq, io, i, j !number of equations and io-status
integer, allocatable:: address(:)
!
neq=sum(nlev)
allocate (xx(neq,neq), xy(neq), sol(neq), indata(neff), address(neff))
xx=0; xy=0
!
open(1, file='data_pr1')
!
do
read(1,*, iostat=io)indata, y
if (io.ne.0) exit
call find_addresses
do i=1, neff
do j=1, neff
xx(address(i), address(j))=xx(address(i), address(j))+
&
weight_cov(i)*weight_cov(j)
enddo
xy(address(i))=xy(address(i))+y *weight_cov(i)
enddo
enddo
!
print*, 'left hand side'
do i=1, neq
print '(100f5.1)', xx(i, :)
enddo
!
print '( "right hand side:" ,100f6.1)', xy
!
call solve_dense_gs(neq, xx, xy, sol) !solution by Gauss-Seidel
print '( "solution:" ,100f7.3)', sol
contains
subroutine find_addresses
    integer :: i
    do i=1,neff
        select case (effecttype(i))
            case (effcross)
                address(i)=sum(nlev(1:i-1))+indata(i)
                weight_cov(i)=1.0
            case (effcov)
                weight Cov(i)=indata(i)
                if (nested cov(i)==0) then
                    address(i)=sum(nlev(1:i-1))+1
                elseif (nested cov(i)>0 .and. nested cov(i).lt.neff) then
                    address(i)=sum(nlev(1:i-1))+indata(nested cov(i))
                else
                    print*, 'wrong description of nested covariate'
                    stop
                endif
            case default
                print*, 'unimplemented effect ',i
                stop
        end select
    end do
end subroutine

end program l sqr

subroutine solve_dense_gs(n,lhs,rhs,sol)
! finds sol in the system of linear equations: lhs*sol=rhs
! the solution is iterative by Gauss-Seidel
    integer :: n
    real :: lhs(n,n),rhs(n),sol(n),eps
    integer :: round
!
    round=0
    do
        eps=0; round=round+1
        do i=1,n
            if (lhs(i,i).eq.0) cycle
            sol new=sol(i)+(rhs(i)-sum(lhs(i,:)*sol))/lhs(i,i)
            eps=eps+ (sol(i) -solnew)**2
            sol(i)=sol new
        end do
        if (eps.lt. 1e-10) exit
        end do
        print*, 'solutions computed in ',round,' rounds of iteration'
    end subroutine
Multiple trait least squares

Asume initially that the same model applies to all traits and that all traits are recorded. The general model is the same:

$$y = X\beta + e, \quad \text{but} \quad V(y) = V(e) = R = R_0 \otimes I$$

and can be decomposed into $t$ single-trait equations with correlated residuals:

$$y_1 = X_1\beta_1 + e_1$$
$$y_2 = X_2\beta_2 + e_2$$
.................
$$y_t = X_t\beta_t + e_t$$

where $t$ is the number of traits, and $V(y) = V(e) = R$

and the normal equations are $XR^{-1}X \beta = XR^{-1}y$

The detailed equations will depend on whether the equations are ordered by effects within traits or by traits within effects. In the first case, assuming all design matrices are equal

$$X_0 = X_1 = X_2 = ... = X_t$$

and all traits are recorded, then:

$$\begin{bmatrix}
X_0'^{-1}X_0 & X_0'^{-1}X_1 \\
. & . \\
X_0'^{-1}X_t & X_0'^{-1}X_t
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
. \\
\beta_t
\end{bmatrix}
= 
\begin{bmatrix}
X_0'^{-1}y_1 + ... + X_0'^{-1}y_t \\
. \\
X_0'^{-1}y_t + ... + X_0'^{-1}y_t
\end{bmatrix}$$

and

$$\begin{bmatrix}
\mathbf{r}_{11} & \mathbf{r}_{1t} \\
. & . \\
\mathbf{r}_{t1} & \mathbf{r}_{tt}
\end{bmatrix}
= R_0^{-1}$$

where $R_0$ is a $t \times t$ matrix of residual covariances between the traits in one observation.
Using the direct product notation that

\[ \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix}
    b_{11}A & \cdots & b_{1n}A \\
    \vdots & \ddots & \vdots \\
    b_{m1}A & \cdots & b_{mn}A
\end{bmatrix} \]

this system of equations could be presented as:

\[
(\mathbf{X}_0' \mathbf{X}_0) \otimes \mathbf{R}_0 \mathbf{\beta} = (\mathbf{X}_0' \otimes \mathbf{R}_0) \mathbf{y}
\]

where as if the equations were ordered by traits within effects,

\[
\mathbf{R}_0 \otimes (\mathbf{X}_0' \mathbf{X}_0) \mathbf{\beta} = (\mathbf{R}_0 \otimes \mathbf{X}_0') \mathbf{y}
\]

Please note that \( \beta \)'s in both examples are not ordered the same way.

In the last case, the rules for creating normal equations are similar to those for single-trait models except that:
- LHS: Instead of adding 1, add \( \mathbf{R}_0^{-1} \)
- RHS: Instead of adding a scalar \( y_i \), add \( \mathbf{R}_0^{-1} y_i \), where \( y_i \) is a \( t \times 1 \) vector of data for observation \( i \).

**Example**

Assume the same data as before but with two traits

<table>
<thead>
<tr>
<th>Data</th>
<th>( i )</th>
<th>( j )</th>
<th>( y_{1ij} )</th>
<th>( y_{2ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

and assume that the correlations between the traits are
The mixed model for the given data:

\[
\mathbf{R}_0 = \begin{bmatrix} 0.4286 & -0.1429 \\ -0.1429 & 0.7143 \end{bmatrix} \quad \text{so that} \quad \mathbf{R}_0^{-1} = \begin{bmatrix} 2.5000 & 0.5000 \\ 0.5000 & 1.5000 \end{bmatrix}
\]

The mixed model for the given data:

\[
\begin{bmatrix} 5 \\ 2 \\ 8 \\ 4 \\ 7 \\ 3 \\ 9 \\ 5 \\ 6 \\ 1 \end{bmatrix} = \left( \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} + \ldots \right) \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \\ b_{11} \\ b_{21} \\ b_{12} \\ b_{22} \\ b_{13} \\ b_{23} \end{bmatrix} + \mathbf{e}
\]

\[
y = (\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \ldots ) \beta + \mathbf{e}
\]

Observation: first second third

For the first two observations:
The bold zeros in the last matrix are in fact 2x2 matrices of zeros.

\[
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
0 & 0 \\
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}
= \begin{bmatrix}
R_0^{-1} & 0 & R_0^{-1} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

The bold zeros are 2x1 vectors of zeros.

\[
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
0 & 0 \\
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
R_0^{-1} \\
\frac{5}{2}
\end{bmatrix}
= \begin{bmatrix}
R_0^{-1} & \frac{5}{2} \\
0 & 0
\end{bmatrix}
= \begin{bmatrix}
13.5 \\
4.5 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

The bold zeros are 2x1 vectors of zeros.

**Rules**

Let address(effect,trait) be a function returning the addresses of level

! repeat these loops for each observation
!
\[
do \text{e1}=1,\text{neffect}
\text{do} \text{e2}=1,\text{neffect}
\text{do} \text{t1}=1,\text{ntrait}
\]
do t2=1,ntrait
  i=address(e1,t1)
  j=address(e2,t2)
  XX_{i,j} = XX_{i,j} + r^{t1,t2}
enddo t2
enddo t1
enddo e2
enddo e1

do e1=1,neffect
  do t1=1,ntrait
    do t2=1,ntrait
      i=address(e1,t1)
      XY_i = XY_i + r^{t1,t2} * y_{t2}
    enddo t1
  enddo e2
enddo e1

**Missing traits**

If some traits are missing, the general rules apply but $R_0^{-1}$ is replaced by $R^m$, a matrix specific for each missing-trait pattern $m$. Let $Q_m$ be a diagonal matrix that selects which traits are present. For example, this matrix selects traits 1, 3 and 4. If some traits are missing, for each pattern of missing traits $m$, replace

$$R_0^{-1} \quad \text{by} \quad R^m = (Q_m R_o Q_m)^{-1}$$

$Q_m R_o Q_m$ can be created by zeroing rows and columns of $R_o$ corresponding to missing traits.

Because then contributions due to missing equations to LHS and RHS are always zero, a computer program may be made more efficient by modifications so that the zero contributions are never added. Also, $R_m$’s can be precomputed for all combinations of missing traits. This is important if the number of traits is large.
Different models per trait

There are two ways of supporting such models:

1. By building the equations within models.
   This is the most flexible way but does not allow for easy utilization of blocks for same effects within traits. Makes iterative methods to solve MME (mixed model equations) slow.

2. By building the equations within traits, i.e., declaring same model for each trait and then selectively nulling unused equations
   This is more artificial way but results in simpler programs and possible more efficient CPU-wise (but not memory-wise).

Example of ordering by models and by traits

Assume the following models for the three beef traits:

Birth weight \[ = cg + \alpha \text{ age}_\text{dam} + \text{an} + \text{mat} + e \]
(150) (1) (1000) (1000)

Weaning weight \[ = cg + \text{an} + \text{mat} + e \]
(100) (1000) (1000)

Yearling weight \[ = cg + \text{an} + e \]
(50) (1000)

where \( \alpha \) is a coefficient of regression on age of dam and \( cg \) is contemporary group, \( \text{an} \) is animal direct and \( \text{mat} \) is animal maternal, and number of levels are in ( ).

Let \( x_{i,j} \) be the \( j \)-th level of \( i \)-th trait of effect \( x \). The order of the equations would be as follows:

**Ordering within traits**

**Ordering within models**
| cg_{1,1} | cg_{1,2} | \text{...} | cg_{1,150} | \alpha | an_{1,1} | an_{1,2} | \text{...} | an_{1,1000} | mat_{1,1} | mat_{1,2} | \text{...} | mat_{1000} | cg_{2,1} | cg_{2,2} | \text{...} | cg_{2,100} | an_{2,1} | an_{2,2} | \text{...} | an_{2,1000} | mat_{1,1} | mat_{1,2} | \text{...} | mat_{1000} | cg_{3,1} | cg_{3,2} | \text{...} | cg_{3,50} | an_{3,1} | an_{3,2} | \text{...} | an_{3,1000} |

Please note that with ordering within models, only existing effects are defined. With ordering within effects, for regularity all 3 traits have 150 levels of cg and all have the maternal effect defined.
Computer program

This program supports multiple-trait least squares with support for missing traits and missing effects. Lines changed from the previous program are highlighted.

program lsqmt
implicit none
!
! As lsqr but with support for multiple traits
!
integer,parameter::effcross=0,& ! effects can be cross-classified
  effcov=1 ! or covaribles
real, allocatable:: xx(:,), xy(:,), sol(:,) ! storage for the equations
integer, allocatable:: indata(:,) ! storage for one line of ! effects
integer,parameter::neff=2,& ! number of effects
  nlev(2)=(/2,3/),& ! number of levels
  ntrait=2,& ! number of traits
  miss=0 ! value of missing trait/effect
real  ::
  r(ntrait,ntrait)=& ! residual covariance matrix
  reshape((/1,2,2,5/),(/2,2/)),& ! values 1, 2, 2 and 5
  rinv(ntrait,ntrait) ! and its inverse
integer :: effecttype(neff)=(/effcross, effcov/)
integer :: nestedcov(neff) =(/0,0/)
real :: weight_cov(neff,ntrait)
real :: y(ntrait) ! observation value
integer :: neq, io, i, j, k, l ! number of equations and i/o-status
integer, allocatable:: address(:, :) ! start and address of each effect
!
neq=ntrait*sum(nlev)
allocate (xx(neq, neq), xy(neq), sol(neq), indata(neff*ntrait,:),
  address(neff, ntrait))
xx=0; xy=0
!
open(1, file='data_pr3')
!
do
  read(1, *, iostat=io) indata, y
  if (io.ne.0) exit
  call find_addresses
  call find_rinv
  do i=1,neff
    do j=1,neff
      do k=1,ntrait !number of equations and i/o-status
        do l=1,ntrait
          xx(address(i,k),address(j,l))=xx(address(i,k),address(j,l))+&
            weight_cov(i,k)*weight_cov(j,l)*rinv(k,l)
        enddo
      enddo
    enddo
  enddo
endo
endo
do k=1,ntrait
do l=1,ntrait
    xy(address(i,k))= xy(address(i,k))+ rinv(k,l)*y(l)*weight_cov(i,k)
  enddo
  enddo
  enddo
!
print*,'left hand side'
do i=1,neq
  print '(100f5.1)',xx(i,:)
enddo
!
print '(" right hand side:",100f6.1)',xy
!
call solve_dense_gs(neq,xx,xy,sol) !solution by Gauss-Seidel
print '(" solution:",100f7.3)',sol
contains

subroutine find_addresses
integer :: i,j,baseaddr
do i=1,neff
  do j=1,ntrait
    if (datum(i,j) == miss) then  !missing effect
      address(i,j)=0  !dummy address
      weight_cov(i,j)=0.0
      cycle
    endif
    baseaddr=sum(nlev(1:i-1))*ntrait+j  !base address (start)
    select case (effecttype(i))
      case (effcross)
        address(i,j)=baseaddr+(datum(i,j)-1)*ntrait
        weight_cov(i,j)=1.0
      case (effcov)
        weight_cov(i,j)=datum(i,j)
        if (nestedcov(i) == 0) then
          address(i,j)=baseaddr
        elseif (nestedcov(i)>0 .and. nestedcov(i).lt.neff) then
          address(i,j)=baseaddr+(datum(nestedcov(i),j)-1)*ntrait
        else
          print*,"wrong description of nested covariable"
          stop
        endif
      case default
        print*,"unimplemented effect ",i
        stop
    end select
  enddo
endo
end subroutine

function datum(ef,tr)
real :: datum
integer :: ef,tr
! calculates the value effect ef and trait tr
datum = indata(ef + (tr-1)*neff)
end function

subroutine find_rinv
! calculates inv(Q R Q), where Q is an identity matrix zeroed for
! elements corresponding to y(i)=miss
integer :: i,irank
real :: w(10)
rinv=r
do i=1,neff
  if (y(i) == miss) then
    rinv(i,:)=0; rinv(:,i)=0
  endif
enddo
call ginv(rinv,ntrait,1e-5,irank)
end subroutine

end program lsqmt

subroutine solve_dense_gs(n,lhs,rhs,sol)
! finds sol in the system of linear equations: lhs*sol=rhs
! the solution is iterative by Gauss-Seidel
integer :: n
real :: lhs(n,n),rhs(n),sol(n),eps
integer :: round
!
round=0
do eps=0; round=round+1
  do i=1,n
    if (lhs(i,i).eq.0) cycle
    solnew=sol(i)+(rhs(i)-sum(lhs(i,:)*sol))/lhs(i,i)
    eps=eps+ (sol(i) -solnew)**2
    sol(i)=solnew
  end do
  if (eps.lt. 1e-10) exit
end do
print*, 'solutions computed in ',round,' rounds of iteration'
end subroutine

Below is a generalized-inverse subroutine that needs to be compiled with the previous program.

subroutine ginv(a,n,tol,rank)
! returns generalized inverse of x(n,n). tol is working zero
! and irank returns the rank of the matrix. rework of rohan fernando's
! f77 subroutine by i. misztal 05/05/87-05/23/00
implicit none
integer n
real a(n,n),w(n),tol,rank
integer i,ii,j
!
rank=n
do i=1,n
do  j=1,i-1
   a(i:n,i)=a(i:n,i)-a(i,j)*a(i:n,j)
enddo
if (a(i,i).lt.tol) then
   a(i:n,i)=0.0
   rank=rank-1
else
   a(i,i)=sqrt(a(i,i))
a(i+1:n,i)=a(i+1:n,i)/a(i,i)
endif
enddo
do i=1,n
if (a(i,i).eq.0.) then
   a(i+1:n,i)=0
else
   a(i,i)=1.0/ a(i,i)
w(i+1:n)=0
   do ii=i+1,n
      w(ii:n)=w(ii:n)-a(ii:n,ii-1)*a(ii-1,i)
      if (a(ii,ii).eq.0.) then
         a(ii,i)=0.
      else
         a(ii,i)=w(ii)/a(ii,ii)
      endif
   enddo
endif
enddo
do j=1,n
   do i=j,n
      a(i,j)=dot_product(a(i:n,j),a(i:n,i))
   enddo
enddo
do i=1,n
   a(i,i+1:n)=a(i+1:n,i)
enddo
end
Mixed models

The traditional notation
\[ y = X\beta + Zu + e, \quad E(y) = X\beta, \quad V(y) = ZGZ' + R, \quad V(u) = G, \quad V(e) = R \]
gives little detail on differences between least squares and mixed models. The mixed model equations (MME) are:
\[
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}Z \\
Z'R^{-1}X & Z'R^{-1}Z + G^{-1}
\end{bmatrix}
\begin{bmatrix}
\beta \\
u
\end{bmatrix}
= 
\begin{bmatrix}
X'R^{-1}y \\
Z'R^{-1}y
\end{bmatrix}
\]
and sometimes are presented as
\[
(W'R^{-1}W + G^{-1}) \hat{t} = W'R^{-1}y
\]
The only difference between mixed and fixed model equations is the inclusion of \( G^{-1} \). Otherwise \( W's \) are created the same way as \( X \) before. The main problem in mixed model equations is figuring out what \( G^{-1} \) is and how to create it.

Usually, \( G^{-1} \) can be decomposed into contributions from each effect or a combination of a few effects. For example, let the random effects be \( u = [p, a, m, h]' \), where \( p \) is permanent environment, \( a \) is animal direct, \( m \) is animal maternal, and \( h \) is herd by sire interaction. Then:
\[
G^{-1} = \begin{bmatrix}
G_p^{-1} & 0 & 0 \\
0 & [G_{aa} \quad G_{am}]^{-1} & 0 \\
0 & 0 & G_h^{-1}
\end{bmatrix}
\]
and contributions to \( G^{-1} \) can be calculated for each individual effect or groups of effects separately.

**Popular contributions to random effects - single trait**

1. Effects are identically and independently distributed (IID)
   \( G_i = I \sigma^2_i \), where \( \sigma^2_i \) is variance component for trait \( i \). Then \( G_i^{-1} = I/\sigma^2_i \). In this case, scalars \( 1/\sigma^2_i \) are added to each diagonal element of LHS corresponding to effect \( i \).

   Popular IID effects are permanent environment and sire effect if relationships among sires are ignored.
2. Numerator relationship matrices
   - animal additive effect
     \[ G_i = A \sigma^2_i, \quad G_i^{-1} = A^{-1} 1/\sigma^2_i \]
   - sire additive effect
     \[ G_i = A_s \sigma^2_i, \quad G_i^{-1} = A_s^{-1} 1/\sigma^2_i \]
   - nonadditive genetic effects

3. Autoregressive structure matrix
   When adjacent levels of a random effect are correlated to the same degree, that effect may
   have an autoregressive structure (Wade and Quaas, 1993):
   \[ G_i = H \sigma^2_i, \quad G_i^{-1} = H^{-1} 1/\sigma^2_i \]
   This structure is useful for smoothing trends, e.g., in genetic groups, or accounting for
   correlations among observations in adjacent management levels.

4. Nonadditive effects.
   Formulas are available to obtain the inverse of the dominance (VanRaden and Hoeschele,
   1991) and additive x additive (Hoeschele and VanRaden, 1991) relationship matrices.

Distributions involving the \( I \) and \( A \) matrices are most popular.

**Numerator relationship matrix \( A \)**

For each individual \( i \) with sire \( s \) and dam \( d \), the following recursive equation applies

\[ a_i = \frac{1}{2}(a_s + a_d) + m_i \]

where \( m_i \) is mendelian sampling and \( \text{var}(m_i) = \frac{1}{2} \sigma^2_s (1-F) \)

If only parent \( p \) is known: \[ a_i = \frac{1}{2}a_p + m_i, \quad \text{var}(m_i) = \frac{3}{4} \sigma^2_s \]
If neither parent is known: \[ a_i = m_i, \quad \text{var}(m_i) = \sigma^2_s \]

In a connected population, the matrix \( A \) has all elements different from 0, and it is impossible to
create directly. For example, \( A \) with 1,000,000 elements would have approximately \( 10^{12} \) nonzero
elements, too many to store anywhere. Fortunately Henderson (1985) has found that \( A^{-1} \) has very
few nonzeros (at most 9 times the number of animals) and that it can easily be created.
For all animals

\[ a = P a + m \]

where \( P \) is a matrix containing \( \frac{1}{2} \) and 0 that relates animals and parents, and \( m \) is the diagonal
matrix containing \( \frac{1}{2}\sigma^2_s, \frac{3}{4} \sigma^2_s \) and \( \sigma^2_s \) dependent on how many parents are known.
\((I-P) a = m, \quad a = (I-P)^{-1} m,\)

\[\text{var}(a) = A \sigma_a^2 = (I-P)^{-1} \text{Var}(m) [(I-P)^{-1}]',\]

and

\[A^{-1} = \sigma^2 (I-P)' \text{Var}(m)^{-1} (I-P)\]

The only matrix left to invert is a diagonal matrix \(\text{Var}(m)\)

Rules to create \(A^{-1}\) directly

\[
A^{-1} = \sum_{\text{animals}} A_i^{-1}
\]

or the whole \(A^{-1}\) can be calculated from separate contribution from each animal in the pedigree. Let \(\text{address}(a_i)\) point to the equation number of animal \(i\), \(\text{address}(a_{sire})\) and \(\text{address}(a_{dam})\) to equation numbers of sire and dam of animal \(i\) if present, and \(\text{address}(a_{parent})\) point to equation number of parent \(p\) if the other parent is unknown. The contributions to \(A^{-1}\) from each animal would be

<table>
<thead>
<tr>
<th>(\text{address}(a_i))</th>
<th>(\text{address}(a_{sire}))</th>
<th>(\text{address}(a_{dam}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(2)</td>
<td>(-1)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(-1)</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(-1)</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
</tbody>
</table>

Please note that these contributions are results of \((I-P) \text{Var}(m)^{-1} (I-P)'\) for one animal:

\[
\begin{bmatrix}
1 \\
-\frac{1}{2} \\
-\frac{1}{2}
\end{bmatrix}
2
\begin{bmatrix}
1 & -\frac{1}{2} & -\frac{1}{2}
\end{bmatrix}
\]

where the coefficient 2 comes from \(\text{var}(m_i)^{-1} = (\frac{1}{2} \sigma^2)^{-1} = 2/\sigma^2\).

