\documentclass[article]{jss}

\usepackage{bm}

\usepackage{graphicx}

\usepackage{amssymb}

\usepackage{amsmath}

\usepackage{amsthm}

\usepackage{multirow}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% declarations for jss.cls %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

\author{Kevin C. Chang\\University of Auckland,\\ New Zealand \And

Richard G. Jarrett\\CSIRO, Adelaide,\\ Australia \And

Chris M. Triggs\\University of Auckland,\\ New Zealand \And

Katya Ruggiero\\University of Auckland,\\ New Zealand }

\title{\pkg{InfoDecompuTE}: an \proglang{R} package for information decomposition of two-phase experiments}

%% for pretty printing and a nice hypersummary also set:

\Plainauthor{Kevin C. Chang, Richard G. Jarrett, Chris M. Triggs, Katya Ruggiero} %% comma-separated

\Plaintitle{InfoDecompuTE: an R package for information decomposition of two-phase experiments} %% without formatting

\Shorttitle{{\small\pkg{InfoDecompuTE}: an \proglang{R} package for information decomposition of two-phase experiments}} %% a short title (if necessary)

%% an abstract and keywords

\Abstract{Studies in which an experimental unit's response to treatment cannot be measured directly are said to be two-phase. In such cases, material harvested from the experimental units requires further processing in a subsequent experiment before measurements can be made. Consequently, each experimental phase introduces different sources of variation and how these interact with one another depends on the experimental designs for each phase, e.g. they may not yield a valid F-test in the analysis of variance.

To assess the properties of competing designs for two-phase experiments, it is necessary to examine their theoretical ANOVA tables, which can be a very time-consuming exercise to perform manually. We will introduce our very flexible R package, \pkg{infoDecompuTE}, which for a given single- or two-phase experiment will quickly construct the ANOVA table, showing any existing strata, expected mean squares for all sources of variation and average efficiency factors, as appropriate.

}

\Keywords{two-phase experiments, experimental design, analysis of variance, \pkg{InfoDecompuTE}}

\Plainkeywords{keywords, comma-separated, not capitalized, InfoDecompuTE} %% without

%% publication information

%% NOTE: Typically, this can be left commented and will be filled out by the technical editor

%% \Volume{13}

%% \Issue{9}

%% \Month{September}

%% \Year{2004}

%% \Submitdate{2004-09-29}

%% \Acceptdate{2004-09-29}

\Address{

Kevin C. Chang\\

Bioinformatics Institute \\

School of Biological Sciences\\

The University of Auckland\\

New Zealand\\

E-mail: \email{kcha193@aucklanduni.ac.nz}\\

}

%% It is also possible to add a telephone and fax number

%% before the e-mail in the following format:

%% Telephone: +43/1/31336-5053

%% Fax: +43/1/31336-734

%% for those who use Sweave please include the following line (with % symbols):

%% need no \usepackage{Sweave.sty}

%% end of declarations %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

\begin{document}

**\section[Introduction]{Introduction}**

Analysis of variance (ANOVA) is a well-established and widely used technique for performing global tests of hypotheses of the differences between three or more population means. In the simplest case of a completely randomised design, for example, it consists of computing the F-statistics, defined as the ratio of the between groups \emph{mean square} (MS), to the within groups MS. When the denominator \emph{expected mean square} (EMS) equals the numerator EMS excluding the non-centrality parameter, then a \emph{valid} F-test is existed. More generally, a valid F-test will exist provided the design is generally balanced and the treatment groups are sufficiently replicated.

For unbalanced designs or designs with complex block structures, obtaining a valid F-test is not always as straightforward, because the denominator EMS may not equals the numerator EMS excluding the non-centrality parameter. In such case, it is sometimes possible to estimate the denominator EMS of the effects of interest either from linear combinations of the available residual EMS across strata \citep{Satterthwaite1946} or by restricted maximum likelihood (REML) \citep{Patterson1971}.

