**Using an example design from an iTRAQ experiment:**

This experiment consists of a completely randomised design with 8 animals and 2 treatments for the first phase, and a 4-by-4 iTRAQ experiment for the second phase.

The following table shows the allocation of disease status (**Con**trol and **Dia**betic) to runs and tags in the iTRAQ experiment. Since both treatments occur exactly twice in every run and are labelled exactly twice with each tag, treatments are orthogonal to both runs and tags.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | Tag | | | |
| 114 | 115 | 116 | 117 |
| 1 | Con | Con | Dis | Dis |
| 2 | Dis | Dis | Con | Con |
| 3 | Dis | Dis | Con | Con |
| 4 | Con | Con | Dis | Dis |

The following table shows the allocation of animals to runs and tags.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | Tag | | | |
| 114 | 115 | 116 | 117 |
| 1 | 1 | 2 | 3 | 4 |
| 2 | 3 | 4 | 1 | 2 |
| 3 | 5 | 6 | 7 | 8 |
| 4 | 7 | 8 | 5 | 6 |

Let denote the abundance of a nominal protein in the proteomic sample from animal with disease status labelled with iTRAQ tag assayed in run . The linear model for the above design can then be written as

where µ denotes the overall mean abundance of the nominal protein, τi and γj denote the fixed effects of disease status *i* and tag *j*, respectively; Rk, Al and εijkl denote the random effects of run *k*, animal *l* and measurement error, respectively. These random effects are assumed to be mutually uncorrelated and normally distributed with mean zero and variances of , and .

The following table shows the theoretical ANOVA corresponding to the above design and linear model .

|  |  |  |  |
| --- | --- | --- | --- |
| **Source of variation** | **DF** | **MS** | **EMS** |
| Between Run |  |  |  |
| Between Animal | 1 |  |  |
| Residual | 2 |  |  |
| Within Run |  |  |  |
| Between Animal |  |  |  |
| Disease status | 1 |  |  |
| Tag | 1 |  |  |
| Residual | 4 |  |  |
| Within Animal |  |  |  |
| Tag | 2 |  |  |
| Residual | 4 |  |  |

Let denote the mean square (MS) which esimates the th pure error EMS in the above ANOVA table, where by pure error it is meant an EMS which contains only those variance components associated with the random effects (i=1,…,4). there are *m* sets of mean squares, denoted by s*i*, from the ANOVA table, these MS are assumed to have a chi-square distribution. Let these MS be denoted by , the distribution can be written as,

where the denotes the expected mean square and is the degrees of freedom for . The log-likelihood function of ??? is then given by

*L* = constant - .

The first derivative with respect to , also known as the *score function*, can then be shown to be

It follows thatthe expectation of the negative of the second derivative is given by

Hence, the negative of the second derivative of the likelihood function can be written as a diagonal matrix with *i*th diagonal element . Note: the MS and DF can be extracted from the ANOVA table based on the experimental data.

We want to estimate each of the variance components in . Hence, we need the coefficients of the variance components from the theoretical ANOVA table to compute the variance components from the expected mean squares. We can construct a matrix, denoted as the G matrix, using the coefficients of the variance components where the row corresponds to the source of variation and the column corresponds to the variance component. For our iTRAQ experiment, the G matrix is written as

.

Hence, the EMS should equal the variance components pre-multiplied by the G matrix. Note only the sources of variation that does not contain the fixed effects are considered. These are the “between animals between runs”, “the residual between runs”, “the residual between animals and within runs” and “the residual within animals and within runs”.

Hence, the score function and the Fisher’s information matrix with respect to the variance component, θ, can be written as

and

**Pseudo code of simulation and Fisher’s scoring algorithm for estimating the variance components**

For this case, 10000 simulated datasets are generated. The variance component estimates are obtained for each simulated data.

reml.VC = matrix(0, nrow=1, ncol = 3) # matrix used to store the variance component estimates from each simulated data set

Repeat 10000 times{

#Simulate a single dataset based on the linear model.

VC.base = variance component of the measure error.

VC.animal = variance component of the animal effects.

VC.run = variance component of the run effects.

Simulated dataset = N(0, VC.base) + N(0, VC. animal) + N(0, VC.run)

#Construct the theoretical ANOVA table based on the experimental design.

G = a matrix consists of coefficients of the variance components obtained from the theoretical ANOVA table.

#Perform ANOVA on the simulated data.

MS = vector of mean squares from ANOVA based the simulated dataset.

DF = vector of degrees of freedom of the corresponding mean square from the ANOVA table based on the simulated dataset

EMS = vector of expected mean squares based on the newly variance component estimates and the G matrix.

newV = c(VC.base , VC.animal, VC.run) # Initialise VCs to their true values, i.e. values used to simulate the dataset

oldV = c(0, 0, 0) # Initialise all VCs to zero

counter <- 1 # Initialise counter

#the convergence tolerance is the differences between the current variance component estimates and the previous variance component estimates are less than 1e-7

while((newV – oldV) >1e-7){

oldV = NewV

EMS = G’ × oldV

score function =

information matrix =

newV = oldV + (information matrix)-1 × (score function)

if ( counter > 1000 or information matrix is invertible)

stop the iteration of the while loop and start a brand new simulation dataset

counter = counter +1

}

reml.VC = rbind(reml.VC, newV) #store the estimates into a matrix

}

apply(reml.VC, 2, mean) #each variance components estimates, i.e. , are then obtained from the means of the variance components estimates from the 10000 simulated datasets