When only one parent is known, the contributions are:
Finally, for animals with no parent known, the contribution is:

\[
\text{address}(a_j) \cdot \text{address}(a_p).
\]

\[
\text{address}(a_j) \cdot 4/3 \cdot -2/3.
\]

\[
\text{address}(a_p) \cdot -2/3 \cdot 1/3.
\]

Finally, for animals with no parent known, the contribution is:

\[
\text{address}(a_j) \cdot 1.
\]

The contributions are done for each and every animal in the pedigree. For example, if one animal is present in the pedigree as a parent only, a separate equation should be created for this animal and a contribution to it added.

Question: explain why the largest contributions are 2, 4/3 and 1!

Unknown parent groups

\[a_i = \frac{1}{2}(a_a + a_d) + m_i\]

If one or two parents are unknown, the merit for the unknown parent is assumed to be 0 and equal for all unknown animals, regardless in what year their progeny was born, sex and origin of genes (domestic or international). Let’s assume that the genetic merit of unknown parent \(p\) is \(g_p\), and that now the animal effect accounts for unequal merit of unknown parents. Now:

\[u_i = \frac{1}{2}(u_a + u_d) + m_i, \quad \text{var}(m_i) = \frac{1}{2} \sigma_a^2 (1 - F_s/2 - F_d/2)\]

where \(F\) are coefficients of inbreeding.

If only parent \(p_1\) is known: \[u_i = \frac{1}{2}(u_{p1} + g_{p2}) + m_i, \quad \text{var}(m_i) = \frac{1}{4} \sigma_a^2\]

If neither parent is known: \[u_i = \frac{1}{2}(g_{p1} + g_{p2}) + m_i, \quad \text{var}(m_i) = \sigma_a^2\]
The contributions with the unknown parent groups are simpler than without the groups:

\[
\begin{array}{cccc}
\text{address}(a) & \text{address}(a_s \backslash V_d) & \text{address}(a_s \backslash V_d) \\
\text{address}(a) & . & 2k & . \\
\text{address}(a_s \backslash V_d) & . & -k & . \\
\text{address}(a_s \backslash V_d) & . & -k & . \\
\text{address}(a_s \backslash V_d) & . & 2k & . \\
\text{address}(a_s \backslash V_d) & . & 2k & . \\
\end{array}
\]

where \( k = \frac{2}{4 \text{ - number of known parents}} \)

Groups have equations in the animal effect but are not treated as animals, i.e., they have no parents and 1 is not added to their diagonal as for animals without known parents.

**Numerator relationship matrix in the sire model**

The derivations below assume the absence of inbreeding, which cannot be calculated accurately in the sire model.

For each bull \( i \) with sire \( s \) and maternal grandsire (MGS) \( t \), the following recursive equation applies:

1) If sire and MGS known: 
   \[ s_i = \frac{1}{2}(s_i + \frac{1}{2}s_d) + m_i, \quad \text{var}(m_i) = \frac{11}{16} \sigma_s^2 \]

2) If only sire \( s \) is known: 
   \[ s_i = \frac{1}{2}s_i + m_i, \quad \text{var}(m_i) = \frac{3}{4} \sigma_s^2 \]

3) If only MGS is known: 
   \[ s_i = \frac{1}{4}s_d + m_i, \quad \text{var}(m_i) = \frac{15}{16} \sigma_s^2 \]

4) If neither parent is known: 
   \[ s_i = m_i, \quad \text{var}(m_i) = \sigma_s^2 \]

The contributions to \( A_s^{-1} \) are done as for \( A^{-1} \) except that the values are:

1) \[ \begin{bmatrix} 16/11 & -8/11 & -4/11 \\ -8/11 & 8/11 & 4/11 \\ -4/11 & 2/11 & 1/11 \end{bmatrix} \]

2) \[ \begin{bmatrix} 4/3 & -2/3 \\ -2/3 & 1/3 \end{bmatrix} \]

3) \[ \begin{bmatrix} 16/15 & -4/15 \\ -4/15 & 1/15 \end{bmatrix} \]

These ratios are the result of multiplication of
where \( k \) is 16/11, 4/3, 16/15 and 1 for cases 1-4, respectively.

**Autoregression**

For \( t \) generated by a first-order autoregressive process:

\[
t = \begin{cases} 
\epsilon_k & \text{for } k = 1 \\
pt_{k-1} + \epsilon_k & \text{for } k > 1 
\end{cases}
\]

with

\[
E(t_k) = 0, \quad \text{var}(t_k) = \sigma_t^2, \quad \text{cov}(t_k, t_{k+1}) = p\sigma_t^2, \quad \text{cov}(\epsilon_k, \epsilon_{k+1}) = 0 \text{ for } k \neq k'
\]

\[
\text{var}(\epsilon_1) = \sigma_t^2, \quad \text{var}(\epsilon_k) = (1-p^2)\sigma_t^2, \text{ for } k > 1
\]

where \(-1 < p < 1\) is coefficient of autocorrelation and \( \epsilon_k \) is error term. Similar to the additive effects, the recurrence equations can be established as:

\[
t = Pt + \epsilon
\]

where \( P \) is a matrix containing 1 and 0 that relate subsequent levels of \( t \). Then:

\[
(I-P) t = \epsilon, \quad t = (I-P)^{-1} \epsilon,
\]

\[
\text{var}(t) = H\sigma_t^2 = [(I-P)^{-1}]' \text{Var}(\epsilon) (I-P)^{-1}
\]

and

\[
H^{-1} = \sigma_t^2 (I-P)' \text{Var}(\epsilon)^{-1} (I-P)
\]

The only matrix left to invert is a diagonal matrix \( \text{Var}(\epsilon) \). While the matrix \( H \) has all nonzero elements:
The autocorrelation structure has been used to smooth effects such as unknown parent groups or year-months within each herd. In the case of year months, there may be observations only in very few year-months. Then, the matrix $H = \{h\}$ can be constructed in such a way so that all unneeded levels are skipped, by using the following rules:

$$H^{-1} = \begin{bmatrix}
1 & -\rho & 0 & \ldots & 0 & 0 \\
-\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & 1+\rho^2 & -\rho \\
0 & 0 & 0 & \ldots & -\rho & 1
\end{bmatrix} \frac{1}{1-\rho^2}$$

The autocorrelation structure has been used to smooth effects such as unknown parent groups or year-months within each herd. In the case of year months, there may be observations only in very few year-months. Then, the matrix $H^{-1} = \{h\}$ can be constructed in such a way so that all unneeded levels are skipped, by using the following rules:

$$h^{11} = 1, \quad h^{nn} = \phi_{n-1}$$

$$h^{k,k} = \phi_k \phi_{k+1} - 1, \quad h^{k,k+1} = \phi_k \psi_{k+1} \quad \text{for } k=2,\ldots,n-1$$

where:

$$\phi_k = \frac{1}{1-\rho^{2d_k}}, \quad \psi_k = -\rho^{d_k}$$

and $d_k$ is the distance between levels $k$ and $k+1$.

### Variance ratios

In single trait MME, the residual variance can be factored out from the equations:

$$(W'R^{-1}W + G^{-1}) \hat{r} = W'R^{-1}y$$
\[
(\sigma^2 \mathbf{W}'\mathbf{W} + G^{-1}) \mathbf{\hat{i}} = \sigma^2 \mathbf{W}'\mathbf{y}
\]

\[
= =
(\mathbf{W}'\mathbf{W} + \sigma^2 \mathbf{G}^{-1}) \mathbf{\hat{i}} = \mathbf{W}'\mathbf{y}
\]

If in the original MME, components of \(G_i\) had the form of 1/\(\sigma_i^2\) \(\mathbf{I}\) or 1/\(\sigma_i^2\) \(\mathbf{A}\) after the premultiplication, they will have the form of \(\sigma_i^2/\sigma_0^2\) \(\mathbf{I}\) or \(\sigma_i^2/\sigma_0^2\) \(\mathbf{A}\), where \(k_i\) are variance ratios. Thus only variance ratios are necessary to set up the MME and not the absolute variances.

**Examples of calculation of the variance ratios**

Assume that there is only one random effect in the model, either the sire effect or the animal effect. Let us calculate the variance ratios given the heritability \(h^2\).

**animal model**

\[
h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} = \frac{1}{1 + \frac{\sigma_e^2}{\sigma_a^2}} = \frac{1}{1 + k_a} \quad \Rightarrow \quad k_a = \frac{1}{h_a^2 - 1}
\]

**sire model**

\[
k_s = \frac{\sigma_s^2 + \frac{3}{4}\sigma_a^2}{\frac{3}{4}\sigma_a^2} = \frac{4(\frac{\sigma_a^2}{\sigma_s^2} + \frac{3}{4})}{\frac{3}{4}} = \frac{4}{h^2 - 3}
\]

For \(h^2 = .25\), \(k_a = 3\) and \(k_s = 13\).
Covariances between effects

Sometimes two random effects are correlated, for example direct additive and maternal additive.

Let the variance and covariances for animal \(i\) between its direct and maternal effects be:

\[
\text{Var} \begin{bmatrix} a_i \\ m_i \end{bmatrix} = \begin{bmatrix} \sigma_{aa} & \sigma_{am} \\ \sigma_{ma} & \sigma_{mm} \end{bmatrix} = \begin{bmatrix} g_{aa} & g_{am} \\ g_{ma} & g_{mm} \end{bmatrix} = G_0
\]

The variance for all animals will be

\[
\text{Var} \begin{bmatrix} a \\ m \end{bmatrix} = A \otimes G_0 = \begin{bmatrix} g_{aa} A & g_{am} A \\ g_{ma} A & g_{mm} A \end{bmatrix}
\]

The inverse will be

\[
\left( \text{Var} \begin{bmatrix} a \\ m \end{bmatrix} \right)^{-1} = A^{-1} \otimes G_0^{-1} = \begin{bmatrix} g_{aa} A^{-1} & g_{am} A^{-1} \\ g_{ma} A^{-1} & g_{mm} A^{-1} \end{bmatrix}, \text{ where } G_0^{-1} = \begin{bmatrix} g_{aa} & g_{am} \\ g_{ma} & g_{mm} \end{bmatrix}
\]

The algorithm to add relationship contributions to these two correlated effects would be as follows:
1. Invert co-variance matrix between the effects \(G_0\),
2. Add \(A^{-1}\) multiplied by:
   a) \(g_{aa}\) to the block corresponding to the direct effect,
   b) \(g_{am}\) to the block corresponding to intersections of direct and maternal effects
   c) \(g_{mm}\) to the block corresponding to the maternal effect

Random effects in multiple traits

Let \(a_i\) be a solution for trait \(i\), \(i=1,\ldots,t\), and let

\[
\text{Var}(a_i) = g_{ii}, \quad \text{Var}(a_i, a_j) = g_{ij}
\]

The variance covariance matrix for \(a\)'s is similar to that of correlated effects above:
The inverse will be

\[
\begin{pmatrix}
  a_1 \\
  a_2 \\
  \vdots \\
  a_i
\end{pmatrix}^{-1} = A^{-1} \otimes G_0^{-1} =
\begin{bmatrix}
  g_{11}A^{-1} & g_{12}A^{-1} \\
  \vdots & \vdots \\
  g_{i1}A^{-1} & g_{i2}A^{-1}
\end{bmatrix}, \text{ where } G_0^{-1} =
\begin{bmatrix}
  g_{11} & g_{12} \\
  \vdots & \vdots \\
  g_{i1} & g_{i2}
\end{bmatrix}
\]

The algorithm to add distribution contributions to correlated traits would be similar to that for correlated effects:
1. Invert co-variance matrix between the traits \( G_0 \),
2. Add \( A \) multiplied by:
   a) \( g^i \) to the block corresponding to trait \( i \),
   b) \( g^j \) to the blocks on intersections of traits \( i \) and \( j \).

**Multiple traits and correlated effects**

For simplicity, assume only two traits: 1 and 2, and two effects: \( a \) and \( b \).

Let \( a_{ij} \) be a vector of random effect \( i, i=a,b,\ldots \) and trait \( j=1,2,\ldots \). Assume that all effects have the same number of levels,
\[
\text{var}(a_{ij})=g_{ij}^iA, \text{ and }\]
\[
\text{var}(a_{ik}, a_{kl})=g_{ik,l}A
\]

The variance covariance matrix for \( a \)'s is similar to that of correlated effects above:
\[
\begin{bmatrix}
  a_{11} \\
  a_{12} \\
  a_{21} \\
\end{bmatrix}
= \mathbf{A} \otimes \mathbf{G}_0
= \begin{bmatrix}
  g_{11,11} \mathbf{A} & g_{11,12} \mathbf{A} & g_{11,21} \mathbf{A} \\
  g_{12,11} \mathbf{A} & g_{12,12} \mathbf{A} & g_{12,21} \mathbf{A} \\
  \vdots & \vdots & \vdots \\
  g_{21,11} \mathbf{A} & g_{21,12} \mathbf{A} & g_{21,21} \mathbf{A} \\
\end{bmatrix}
\]

The remaining steps are similar as in the previous two sections.

Please note that for \( t \) traits and \( f \) correlated effects, the size of \( \mathbf{G}_0 \) will be \( tf \times tf \). For example for \( t=4 \) and \( f=2 \), the dimension will be \( 8 \times 8 \) or 36 variance components.
**Computer pseudo-code**

Let $\text{add}(i,j,k)$ be address of an equation of $i$-th effect, $j$-th level and $k$-th trait.

Let $a_{ij}$ be the value of $A^{-1}$ for the $i$-th row and $j$-th column.
Let $g_{ijk}$ be the value of $G_{0}^{-1}$ corresponding to $i$-th effect and $j$-th trait.

The following code would create corresponding contributions to the LHS of the MME.

```fortran
do t1=1,ntrait
  do t2=1,ntrait
    do e1="correlated effects"
      do e2="correlated effects"
        LHS(block t1,t2,e1,e2) = LHS(block t1,t2,e1,e2) + A^{-1} \cdot g_{t1,t2}^{e1,e2}
      enddo
    enddo
  enddo
enddo
```

Below is a more detailed program for adding the animal relationship matrix. “Number of correlated effects” will be 1 for effects uncorrelated with other effects, 2 if the currently processed effect is correlated with the next one, etc.

```fortran
real, parameter::w=(/1, -.5, -.5/) !matrix generating contributions to $A^{-1}$
...........................
do 
  read animal, sire, dam and npar !npar=number of parents known
    if end of file then quit loop
    p(1)=animal; p(2)=sire; p(3)=dam
    do i=0,number of correlated effects - 1
      do j=0,number of correlated effects - 1
        do t1=1,ntrait
          do t2=1,ntrait
            do k=1,3
              do l=1,3
                m=address(effect+i,t1,p(k))
                n=address(effect+j,t2,p(l))
                xx(m,n)=xx(m,n)+g(effect+j,t1,effect+k,t2)* w(k)*w(l)*4./(4.-npar)
              enddo
            enddo
          enddo
        enddo
      enddo
    enddo
  enddo
```


Computer program

Program LSQMT was modified to Program BLUP by including contributions for random effects. The following conventions apply:

- \( g(i,:,:) \) contains variances for effect \( i \)
- \( \text{randomtype}(i) \) contains types of random effects, as shown below
- \( \text{randomnumb}(i) \) shows how many successive random effects are correlated. For example, if \( \text{randomnumb}(3)=2 \), then contributions will also be made to effect 4 and blocks of effects 3 and 4. If \( \text{randomnumb}(i)=p>1 \), then \( g(i,:,:) \) should have a rank of \( p*n\text{trait} \); the ordering of \( g(i,:,:) \) is then within traits and by effects.

program BLUP1

! As lsqmt but with support for random effects

........................
........................
! Types of random effects
integer, parameter :: g_fixed=1,& ! fixed effect
g_diag=2, & ! diagonal
g_A=3, & ! additive animal
g_As=4 ! additive sire

integer :: randomtype(neff),& ! status of each effect, as above
randomnumb(neff) ! number of consecutive correlated effects
real :: g(neff,20,20)=0 ! The random (co)variance matrix for each trait
! maximum size is 20 of trait x corr. effect combinations

........................
........................
! Assign random effects and \( G \) and \( R \) matrices
randomtype=(/g_fixed, g_A/) ! types of random effect
randomnumb=(/0, 1/) ! number of correlated effects per effect
g(2,1,1)=2; g(2,1,2)=-1; g(2,2,1)=g(2,1,2); g(2,2,2)=1
r(1,1)=1; r(1,2)=0; r(2,1)=0; r(2,2)=1
call setup_g ! invert \( G \) matrices
open(2, file=’pedname’) !pedigree file
open(1, file=’data_pr3’) !data file
!
........................................
........................................
! Random effects' contributions
do i=1,neff
select case (randomtype(i))
case (g_fixed)
  continue ! fixed effect, do nothing
case (g_diag)
call add_g_diag(i)
case (g_A, g_As)
call add_g_add(randomtype(i),i)
case default
   print*, 'unimplemented random type', randomtype(i)
endselect
enddo

........................
........................

function address1(e,l,t)
! returns address for given level l of effect e and trait t
integer :: e,l,t, address1
address1= sum(nlev(1:e-1))*ntrait+(l-1)*ntrait+t
end function

........................
........................

subroutine setup_g
! inverts g matrices
real :: w(20)
integer :: rank
do i=1,neff
   if (randomnumb(i).ne.0) then
      call printmat('original G', g(i,:, :), 20, randomnumb(i)*ntrait)
      call ginv(g(i,:,:), 20, 1e-5, rank)
      call printmat('inverted G', g(i,:, :), 20, randomnumb(i)*ntrait)
   endif
enddo
end subroutine

subroutine add_g_diag(eff)
! adds diagonal (IID) contributions to MME
integer :: eff, i,j,l,m,t1,t2
do i=1,nlev(eff)
   do j=0,randomnumb(eff)-1
      do k=0,randomnumb(eff)-1
         do t1= 1,ntrait
            do t2 =1,ntrait
               m=address1(eff+j,i,t1); l=address1(eff+k,i,t2)
               xx(m,l)=xx(m,l)+g(eff,t1+j*ntrait,t2+k*ntrait)
            enddo
         enddo
      enddo
   enddo
enddo
end subroutine

subroutine add_g_add(type,eff)
! generates contributions for additive sire or animal effect
integer :: type,eff,i,j,t1,t2,k,l,m,n,io,animal,sire,dam,par_stat,p(3)
real ::w(3), res(4)
!
select case (type)
case (g_A)
   w=(/1., -.5, -.5/)
   res=(/2., 4/3., 1., 0./)
case (g_As)
end select

w=(/1., -.5, -.25 /)
res=(/16/11., 4/3., 15/15., 1./)
end select

do
read(2,* ,iostat=io) animal, sire, dam, par_stat ! status of parents
if (io /= 0) exit
p(1)=animal
p(2)=sire
p(3)=dam
print*,p
do i=0,randomnumb(eff) - 1
  do j=0,randomnumb(eff) - 1
    do t1=1,ntrait
      do t2=1,ntrait
        do k=1,3
          do l=1,3
            if (p(k)/=0 .and. p(l)/=0) then
              m=address1(eff+i,p(k),t1)
              n=address1(eff+j,p(l),t2)
              xx(m,n)=xx(m,n)+g(eff,t1+i*ntrait,t2+j*ntrait)*&
                       w(k)*w(l)*res(par_stat)
            endif
          enddo
        enddo
      enddo
    enddo
  enddo
enddo
end subroutine

end program blup1

.....................

.....................
A new function address1 that returns an address of an equation for effect e, level l and trait t allowed for a large reduction of complexity of subroutine find_addresses although the functionality of that subroutine remained unchanged.

```fortran
subroutine find_addresses
integer :: i,j,baseaddr
do i=1,neff
    do j=1,ntrait
        if (datum(i,j) == miss) then ! missing effect
            address(i,j)=0 ! dummy address
            weight_cov(i,j)=0.0
            cycle
        endif
        select case (effecttype(i))
        case (effcross)
            address(i,j)=address1(i,datum(i,j),j)
            weight_cov(i,j)=1.0
        case (effcov)
            weight_cov(i,j)=datum(i,j)
            if (nestedcov(i) == 0) then
                address(i,j)=address1(i,1,j)
            elseif (nestedcov(i)>0 .and. nestedcov(i).lt.neff) then
                address(i,j)=address1(i,datum(nestedcov(i),j)
            else
                print*, 'wrong description of nested covariable'
                stop
            endif
        case default
            print*, 'unimplemented effect ',i
            stop
        end select
    enddo
enddo
end subroutine

function address1(e,l,t)
! returns address for given level l of effect e and trait t
integer :: e,l,t, address1
address1= sum(nlev(1:e-1))*ntrait+(l-1)*ntrait+t
end function
```
Storing and Solving Mixed Model Equations

Let \( n \) be the number of mixed model equations. The straightforward way to store the mixed model equations would be in a square matrix, and the straightforward way to solve them would be to use one of general methods such as Gaussian elimination or LU decomposition. These methods would cost:

- Memory: \( n^2 \) of double precision (8 bytes) numbers
- Arithmetic operations: \( n^3 \)

Assuming a large 800 Mbyte memory, the memory limit would be approached at \( n \approx 10,000 \). Computing the solutions would take arithmetic operations \( \approx 10^{12} \), or 11 hrs for a computer with a speed of 25 MFLOPS (million of floating point operations per second). For every problem two times larger, the memory requirements would quadruple and the computer time demands would increase 8 times!

Let count the number of nonzero elements in MME. Let the model has \( e \) effects, \( t \) traits, and let the only random effects be additive animal direct and maternal. Let the number of records be \( r \), and let the total number of animals including ancestors be \( a \). The number of nonzero contributions to the LHS of MME \( N \) would be:

\[
N < (et)^2r + 4*9 a t^2
\]

Thus the number of nonzero coefficients depends linearly on the number of animals and records. Assume that the number of records is equal to the number of animals:

\( r = a \)

and the number of equations is a function of the number of animals, say \( n = 4at \)

Then, the number of nonzero elements per one row of LHS would be:

\[
N/n < (36 + e^2)\bar{a}^2/ (4at) = t (e^2/4 + 9)
\]

For \( e=5 \) and \( t=3 \), \( N/n < 45.75 \). Assuming that each nonzero number is stored as three numbers: \( (a_{ij}, i, j) \) or in 16 bytes assuming that \( a_{ij} \) is double precision, the 800 Mbyte computer can now store over 1 million equations, an increase of 100 fold! Further storage optimizations can increase that number of few times.