An example of the designs with the complex block structure is a \emph{two-phase experiment} in which the responses of experimental units to treatments, i.e. Phase 1 experiment, cannot be measured directly, so the experimental units require subsequent processing, i.e. Phase 2 experiment, in order for the measurements to be made \citep{McIntyre1955}. The second phase experiment introduces additional sources of variation that may interact with the sources of variation introduced at Phase 1. Hence, it is important to consider the sources of variation that are introduced at each phase when designing two-phase experiments and ensure a valid F-test can still be conducted.

%Biological experiments involving high-throughput biotechnologies are a good example of the situation in which two-phase experiments arise. The cellular materials from living organisms subjected to a set of treatments (Phase 1 experiment), in the first phase experiment, while the second phase involves making measurements (e.g. gene expression or protein abundance) on the material using laboratory-based biotechnologies.

Two-phase experiments were introduced by \citeauthor{McIntyre1955} in \citeyear{McIntyre1955}, where he investigated the effects of four light treatments on the synthesis of tobacco mosaic viruses in the leaves of tobacco plants. The first phase experiment consisted of four light treatments assigned to eight plants and four leaf positions. The effectiveness of the light treatments were then estimated by injecting sap expressed from each of the first phase leaves into half-leaves of specific assay plants in the second phase experiment. Since then, efforts have been made towards developing a general theory for the design of two-phase experiments \citep{Brien1983, Wood1988, Brien1999, Jarrett2008}. with substantial progress being made in recent years by \cite{Brien2006b, Brien2009, Brien2010}.

The construction of ANOVA tables with the computation of the EMS for two-phase experiments is valuable in comparing the properties of different two-phase experimental designs. Methods for computing the EMS have been studied \citep{Hartley1967, Gaylor1970, Goodnight1980}. One of the better known methods, before the improvement of computational power, is the \emph{Cornfield-Tukey method} \citep{Cornfield1956}. This method first constructed a table where the rows are associated with the sources of variation and columns are associated with all the subscripts of the linear model. The coefficient of the EMS are then obtained by following a set of rules \citep{Cornfield1956}. This method is easy to understand, but can be difficult to implement when the number of block and treatment factors increases, as the dimensions of the table also increase. In addition, this method can only be performed on balanced designs. Currently, there is no freely available statistical software that can compute the EMS for two-phase experiments directly. The \code{AMTIER} procedure within \proglang{Genstat}, developed by \cite{Brien2006a}, was designed to analyse two-phase experiments. The \code{AMTIER} procedure can also produce the structure of an ANOVA including the decomposition of degrees of freedom (DF), but it does not gives the EMS. The \code{GLM} and \code{MIXED} procedures in \proglang{SAS}, and \code{ANOVA} command in \proglang{Minitab}, have the ability to compute the EMS but neither program can easily cope with two-phase experiments.

For even a small two-phase experiments, the construction of ANOVA tables can still be laborious and need to be performed manually. Therefore, we introduce an \proglang{R} package called \pkg{infoDecompuTE} which stands for {\bf info}rmation {\bf decomp}osition of {\bf t}wo-phase {\bf e}xperiments. It allows statistical researchers to quickly check the properties of the ANOVA table with EMS by entering any single or two-phase experimental design. Therefore, this package will not only allow researchers to determine whether or not a valid F-test can be conducted, it will also enable them to study the decomposition of the raw data into different sources of variation.

In this article, we aim to demonstrate the concepts and the method underlying \pkg{infoDecompuTE}. Section~\ref{sec:infoDecomp} explains the general methods of behind \pkg{infoDecompuTE}. Section~\ref{sec:exampleTwoPase} uses a simple two-phase experiment example to further illustrate the methods. Section~\ref{sec:package} explains the arguments of the \proglang{R} function in \pkg{infoDecompuTE}. Finally, section~\ref{sec:example} will show \pkg{infoDecompuTE} accurately reproduces the ANOVA table with the EMS of a rather complex two-phase viticulture-sensory evaluation experiment by \cite{Brien1999}.