Efficient computation of solutions to MME involves storing only nonzeros and doing arithmetic operations on nonzeros only.
Storage of sparse matrices

Low storage requirements for MME can be achieved in several ways:
- Exploiting the symmetry by storing upper- or lower-diagonal elements only (savings \( \approx 50\% \))
- Storing only nonzeros (savings in the thousands)
- Coding of repetitive fragments of MME

The following storage structures are popular:
1. Ordered or unordered triples: \((a_{ij}, i, j)\)
   
   For a sample matrix of
   
   \[
   \begin{bmatrix}
   101 & 20 & 30 \\
   20 & 102 & 40 \\
   30 & 40 & 103
   \end{bmatrix}
   \]
   
   the lower-diagonal storage by triples could be stored in a 3-column real matrix or one real vector and 2-column integer matrix.

   \[
   \begin{array}{ccc}
   i & j & a_{ij} \\
   1 & 1 & 101 \\
   1 & 2 & 20 \\
   1 & 3 & 30 \\
   2 & 2 & 102 \\
   2 & 3 & 40 \\
   3 & 3 & 103
   \end{array}
   \]

   Triples is the simplest form of storing sparse matrices. It is not the most memory efficient, and matrix organization cannot be easily deduced.

2. IJA where i-indices would point to elements with new rows.
Please note that now rows must be ordered.

In large matrices, IJA storage reduces the amount of memory necessary to store the row information to a negligible amount.

\[
\begin{array}{ccc}
\text{iA} & \text{ja} & a_{ij} \\
1 & 1 & 101 \\
4 & 2 & 20 \\
6 & 3 & 30 \\
7 & 2 & 102 \\
3 & 40 \\
3 & 103
\end{array}
\]

IJA is an efficient method with fast access to rows although searching for a particular column could be time consuming. Insertion of new elements is not easy with IJA and therefore this storage would mainly be used as a final form of storage, after conversion from other formats.

3. Linked lists
   In linked lists, the row pointer points to the first element of each row, and each element points to the next element, or “null” if it is the last element.
Linked list can be implemented by “pointers”, i.e., using real dynamic memory allocation, or by using vectors as below

<table>
<thead>
<tr>
<th>Element</th>
<th>Row Pointer</th>
<th>Column</th>
<th>Value</th>
<th>Next element pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>101</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>102</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>103</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

Linked list is not as memory efficient as IJA but it allows for easy insertion of new elements.

4. Trees
Trees allow for fast access to particular element but are more complicated. Each node of a tree contains three items:
  a) Data point
     - index
     - data
Tree is a flexible structure where extra elements can easily be added or eliminated. A disadvantage is memory overhead for pointers, and more complicated programming.

**Summing MME**

The loops to generate the MME create LHS elements unordered. At least 50% of all the elements are repeated. Solving algorithms mostly require the repeated coefficients summed and then sorted by row. The simplest way to do it is by sorting and summing:
- store all the unsummed and unordered coefficients on disk,
- sort by row and column
- sum repeated contributions and store them in a way convenient for obtaining the solutions.

The number of unsummed contributions can be a few times larger than the number of summed coefficients. Sorting requires scratch space of 2-3 times larger than the sorted file. Therefore this approach may require a large temporary disk space up to 10 times the final coefficient file. Also it can be slow because sorting of large files on disk is slow. Therefore this approach should be avoided if alternatives are available.

If LHS is created directly in memory in a sparse form, for each k-th contribution to row i and column j $c_{ijk}$ one needs a method to quickly search whether the storage for i-j was already created. If yes, that contribution is updated. If no, it is created. Therefore, fast searching algorithms are critical to the efficiency of setting up the MME.
Searching algorithms

For simplicity, searching methods below would search for one element a out of n element vector b.

1. Linear unsorted list

Algorithm: check each entry sequentially
Average cost: n/2 searches

2. Linked list

Linked list as shown above is many unsorted lists.
The cost would be as above, but the n would be smaller (the average number of nonzero columns)

3. Binary search

Done on a sorted vector. One divides each vector in half, and finds which half has the searched element. The division continues until the final half has only 1 element. The average cost is log₂(n), but one cannot insert new elements easily.

4. Tree

Properties of the trees are like those of a sorted vector but with capability of insertion and deletion.
The best case cost (balanced tree) is log₂(n). The worst case cost (completely unbalanced tree) is n/2. Trees have higher overhead.

5. Hash

There are many hash methods. The most popular one is presented here.
Let hash(a) be a function that:
   a) returns values being addresses to elements of vector b
   b) for similar arguments returns very different addresses

The hash method works in the following way:
   a) calculate hash(a)
   b) check if that location in vector b contains the searched item
      - if yes, exit
      - if that location is empty, return that location with status “empty”
      - if that location contain another item, add p and return to b)

The average cost of hash search is 1/η, where η is fraction of already filled elements in b. The
cost is independent of the number of elements. Hash allows for fast addition of elements but not for deletions.

For a scalar argument, the hash function can have a form

\[ \text{hash}(a) = \text{mod}(a \cdot \alpha, n) + 1, \]

where \( \alpha \) is a large prime number. In this function, a unit of change in \( a \) causes an \( \alpha \) change in hash. For multidimensional \( a \), \( \alpha \) would be a vector of large primes.
Numerical methods to solve a linear system of equations

Let \( Ax = b \) be the system of \( n \) linear equations.

This section will assume, wherever applicable, that \( A \) is LHS of MME and therefore symmetric and semipositive definite, i.e., all eigenvalues of \( A \) are nonnegative.

The methods can be partitioned into:

1. Finite, where exact solutions are obtained in finite stages
   - dense matrix
   - sparse matrix
2. Iterative, where each similar step improves the accuracy of all solutions

**Gaussian elimination**

In Gaussian elimination the number of equations is reduced one by one, until there is only one left. After solving this equation, the remaining solutions are calculated by backsolving. Gaussian elimination is non competitive with the other methods and is rarely utilized. In the past, it was used to eliminate one (and sometimes two) effects with the most numerous levels (CG) so that only a handful of equations could be solved. Absorption was useful with the sire model when the number of sires was much lower than the number of contemporary groups and when memory was limited but with the animal model the savings are limited and sometimes negative (because the number of nonzero coefficient can increase --see fill-in sparse methods). One specific application of Gaussian elimination was *absorption*, i.e., Gaussian elimination of one effect when setting up MME.

Assume a single trait model

\[
y = Hf + Wt + e
\]

where \( f \) is vector of a fixed crossclassified effect to be absorbed and \( t \) is vector of all the remaining effects. The MME are:

\[
\begin{bmatrix}
H' H & H' W \\
W' H & W' W + \mathbf{G}^{-1}
\end{bmatrix}
\begin{bmatrix}
h' \\
w'
\end{bmatrix}
= 
\begin{bmatrix}
h' y \\
w' y
\end{bmatrix}
\]

Since \( H' h + H' W \hat{t} = H' y \), \( \hat{t} = (H' H)^{-1} (H' y - H' W \hat{t}) \)

the equation after the absorption is:
\[ [W'W - W'H(H'H)^-1H'W + G^{-1}]^{-1} = [W' - W'(H'H)^-1H'\delta]y \]

The computations are as follows:

a) sort the data by effect to be absorbed,

b) read the data one level effect \( f \) at a time creating \( W'W \), \( W'y \), and accumulating:

\[
\begin{align*}
d &= (H'H)_{ii} - \text{number of observations in the current level } i \\
y_t &= (H'y)_i - \text{total for all observations in level } i \\
h_w &= (H'W)_i - \text{vector of counts between } f \text{ and } t
\end{align*}
\]

c) subtract

\[
(\text{H}'W)'\frac{y_t}{d} \text{ from } W'y \\
(\text{H}'W)'\frac{(\text{H}'W)/d}{d} \text{ from } W'W
\]

d) continue to b) until end of data

e) add \( G^{-1} \) and solve

**LU decomposition (Cholesky factorization)**

Decompose \( A = LU \), where \( L \) is lower and \( U \) is upper diagonal, and solve in 2 stages:

\( LU \ x = b \) as \( Ly = b \) and \( Ux = y \)

Solving triangular systems of equations is easy.

For symmetric matrices, \( L = U' \) such that \( LL = A \) is called the Cholesky factorization.

\( L \) is easy to derive from the recursive formulas.

\[
\frac{1}{l_{i,i}} = \sqrt{x_{i,i} - l_{i,i-1}^2} \frac{1}{l_{i,i-1}} \\
\frac{1}{l_{i,i}} = \frac{x_{i,i} - l_{i,i-1}^2}{l_{i,i-1}^2} \frac{1}{l_{i,i-1}}
\]
The computations may be done by columns or by rows

!by columns

do i=1,n

\[ l_y = \sqrt{x_{ii} - l_{ii}^i - 1} \]

do j=i+1,n

\[ l_{ji} = \frac{x_{ji} - l_{ji}^j - 1}{l_{ii}} \]

end do
end do

!by rows

do i=1,n
do j=1,i-1
    $l_{ij} = \frac{x_{ij}-l_{ij-1}l_{ij-1}}{l_{jj}}$
end do

end do

If L is replaced by X, contents of X are replaced by L (only lower diagonal).

The computations above fail when the system of equations is less than full rank. To make it work, $l_{ii}$ would be set to zero if the expression inside the square root is zero or close to zero. Computer programs that implement the Cholesky decomposition are available with the other programs as chol.f90 (by column) and cholrow.f90 (by row).

The cost of the Cholesky factorization is approximately $n^3/3$.

Cholesky factorization can be presented in a form where $l_{ii}=1$

$$A = LDL'$$

where the root-squaring operation is not necessary.

The factorization can be used to obtain the determinant of a matrix

$$|A| = |LL'| = |L||L| = \prod l_{ii}^2$$

and the inverse by solving $n$ systems of equations, each with same LHS:

$$LL' A^{-1} = I$$

Storage

Triangular
Because in the Cholesky factorization only the lower diagonal elements are used and updated, storage can be saved by using only the “triangular” storage. Unfortunately the matrix can no longer be addressed plainly as $l(i,j)$ but as $a[(i-1)i/2+j]$. Despite a multiplication and the addition in the indices, this would not slow the computations much if access is by rows and $(i-1)i/2$ is computed just once per row.

Sparse
If the matrix $A$ is sparse, some of the $L$ elements can be sparse too.
Theorem
If \(a(i,j) \neq 0\) and \(a(i,k) \neq 0\) for \(j,k > i\), then \(l(j,k) \neq 0\) even if \(a(j,k) \neq 0\).

An element nonzero in \(L\) but not in \(A\) is called *fill-in*. To minimize the number of fill-ins, equations can be reordered.

Example

Calculate \(L\) for a random model for 5 sires, each with 2 observation; variance ratio = 2.

Model I \[ y_{ij} = \mu + s_i + e_{ij} \]
Model II \[ y_{ij} = s_i + \mu + e_{ij} \]

The LHS denotes as \(A\) and the corresponding \(L\) are shown below

Both models resulted in LHS with the same number in zero. However, \(L\) in model 1 was completely filled because of many fill-ins, while \(L\) in model 2 had no fill-ins and remained sparse.

In sparse matrix factorization, the following steps are performed:

a) Finding the ordering \(Q\) so that minimizes the number of fill-ins in \(QAQ'\),
b) Preparing storage structures for \(L\)
c) Computing the factorization $LL' = QAQ'$
d) Solving the system $QAQ'Qx = Qb$ as $LL'y = c$ and $x = Q'b$

Matrix $Q$ is a permutation of the identity matrix. Thus it can be stored as a vector, and all vector multiplications will involve a vector.

Steps a) and b) need to be done just once for each nonzero structure of $A$. If only values of $A$ change, only step c) and d) need to be repeated. If only RHS changes, only step d) need to be repeated.

$$\text{cost(sparse } L) = n \left[ z(1 + f) \right]^2$$
where $z$ is the average number of nonzeros per row, and $f$ is the fraction of fill-ins. For animal model problems, sparse factorization allows to work with matrices about 10-50 times larger than the dense-matrix factorization.

Ordering is a complicated process. Excellent coverage of sparse methods is in books by George and Liu (1981) and Duff et al. (1989). Packages for sparse matrix operations include commercial SPARSPAK, SMPAK, MA 28, or public domain FSPAK (anonymous ftp at num.ads.uga.edu, directory FSPAK).

**Sparse matrix inversion**

Even if the LHS is sparse, usually the inverse of the LHS is dense. Therefore, storage of such a complete matrix would be impossible for dimensions over 10,000-20,000! (why?). For operations such as variance components by REML or accuracies (PEV), only selected animals of the inverse are required. If sparse factorization is available, one can obtain inverse elements corresponding to nonzeros of the factorization at a cost of twice the factorization, using the Takahashi algorithm.

Clarification: for the purpose of the inversion, nonzeros in L mean elements that were created but could have values of zero due to numerical cancellation.

For the diagonal factorization: $A = LDL'$

$$A^{-1} = D^{-1}L^{-1} + (I-L')A^{-1}$$

By starting the computations from the last row/column, it is possible to avoid the computations of $L^{-1}$ explicitly.

The drawing below shows the progress of computations of a sparse inverse. The computation requires a access to L by columns or row access to L'. Please note that the computations start with the last diagonals and progress up to the first row. The sparse inverse, as computed, replaces the factorization without the need for much extra memory.
The operations on rows of sparse matrices require sparse vector by sparse vector multiplication. Since two sparse vectors generally need not share the same nonzero structure, the multiplication is hard to do directly. Therefore this is done in several steps
a) zero a work dense vector,
b) uncompress the first sparse vector to the dense vector
c) multiply the second sparse vector by the dense vector (easy),
d) zero elements of the dense vector changed in step b)
Example: define the sparse vector $p$ as triple $(p_n \ pia \ pa)$
- $p_n$ - size of vector
- $pia$ - vector of nonzero columns
- $pa$ - vector of nonzero values
Let the work dense vector be $w$. This program would multiply $x$ by $y$

! sum=x*y
! uncompress x
do i=1,xn
  w(xia(i))=xa(i)
enddo
! w*y
sum=0
do i=1,yn
  sum=sum+w(yia(i))*ya(i)
enddo
! nullify changed elements of w
do i=1,xn
  w(xia(i))=0
enddo

**Numerical accuracy**

Usually, dense matrix operations use double-precision numbers to avoid numerical inaccuracies. Errors usually propagate in finite methods, i.e., an error done during the computations of one matrix element is usually compounded during the computations of the following elements. One critical point in Cholesky factorization of matrices of not full rank is the detection of unnecessary equations, which may be less than perfect. Making LHS full-rank by eliminating the unnecessary rows explicitly solves the problem. Sparse factorization results in more accurate computations because of fewer computations and therefore fewer rounding errors.
Iterative methods

In iterative methods, a sequence of solutions is produced

\[ x^0, x^1, x^2, \ldots, x^i, \]

where each subsequent solution is more accurate than the previous one. The two simplest iterations are Jacobi and Gauss-Seidel. They originate from a family of stationary iteration methods where iteration parameters stay constant during all rounds of iteration.

Decompose \( A \) into a part that can easily be inverted \( M \) and the remainder \( N \)

\[ A = M + N \]

and write

\[ Ax = b, \]
\[ (M+N)x = b, \]
\[ Mx = -Nx + b \]

In the iteration:

\[ Mx^{(n+1)} = -Nx^{(n)} + b \]

the iterative solution is obtained as

\[ x^{(n+1)} = M^{-1}(-Nx^{(n)} + b) \]

Let \( A \) be decomposed as

\[ A = L + D + U \]

where for symmetric matrices

\[ U = L' \]
Jacobi
In the Jacobi method, $M = D$, and $N = L + U$

$$x^{(n+1)} = x^{(n)} + D^{-1} (b - A x^{(n)})$$

where $D = \text{diag}(A)$ is a diagonal matrix

or

$$\text{do } i=1,n$$

$$\quad x_i^{(n+1)} = x_i^{(n)} + \frac{b_i - \sum_{j=1}^{m} a_{ij} x_j^{(n)}}{a_{ii}}$$

$$\text{enddo}$$

In the Jacobi iteration, newly computed solutions are not used until all the new solutions are known.

Gauss-Seidel and SOR
In Gauss-Seidel the computations are similar to Jacobi but the newly computed solutions are used immediately

$$\text{do } i=1,m$$

$$\quad x_i^{(n+1)} = x_i^{(n)} + \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{m} a_{ij} x_j^{(n)}}{a_{ii}}$$

$$\text{enddo}$$

The GS iteration can be rearranged to

$$\sum_{j=1}^{i} a_{ij} x_j^{(n+1)} = b_i - \sum_{j=1}^{m} a_{ij} x_j^{(n)} \rightarrow (L + D)x^{(n+1)} = -Ux^{(n)} + b$$

Thus, this iteration is equivalent to obtaining $(L + D)^{-1}$ which better approximates $A^{-1}$ than $D^{-1}$ in the Jacobi iteration and therefore can be expected to have better properties.

The computer program for the Gauss-Seidel iteration is very easy, assuming that new solutions
immediately replace the old solutions

\[
\begin{align*}
\text{do } i=1,n \\
\quad x_i &= x_i + \frac{b_i - \sum_{j=1}^{n} a_{ij} x_j}{a_{ii}} \\
\text{enddo}
\end{align*}
\]

or

\[
\begin{align*}
\text{do } i=1,n \\
\quad \text{diff} &= b(i) - \text{matmul}(a(:,i), x) \\
\quad x(i) &= x(i) + \text{diff} / a(i,i) \\
\text{enddo}
\end{align*}
\]

The successive overrelaxation (SOR) is modified GS

\[
\begin{align*}
\text{do } i=1,n \\
\quad x_i &= x_i + \omega \frac{b_i - \sum_{j=1}^{n} a_{ij} x_j}{a_{ii}} \\
\text{enddo}
\end{align*}
\]

where \( \omega \) is the relaxation factor. For complicated LHS, a good choice of \( \omega \in (1,2) \) results in a better convergence rate.

GS can be implemented with half- upper-stored A at a cost of storing an adjusted right hand side. When solution i is computed, the right hand side for equations i+1,..., can be adjusted for that solution. The program (one round only) would be

\[
\begin{align*}
c &= b \\
\text{do } i=1,n \\
\quad \text{diff} &= c(i) - \text{matmul}(a(:,i), x) \\
\quad x(i) &= x(i) + \text{diff} / a(i,i) \\
\quad c(i+1:n) &= c(i+1:n) - \text{matmul}(a(i,i+1:n), x(i+1:n)) \\
\text{enddo}
\end{align*}
\]
Preconditioned conjugate gradient

The method of preconditioned conjugate gradient (PCG), as used in ITPACK, Berger et al. (1989) and by Lidauer et al. (1999), converges much faster for complicated models than either Gauss-Seidel, SOR or Jacobi. Good overview of the theory of PCG is in Meurant (1999).

PCG has some similarities to the second-order Jacobi, which is:

\[ x^{(n+1)} = x^{(n)} + \alpha(x^{(n)} - x^{(n-1)}) + \beta \, D^{-1} \, (b - A \, x^{(n)}) \]

PCG uses a similar formula except that \( \alpha \) and \( \beta \) are recomputed every round for orthogonal \( (b - A \, x^{(n)}) \), and \( D \) is replaced by \( M \), which is now called the preconditioner. \( M \) may equal to \( \text{diag}(A) \) or any other approximation of \( A \) that is easily invertable.

The complete pseudo code for the PCG iteration is:

\[ x=0 \ ; \ r=b-Ax; \ k=1 \]

\[ \text{do while (} r' \ r \text{ "not sufficiently small")} \]

\[ z=M^{-1} \ r \]

\[ \tau_{k,1}=z' \ r \]

\[ \text{if (} k==1 \text{) then} \]

\[ \beta=0; \ p=z \]

\[ \text{else} \]

\[ \beta=\tau_{k,1}/\tau_{k,2}; \ p=z+\beta \, p \]

\[ \text{endif} \]

\[ w=Ap \]

\[ \alpha=\tau_{k,1}/(p'w) \]

\[ x=x+\alpha p \]

\[ \text{if (} \text{mod}(k,100) /=0) \text{ then} \]

\[ r=r-\alpha w \]

\[ \text{else} \]

\[ r=b-Ax \]

\[ \text{endif} \]

\[ k=k+1 \]

\[ \text{enddo} \]

Despite a seemingly complicated code, PCG is very easy to implement because the only compute-intensive operations are \( Ax \) or \( Ap \). As presented, the PCG algorithm needs 7 variables, but two of them: \( b \) and \( z \) can be eliminated, for a total of 5. This algorithm lacks the self-correcting properties of Gauss-Seidel, SOR or second-order Jacobi iterations, and may diverge with large data sets if the variables are not in double precision.
Other methods

For more information on a number of iterative methods, see a book available on line at http://netlib.org/templates/Templates.html. A number of popular iterative methods are implemented in a public-domain package ITPACK. This package not only calculates the solutions but also determines optimal parameters and approximate accuracy of solutions.

For the MME, variants of Jacobi and GS have very desirable properties for MME including fast convergence rate for some models and small memory use. These basic iteration schemes can be “improved” or “accelerated” for increased convergence. However, there is increased evidence that other methods, especially the PCG, while less obvious and more memory intensive, can provide faster and more reliable convergence for a larger set of models.

Convergence properties of iterative methods

A “good” iterative method would converge fast and would have small computational requirements.

The “true” convergence criterion $C_i$ can be described by a measure of distance from converged solutions

$$C_i^{(b)} = \frac{\|x^{(b)} - x\|}{\|x\|}, \quad \text{where} \quad \|x\| = \sqrt{\sum_i x_i^2}$$

Because converged solutions are not available during the iteration process, the following criteria are used instead:

- based on differences between consecutive solutions

$$C_i^{(d)} = \frac{\|x^{(b)} - x^{(b-1)}\|}{\|x^{(b)}\|}$$

- based on differences between the RHS and LHS

$$C_i^{(c)} = \frac{\|b - Ax^{(b)}\|}{\|b\|}$$

For convergence, $C^{(i)} < C^{(i+1)}$. Usually, $C=10^{-k}$ corresponds to changes on the k-th significant digit.

The typical convergence curves for stationary methods are:
In a typical convergence curve, a seemingly rapid convergence is followed by a steady rate of convergence, finally changing into random fluctuations. The convergence criteria based on differences between the consecutive solution show much better convergence than the “true” criterion. However, both are parallel during the period of steady convergence. Let’s determine the necessary conditions for convergence and calculate the convergence rate based on a particular method of iteration.

A typical curve for a solution is
Let’s determine why solutions are converging so smoothly in later rounds of iteration.

The stationary iteration can be presented in a form

\[ x^{(n+1)} = B x^{(n)} + c \]

Decompose \( B \) into eigenvalues and eigenvectors. For nonsymmetric \( B \), these are complex

\[ B = \mathbf{V} \mathbf{E}' \]

where \( \mathbf{E} \) is diagonal, \( e_i \) are eigenvalues, and rows of \( \mathbf{V} \): \( \mathbf{v}_i \) are eigenvectors, which are orthogonal and normalized only for symmetric \( B \).

The successive stages of the iteration can be presented as:

\[
\begin{align*}
    x^{(1)} &= B x^{(0)} + c \\
    x^{(2)} &= B x^{(1)} + c = B (B x^{(0)} + c) + c = B^2 x^{(0)} + Bc + c \\
    x^{(3)} &= B x^{(2)} + c = B (B^2 x^{(0)} + Bc + c) + c = B^3 x^{(0)} + B^2 c + Bc + c \\
    \vdots \\
    x^{(n)} &= B^n x^{(0)} + B^{n-1} c + \ldots + Bc + c
\end{align*}
\]

and, assuming that \( x^{(0)} = 0 \)

\[ x^{(n)} - x^{(n-1)} = B^{n-1} c \]

Because the product \( B^n \) is

\[ \mathbf{V} \mathbf{E} \mathbf{V}' \mathbf{V} \mathbf{E} \mathbf{V}' \ldots \mathbf{V} \mathbf{E} \mathbf{V}' = \mathbf{V} \mathbf{E}^n \mathbf{V}' \]
therefore

\[
B^n c = V E^n V^{-1} c = \sum_i e_i^n v_i v_i^t c = e_1^n \sum_i \left( \frac{e_i}{e_1} \right)^n v_i v_i^t c
\]

Arrange \( V E V \) so that the absolute values of the eigenvalues are ordered in the descending order

\(|e_1| \geq |e_2| \geq ... \geq |e_m|\)

If we assume \(|e_1| > |e_2|\) then for sufficiently large \( n \)

\[
\left| \frac{e_i}{e_j} \right|^n \approx 0 \text{ for } i \neq j
\]

and

\[
B^n c \approx e_1^n v_1 v_1^t c = |e_1|^n f
\]

Then

\[
x^{(n)} = x^{(n-1)} + B^{n-1} c = x^{(n-1)} + |e_1|^{n-1} f
\]

The last expression shows for convergence \(|e_1^{n+1}| < |e_1^n| \Rightarrow |e_1| < 1\), or that the largest eigenvalue must be smaller than one.

It also shows that for large \( n \) the changes in the solutions will be predictable, where the same vector \( f \) is added with a coefficient becoming \( e_i \) times smaller every next round. That predictable change is the reason for a relatively flat convergence curve for larger \( n \). We can rewrite the formula for convergence criterion based on differences between the consecutive solutions

\[
C_{d}^{(b)} = \frac{\|x^{(b)} - x^{(b-1)}\|}{\|x^{(b)}\|} \approx \frac{|e_1^{n-1}| f}{\|x^{(b)}\|} \propto |e_1|^{n-1}
\]

and

\[
\log(C_{d}^{(b)}) = n \log(|e_1|) + \text{const}
\]

Thus, the slope of the convergence curve for larger \( n \) is constant and determined by the largest eigenvalue of the iteration matrix.

Since for large \( n \) the solutions increase predictably, one can approximate \( x^{(n)} \). From

\[
x^{(n)} = x^{(n-1)} + B^n c = x^{(n-1)} + |e_1|^{n-1} f
\]

the formula for \( x^{(n)} \) will be
\[
\mathbf{x}^{(n)} = \mathbf{x} = \mathbf{x}^{(n-1)} + \mathbf{f} \sum_{j=n-1}^{\infty} |e_j| = \mathbf{x}^{(n-1)} + \frac{\mathbf{f} |e_1^{n-1}|}{1-|e_1|}
\]

We can estimate \(e_1\) and \(e_0\) \(f\) directly from the iteration

\[
|e_1| = \frac{C_d^{(0)}}{C_d^{(n-1)}},
\]

and

\[
|e_1^{n-1}| f = x^{(n)} - x^{(n-1)}
\]

so the formula for \(x^{(n)}\) becomes

\[
x^{(n)} = \mathbf{x} = \mathbf{x}^{(n-1)} + \frac{x^{(0)} - x^{(n-1)}}{C_d^{(0)} C_d^{(n-1)}}
\]

This formula may not work when eigenvalues are too close or when \(x^{(n)} - x^{(n-1)}\) is very small and therefore contains large rounding errors. However, it is usually accurate enough for determination of the true convergence criterion

\[
C_d^{(n)} = \frac{|x^{(n)} - x|}{\|x\|} \approx \left[1 - \frac{C_d^{(0)}}{C_d^{(n-1)}}\right]^{-1} C_d^{(0)}
\]

**Example**

Let \(C_d^{(100)} = 1.105 \times 10^{-3}\), \(C_d^{(101)} = 1.008 \times 10^{-3}\), indicating three significant digits. The real accuracy is close to

\[
C_d^{(101)} \approx 1/(1-1.008/1.105) C_d^{(101)} \approx 11 C_d^{(101)}
\]

so it is 11 times lower than that indicated by \(C_d^{(101)}\).

**Conclusions**

1. Initial convergence is fast because all eigenvalues contribute to the convergence,
2. In later rounds, the convergence is determined by the largest eigenvalue of the iteration matrix,
3. After many rounds, the convergence reaches the limit of accuracy of computer floating point numbers.
4. Good convergence rate is determined by steep slope of the convergence criterion and not by its absolute value,
5. The criterion based on differences in solutions between consecutive rounds of iteration can grossly overstate the accuracy.

Properties of the Jacobi iteration

The Jacobi iteration leads to very simple programs because the expression

\[ b - Ax \]

is easy to obtain for any format of \( A \). In general, Jacobi converges for matrices diagonally dominant, where \( a_{ii} > a_{ij} \), but usually the MME do not satisfy this condition. In practice, Jacobi diverges with more than one fixed effect and with animal-additive or dominance effect.

To converge with many fixed effects, the following modifications are generally successful:

- impose constrains on fixed effects, for example by forcing the sum of solutions for each fixed effect except the first one to 0; the constrain by setting unnecessary equations to zero results in poor convergence, or
- update only one effect at a time

To converge with animal-additive or dominance effect, extend Jacobi to second-order Jacobi (also called Chebyshev-accelerated)

\[
 x^{(n+1)} = x^{(n)} + \alpha (x^{(n)} - x^{(n-1)}) + \beta D^{-1} (b - A x^{(n)})
\]

where \( \alpha \) and \( \beta \) are iteration parameters, which are adjusted for the best convergence. In the animal model, the good parameters are \( \alpha = .7-.9 \) and \( \beta = 1 \). In the dominance model, additionally \( \beta = .4-.6 \).

The Jacobi iteration is of interest only when the LHS is not available explicitly, as with iteration on data.

Properties of the Gauss-Seidel and SOR iterations

The GS and SOR iterations converge for all symmetric semipositive-definite matrices. No constrains are necessary for redundant equations or fixed effects. The disadvantage is that elements of \( A \) must be available sequentially, row by row. As will be shown later, SOR has advantage over GS only for effects with nondiagonal blocks.

Practical convergence issues
Predictions of breeding values for genetic evaluation need to be calculated with very limited precision, no more than 2 decimal digits. Such precision is obtained with 5-15 rounds in a single-trait sire model, 30-100 rounds in a single-trait animal model. With models with multiple traits and the maternal effect, the number of rounds can increase to > 500. One way of increasing the convergence rate for multiple traits is to set the MME within traits and decompose

\[ A = L + D + L' \]

so that \( D \) would include diagonal blocks of traits, and possible blocks of direct and maternal effects. Such a modification described by Dwyer and VanVleck (JDS, 1985?). Then, the convergence rate with multiple traits is comparable to that with a single-trait.

Jacobi by effects, where only one effect is solved at a time, would be equivalent to a mixed iteration where the outer iteration is GS and the inner iteration is Jacobi. Note that for fixed effects where the diagonal blocks for effects are diagonal, Jacobi by effects is equivalent to Gauss-Seidel. A disadvantage of block iteration is increased storage for diagonal elements.

MME are usually not of full rank. While regular Jacobi iteration requires “sum to zero” for all fixed effects but one, Jacobi-by-effect, GS and SOR handle the MME well and do not need any constrains.

**Determination of best iteration parameters**

Some iterations, including SOR and second order Jacobi, require the choice of quasi-optimal iteration parameters. Those parameters can be selected in a few ways, by

1. setting the iteration matrix \( B \) for several parameters, calculating eigenvalues, and selecting parameters corresponding to the smallest \( e_i \); hard to do for large systems of equations
2. running the iteration many times with different parameters and selecting the parameter with the best convergence rate in later rounds of iteration (lowest \( e_i \)),
3. updating the parameters during the course of iteration, as in Hageman et al. (1981?) or ITPACK.

Usually parameters quasi-optimal for one data set remain quasi-optimal for similar data sets.

**Strategies for iterative solutions**

**Solving by effects**

Consider a model

\[ y = \sum_{i} f_i x_i + \epsilon \]

where \( f_i \) may be either random or fixed effect. The MME assuming for simplicity \( R=I \) are
This system of equation is equivalent to block equations

\[
\begin{bmatrix}
F_i'F_i + G_i^{-1} & F_i'F_2 & F_i'F_3 \\
F_2'F_i & F_2'F_2 + G_2^{-1} & F_2'F_3 \\
F_3'F_i & F_3'F_2 & F_3'F_3 + G_3^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_i \\
\hat{\beta}_2 \\
\hat{\beta}_3
\end{bmatrix}
= \begin{bmatrix}
F_i'y \\
F_2'y \\
F_3'y
\end{bmatrix}
\]

where \(y_i^{(e)}\) is vector of observations adjusted for all effects but \(i\). Instead of solving the full MME iteratively, it is possible to solve each block separately by any method, and repeat the solving by blocks until convergence.

Creating the adjusted observations is simple. For example, if a model for one observation is

\[5.2 = c_i + a_{20} + p_{15} + e\]

and if \(\hat{c}_1=1.4, \hat{a}_{20}=5.2, \hat{p}_{15}=3.8\), the value of \(y\) adjusted for all effects except \(a\) would be

\[5.2 - 1.4 - 3.8 = 0.0\]

**Iteration on data (matrix-free iteration)**

If MME cannot be stored in memory but can be created on disk, the system of equations can be solved by repeatedly reading the MME from disk. However, reading from disk is relatively slow. The size of the LHS of MME can be 10 times or larger than the size of data, with difference increasing with the number of traits and effects. Therefore instead of reading the MME from disk it may be faster (and simpler) to read the much smaller data from disk and recreate necessary parts of MME each round.

Iteration on data (I.D.) is very simple when the iteration method involves \(A\) only in the expression
This expression can be created as follows:

<table>
<thead>
<tr>
<th>Original program</th>
<th>I.D. program</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>do round=1,p</td>
</tr>
<tr>
<td></td>
<td>AX=0; D=0</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>A(i,j)=A(i,j)+z</td>
<td>if (i == j) D(i)=D(i)+z</td>
</tr>
<tr>
<td>...</td>
<td>AX(i)=AX(i)+z*x(j)</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>! This is for Jacobi iteration only</td>
</tr>
<tr>
<td>call solve(A,b,x)</td>
<td>do i=1,nef</td>
</tr>
<tr>
<td></td>
<td>x(i)=x(i)+(b(i)-AX(i))/D(i)</td>
</tr>
<tr>
<td></td>
<td>end do</td>
</tr>
<tr>
<td></td>
<td>end do</td>
</tr>
</tbody>
</table>

The statements to set-up D and AX are added for parts of the program corresponding to contributions due to records and to contributions due to relationships.

When the iteration updates only one effect in one round, the modifications would be

```fortran
       do round=1,p
          AX=0; D=0
          do e1=1,nef
             ...
             if (i == j) D(i)=D(i)+z
             AX(i)=AX(i)+z*x(j)
             ...
             ! This is for Jacobi iteration only
             do i=sum(nlev(1:e1-1)+1,i=sum(nlev(1:e1)
                x(i)=x(i)+(b(i)-AX(i))/d(i)
             enddo
          end do
        end do
```
Please see that the program statements related to relationship contributions are not shown above separately. For the iteration that needs only the matrix-vector product $Ax$, the data and pedigree files need not be in any particular order. For multiple-trait block iteration, the programs would be modified to create block-diagonal $D$.

The same strategy can be used in PCG except that the product to derive can be either $Ax$ or $Ap$.

The implementation of the Gauss-Seidel or SOR iteration is more difficult because LHS needs to be available one row at a time. To do so:

- for $e$ effects, create $e$ copies of the data file, each sorted for different effect
- for pedigree file, create 3 copies, sorted by animal, sire and dam, respectively.
- solve as follows:

```fortran
do e=1,neff
   rewind i
   ..... 
   ..... 
   do l=1,nlev(i)
      d=0
      ax=0
      current=address(e,l)
      ..... 
      ..... 
      do
         "read file sorted by level l until the end of level l"
         ..... 
         ..... 
         if (i == current) then
            if (i==j) d=d+z
            ax=ax+z*x(j)
         endif
      end do !of reading the level l
   !
   if (e is animal effect) then
      do
         read pedigree line with animal l &
         as either animal, sire or dam
         ..... 
         ..... 
         if (i == current) then
            if (i==j) d=d+z
            ax=ax+z*x(j)
         endif
      end do
   end if
end do e=1,neff
```


if (e is other random effect then)  
    ‘create other random effect contributions’
endif

! solve for level 1 of effect e
x(current)=x(current)+ (b(current)-ax)/d
......
endo
dono
......
dono

Schaeffer and Kennedy (1996) have shown a technique where for e effects one needs only e-1 sorted files. Another technique that reduces the need for sorted files is to solve small effects by a finite method. This technique is used in PEST. For effects where diagonal blocks for effect are diagonal matrices, Jacobi_by_effect is equivalent to GS, and no sorting is required.

**Indirect representation of mixed-model equations**

The mixed model equations, especially those simpler ones, can be solved without the use (and knowledge) of any matrices (Misztal and Gianola, 1987). The indirect iteration gives insight into the workings of GS and J methods.

Assume a single-trait model

\[ y_{ijkl} = h_i + g_j + s_k + e_{ijkl}, \quad \sigma_e^2 / \sigma_s^2 = \alpha \]

where \( y_{ijkl} \) is a yield of a cow in herd i, genetic group j and sire k. The contribution of one record \( y_{ijkl} \) to LHS and RHS is
All the contributions corresponding to the row of $h_i$ result in an equation

$$\sum_{jk} (h_i + g_j + s_k) = \sum_{jk} y_{ijkl}$$

The GS iteration for $h_i$ would be

$$h_i = h_i + \frac{\begin{bmatrix} h \\ g \\ s \end{bmatrix} \cdot \begin{bmatrix} \text{RHS} - \text{LHS} \end{bmatrix}}{n_h}$$

or

$$h_i = h_i + \frac{\sum_{jk} (y_{ijkl} - h_i - g_j - s_k)}{n_h}$$

The solution for $h_i$ is an average of observations that include $h_i$ adjusted for all the other effects.

Similarly, we can derive equations for $g_j$ and $s_k$. 

\[ g_j = \frac{y_{ijkl} - h_i - s_k}{n_k} \]

\[ s_k = \frac{y_{ijkl} - h_i - g_j}{n_k + g} \]

The regression towards the mean in the last equation is due to \( s_k \) being the random effect.

In the GS iteration, \( h_i \) will be solved first, followed by \( g_j \) and then by \( s_k \). In the Jacobi iteration, all effects will be solved simultaneously.

**Example**

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>k</th>
<th>( y_{ijkl} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Assume that \( \alpha = 8 \) and that the initial solutions are 0. The first round solutions by GS are

\[
\begin{align*}
    h_1 &= \frac{(17+10+12+7)}{4} = 11.5 \\
    h_2 &= \frac{(14+9+11+4)}{4} = 9.5 \\
    g_1 &= \frac{\left[(17-11.5) + (10-11.5) + (14-9.5) + (9-9.5)\right]}{4} = 2 \\
    g_2 &= \frac{\left[(17-11.5) + (7-11.5) + (11-9.5) + (4-9.5)\right]}{4} = -2 \\
    s_1 &= \frac{4}{(4+8)} \left[ (17-11.5-2) + (12-11.5-(2)) + (14-9.5-2) + (11-9.5-(-2)) \right] = 1 \\
    s_2 &= -1
\end{align*}
\]

The solutions will remain identical the next round so the convergence for this special balanced case was reached in 1 round.

In Jacobi
\[ h_1 = \frac{(17+10+12+7)}{4} = 11.5 \]
\[ h_2 = \frac{(14+9+11+4)}{4} = 9.5 \]
\[ g_1 = \frac{(17+10+14+9)}{4} = 12.5 \]
\[ g_2 = \frac{(17+7+11+4)}{4} = 8.5 \]
\[ s_1 = \frac{4}{(4+8)} \frac{(17+12+14+11)}{4} = 4.5 \]
\[ s_2 = \frac{4}{(2+8)} \frac{(10+7+9+4)}{4} = 2.5 \]

During the next round the solutions will diverge. However, J will converge if the system of equations is made full rank by setting the sum of all g solutions to 0. Id additionally the sum of s solutions is set to 0, J will converge in one round.

That convergence in one round is a special property of GS and constrained J, and it is missing from other methods. This is why J and GS and their modifications are naturally suited to MME. Note that second-order Jacobi and SOR won’t converge in one round in the example. One way to make them do it is not to use any relaxation factors for fixed effects and for random effects with the diagonal (co)variance matrix.

Animal model

In the model with groups, the contributions to the left hand side are:

\[
\begin{array}{cccc}
\text{LHS} & \text{a}_i & \text{d}_i \ (\text{or g}_q) & \text{s}_i \ (\text{or g}_s) \\
\hline
\text{a}_i & 2\alpha v_i & -\alpha v_i & -\alpha v_i \\
\text{d}_i \ (\text{or g}_s) & -\alpha v_i & \alpha v_i/2 & \alpha v_i/2 \\
\text{s}_i \ (\text{or g}_q) & -\alpha v_i & \alpha v_i/2 & \alpha v_i/2 \\
\end{array}
\]

\[ a_i, s_i \text{ and } d_i \text{ are animal, and sire and dam solutions, respectively, g’s are unknown parent groups if sire or dam are missing, } \alpha \text{ is variance ratio and } v_i = 2/(4- \text{number of known parents}) \]

Calculate the quantity
\[ z = - \alpha v_i (2a_i - d_i - s_i) \]

The contribution to the adjusted right hand side (b-AX) from each pedigree line is:

to animal
\[ - \alpha v_i (2a_i - d_i - s_i) = z \]
to sire (or sire unknown parent group)
\[ \alpha v_i (2a_i - d_i - s_i)/2 = -z/2 \]
to dam (or dam unknown parent group)
\[ \alpha v_i (2a_i - d_i - s_i)/2 = -z/2 \]

The corresponding changes to the diagonal would be:

- to animal
  \[ 2 \alpha v_i \]
- to sire (or sire unknown parent group)
  \[ \alpha v_i / 2 \]
- to dam (or dam unknown parent group)
  \[ \alpha v_i / 2 \]

Example

Consider the “dairy” repeatability model

\[ y_{ijkl} = m_i + p_j + a_k + e_{ijkl} \]

where \( y_{ijkl} \) is yield of cows in management groups \( m_i \), with additive value \( a_k \) and permanent environment \( p_j \), and

\[ \text{var}(p) = I \sigma_p^2, \quad \text{var}(a) = A \sigma_a^2, \quad \sigma_e^2/\sigma_s^2 = \alpha, \quad \sigma_e^2/\sigma_p^2 = \beta \]

The matrix-free equations are

\[ m_i = \frac{y_{ijkl} - p_j - a_k}{n_{ijkl}} \]
\[ p_j = \frac{y_{ijkl} - m_i - a_k}{n_{ijkl}} \frac{n_{ijkl}}{n_{ijkl} + \beta} \]
\[ a_k = \frac{\sum_{\text{progenies of } k} (y_{ijkl} - m_i - p_j) - \alpha v_k (d_k - s_k) - \alpha \sum_{i=1}^{\text{progenies of } k} v_{kl} (-p_{rkl} + \text{mate}_{kl})}{n_k + 2\alpha v_k + \alpha/2 \sum_{i=1}^{\text{progenies of } k} v_{kl}} \]

where \( p_{rkl} \) and \( \text{mate}_{kl} \) are the progeny and mate of the \( l \)-th mating of the \( k \)-th animal, respectively,
and \( v_{kl} \) is the corresponding residual (=.5 if mate is known and 1/3 if mate is replaced by unknown parent group).

The extension of these rules to multiple traits, maternal effects etc. can be done following the above methodology. For GS iteration, it was done by Kennedy and Schaeffer (1986, WCGALP).

**Summary of iteration strategies**

The GS and SOR iterations are the methods of choice when the coefficient matrix can be stored in memory and simplicity is important. This would usually require memory of 100-1000 bytes/equations. The iteration on data by modifications of Jacobi would be the next choice for simpler models, with memory requirements of 12-16 bytes/equation. However, the method of choice nowadays seems to be PCG. Despite a seemingly complicated code, PCG is very easy to implement because the only compute-intensive operations are \( Ax \) or \( Ap \), especially with Fortran 90 with vector instructions. PCG with diagonal preconditioner converged for all attempted models, including multiple trait, random regressions, with maternal effect and dominance. The convergence rate was as good or better as J or SOR. Also, there are no acceleration parameters to determine. One disadvantage of PCG is that it requires more memory: 36-56 bytes/equations, but this is not as much of a problem as it used to be.
Variance components

Out of very many methods for estimating variance components, three types are in popular use today (1998). They are:

1. REML
2. Bayesian methods via Gibbs-sampling, also called Markov Chain Monte Carlo (MCMC),
3. Method R.

REML is popular because of resistance to selection bias and the estimates are always within the parameter space. Its disadvantages include high cost for large models and relatively difficult programming especially for nonstandard models although various implementations of REML differ widely in their computing properties.

Bayesian methods via Gibbs-sampling result in simple programs even for complicated models, and have good theoretical properties. Their disadvantage include slow convergence because of the Monte-Carlo origin and therefore very long running time for larger problems.

Method R is applicable to very large systems of equations and is simple to implement for common models. However, the sampling error is higher than in the above methods, and incorrect fixed effects can result in drastically biased estimates. Also, for multiple traits there are only approximations.

REML

Let the model be
\[ y = X\beta + e \]
where the distribution of \( y \) is multivariate normal with mean \( X\beta \) and variance \( V \),
\[ y \sim MVN(X\beta, V) \]
or
\[ \text{Var}(e) = V \quad \text{and} \quad E(y) = X\beta \]

Following Quaas (1990), the likelihood of \( \beta \) and \( V \) is
\[
L(\beta, \phi; y) = \frac{1}{\sqrt{2\pi|V|}} e^{-\frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta)}
\]
where \( \phi \) is a vector or a matrix of variance components in \( V \). This can be simplified to
\[
\ln L(\beta, \phi; y) = -\frac{1}{2} \left[ \ln(2\pi) + \ln|V| + (y-X\beta)^T V^{-1}(y-X\beta) \right]
\]
The maximum likelihood (ML) estimates can be obtained by maximizing ln L. The ML estimates of variance components depend on fixed effects because the degrees of freedom due to estimating the fixed effects are not taken into account. In restricted maximum likelihood, y is replaced by My such that

$$MX = 0 \Rightarrow My = Me$$

which is equivalent to integrating the fixed effects from the ML likelihood. In subsequent derivations M cancels out, and the log of the restricted maximum likelihood becomes

$$\ln L(\Phi; y) = -\frac{1}{2} \left\{ \ln(2\pi) + \ln|V| + \ln|X'V^{-1}X| + y' [V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}] y \right\}$$

Most of the research in REML went into simplifying the above likelihood for particular models and into maximizing that likelihood. For mixed models, the log likelihood is proportional to

$$-\ln L(\Phi; y) \propto \ln|R| + \ln|G_j| + \ln|C^*| + y' R^{-1} (y - X\hat{\beta} - Z\hat{a})$$

where $C^*$ is the coefficient matrix of the MME converted to full-rank; the determinant of a singular matrix is 0.

For

$$R = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \\ \vdots & \vdots \end{bmatrix}, \quad G = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \\ \vdots & \vdots \end{bmatrix}$$

where $R_i$ is the residual (co)variance matrix for observation $i$, and $G_j$ is the (co)variance matrix for random effect $j$, the components of the likelihood can be further simplified to

$$\ln|R| = \sum_i \ln|R_i|, \quad \ln|G| = \sum_j \ln|G_j|$$

In particular, for specific $G$’s,

$$\ln|G_{ab} \otimes A| \propto n_a \ln|G_{oa}|, \quad \ln|G_{ab} \otimes I| \propto n_b \ln|G_{ib}|$$

where $n_a$ and $n_b$ are the number of levels.

The REML estimates are derived by maximization

$$\hat{\Phi} : \max_{\Phi} \ln L(\Phi; y)$$
by one of the many maximization algorithms for nonlinear functions. REML is relatively easy to evaluate. Matrices $R_i$ and $G_m$ are small. $\ln|C^*|$ can be calculated from the Cholesky decomposition of LHS. Also, solutions can be obtained by Cholesky decomposition.

The simplest maximization is derivative-free:

a) choose a derivative-free maximization subroutine from a library (IMSL, "Numerical Recipes", "www.netlib.org", ...)

b) write a function $L$(variance components)

c) apply a) to b)

The computational constrain in derivative free methods would be time to calculate the determinant.

The alternate way to maximize the REML would be by using the first derivative,

$$
\phi: \frac{\delta \ln L(\phi; y)}{\delta \phi} = 0
$$

which is a problem of finding zeroes of a multidimensional nonlinear function. In single traits, this leads to the following equations

$$
\delta_i^2 = \frac{\hat{u}_i^T A_i^{-1} \hat{u}_i + \delta_i^2 \text{tr}(A_i^{-1} C^{ii})}{n_i}
$$

where $n_i$ is the number of levels of random effect $i$, $\sigma_i^2 A_i$ is its (co)variance matrix, and $C^{ii}$ is a diagonal block of the inverse of $C$ corresponding to effect $i$. This formula can be derived differently as EM REML. A faster converging version of the same formula is

$$
\delta_i^2 = \frac{\hat{u}_i^T A_i^{-1} \hat{u}_i}{n_i - \text{tr}(A_i^{-1} C^{ii}) \delta_i^2}
$$

where $n_i$ is the number of levels in the $i$-th random effect. The residual is estimated as

$$
\delta_e^2 = \frac{y^T (y - X\hat{\phi}) - \sum_i Z_i \hat{u}_i}{n_r - \text{r}(X'X)}
$$

where $n_r$ is the number of records and $r(X'X)$ is the rank of the fixed-effect part of the MME. In the first-derivative REML, the largest expense is in calculating the inverse of $C$. The trace can be computed as
\[
\text{tr}(A_i^{-1}C^{-1}) = \text{sum}(A_i^{-1} \# C^{-1})
\]
or as the sum of \(A_i^{-1}\) and \(C^{-1}\) multiplied directly, element by element. When \(A_i^{-1}\) is sparse, one does not require all elements of \(C^{-1}\) but only those that are nonzero in \(C^{-1}\), or a small fraction of all inverse elements of \(C\). The sparse inverse described before contains all the inverse elements nonzero in \(A_i^{-1}\). Computing that inverse from sparse factorization costs twice more than computing the factorization alone, and almost no extra memory. Thus, the most compute-intensive operation in one step of first-derivative REML is three times more expensive than a similar operation in derivative-free REML.

For multiple traits, denote \(G\) as

\[
G = \begin{bmatrix}
\sigma^2_{ijkl} A_{ijkl}
\end{bmatrix}
\]

and let the system of equations be

\[
(W'R^{-1}W + G^{-1}) \theta = W'R^{-1}y, \quad W = [X, Z], \quad \theta = \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix}
\]

where \(A_{ijkl}\) can either be a numerator relationship matrix or an identity matrix for a block corresponding to effects i and j and traits k and l. Denote \(A_{ijkl}^{-1} = A_{ijkl}\)

In multiple traits and the first-derivative REML, computations are similar to single-traits for random effect variances

\[
\delta^2_{ijkl} = \frac{\theta_{ik} A_{ijkl} \theta_{jl} + \text{tr}(A^{ijkl}C^{ijkl})}{n_{ik}}
\]

where \(\delta_{ij} = 1\) if \(i=j\) and 0 otherwise, and \(n_{ik}\) is the number of levels for effect i and trait k \((n_{ik} = n_j)\).

The formula for the residual variances when all traits are recorded is

\[
\delta^2_{ij} = \frac{\text{tr}(C^{ij}W_iW_j)}{n_t} \quad \text{where} \quad \epsilon_i = y_i - X_i \hat{\beta} + Z_i \hat{a}
\]

where \(n_t\) is the number of records. With missing traits, the formulas are more complicated to write
Let

\[ \mathbf{R}_k^{-1} = (\mathbf{P}_k \mathbf{R}_0 \mathbf{P}_k)^{-1} \]

be a matrix of residual variances for observation \( k \), where \( \mathbf{P}_k \) is a diagonal matrix with 1 in diagonals corresponding to present traits and 0 otherwise. Define \( \mathbf{E} \) as matrix with 1 in position \((i,j)\) and \((j,i)\) and 0 elsewhere. Then compute the following:

\[
\mathbf{Q}_{k,j} = \mathbf{R}_k^{-1} \mathbf{E}_{j} \mathbf{R}_k^{-1}, \quad i=1,...,t, \quad j=1,...,t
\]

\[
\mathbf{Q}_j = \sum_k \mathbf{Q}_{k,j}, \quad \mathbf{E}\mathbf{Q} = \sum_i \mathbf{e}_i \mathbf{Q}_{k,j} \mathbf{e}_i, \quad \mathbf{TRQWCW}_j = \sum_i \text{tr} (\mathbf{Q}_{k,j} \mathbf{W}_k \mathbf{CW}_k)
\]

Using the notation:

\[
\mathbf{Q}_{ij} = \{q_{ij,km}\}
\]

next round estimates are obtained by solving the system of equations:

\[
\begin{bmatrix}
q_{11,11} & 2q_{11,12} & \ldots & 2q_{11,1t} & q_{11,22} & \ldots & q_{11,tt} \\
q_{12,11} & 2q_{12,12} & \ldots & 2q_{12,1t} & q_{12,22} & \ldots & q_{12,tt} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
q_{1t,11} & 2q_{1t,12} & \ldots & 2q_{1t,1t} & q_{1t,22} & \ldots & q_{1t,tt} \\
q_{22,11} & 2q_{22,12} & \ldots & 2q_{22,1t} & q_{22,22} & \ldots & q_{22,tt} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
q_{tt,11} & 2q_{tt,12} & \ldots & 2q_{tt,1t} & q_{tt,22} & \ldots & q_{tt,tt}
\end{bmatrix}
\begin{bmatrix}
\sigma^2_{e11} \\
\sigma^2_{e12} \\
\vdots \\
\sigma^2_{elt} \\
\sigma^2_{e21} \\
\sigma^2_{e22} \\
\vdots \\
\sigma^2_{ett}
\end{bmatrix} = 
\begin{bmatrix}
eq_e 11 + \text{tr} w_1 w_1 \\
eq_e 12 + \text{tr} w_2 w_2 \\
\vdots \\
eq_e tt + \text{tr} w_t w_t
\end{bmatrix}
\]

Caution: blocks of inverse \( \mathbf{C}^{ij} \) and \( \mathbf{C}^{ijkl} \) for covariance are not necessarily symmetric. This may create computing problems if sparse matrix software operates on symmetric matrices only.

Second derivatives of REML or their expectation cannot be easily computed using sparse matrix computations because they require computations of off-diagonal blocks of \( \mathbf{C} \).
Jensen et al. (1996-7) have shown that the average of second derivatives and their expectations results in cancellation of terms that are hard to compute. Their method based on the ideas of R. Thompson was called “Average Information” REML. In this method, the most important computation is of matrix $F$ defined by $n \times p$, where $n$ is the number of observations and $p$ is the number of different variance components including the residuals. Matrix $F$ is composed of weighted solutions or residuals. Then the matrix of Average Information, which is used later as the approximation of the matrix of second derivatives, is:

$$I_a = F'R^{-1}F - T'WR^{-1}F,$$

$$T = C^{-1}WR^{-1}F$$

where $T$ is solution to the mixed model equations with cluster of $T$ substituting for $y$. Although this algorithm is fast when it converges, it requires heuristics when $I_a$ is not positive definite and its implementation is quite complex.
General Properties of Maximization Algorithms

The problem is to find

\[ \mathbf{x} : \max_{\mathbf{x}} f(\mathbf{x}) \]

where \( f(\mathbf{x}) \) is a nonlinear function. The algorithms for the maximization can be categorized by their use of derivatives of \( f(\mathbf{x}) \):

- Derivative-free - Only function values available, no derivatives available.
- First-derivative - \( \frac{df(\mathbf{x})}{d\mathbf{x}} \) available
- Second-derivative - \( \frac{d^2f(\mathbf{x})}{d\mathbf{x}^2} \) and \( \frac{df(\mathbf{x})}{d\mathbf{x}} \) available

The underlying mechanisms of these three methods are presented below for a one-dimensional case. It is assumed that the function to be maximized has exactly one maximum.

The example below shows two points that could be made available in a derivative-free method. Each point provides the value of the function but does not indicate the direction of the maximum. Two points can show the direction of the increase, and three points can bracket the maximum.

In the first-derivative method, each derivative shows the direction of the maximum. Two evaluation of the derivatives may bracket the maximum.
In a second-derivative method, each point contains the quadratic approximation of the function from that point. Thus one point is sufficient to obtain the first approximation of the maximum.
Unless the function is quadratic, in which case the second-derivative methods converge in one step, all methods need to be applied iteratively.

In general, the more derivatives are available, the faster the method. However, if the function is far from quadratic and a second-derivatives are used, next-round estimates may be far from optimum or even out of bounds.

Derivatives may be hard to calculate. In that case, the higher derivative methods can be derived by numerical differentiation

\[
\frac{\delta f(x)}{\delta x_i} = \frac{f(x+\Delta p_i) - f(x)}{\Delta}
\]

and

\[
\frac{\delta^2 f(x)}{\delta x_i \delta x_j} = \frac{\delta f(x+\Delta p_j)}{\delta x_i} \frac{\delta f(x)}{\delta x_j} \frac{1}{\Delta}
\]

where \(p_i\) is the i-th column of the identity matrix. Unfortunately, the derivatives thus obtained have a lower accuracy than those obtained analytically. According to Press et al. (1986), the
number of significant digits in the derivatives calculated numerically cannot be higher than half the number of significant digits in the original function. In general the selection of $\Delta$ is not an easy one.

**Some popular maximization algorithms**

**Downhill simplex (derivative-free)**

One of the most popular derivative-free maximization algorithms is downhill simplex. It is described in the animal breeding context by Boldman et al (1995) in the MTDFREML manual. Let $n$ be the number of parameters to find. Initially create $n+1$ vectors $x_0, x_1, \ldots, x_n$ and calculate the functions $f(x_i) = f_i$ ordered so that $f_0, f_1, \ldots, f_n$

Choices on the maximization for one dimension are shown below.
The general algorithm is

**do until convergence**

Calculate the mean point:

\[ x_m = \text{ave}(x_1, \ldots, x_m) \]

Calculate a reflection of the worst point against the center point:

\[ x_r = x_m + \alpha(t_m - t_p) \]

**case**

**The new point is best!**

\[ f_r > f_0 \]

Try even a better point

\[ x_b = x_m + \gamma(x_r - x_m) \]

\[ x_n = \text{better}(x_r, x_b) \]

**The new point is neither worst nor best**

\[ f_0 > f_r > f_p \]

Replace the worst point

\[ x_p = x_r \]

**The new point is almost worst**

\[ f_{p-1} > f_r > f_p \]

Contract to the opposite side of the worst point

\[ x_p = x_m + \beta(x_r - x_m) \]

**The new point is worst or equal**

\[ f_r \leq f_p \]

Try contracting first to the worst point

\[ x_r = x_m + \beta(x_p - x_m) \]

**New point better**

\[ f_r > f_p \]

\[ x_p = x_r \]

**New point worse**

\[ f_r \leq f_p \]

Contract in every direction towards the best point

\[ x_i = (x_i + x_0) / 2 \] for \( i = 1, \ldots, n \)

**end case**

**end do**

Sample constants: \( \alpha = 1; \beta = .5; \gamma = 2 \)

**Powell (derivative-free)______**
Set the initial point $x_{\text{base}}$

do until convergence

Find the “best” direction $d$ by probing $n$ dimensions

Find $\alpha$ such that:

$$\max_{x} F(x_{\text{base}} + d\alpha)$$

$$x_{\text{new}} = x_{\text{base}} + d\alpha$$

enddo

Fixed point (first derivative)

Rearrange the equation

$$df(x)/dx = 0$$

to

$$x = g(x)$$

and solve as

$$x^{(n+1)} = g(x^{(n)})$$

The choice of $g$ is critical to convergence. For example, the rearrangement of

$$\sqrt{x} - \frac{x}{\sqrt{x}} = 0 \Rightarrow (\sqrt{x})^{(n+1)} = \frac{x}{\sqrt{x}^{(n)}}$$

results in divergence, but a similar expression converges rapidly

$$2\sqrt{x} - \sqrt{x} - \frac{x}{\sqrt{x}} = 0 \Rightarrow (\sqrt{x})^{(n+1)} = \frac{\sqrt{x}^{(n)} + \frac{x}{\sqrt{x}^{(n)}}}{2}$$

Newton (second derivative)

$$x^{(n+1)} = x^{(n)} - \left[\frac{d^2f(x)}{(dx)^2}\right]^{-1} df(x)/dx$$

Quasi-Newton (first derivative)

When second derivatives are difficult to compute, they may be approximated from first
derivatives. A method that approximates the second derivatives by finite differences is called
secant. For more information, see More and Wright (1993) or Woodford (1992)
Acceleration to EM (fixed point) estimates

The convergence rate in the EM algorithm can be slow, especially when the number of equations relative to records is high. Denote the EM algorithm as

\[ g^{(n+1)} = \text{em}(g^{(n)}) \]

In the single trait, the iteration would be similar to

And the convergence curve would be similar to
Thus in later rounds, the convergence would be predictable, similar to convergence of stationary methods in the linear system of equations. Little (1987) has shown that the convergence rate in the EM algorithm is inversely proportional to the fraction of missing data. That fraction like the largest eigenvalue of the iteration matrix is hard to compute analytically but can be approximated during the progress of iteration.

For a single component, Schaeffer (1979) showed how results from two separate cycles of the EM iteration can be used to extrapolate almost to the converged solutions. Misztal and Schaeffer (1986) have shown that the \( \text{em}(x) \) approximately follows the geometric progression, and that the final value can be extrapolated from estimates of successive rounds of EM iteration. Laird et al. (1987) presented similar ideas for multiple dimensions.

Function \( \text{em}(x) \) is locally linear near the convergence point. Therefore, for large \( n \), we can write

\[
\text{g}^{(n+1)} = \text{g}^{(n)} + \alpha [\text{g}^{(n)} - \text{g}^{(n-1)}]
\]

Then
and finally

\[ g^{(n+2)} = g^{(n)} + (\alpha + \alpha^2) \left[ g^{(n)} - g^{(n-1)} \right] \]

The extrapolation is beneficial only when the expression

\[ \alpha^n = \frac{g^{(n+1)} - g^{(n)}}{g^{(n)} - g^{(n-1)}} \]

becomes stable. Otherwise the extrapolation can deteriorate the convergence. Various types of accelerations of the EM algorithm have been used successfully in packages DMUEM, and REMLF90, with the reduction in the number of rounds by 2-10 times. Quasi-Newton implementation in VCE, where only the first derivatives of the restricted likelihood are computed, can be regarded as either approximated second-derivative algorithm, or accelerated EM.
Convergence properties of maximization algorithms in REML

Examine the cost of derivative and derivative-free algorithms in REML following Misztal (1995). Convergence rate can be described in terms of accuracy as a function of the number of rounds of iteration (Minoux 1986). Linear convergence means linear relationship between accuracy and round of iteration, i.e., constant number of rounds is required for a gain of one extra digit of accuracy. The fixed point iteration (EM algorithm) seems to have a linear convergence. In a superlinear convergence, each round of iteration results in increasingly larger gains in accuracy. Finally, in a n-step convergence, n steps are required to implement one round of iteration.

The convergence rate of maximization algorithms for general functions cannot be derived easily. However, it could be derived for quadratic functions, and any twice-differentiable function, including L, is locally quadratic. Then, the convergence of better derivative-free (DF) algorithms such as Powell or Rosenbrock is n-step superlinear (Minoux 1986), and is dependent on n, the dimension of the function being maximized. For better derivative (D) algorithms, such as quasi-Newton (first-derivative) or Newton-Raphson (second-derivative), the convergence is superlinear and does not depend on n (Bazaraa 1993; Minoux 1986). This leads to conjecture that better DF algorithms converge, in general, n times slower than better D algorithms. N can be expressed in terms of number of random effects, \( n_r \), and number of traits, \( n_t \), as:

\[
n = (n_r + 1)n_t(n_t+1)/2
\]

where the extra 1 is for the residual.

This difference in convergence rate is only approximate because differences exist within better D or DF algorithms, and L is not quadratic. Also, more sophisticated methods may fail when the starting point is far away from the maximum.

Cost of one step of D and DF in multiple traits
Let \( C \) be a coefficient matrix of the mixed model equations. Assume that in DF, all computing resources are spent in computing \( |C| \). Also assume that in D, all computing resources are spent in finding elements of \( C^{-1} \) corresponding to nonzero elements in \( C \), or a sparse inverse of \( C \). In dense matrices, computing the inverse requires 3 times more arithmetic operations than computing the determinant alone (Duff et al. 1989). Both operations require an equal amount of storage. Approximately the same applies to the number of arithmetic operations in sparse matrices (Misztal and Perez-Enciso 1993). The amount of storage required is 3 times larger for sparse inversion if the inverse elements are computed by columns. Both options are supported under FSPAK.

In multiple traits, assuming that almost all traits are recorded, the coefficient matrix of the mixed model equations has \( n_t \) times more equations and an average equation has \( n_r \) times more nonzero elements than in single traits. In dense matrices, the memory and computing requirements for inversion or computing the determinant for a matrix \( n_t \) larger increase as \( n_t^2 \) and \( n_t^3 \), respectively. The same applies to sparse matrices, where memory and computing requirements
increase approximately as pq and pq^2, where p is the number of equations and q is the average number of nonzeros per equation (Duff et al. 1989).

**Cost of one unit of convergence relative to single trait estimation**

Let C_{d} be the cost of one step of D in a single trait estimation. Let us compute costs of the same level of convergence in n traits for the DF and D algorithms: C^n_{df} and C^n_{d}. If the DF convergence is n times slower, computations for the matrix operations increase as n^3, and computing the determinant costs a third of computing the inverse, the following formulas can be derived:

\[
C^n_d = n^3 \times C^1_d
\]  

\[
C^n_{df} = \frac{1}{6} (n_t +1) n_t(n_t+1)/2 \times n^3 \times C^1_d = (n_t +1) n_t^4(n_t+1) \times C^1_d / 6
\]  

and the relative costs are:

\[
C^n_d / C^1_d = n_t^3
\]  

\[
C^n_{df} / C^1_d = (n_t +1) n_t^4(n_t+1) / 6 
\approx (n_t +1) n_t^5 / 6
\]  

\[
C^n_{df} / C^n_d = (n_t +1) n_t(n_t+1) / 6 
\approx (n_t +1) n_t^2 / 6
\]

According to equation (3), the number of numerical operations in D increase cubically with the number of traits. In equation (4), the cost of DF increases with the fifth power of the number of traits. From equation (1) one can find that the costs of DF and D in models with 2 random effects are similar in single trait. DF is less expensive with 1 random effect, and D is less expensive with more than 2 random effects. In multiple traits, DF is n_t^2 more expensive than D.

Relative costs of multitrait DF REML evaluation using DF and D algorithms, computed with formulae (3) and (4) are presented below

<table>
<thead>
<tr>
<th>Number of traits</th>
<th>Number of arithmetic operations</th>
<th>Memory requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>162</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>640</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>1875</td>
<td>125</td>
</tr>
</tbody>
</table>
The table below shows time to calculate multiple trait REML estimates assuming that D and DF both took a minute in single traits.

<table>
<thead>
<tr>
<th>Number of traits</th>
<th>Approximate CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
</tr>
<tr>
<td>1</td>
<td>1 min</td>
</tr>
<tr>
<td>2</td>
<td>½ hr</td>
</tr>
<tr>
<td>3</td>
<td>2½ hrs</td>
</tr>
<tr>
<td>4</td>
<td>11 hrs</td>
</tr>
<tr>
<td>5</td>
<td>&gt; 1 day</td>
</tr>
<tr>
<td>6</td>
<td>3 days</td>
</tr>
</tbody>
</table>

**Accuracy of the D and DF algorithms**

In DF, if $L(\Theta)$ is computed with $r$ significant digits, then $\Theta$ can be bracketed with at most $r/2$ significant digits (Press et al. 1989). Such a limit does not exist in D. Worse numerical accuracy of the DF maximization is illustrated with in the Figure below, which shows a quadratic function and its derivative.
The function is very flat, and the maximum of the function cannot be determined accurately by looking at the function alone. The maximum can be determined much more precisely by finding a zero of the function's first derivative.

Loss of accuracy in DF does not appear a problem at first. Most computations are done with double precision, which corresponds to 15-17 significant decimal digits, and estimates with 1% or 2 significant digits of accuracy are considered sufficiently accurate in practice. However, the likelihood function can have low accuracy for a variety of reasons. Some accuracy is lost due to a roundoff error when summing all components of L. Components of that function may have reduced accuracy. This particularly applies to the determinant, where the loss of accuracy could result from lack of pivoting in Cholesky decomposition-based packages, poor conditioning of the coefficient matrix and rounding errors associated with computing in large matrices. In multiple traits, the coefficient matrix of the mixed model equations is composed of $R_0^{-1}$ and $G_0^{-1}$ - covariance matrices between traits for residual and random effect i, respectively. Poor conditioning of these matrices results in poor conditioning of W and subsequently low accuracy of determinants and traces. $R_0^{-1}$ and $G_0^{-1}$ would be poorly conditioned numerically when traits are highly correlated or linearly dependent. The loss of accuracy due to poor conditioning of C is also present in derivative maximization, but it has smaller effect on the estimates because of better numerical properties of the derivative maximization.
Another source of inaccuracy could arise in algorithms where derivatives are obtained numerically by differentiation. For example, the Quasi-Newton algorithm can be so implemented in DF, and one can regard other DF algorithms as using the numerical differentiation implicitly. The accuracy of such differentiation is dependent on the step side and could be very low for steps too large or too small. Subsequently, the accuracy would be dependent on parameters that define the step size, and in particular could be good for some problems but poor for others.
Method R

The Method R was proposed by Reverter et al. (1994) and can be summarized as follows.

Let \( u_i \sim N(0, G_i \sigma^2_i) \).

Let \( \hat{u}_i \) be a BLUP of \( u_i \) based on all the data available, and let \( \hat{u}_i^p \) be a BLUP of \( u_i \) based on a partial data. Then

\[
\text{Cov}(\hat{u}_i^p', \hat{u}_i) = \text{Var}(\hat{u}_i^p)
\]

Define \( r_i \) as the regression of predictions based on full data from those based on a partial data:

\[
r_i = \frac{\hat{u}_i^p' G_i^{-1} \hat{u}_i}{\hat{u}_i^p' G_i^{-1} \hat{u}_i^p}
\]

Define \( r = \{ r_i \} \) and \( s = \{ \sigma^2_e / \sigma^2_i \} \), where \( \sigma^2_e \) is residual variance and \( s \) is vector of variance ratios. Reverter et al. showed that for true variances,

\[
E[r(s)] = 1.
\]

For arbitrary \( x \) and one component

\[
r_i(x) > 1 \quad \Rightarrow \quad \text{too small variance}
\]
\[
r_i(x) < 1 \quad \Rightarrow \quad \text{too large variance}.
\]

For several components, the relationships between \( s \) and \( r \) are more complex. The problem of finding such \( s \) becomes a problem of solving a system of a nonlinear equations. It can be solved by a fixed point iteration, Newton or quasi-Newton methods.

Method R can be generalized to multiple traits. Let \( \hat{u}_{ij} \) be a BLUP of \( u_{ij} \) based on all the data available, and let \( \hat{u}_{ij}^p \) be a BLUP of \( u_{ij} \) based on a partial data, where \( i \) spans the random effects and \( j \) spans different traits.

Define \( r_{ijkl} \) as the regression of predictions based on full data from those based on a partial data for effects \( i \) and \( j \) and traits \( k \) and \( l \), using terminology similar as in formula for multivariate REML.
For arbitrary $x$ and one component

$$r_{ijkl}(x) > 1 \Rightarrow \sigma^2_{ijkl} \text{ too small}$$

$$r_{ijkl}(x) < 1 \Rightarrow \sigma^2_{ijkl} \text{ too large.}$$

No exact formulas are available for the residual variance but quasi-REML approximations were used by Reverter et al. (1994) as follows

$$\hat{\sigma}^2_{ij} = \frac{\hat{\varepsilon}_i'Y_i}{n_i - \text{rank}(X_i)}, \quad \text{where} \quad \hat{\varepsilon}_i = y_i - X_i\hat{\beta}_i - Z_i\hat{u}_i$$

where $n_i$ is the number of records in trait $i$, and

$$\hat{\sigma}^2_{ij} = \frac{\hat{\varepsilon}_i'Y_j + \hat{\varepsilon}_j'Y_i}{(n_{ij} - \text{rank}(X_i'X_j))}$$

where $n_{ij}$ is the number of records with both traits $i$ and $j$ recorded.

Method R is potentially less expensive than other methods because it only requires solutions to the mixed model equations. Aside from being feasible computationally, Method R has several other desirable properties. First, at least in single traits it returns estimates within the parameter space. Second, it is resistant to bias caused by assortative mating because it considers the relationship matrices. Also, it has been shown to be as resistant as REML to selection caused by non-random replacements (Snelling, 1994) and incomplete reporting (Kaiser et al., 1996); Cantet and Birchmeier (1998) initially reported bias under selection but later discovered that it was due to an error in programming. Disadvantages of Method R would include a larger sampling variance than other methods for the same data set, however, this disadvantage can be compensated by the ability of Method R to consider much larger data sets.

Recently, Druet et al. (2001) has found that $r_{ijkl}$ for covariances has a discontinuity because both its nominator and the denominator can have a value of 0. They suggested alternative formulas. Exact formulas for Method R in multiple traits are unknown.

**Selection of partial data**

The regressions do not specify which subset of data is selected. If subset is either a very small or a very large percentage of the complete data set, $u_i$'s will either be close to 0 or to $u$'s, and the regressions will have very large sampling error. Thus a size of 50% seems to be the a good choice. Also by inappropriate selection of records, one can create bias for the model with the reduced data.
set. A few types of subsets can be considered:
   a) random,
   b) by truncation over time
   c) by truncation over later records with repeated records.
Random selection seems to be attractive, especially that subsets can be easily changed by
changing seed in a random number generator, and sampling variance can be approximated as
variance of the samples.

Method R can have poor properties when the model does not contain all fixed effects. For
example, in the study by Misztal et al. (1997), heritability for stature with single record per animal
was 44%. When multiple records were considered, the heritability has increased to almost 70%.
After adding the adjustment for age, the heritability dropped to 45%.

In simple models (Duangajinda et al., 2001), Method R has been shown to be mostly unbiased.
Biases were present with small data sets and with phenotypic selection across but not within
contemporary groups.

**Numerical considerations**

When the problem is small enough so that LHS can be factorized, REML would provide better
estimates than Method R. Therefore the value of Method R is in larger problems where solutions
can be obtained only by iteration. For calculation of breeding values, the accuracy of solutions is
usually not critical, and an accuracy of 2 decimal digits is more than satisfactory. For method R,
convergence to
\[ r = 1 \pm 0.0001 \]
was found to be corresponding to the accuracy of the estimates of
\[ s \pm 5\% \]
Thus, \[ r = 1 \pm 0.0001 \] or calculated with at least 4 decimal significant digits should be considered the
minimum. To get that accuracy, the solutions to MME should be calculated with even higher
accuracy, 5 to 6 significant digits. Otherwise convergence of Method R may not be achieved.

Because Method R results in a problem of solving the nonlinear system of equations, the
convergence of this method is more similar to D than to DF.

**Solution of the nonlinear system of equations by Secant**

The problem of finding
\[ s : r(s) = 1 \]
was solved by the secant method, which initially is equivalent to the Newton method, where
derivatives are computed by numerical differentiation. This algorithm is equivalent to assuming that
Consequently, approximate solutions are obtained setting $r = 1$

$$\mathbf{1} = \mathbf{A}\mathbf{b} + b \quad \Rightarrow \quad \mathbf{b} = \mathbf{A}^{-1}(\mathbf{1} - b)$$

$A$ and $b$ can be found when $n + 1$ pairs of $\{s^{(i)}, r(s^{(i)})\}$ are available. Then,

$$r(s^{(i+1)}) - r(s^{(i)}) = \mathbf{A}[s^{(i+1)} - s^{(i)}]$$

denoting

$$\Delta s = [s^{(1)} - s^{(0)}, s^{(2)} - s^{(1)}, \ldots, s^{(n)} - s^{(n-1)}]$$

$$\Delta r = [r(s^{(1)}) - r(s^{(0)}), r(s^{(2)}) - r(s^{(1)}), \ldots, r(s^{(n)}) - r(s^{(n-1)})]$$

the equation can be written as

$$\Delta \mathbf{r} = \mathbf{A}\Delta s$$

Thus

$$\mathbf{A} = \Delta \mathbf{r} (\Delta \mathbf{s})^{-1}$$

$\mathbf{b}$ can be computed from the last point

$$\mathbf{b} = r(s^{(n)}) - \mathbf{A}s^{(n)}$$

which will result in

$$s^{(n+1)} = \mathbf{A}^{-1}[1 - r(s^{(b)}) + \mathbf{A}s^{(b)}] = s^{(b)} + \Delta \mathbf{s}(\Delta \mathbf{r})^{-1}[1 - r(s^{(b)})]$$

The algorithm contains heuristics to ensure the convergence of the algorithm that would otherwise result in estimates out of the parameter space or in a slower convergence rate.

1. Choose an initial guess $s^{(0)}$;
2. Assign $n$ extra points as follows:
   $$s^{(1)} = s^{(0)} + [\delta_1, 0, \ldots, 0]'$$
   $$s^{(2)} = s^{(1)} + [0, \delta_2, 0, \ldots, 0]'$$
   .....
   $$s^{(n)} = s^{(n-1)} + [0, 0, \ldots, \delta_n]'$$
3. Compute
   $$\{r(s^{(i)})\}, \ i = 0, n;$$
4. Compute differences:
   $$\Delta s = [s^{(1)} - s^{(0)}, s^{(2)} - s^{(1)}, \ldots, s^{(n)} - s^{(n-1)}]$$
   $$\Delta r = [r(s^{(1)}) - r(s^{(0)}), r(s^{(2)}) - r(s^{(1)}), \ldots, r(s^{(n)}) - r(s^{(n-1)})].$$
5. Compute next-round solutions
   $$s^{(n+1)} = s^{(n)} + \Delta s (\Delta r)^{-1} [1 - r(s^{(n)})]$$
6. Limit changes if too large:
maxchange = \max_i \left( \frac{|s_i^{(n+1)} - s_i^{(n)}|}{|s_i^{(n)}|} \right)

if maxchange > \gamma then \quad s^{(n+1)} = s^{(n)} + \gamma/\text{maxchange} \, (s^{(n+1)} - s^{(n)})

6. If change large, compute n new points as before; if changes small, use all but one last point.

if \max_i |r_i(s^{(n+1)}) - 1| > \epsilon then

\quad \text{if } \max_i |s_i^{(n+1)} - s_i^{(n)}| / |s_i^{(n)}| > \lambda

\quad \text{set } s^{(0)} = s^{(n+1)}
\quad \text{goto 2}

\quad \text{else}

\quad s^{(0)} = s^{(1)}, s^{(1)} = s^{(2)}, ..., s^{(n-1)} = s^{(n)}, s^{(n)} = s^{(n+1)}
\quad \text{goto 3.}

7. Optionally, refine s using previously computed matrices

\quad s^{(n+1)} = s^{(n+1)} + \Delta s \, (\Delta r)^{-1} [1 - r(s^{(n+1)})].

In program JAADOM, changes between samples \delta_i were chosen to correspond to a change of variance of \sigma_i^2/200, maximum relative changes were limited to \gamma = 0.7, the stopping criterion was \epsilon = 0.0001, and changes were regarded as small for \lambda = 0.1.

\textbf{Numerical example}

\quad s^{(0)} = [0.6, 0.9]^t
\quad s^{(1)} = s^{(0)} + \{-0.0017946, 0, ..., 0\} = [0.5982054, 0.9]^t
\quad s^{(2)} = s^{(1)} + \{0, -0.00403189, 0, ..., 0\} = [0.5982054, 0.89596811]^t

\quad r(s^{(0)}) = [1.008687, 1.009935]
\quad r(s^{(1)}) = [1.008587, 1.009897]
\quad r(s^{(2)}) = [1.008552, 1.009800]

\quad \Delta s = \begin{bmatrix} -0.0017946 & 0.0000 \\ 0.0000 & -0.00403189 \end{bmatrix}

\quad \Delta r = \begin{bmatrix} -0.000100 & -0.000035 \\ -0.000038 & -0.000097 \end{bmatrix}

\quad s^{(n+1)} = [0.4940763, 0.5808855]^t
\quad r(s^{(n+1)}) = [1.000021, 0.9973247]^t.

Because
1 - 0.9973247 = 0.0026753 > 0.0002, \( s^{(0)} \) is set to \( s^{(n+1)} \), and computations are repeated from step 2. After that, the procedure converges at
\[
\mathbf{s}^{(n+1)} = [0.4791, 0.6592]'
\]
where
\[
r(\mathbf{s}^{(n+1)}) = [1.00008, 0.99981]
\]
It would be worthwhile to find out whether better performance could be obtained by quasi-Newton methods.
Bayesian methods and Gibbs sampling

Following Wang et al. (1993), in the single-trait model
\[ \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i} \mathbf{Z}_{i}\mathbf{u}_{i} + \mathbf{\epsilon} \]
the conditional distribution which generates the data is
\[ \mathbf{y|}\boldsymbol{\beta},\mathbf{u}_{1},\ldots,\mathbf{u}_{c}\mathbf{,}\sigma_{e}^{2} \sim N(\mathbf{X}\boldsymbol{\beta} + \sum_{i}^{c} \mathbf{Z}_{i}\mathbf{u}_{i}, \mathbf{R}\sigma_{e}^{2}) \]
with the density
\[ p(\mathbf{y} \mid \boldsymbol{\beta},\mathbf{u}_{1},\ldots,\mathbf{u}_{c}\mathbf{,}\sigma_{e}^{2},\sigma_{u}^{2},\sigma_{v}^{2},\sigma_{d}^{2}) \propto \frac{1}{(\sigma_{e}^{2})^{n/2}} \exp \left\{ -\frac{1}{2\sigma_{e}^{2}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i}^{c} \mathbf{Z}_{i}\mathbf{u}_{i})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{i}^{c} \mathbf{Z}_{i}\mathbf{u}_{i}) \right\} \]

One can specify distributions for \( \beta \), \( \mathbf{u} \)'s and \( \sigma \)'s. The distribution of \( \beta \) is flat
\[ p(\beta) \propto \text{constant} \]
The distribution for \( \mathbf{u} \)'s is multivariate normal
\[ \mathbf{u}_{i} \mid \mathbf{G}_{i}\sigma_{u}^{2} \sim N(\mathbf{0}, \mathbf{G}_{i}\sigma_{u}^{2}) \]
with the density
\[ p(\mathbf{u}_{i} \mid \mathbf{G}_{i}\sigma_{u}^{2}) \propto \left[ \frac{1}{(\sigma_{u}^{2})^{2}} \exp \left\{ -\frac{1}{2\sigma_{u}^{2}} (\mathbf{u}_{i} / \mathbf{G}_{i}^{-1}\mathbf{u}_{i}) \right\} \right] \]
and for the variance components, it can be scaled inverted \( \chi^{2} \)
\[ p(\sigma_{e}^{2}) \propto (\sigma_{e}^{2})^{-\nu_{e}/2} \exp \left\{ -\frac{1}{2} \frac{\nu_{e}}{\sigma_{e}^{2}} \right\} \]
\[ p(\sigma_{u}^{2}) \propto (\sigma_{u}^{2})^{-\nu_{u}/2} \exp \left\{ -\frac{1}{2} \frac{\nu_{u}}{\sigma_{u}^{2}} \right\} \]
where \( s_{e} \) is the prior value and \( \nu_{e} \) can be viewed as the degree of belief (uniformative prior for \( \nu_{e}=0 \)). The joint posterior density with flat priors is
\[ p(\boldsymbol{\beta},\mathbf{u}_{1},\ldots,\mathbf{u}_{c}\mathbf{,}\sigma_{u}^{2},\sigma_{v}^{2},\sigma_{d}^{2}) \mid \mathbf{y} \propto p(\mathbf{y} \mid \boldsymbol{\beta},\mathbf{u}_{1},\ldots,\mathbf{u}_{c}\mathbf{,}\sigma_{u}^{2},\sigma_{v}^{2},\sigma_{d}^{2}) \sum_{i=1}^{c} p(u_{i}\sigma_{u}^{2}) p(\sigma_{u}^{2}) \]
or
Information about variance components could be obtained from marginal distributions by integrating all unnecessary effects/parameters out

\[ p(\sigma^2 | y) = \int p(\beta, u, \sigma^2 | y) \, d\beta \, du \, d\sigma^2 \]

\[ p(\sigma^2 | y) = \int p(\beta, u, \sigma^2 | y) \, d\beta \, du \, d\sigma^2 \]

From the marginal posterior distribution one can calculate mean, mode, standard deviation or confidence intervals.

For realistic models, the analytical integration is hard or impossible.

Gibbs sampling (Geman and Geman, 1984) is an easy although computationally-intensive method of numerical integration of likelihoods whenever conditional distributions are easy to compute. The method would progress as follows:

1. Assign priors to \( \beta, u, \sigma^2 \)

2. Sequentially calculate samples from conditional distributions

\[ \beta | u, \sigma^2 \]

\[ u | \beta, \sigma^2 \]

\[ \sigma^2 | \beta, u \]

\[ \sigma^2 | \beta, u \]

\[ \sigma^2 | \beta, u \]

\[ \sigma^2 | \beta, u \]
Elements of $\beta$ or $u$ can be sampled either as by vector or by individual elements.

3. Calculate 2. repeatedly. After a period of burn-in, samples generated above will be from their respective marginal distributions.

The conditional distributions are constructed by assuming the conditional parameters fixed and by dropping terms from the posterior density that are dependent on conditional parameters only. Thus, the conditional posterior density for $\beta$ is

$$p(\beta \mid u_1, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y) \propto \exp \left\{ \frac{1}{2\sigma^2_\epsilon} \left( y - X\beta - \sum_{i=1}^c Z_iu_i \right) \left( y - X\beta - \sum_{i=1}^c Z_iu_i \right)' \right\}$$

This corresponds to the distribution

$$X'X\beta \mid u_1, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y \sim N \left( (y - \sum_{i=1}^c Z_iu_i)'(y - \sum_{i=1}^c Z_iu_i), \frac{\sigma^2_\epsilon}{\sigma^2_y} \right)$$

or

$$X'X\beta \mid u_1, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y \sim N \left( X'(y - \sum_{i=1}^c Z_iu_i), X'X\sigma^2_\epsilon \right)$$

and finally

$$\beta \mid u_1, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y \sim N \left( (X'X)^{-1}X'(y - \sum_{i=1}^c Z_iu_i), (X'X)^{-1}\sigma^2_\epsilon \right)$$

Similarly, for the $i$-th random effect $u_i$, the distribution is

$$u_i \mid \beta, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y \sim N \left( P^{-1}_iZ_i(y - X\beta - \sum_{j=1, j\neq i}^c Z_ju_j), P^{-1}_i\sigma^2_\epsilon \right)$$

where

$$P_i = Z_i'Z_i + G_i^{-1}\frac{\sigma^2_\epsilon}{\sigma^2_{ui}}$$

For the residual variance, the full conditional density is a scaled inverted $\chi^2$.  

$$\sigma^2_\epsilon \mid \beta, u_1, \ldots, u_n, \sigma^2_{\epsilon 11}, \ldots, \sigma^2_{\epsilon c}, \sigma^2_y \sim \chi^2 \left( \frac{\sigma^2_\epsilon}{\sigma^2_y} \right)$$
\[ p(\sigma_c^2 \mid \beta, u_1, \ldots, u_c, \sigma_{u1}, \ldots, \sigma_{uc}, y) \propto \frac{1}{(\sigma_c^2)^{n_t/2-1}} \exp \left\{ -\frac{1}{2\sigma_c^2} (y - X\beta - \sum_i^c Z_i u_i)'(y - X\beta - \sum_i^c Z_i u_i) \right\} \]

with the distribution

\[ \sigma_c^2 \mid \beta, u_1, \ldots, u_c, \sigma_{u1}, \ldots, \sigma_{uc}, y \sim \nu_0^2 \chi_0^2 \]

where

\[ \nu_0 = n_t + \nu \]

and

\[ \mathbf{s}_c^2 = \frac{(y - X\beta - \sum_{i=1}^c Z_i u_i)'(y - X\beta - \sum_{i=1}^c Z_i u_i)}{\nu_0} = \frac{e' e}{\nu_0} \]

For the other variance components, the distributions are similar

\[ p(\sigma_{ui}^2 \mid \beta, u_1, \ldots, u_c, \sigma_{u1}, \ldots, \sigma_{uc}, \sigma_{e}^2, y) \propto \frac{1}{(\sigma_{ui}^2)^{q_{ui}/2-1}} \exp \left\{ -\frac{e_i G_i^{-1} u_i}{2\sigma_{ui}^2} \right\} \]

with the distribution

\[ \sigma_{ui}^2 \mid \beta, u_1, \ldots, u_c, \sigma_{u1}, \ldots, \sigma_{uc}, \sigma_{e}^2, y \sim \nu_{ui} \chi_{0,ui}^2 \]

where

\[ \nu_{ui} = q_i + \nu_{ui} \]

and

\[ s_{ui}^2 = \frac{e_i G_i^{-1} u_i}{\nu_{ui}} \]

A multivariate distribution for a complete fixed or random effect is hard to obtain. However, a distribution for a single level of a fixed or random effect is simple. Let

\[ \beta = \{\beta_i\}, i=1,m \]

and

\[ \mathbf{W} = [\mathbf{X} \ Z], \ \mathbf{\Theta} = \begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} \]

The conditional marginal density for \( \beta_i \) is

\[ p(\beta_{i1}, \ldots, \beta_{i,1}, \ldots, \beta_{im} \mid \beta_{-i1}, \ldots, \beta_{-im}, u_{i1}, \ldots, u_{ic}, \sigma_{u1}, \ldots, \sigma_{uc}, \sigma_e^2, y) \propto \]
where \(y^{(bi)}\) are records containing \(\beta_i\) only, and \(Z^{(bi)}_j\) corresponds to \(y^{(bi)}\). The corresponding distribution is

\[
\sum_j x_{ji} \beta_i \mid \beta_{i-1}, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_m, \mu, \sigma_{\epsilon}^2, \sigma_{\mu}^2, \sigma_{\epsilon}^2 y \sim N \left( \left( y^{(bi)} - \sum_{j,k \neq i} x_{jk} \beta_k - \sum_{j=1}^c Z^{(bi)}_j \mu \right), \sigma_{\epsilon}^2 \right)
\]

or

\[
\beta_i \mid \beta_{i-1}, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_m, \mu, \sigma_{\epsilon}^2, \sigma_{\mu}^2, \sigma_{\epsilon}^2 y \sim N \left( \left[ \left( y^{(bi)} - \sum_{j,k \neq i} x_{jk} \beta_k - \sum_{j=1}^c Z^{(bi)}_j \mu \right) \frac{1}{x_i x_{ji}}, \frac{\sigma_{\epsilon}^2}{x_{ji}} \right], \sigma_{\epsilon}^2 \right)
\]

which can be further simplifying by expressing the last equation in terms of RHS and LHS of the MME

\[
\beta_i \mid (\text{everything else}) \sim N \left( \frac{\text{RHS}_{\beta_i} - \sum_{j \neq i} \text{LHS}_{\beta_i} \beta_j - \sum_{j,k \neq i} \text{LHS}_{\beta_i \beta_k} \mu_{jk}}{\text{LHS}_{\beta_i}}, \frac{\sigma_{\epsilon}^2}{\text{LHS}_{\beta_i \beta_j}} \right)
\]

The expression for the mean is a solution in the Gauss-Seidel iteration. A similar distribution can be obtained for \(j\)-th level of random effect \(i\): \(u_{ij}\)

\[
u_{ij} \mid (\text{everything else}) \sim N \left( \frac{\text{RHS}_{u_{ij}} - \sum_{k \neq i} \text{LHS}_{u_{ij}} \theta_k}{\text{LHS}_{u_{ij}}}, \frac{\sigma_{u_j}^2}{\text{LHS}_{u_{ij}} \text{LHS}_{\theta_k}} \right)
\]
and the data

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>( y_{ijk} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

Assume priors:
\( \mu = 0, \sigma^2 = 1, \sigma^2_e = \frac{1}{1} \)

The mixed model equations are:

\[
\begin{bmatrix}
2 & 0 & 2 & 0 \\
0 & 3 & 2 & 1 \\
2 & 2 & 14 & 0 \\
0 & 1 & 0 & 11
\end{bmatrix}
\begin{bmatrix}
\hat{g}_1 \\
\hat{g}_2 \\
\hat{s}_1 \\
\hat{s}_2
\end{bmatrix}
=
\begin{bmatrix}
7 \\
18 \\
18 \\
7
\end{bmatrix}
\]

The initial distributions are:

\( g_1 | \ldots \sim N(7/2, 1/2) \), for example sampled as 4
\( g_2 | \ldots \sim N(18/3, 1/3) \), for example sampled as 6
\( s_1 | \ldots \sim N([18-2*4 -2*6]/14, 1/14) \), for example sampled as -.2
\( s_2 | \ldots \sim N([7-1*6]/11, 1/11) \), for example sampled as .3

\[
e' = [ (3-4- -.2), (4-4- -.2), (5-6- -.2), (6-6- -.2), (7-6-.3)] = (-.8, .2, -.8, .2, .7)
\]

\[
e'e=1.85
\]

\[
\sigma^2_{e} | \ldots \sim 5 * 1.85/5 \chi^2, \text{ for example sampled as } 1.1
\]

\[
\sigma^2_{e} | \ldots \sim 2 * .13/2 \chi^2, \text{ for example sampled as } .05
\]

The next round distributions will be:

\( g_1 | \ldots \sim N(7-2*-.2)/2, 1.1/2) \), for example sampled as 3.5
\( g_2 | \ldots \sim N((18-2*-3)/3, 1.1/3) \), for example sampled as 5.5
\( s_1 | \ldots \sim N((18-2*3.5 -2*5.5)/15, 1/15) \), for example sampled as .1
\( s_2 | \ldots \sim N((7-5.5)/12, 1/12) \), for example sampled as .2
Using the Gibbs sampler

Typically, samples from the Gibbs sampler would show an initial period of burn-in, followed by a repetitive pattern. This is shown in the idealized chart below; the real curve would not be as smooth.

The burn in period lasts usually between 500 and 10000 samples, although for slowly converging sampler that number could be considerably higher. One way to determine the burn-in period is to start two chains with the same random number generators but different priors.
Convergence of the two chains indicate the end of the burn-in. Repetitive pattern that follows the burn-in period is due to the largest eigenvalue of the iteration matrix, and its length is dependent on the model and data structure. Because of repetitions, the subsequent samples are correlated, and in order to make inferences about parameters of the model, it makes sense to analyze only every n sample, where n can be from 10 to 1000. Selected samples can be used to create an approximate marginal density, where the accuracy of the approximation will depend on the number of samples and on the correlation of subsequent samples. The approximate density could look as follows.

The total number of samples necessary to plot a density could be as high as 1 million. Fewer samples are needed to estimate a mean or a mode of the distribution.
Multiple-trait

In multiple traits, the marginal distributions for \( u \)’s and \( \beta \)’s can be obtained as in single traits, by using an analogous of the Gauss-Seidel iteration. For variances, a univariate inverted chi-square distribution becomes a multivariate inverted Wishart distribution.

For simplicity, assume a \( t \) trait model with a single random effect. Denote \( u \) and \( e \) as \( i \)-th trait random samples for the random effect and the residual, respectively. The distribution for the variance components of the random effect is

\[
G_e \sim \text{IW}[n_g + v_g \cdot (S_g + P_e)^{-1}]
\]

where \( n_g \) is number of levels in the random effect, \( v_g \) is degrees of belief (equal to \(-t-1\) for flat priors), and

\[
P_e = \begin{bmatrix}
u_1'G^{11}u_1 & u_1'G^{12}u_2 \\
u_2'G^{21}u_1 & u_2'G^{22}u_2 \\
\end{bmatrix}
\]

The distribution for the residual variance components is

\[
R_e \sim \text{IW}[n + v_e \cdot (S_e + P_e)^{-1}]
\]

where \( n \) is the maximum number of records (including missing), \( v_e \) is degrees of belief (equal to \(-t-1\) for flat priors), and

\[
P_e = \begin{bmatrix}e_1'e_1 & e_1'e_2 \\
e_2'e_1 & e_2'e_2 \\
\end{bmatrix}
\]

If some records are missing, they are predicted from the distribution “missing records”|\( u, \beta, R, G, \)”known records” each round and then used as if they were known. Let subscript \( m \) denote quantities associated with missing traits, and \( p \) associated with present traits. The missing records can be predicted as:

\[
y_m \sim \text{MVN}[E(y_m|\ldots), \text{Var}(y_m|\ldots)]
\]

with
Convergence properties

The Gibbs sampler’s convergence is tied to the convergence of the method of Gauss-Seidel, and to priors. When the corresponding convergence of Gauss-Seidel is slow, the convergence of Gibbs sampler will be slow too. One way to increase the convergence speed in multiple traits is to use block iteration, where samples are generated by block of traits.

The Gibbs sample may diverge with certain priors. For example, uninformative priors for variance components ($v=0$) are often worse than flat priors ($v=-t-1$).

Computing implementations

The Gibbs sampler require a large number of rounds. Therefore, its efficiency is quite important. In the most obvious implementation, the LHS is recreated each round, with procedures to set up the equations taking perhaps as much as 95% of computing time. Also, memory requirements may be fairly steep. Several algorithms to cut time and memory are possible although some of them are hard to implement for complicated models. In all cases, the data can be stored in memory or read by fast routines.

Single trait models

In single traits, when working with variance ratios, only the part of the LHS due to (co)variance matrices changes. If the part of LHS due to records and the relationship matrices are kept in memory separately, each equation can be reconstructed by combining these matrices and the current variance ratios, without recomputing the LHS.

Multiple trait models with same model per trait

In this case, only the structure of the LHS for single trait needs to be stored, and the multiple-trait LHS is reconstructed dynamically using the current (co)variance matrices, as in the previous case. If some traits are missing, the data needs to be read every round to recreate the missing traits, but the LHS is not affected.

Other multiple trait models

In large multitrait models, contribution from relationship matrices can be much larger than contributions from records because single-trait numerator relationship matrices need to be
duplicated over for all combinations of traits and possible sets of correlated effects. To reduce storage and computations, only the contributions from records can be stored, and contributions from pedigrees can be recreated from a single-trait copy of the numerator relationship matrix.

Alternately, the model can be redefined so as to have the same model per trait. For example, let the model be identical except for different fixed management effect $m$ in the $i$-th trait:

$$y_i = X_i m_i + ...$$

Redefine that model so that for each trait management effects from all traits are present:

$$y_i = X_i m_{ii} + X_{2i} m_{2i} + X_{3i} m_{3i} + ...$$

and treat the effects as random with (co)variances:

$$\text{cov}(m_{ij}, m_{lk}) = \begin{cases} a\rightarrow\infty \text{ if } i=j=k \\ b\rightarrow0 \text{ otherwise} \end{cases}$$

Variance close to 0 cause an effect to disappear while variance of infinity makes an effect fixed; In practical implementation, values of $a$ should be very high and $b$ very small but in a range so that numerical errors can be avoided. Also, the Gibbs sampler need to be modified not to sample variances of $m$.

Large models

One possibility with large data is by iteration-on-data by Gauss-Seidel, as discussed in an earlier chapter. Several copies of the data and relationship files may be needed. Programming may become quite complicated for complex models but is simple for simple models.

General models

Updating MME in linked-list form may be time consuming. Matrices in HASH format are faster to set up, but they need to be converted to IJA format so that rows of the LHS are available sequentially. Both creation of the hash matrix for the first time and conversion to the IJA format take time. Schnyder (1998, personal communication) noticed that, after the first round, the structure of the matrices in the IJA and HASH does not change; only values change. He suggested the following algorithm:

- a) after the first round, create links to facilitate fast conversion between HASH and IJA structures,
- b) before every consecutive round, only zero HASH values but not indices.

In the end, the creation of the LHS can be 2-3 times faster without sacrificing any modeling
flexibility.
**Ways to reduce the cost of multiple traits**

In $t$ traits, the computing costs increase as follows

<table>
<thead>
<tr>
<th>Approximate Cost</th>
<th>Arithmetic Operations</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>solutions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iterative</td>
<td>$&gt;t^2$</td>
<td>$t^2$</td>
</tr>
<tr>
<td>finite</td>
<td>$t^3$</td>
<td>$t^2$</td>
</tr>
<tr>
<td>REML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derivative-free</td>
<td>$t^5$</td>
<td>$t^2$</td>
</tr>
<tr>
<td>Derivative</td>
<td>$t^3$</td>
<td>$t^2$</td>
</tr>
</tbody>
</table>

The cost of obtaining iterative solutions includes the cost of increased number of rounds because of a lower convergence rate, which is especially low when traits are missing.

**Canonical transformation**

Let the $t$-trait model be

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \mathbf{e}$$

with

$$\text{Var}(\mathbf{u}) = \mathbf{G}_0 \otimes \mathbf{G}_0, \quad \text{Var}(\mathbf{e}) = \mathbf{R}_0 \otimes \mathbf{I}$$

where $\mathbf{G}_0$ and $\mathbf{R}_0$ are the $t \times t$ variance-covariance matrices. For individual trait $i$, the model is

$$\mathbf{y}_i = \mathbf{X}^*\beta_i + \mathbf{Z}^*\mu_i + \mathbf{e}_i$$

where $\mathbf{X}^*$ and $\mathbf{Z}^*$ are single-trait design matrices, same for all traits. The covariances between the traits are

$$\text{Cov}(\mathbf{u}_i, \mathbf{u}_j) = (\mathbf{g}_0)_i^j \mathbf{G}_0, \quad \text{Cov}(\mathbf{e}_i, \mathbf{e}_j) = (\mathbf{r}_0)_i^j \mathbf{I}$$

For any symmetric semi-positive-definite $\mathbf{G}_0$ and $\mathbf{R}_0$ there is a transformation $\mathbf{M}$ such that

$$\mathbf{M} \mathbf{G}_0 \mathbf{M}' = \mathbf{D}$$

and

$$\mathbf{M} \mathbf{R}_0 \mathbf{M}' = \mathbf{I}$$
where \( D \) is the diagonal matrix. For each record \( j \), the vector of observation is 
\[
(y_{{1j}}, y_{{2j}}, \ldots, y_{{qj}})
\]
Multiply each vector of observation by \( M \)
\[
M(y_{{1j}}, y_{{2j}}, \ldots, y_{{qj}}) = (y_{{1j}}^m, y_{{2j}}^m, \ldots, y_{{qj}}^m)
\]
which for all observations can be written as
\[
y = (M \otimes I)y = y^m
\]
Then, the model becomes
\[
y^m = X\beta^m + Zu^m + e^m
\]
with variances
\[
\text{Var}(u^m) = D \otimes G, \quad \text{Var}(e^m) = I
\]
where the transformed traits are uncorrelated. For each transformed trait \( i \)
\[
y_i^m = X_i^* \beta_i^m + Z_i u_i^m + e_i^m
\]
with covariances between the traits
\[
\text{Cov}(u_i^m, u_j^m) = \begin{cases} d_{ij} G, & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}, \quad \text{Cov}(e_i^m, e_j^m) = \begin{cases} 1, & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}
\]
and the mixed model equation for each trait \( i \) are
\[
\begin{bmatrix}
X_i^m'y_i^m \\
Z_i^m'y_i^m + \frac{1}{d_{ii}} G^{-1} u_i^m
\end{bmatrix}
\begin{bmatrix}
\beta_i^m \\
u_i^m
\end{bmatrix}
= \begin{bmatrix}
X_i^m'y_i^m \\
Z_i^m'y_i^m + \frac{1}{d_{ii}} G^{-1} u_i^m
\end{bmatrix}
\]
Thus, the single-trait equations after the canonical transformation are uncorrelated and can be solved by single trait procedures. To compute the solutions on the original scale, back transform
\[
u = (M^{-1} \otimes I)u^m
\]
Note that the canonical transformation can be performed only when the following apply:
- same model for all traits
- no missing traits for any observation
- single random effect

The transformation matrix can be constructed as follows. Decompose \( R_o \) into eigenvectors and eigenvalues
\[
R_o = V D_r V'
\]
Then decompose again
\[ D_r^{-\frac{1}{2}} V' G_0 V D_r^{-\frac{1}{2}} = W D W' \]

The transformation matrix is
\[ M = W' D_r^{-\frac{1}{2}} V' \]

For verification
\[ M G_0 M' = W' D_r^{-\frac{1}{2}} V' G_0 V D_r^{-\frac{1}{2}} W = W' W D W' W = D \]
\[ M R_0 M' = W' D_r^{-\frac{1}{2}} V' V D_r V' V D_r^{-\frac{1}{2}} W = W D_r^{-\frac{1}{2}} D_r D_r^{-\frac{1}{2}} W' W' = I \]

Canonical transformation and REML

Estimation of REML by canonical transformation can be done by the following steps:

a) Assign priors for \( R_0 \) and \( G_0 \)
b) Find transformation and transform the data
c) Estimate single-trait variance components
d) Combine the estimates; the formulas below assume the first-derivative algorithm

\[ r_{\hat{u}}^m = \frac{\beta_1^m X' y_i^m - \mathbf{u}_1^m Z' y_i^m + \mathbf{u}_1^{m'} G^{-1} \mathbf{d}_i^m}{d_{ij}} \]

\[ r_{\hat{u}}^m = \left( \frac{(y_i^m y_i^m - \beta_1^m X' y_j^m - \mathbf{u}_1^m Z' y_j^m)}{n_r} \right) \text{ for } i \neq j \]

\[ g_{\hat{u}}^m = \frac{\mathbf{u}_1^{m'} G^{-1} \mathbf{d}_i^m}{n_k - \frac{\text{tr}(G^{-1} C^{m,i})}{d_{ij}}} \]

\[ g_{\hat{u}}^m = \frac{\mathbf{u}_1^{m'} G^{-1} \mathbf{d}_j^m}{n_k} \]

where \( n_r \) is number of records, \( n_k \) is number of levels in the random effect, and \( C^{m,i} \) is the
block of the inverse of the coefficient matrix corresponding to random effect i. Form

\[ \hat{R}_0^{\text{na}} = \{\tilde{x}_{ij}\}, \quad \hat{G}_0^{\text{na}} = \{\tilde{g}_{ij}\} \]

The new estimate of (co)variance components on the original scale are

\[ \hat{R}_0 = M^{-1}\hat{R}_0^{\text{na}}M^{-1'}, \quad \hat{G}_0 = M^{-1}\hat{G}_0^{\text{na}}M^{-1'} \]

e) If no convergence, repeat from b)

Contrary to general model REML, the cost of canonical transformation REML is only linear with the number of traits, and there is no increase in memory over single-trait. However, all limitations of the canonical transformation apply.

**Numerical properties of canonical transformation**

With many traits, some traits are likely to become almost linear functions of the other traits. Therefore, matrices \( R_0 \) and \( G_0 \) may become almost singular, and the whole coefficient matrix can become almost singular. This can easily be shown for a mixed model. The condition number of a matrix is defined as

\[ \text{cond}(B) = |B| |B^{-1}| \]

where higher condition numbers are associated with poorer numerical properties. In a general multiple-trait fixed model, the coefficient matrix is

\[ R_0 \otimes X'X \]

with the condition number

\[ \text{cond}(R_0 \otimes X'X) = \text{cond}(R_0) \text{cond}(X'X) \]

Similar increase in condition number occurs in mixed models, where that condition is also a function of \( G_0 \) but it is more difficult to show analytically. High condition number of a matrix leads to numerical errors in the factorization, or to extremely low convergence rate in iterative methods. In variance component estimation, the computed rank of the coefficient matrix may change from round to round.

In the canonical transformation, the coefficient matrix in the fixed model is

\[ X'X \]

and thus its condition is independent on \( R_0 \).

In canonical transformation, the near singularity \( R_0 \) and \( G_0 \) shows in elements of \( D \) and \( D' \) being almost 0. Computer program MTC contains the following heuristics with respect to the canonical
transformation. For the residual eigenvalues:

\[ \text{if } (D_r)_{ii} = 0 \text{ then set } (D_r^{-1})_{ii} = 0 \]

and for the “random effect eigenvalues”:

\[ d_{ii} = \max(2, d_{ii}), \quad d_{ii} = \min(d_{ii}, 100,000) \]

While general-model REML by any algorithm for more than 3-6 traits isn’t usually stable numerically, the canonical transformation REML was known to be successful for 32 traits!

**Extensions in canonical transformation**

**Multiple random effects**

Extend the model to \( p \) multiple random effects

\[ y = X\beta + \sum_k Z_k u_k + \epsilon \]

with

\[ \text{Var}(u_k) = G_{0k} \otimes G_k \]

We need such a transformation that

\[ M R_o M' = I \]

and

\[ M G_{0k} M' = D_k, \quad k=1,p \]

where \( D_k \) are diagonal matrices. In general, such a transformation does not exist for \( p > 1 \). Lin and Smith (4) hypothesized that in practical situations an approximate equation can hold

\[ M G_{0k} M' \approx D_k \]

The algorithm to calculate \( M \) would be similar as before

\[ R_o = VD_r V' \]

\( W \) and \( D_r \) in the decomposition

\[ D_r^{-\frac{k}{2}} V' G_{0k} VD_r^{-\frac{k}{2}} = WD_i W' \]

can be obtained approximately by the FG algorithm (Flury, 1988)

The transformation matrix is

\[ M = W D_r^{-\frac{k}{2}} V' \]

For a good approximation, matrices \( G_{0k} \) must be alike. In general multiple traits with \( t \) traits and \( p \)
random effects, the total number of parameters to estimate is:
\[ p t (t+1)/2 \]
With the approximation, the number of parameters would be
\[ (p-1) t + t(t+1)/2 \]
For 2 random effects and 4 traits, the number of parameters is 20 in the first case and 14 in the second case.

Misztal et al. (1993) has obtained very good results with the approximation for type and milk traits. This indicates that the multiple-trait variance-covariance structure in several random effects is similar. Reversely, in that case, general REML would estimate more parameters than necessary.

**Missing traits**

Ducrocq and Besbes (1993) have developed a method to apply the canonical transformation when records on some traits are missing. Let \( y_m \) be the \( m \)-th observation partitioned as \((y_{ma}, y_{mb})\), where \( y_{ma} \) was recorded but \( y_{mb} \) is missing. Denote the model for each part as
\[
\begin{align*}
y_{ma} &= x_{ma}^t \alpha + z_{ma}^t u + e_{ma} \\
y_{mb} &= x_{mb}^t \beta + z_{mb}^t u + e_{mb}
\end{align*}
\]
Let (k) denote the current round of iteration. Missing record \( y_{mb} \) can be predicted using the equation
\[
\hat{y}_{mb}^{(k)} = x_{mb}^t \beta^{(k)} + z_{mb}^t u^{(k)} + E(e_{mb} | y_{ma}^{(k)}, \beta^{(k)}, u^{(k)})
\]
\[
= x_{mb}^t \beta^{(k)} + z_{mb}^t u^{(k)} + R_{0,\beta} R_{\alpha,\alpha}^{-1} (y_{ma}^{(k)} - x_{ma}^t \alpha^{(k)} + z_{ma}^t u^{(k)})
\]
where \( R_{0,\alpha} \) is a block of \( R_0 \) corresponding to traits \( \alpha \) and \( \beta \), and \( R_{0,\alpha} \) is a square block of \( R_0 \) corresponding to trait \( \alpha \).

To solve the mixed model equations, in each round predict missing records and use them for one round of iteration as if they were known. Then, using the new estimates, predict the missing records again. Continue until convergence.

Canonical transformation with missing traits requires that in each round the previous round estimates be untransformed, data read, the missing records predicted, and all the records transformed.

Formulas to apply the modification for missing records to REML has not been derived. Thus, the procedure is applicable only for obtaining the solutions.
When a random effect is absent in one or more traits, one can still define that effect for all traits but set the variance(s) for these traits very high so that this effect is practically eliminated. If there are many random effects and the approximate diagonalization is being used, this could have a side effect of poor diagonalization.

There are two methods to handle different fixed effects per trait. In a simple but approximate method by Gengler and Misztal (1995), one creates a model containing all fixed effects. Then clusters of traits with identical models are created. Then, such clusters are augmented to all traits by predicting the missing observations, and the missing-record procedures are applied. The model for each cluster includes all fixed effects, but effects undesirable in the given cluster are assigned “dummy” levels. Consider the data

<table>
<thead>
<tr>
<th>Cow</th>
<th>Fixed effects</th>
<th>Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Herd-year-season</td>
<td>Herd-time of classification</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

where herd-year season applied to milk and fat, and heard-time-of-classification applied to stature and udder width.

The data can be restructured to

<table>
<thead>
<tr>
<th>Cow</th>
<th>Fixed effects</th>
<th>Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Herd-year-season</td>
<td>Herd-time of classification</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
<td>1</td>
</tr>
</tbody>
</table>
where M denotes missing trait and D denotes the dummy effect. For example, if herd year season had 100 levels, and herd-time-classification had 1000 levels, D=101 for herd-year-season and D=1001 for herd-time-classification. Because the herd effects do not overlap, they can be joined to reduce storage.

This method may be an approximation because the residuals of different clusters are assumed uncorrelated.

Another method was proposed by Ducrocq and Chapuis (1997). Define all effects for all traits. Let \( \beta \) be solutions to all the fixed effects where all fixed effects are applied to all traits, and let \( \beta^c \) be solutions constrained so that solutions to fixed effects that should be absent for certain traits are 0.

Find such a transformation \( P \) so that

\[
\beta^c = P\beta
\]

and then find an equivalent transformation on the canonical transformation side

\[
\beta^c_{\text{can}} = P_{\text{can}}\beta_{\text{can}}
\]

Solve iteratively using all the fixed effects, and the end of each round apply the above transformation, either on the original or on the transformed scale.
Implicit representation of general multiple-trait equations

Tier and Graser (1991) presented a way to drastically reduce the amount of storage for multiple-trait mixed model equations. Initially assume that all traits follow the same model but some traits are missing, and that there is only one random effect. For each missing trait combination $i$, the mixed model can be written as

$$ y_{i} = W_{i}\theta + e_{i} $$

The left hand side of the mixed point equations can be written as

$$ R_{o}^{(i)}W_{i}^{\top}W_{i} + R_{o}^{(2)}W_{2}^{\top}W_{2} + \ldots + R_{o}^{(q)}W_{q}^{\top}W_{q} + G_{o}^{-1}A^{-1} $$

$$ = R_{o}^{(1)}C^{(i)} + R_{o}^{(2)}C^{(2)} + \ldots + R_{o}^{(q)}C^{(q)} + G_{o}^{-1}C^{(q)} $$

where $R_{o}^{(i)}$ is the inverse of the residual (co)variance matrix for a missing trait combination $i$. For $t$ traits, the memory needed to store the coefficient matrix can be close to $t^2$ greater than that for single traits because scalars in single trait are being expanded to txt blocks. Each txt block corresponding to the $jk$-th single-trait equation can be computed as

$$ R_{o}^{(1)}c_{jk}^{(1)} + R_{o}^{(2)}c_{jk}^{(2)} + \ldots + R_{o}^{(q)}c_{jk}^{(q)} + G_{o}^{-1}c_{jk}^{(q)} $$

$$ = \begin{bmatrix} c_{jk}^{(1)} & c_{jk}^{(2)} & \ldots & c_{jk}^{(q)} & c_{jk}^{(q)} \end{bmatrix} \begin{bmatrix} R_{o}^{(1)} \\ R_{o}^{(2)} \\ \vdots \\ R_{o}^{(q)} \\ G_{o} \end{bmatrix} $$

$$ = c_{jk} R^{*} $$

Instead of storing the complete coefficient matrix, it is sufficient to store $C=\{c_{jk}\}$. Each vector $c_{jk}$
contains counts of “kernels”, and it has as many elements as there are combinations of missing effects in the data + 1. In a 5-trait situation, where there are five patterns of missing traits, each txt block would require the storage of 15 elements, and the corresponding vector \( c \) only 6 elements. The storage requirements can drop further if counts are kept by 2-byte or even 1-byte numbers as opposed to 4- or 8-byte real numbers for the coefficients.

The method can be further refined if one creates a list of all txt blocks that are present in the coefficient matrix, and each entry in the coefficient matrix is a pointer to an element in this list.

**Example**

Consider the model

\[ y = Hh + Za + e \]

where \( h \) is the herd effect, \( a \) is the animal effect, and the data is

<table>
<thead>
<tr>
<th>animal</th>
<th>sire</th>
<th>dam</th>
<th>herd</th>
<th>milk</th>
<th>protein</th>
<th>somatic cell score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>10,000</td>
<td>300</td>
<td>300,000</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>-</td>
<td>1</td>
<td>12,000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>-</td>
<td>2</td>
<td>13,000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>13,000</td>
<td>350</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>11,000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The data contains 3 classes of missing effects:

<table>
<thead>
<tr>
<th>class</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td>2</td>
<td>somatic cell score</td>
</tr>
<tr>
<td>3</td>
<td>protein and somatic cell score</td>
</tr>
</tbody>
</table>

The matrix \( C \) would contain the following vectors when shown upper-stored

<table>
<thead>
<tr>
<th>1 0 1 0</th>
<th>1 0 0 0</th>
<th>0 0 1 0</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 0</td>
<td>0 0 1 0</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0 0 2</td>
<td>0 0 0 -1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0 2 1.33</td>
<td>0 0 0 -0.66</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 1 1 1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0 0 0 1.86</td>
</tr>
</tbody>
</table>
Because many vectors would appear more than once, one can store the matrix $C$ as links and a list.

Most likely, only a few elements in the list will occur in most of the entries. Solutions by block iteration would involve Cholesky decomposition of diagonal blocks $c_{ij}R$. Because some of the blocks can occur many times, computer time may be saved by precomputing the Cholesky decomposition for all types of blocks on the diagonal of $C$.

The above technique complicates when models for each trait are different. For special considerations in such a case, see the work by Tier (1992).
Data manipulation

So far we assumed that the data are available in an ordered way, with all effects numbered consecutively from 1. A substantial part of data analysis is selection of records, data checks and renumbering.

Renumbering

Data for mixed model programs require consecutive coding. Each effect should be coded from 1 to the last level, without gaps. One of the more difficult renumbering is for the animal effect. The purpose of that ordering is:

1. Detection of discrepancies in pedigrees
   - parents too young or too old than progenies
   - animal as both male and female,
   - too many progenies per female

2. Elimination of unnecessary animals
   - animals without records and
     - without progeny, or
     - with one progeny and no parent

3. Assignment of unknown parent groups (if desired)

4. Actual renumbering in the format
   \text{animal code; sire code; dam code; parents' code}.
   The parents' code tells how many animals have been replaced by unknown parent groups. It is desirable to code animals with records first because in that case the permanent environmental effect (in the repeatability model) can share coding with the animal effect.

5. Renumbering the animal id in the performance file. If the maternal effect is present, it needs to be added to the performance file and a code for missing dams assigned for a record of an animal whose maternal information is unavailable.

Memory consideration

If one prepares a subset of the data, e.g., for a REML analysis, the number of animals in the pedigree file can be much larger than the number of animals in the performance file because of many noncontributing animals. The size of pedigrees can be limited if one considers animals in the following order:
- animals with records
- parents of above animals
- parents of above animals
- .....................................

or from the youngest animal to the oldest. The removal of unnecessary animals can then proceed in
the reverse direction, from the oldest to the youngest.

There are two basic methods of renumbering:
   A. by look-up tables
   B. by sorting and merging

A. Look-up tables

Pseudo-program

Open performance file
Load all animals with records into memory
Sort pedigree file by year of birth descending
For each pedigree record
   If animal in memory, load parents, year of birth, and possibly other info; accumulate #
   parents and #progeny
Sort animals in memory by year of birth ascending
For each animal
   If noncontributing
      Remove,
      Eliminate contributions of current animal to parents and progeny
If desired, calculate inbreeding and/or unknown parent groups
Write renumbered pedigree file
Open performance file
For each record
   Substitute original animal id by consecutive number
   If desired, add inbreeding and/or maternal fields
   Write record

For fast search, the "hash" is about the most efficient.

Look-up approach is fast but memory intensive:
   10-50 bytes of memory per animal,
   10-50 Mbytes per 1 mln animals
B. Sort and merge

1. From pedigree and data files, extract animal id with estimated year of birth and code, which is one of a) has record, b) has pedigree record, c) has progeny

2. Sort by animal id

3. Summarize single entry per animal with:
   - number of records,
   - number of parents
   - number of progeny

4. Remove noncontributing animals
5. Assign consecutive numbers, first to animals with records
6. Sort data file by animal,
7. Sort pedigree file by animal; merge in animal consecutive number
8. Sort pedigree file by dam, merge in dam consecutive number
9. Sort pedigree file by sire; merge in sire consecutive number

Sort and merge renumbering:
   - need very little memory but plenty of disk space
   - speed dependent on sorting!

Examples
animal a2 is redundant!

<table>
<thead>
<tr>
<th>animal</th>
<th>dam</th>
<th>sire</th>
<th>year of birth</th>
<th>record</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>a0</td>
<td>-</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td>a4</td>
<td>a1</td>
<td>a3</td>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>a5</td>
<td>a4</td>
<td>-</td>
<td>90</td>
<td>12</td>
</tr>
</tbody>
</table>

sire pedigree file

<table>
<thead>
<tr>
<th>bull</th>
<th>dam</th>
<th>sire</th>
<th>year of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2</td>
<td>-</td>
<td>-</td>
<td>75</td>
</tr>
<tr>
<td>a3</td>
<td>a0</td>
<td>a2</td>
<td>80</td>
</tr>
</tbody>
</table>

Required unknown parent group allocations

<table>
<thead>
<tr>
<th>year of birth of youngest progeny</th>
<th>unknown parent group</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;85</td>
<td>1</td>
</tr>
<tr>
<td>≥85</td>
<td>2</td>
</tr>
</tbody>
</table>

Contents of look-up table after loading animals with records (order dependent on search algorithm)
Contents of look-up table after loading parents and parents of parents of animals with records

<table>
<thead>
<tr>
<th>animal</th>
<th>#parents</th>
<th>#progeny</th>
<th>#records</th>
<th>birth</th>
<th>#number</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>a4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>85</td>
<td>2</td>
</tr>
<tr>
<td>a5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>90</td>
<td>3</td>
</tr>
</tbody>
</table>

The last animal is noncontributing and can be eliminated. Renumbering of the pedigree done by one pass through the pedigree file. Unknown parents initially are assigned the birth year of their oldest progeny

animal | dam | sire | year of birth
-------|-----|------|----------------
1      | 4   | (80) | 80             
2      | 1   | 5    | 85             
3      | 2   | (90) | 90             
5      | 4   | (80) | 80             

4 (75) (75) 80 year of birth predicted as year of birth of progeny - 5

Unknown parent groups are replaced by consecutive numbers after the highest animal number, and the code shows the number of unknown parents

animal | dam | sire | code
------|-----|------|-----
1      | 4   | 6    | 1   
2      | 1   | 5    | 0   
3      | 2   | 7    | 1   
5      | 4   | 6    | 1   
The records’ file with animal id and records is created by reading the records file again and merging the animal id with the animal consecutive number.

<table>
<thead>
<tr>
<th>animal</th>
<th>record</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

**Sort and Merge approach**

First file:

<table>
<thead>
<tr>
<th>animal</th>
<th>type of entry</th>
<th>year of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>record&amp;ped</td>
<td>80</td>
</tr>
<tr>
<td>a0</td>
<td>dam</td>
<td>80</td>
</tr>
<tr>
<td>a4</td>
<td>record&amp;ped</td>
<td>85</td>
</tr>
<tr>
<td>a1</td>
<td>dam</td>
<td>85</td>
</tr>
<tr>
<td>a3</td>
<td>sire</td>
<td>85</td>
</tr>
<tr>
<td>a5</td>
<td>record&amp;ped</td>
<td>90</td>
</tr>
<tr>
<td>a4</td>
<td>dam</td>
<td>90</td>
</tr>
<tr>
<td>a2</td>
<td>sire_ped</td>
<td>75</td>
</tr>
<tr>
<td>a3</td>
<td>sire_ped</td>
<td>80</td>
</tr>
<tr>
<td>a0</td>
<td>dam</td>
<td>80</td>
</tr>
<tr>
<td>a2</td>
<td>sire</td>
<td>80</td>
</tr>
</tbody>
</table>

After sorting by type of entry within animal:
(rec = record, pedx = pedigree record with x known parents)

<table>
<thead>
<tr>
<th>animal</th>
<th>type of entry</th>
<th>year of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0</td>
<td>dam</td>
<td>80</td>
</tr>
<tr>
<td>a0</td>
<td>dam</td>
<td>80</td>
</tr>
<tr>
<td>a1</td>
<td>dam</td>
<td>85</td>
</tr>
<tr>
<td>a1</td>
<td>rec&amp;ped1</td>
<td>80</td>
</tr>
<tr>
<td>a2</td>
<td>sire</td>
<td>80</td>
</tr>
<tr>
<td>a2</td>
<td>sire_ped0</td>
<td>75</td>
</tr>
<tr>
<td>a3</td>
<td>sire</td>
<td>85</td>
</tr>
<tr>
<td>a3</td>
<td>sire_ped1</td>
<td>80</td>
</tr>
<tr>
<td>a4</td>
<td>dam</td>
<td>90</td>
</tr>
<tr>
<td>a4</td>
<td>reco&amp;ped2</td>
<td>85</td>
</tr>
<tr>
<td>a5</td>
<td>rec&amp;ped1</td>
<td>90</td>
</tr>
</tbody>
</table>
At this point entries can be read and analyzed, one animal at a time.

The renumber file would first be constructed with negative numbers for animals without records, because the maximum number of animals with records is not known until the file end:

<table>
<thead>
<tr>
<th>animal</th>
<th>#consecutive</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0</td>
<td>-1</td>
</tr>
<tr>
<td>a1</td>
<td>1</td>
</tr>
<tr>
<td>a3</td>
<td>-2</td>
</tr>
<tr>
<td>a4</td>
<td>2</td>
</tr>
<tr>
<td>a5</td>
<td>3</td>
</tr>
</tbody>
</table>

and in another pass, the same animals with records will have consecutive numbers corrected:

<table>
<thead>
<tr>
<th>animal</th>
<th>#consecutive</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0</td>
<td>4</td>
</tr>
<tr>
<td>a1</td>
<td>1</td>
</tr>
<tr>
<td>a3</td>
<td>5</td>
</tr>
<tr>
<td>a4</td>
<td>2</td>
</tr>
<tr>
<td>a5</td>
<td>3</td>
</tr>
</tbody>
</table>

The cow file is sorted by animal and merged with the animal file. The result is the data file:

<table>
<thead>
<tr>
<th>animal</th>
<th>record</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

and pedigree files:

1 a0 (80)  
(80) means unknown with animal born in 1980!
2 a1 a3
3 a4 (90)

The sire pedigree file is sorted by sire, merged with the animal file, and added to the previous pedigree file:

1 a0 (80)
2 a1 a3
3 a4 (90)
5 a0 a2

Then the file is sorted and merged, first for dam, second for sire:
After that, pedigree entries are added for animals without pedigree records (a0):

```
1 4 (80)
2 1 5
3 5 (90)
5 4 (80)
4 (80) (80)
```

Finally, unknown parents are replaced by unknown parent groups. In this case, born before 1985 are assigned to group 6, the remaining animals assigned to group 7:

```
1 4 6
2 1 5
3 5 7
5 4 6
4 6 6
```
**Tools for data manipulation**

A look-up approach requires a special program. Such a program is likely to be very fast but can be quite complicated because of a large number of details. On the contrary, a sort-and-merge type program can utilize some standard programs and therefore it could be easier to program although the CPU time would likely to be higher. CPU time is not always important if the program is run seldomly. What counts is cost of writing the program and ease of modification, which is directly related to program’s complexity.

In pedigree numbering the following standard operations (“primitives”) would be almost sufficient:

1. Select records from a file based on conditions
2. Select fields from a file with many fields
3. Sort based on one or more keys
4. Join two or more files
5. Retain unique records (or delete duplicate records)
6. Merge (catenate) two or more files
7. Number records consecutively

Consider the following data

<table>
<thead>
<tr>
<th>file ped</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
</tr>
<tr>
<td>dam</td>
</tr>
<tr>
<td>sire</td>
</tr>
<tr>
<td>code</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>file rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
</tr>
<tr>
<td>record</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

Create a new file recm that contains all fields from rec and field dam from ped so that the data can be analyzed with the maternal effect. In the most abstract way, this can be specified as

\[
\text{select animal, record, dam from ped,rec store in recm} \quad (1)
\]

Alternatively, using the primitives
sort ped by animal into peds  (2)
sort rec by animal into recs
join rec and ped by animal into recsm
select animal record dam from recsm into recmat
delete files peds, recs, recsm

The pseudo-programs above assume that the “system” knows names of fields in file. If the system
knows the consecutive numbers of fields only, the same program would be

sort ped by 1 into peds  (3)
sort rec by 1 into recs
join rec by 1 and ped by 1 into recsm
select 1 2 4 from recsm into recmat
delete files peds, recs, recsm

If one needs to specify formats for each file separately, then the length of the program increases
drastically and so does the chance of programming errors. For example, the line

sort ped by 1 into peds

could be equivalent to

infile ped
outfile peds
input
   animal 1-10
   dam 11-20
   sire 21-30
   code 31
sort by animal

Please note that a change in the format in one file may require corresponding changes in many other
files.

Data manipulation options

Command in pseudo-program (1) would be typical for better databases (Mattison, 1993). In a
database, each file is defined only once. Therefore the chance of making a mistake is greatly
reduced. By using a high definition data manipulation language, e.g., SQL, details of internal
computations need not be specified. Commercial databases may be expensive and require a good
deal of knowledge to operate smoothly. Some elements of the database are present in SAS.
Commands in (3) can be emulated by custom programs (see a section on MANIP) or by Unix utilities.

**Unix utilities**

Let fields in a file be separated by spaces.

The sort command works as

\[ \text{sort } f1 \ -o \ f2 \]

or

\[ \text{sort } f1 \ >f2 \]

sorts file \( f1 \) to \( f2 \) by all fields starting from the first one. The option -o allows for the same input and output file. Other options allow to sort by multiple keys and in a reverse order

\[ \text{sort } +3 \ -4 \ +1 \ -2 \ -r \ f1 \ -o \ f2 \]

sorts by fields 4 (skip 3 end at 4) and 2 (skip 1 end at 2) in the reverse order (-r). The sort utility can also be used to merge sorted files, and to test whether a file is sorted.

Repeated fields can be removed by the uniq command

\[ \text{uniq } f1 \ >f2 \]

An option -c prints not only the unique records but also counts of repeated lines.

Command join joins two files based on selection criteria and retains only specified fields

\[ \text{join } -j1 \ 1 \ -j2 \ 4 \ -o \ 1.5 \ 1.6 \ 2.3 \ 2.4 \ f1 \ f2 \ >f3 \]

joins files \( f1 \) and \( f2 \) into \( f3 \) based on field 1 in \( f1 \) and field 4 in file 2. File \( f3 \) contains fields 5 and 6 from file \( f1 \) and fields 3 and 4 from file 2.

A selection of fields and records can be done using AWK (GAWK, NAWK). The command

\[ \text{awk} \ \$1>125 \ f1 \ >f2 \]

selects records from \( f1 \) where the field 1 is greater than 125. Conditions may be more complicated

\[ \text{awk} \ \$1>125 \ \&\& \ \$2!="" \ | \ || \ $1==0 \ f1 \ >f2 \]

where either filed 1 is greater than 125 and filed 2 not blank, or field one is zero. Another variant of awk
awk '{print $1,$5,$6}' f1 >f2

selects only fields 1, 5 and 6 from file f1. There are many special variables in AWK. NR denotes the consecutive record number, and $0 denotes the complete record. Thus

awk '{print NR, $0}' f1 >f2

precedes each record with its consecutive number. The two variants of AWK can be used together.

awk ' $5 == “May” {print $3, $4}' f1 >f2

AWK is also a general programming language that can be used for tasks much more complicated than shown above. Books on Unix utilities including AWK are available, e.g., Swartz (1990).

The utilities use space as a standard delimiter between fields. If ID’s contain spaces, the utilities have options for other delimiters, e.g., “:”.

Perl is a tool more general that awk. It contains constructs like built-in hash storage. There are reports of Perl being successful in pedigree preparation large national populations.

Unix utilities are available from public domain also for Windows.
Sources and References


Hoeschele I. 1997. STAT 5554 Variance Components. Class notes, Virginia Polytechnic Insitute,


Misztal, I., T.J. Lawlor, and K. Weigel. 1995. Approximation of estimates of (co) variance
components with multiple-trait Restricted Maximum Likelihood by maximum diagonalization for more than one random effect. J. Dairy Sci. 78:1862-1872


Swartz R. 1990. Unix applications programming: mastering the shell. Sams, Carmel, IN.