%The purpose of \emph{analysis of variance} (ANOVA) is comparing for the differences between groups statistically. An ANOVA table is constructed aiming to determine whether a valid test for differences between groups is possible. The test statistics is a ratio of the mean squares that is chosen such that expected value of the numerator mean square differ from the expected value of the denominator mean square only by the variance components or the fixed effects in which we are interest. Thus, a valid statistical test is present only when the \emph{coefficients of variance components} from these two \emph{expected mean squares} (EMS) are identical. For a simple experimental design, with a single block factor, the coefficients of variance component of blocks is the replication number. The \emph{block factors} as the factors that reduce the variability transmitted from nuisance factors and \emph{treatment factors} as the factors administered to experimental units. However, as the design become more complex, i.e., made up of more block and treatment factors, the computation of coefficients is less straightforward.

**\section[Information Decomposition]{General methods of behind** \pkg{infoDecompuTE}} \label{sec:infoDecomp}

Once the data are collected from any experiment, it will likely contain many different sources of variation, i.e. biological and technical variation. The process of \emph{information decomposition} breaks down the raw data and separates the variation into orthogonal pieces according to their strata and sources. The stratum is formally defined as the eigenspace of the covariance matrix \citep{Bailey2008}. A covariance matrix is a matrix which contains the covariance between any pair of random variables. Package \pkg{infoDecompuTE} uses the methods of information decomposition to breaks down the raw data.

This section will explain the methods, using a general mixed model, of information decomposition on how the block and treatment information is decomposed into its constituent components based on the relationships between the experimental units, or \emph{block structure}, and the nature and assignment of treatments to the experimental units, or \emph{treatment structure}. In additional, this section will also demonstrate the mathematical process behind \pkg{infoDecompuTE}.

**\subsection{General mixed model}\label{subsec:decompRCD}**

A mixed model is a statistical model containing both fixed and random effects. Consider a general mixed model with $t$ treatments and $b$ blocks, this model in matrix notation is given by

\begin{equation}\label{eq:matrix}

\bm{y} = \bm{1}\mu + X\bm{\tau} + Z\bm{\beta} + \bm{\epsilon},

\end{equation}

where $\bm{y}$ is a $bt \times 1$ vector of responses, $\bm{1}$ is also a $bt \times 1$ vector with all elements unity, $\mu$ denotes the grand mean, and $\bm{\epsilon}$ is a $bt \times 1$ vector of unobserved random errors. The $Z$ is the $bt \times b$ binary matrix describes the block design, this matrix indicates which block contains the observations, where the columns correspond to the blocking groups and rows correspond to the observations. The $X$ is the $bt \times t$ binary matrix describes the treatment design, where the columns correspond to every combination of all the treatment factors and row correspond to the observations, hence this matrix is also known as \emph{treatment incidence matrix}. The parameters $\bm{\tau}$ and $\bm{\beta}$ are $t \times 1$ and $b \times 1$ vectors of block and treatment parameters. Furthermore, the grand mean and the treatment effects are fixed effects. The block effect and experimental error are random effects, and are assumed to be independently and normally distributed with mean zero and variances of $\sigma\_{B}^2$ and $\sigma^2$, respectively.

The expected sum of squares (ESS) from model~(\ref{eq:matrix}) is given by

\begin{equation}\label{eq:ESS}

E(\bm{y}'\bm{y}) = \mu' \bm{1}' \bm{1} \mu + \bm{\tau} ' X'X \bm{\tau} + \bm{\beta}' Z'Z \bm{\beta} + \bm{\epsilon}' \bm{\epsilon},

\end{equation}

where $\bm{y}'\bm{y}$ denotes unadjusted total sum of squares (SS), the remaining terms $\mu' \bm{1}' \bm{1} \mu$, $ \bm{\tau} ' X'X \bm{\tau}$, $\bm{\beta}' Z'Z \bm{\beta}$ and $\bm{\epsilon}' \bm{\epsilon}$ are sums of squares of grand mean, treatment effects, block effects and experiment error. Hence, Equation~(\ref{eq:ESS}) gave a simple illustration on the purpose of the information decomposition which is to separate random and fixed effects from the raw data.

**\subsection{Information decomposition using projection}\label{subsec:infoDecompProj}**

The raw data vector from Equation~(\ref{eq:matrix}), denoted by $\bm{y}$, spans a $bt$-dimensional space, because this raw data vectors is made up of $bt$ observations. The first step is to sweep the grand mean from the raw data, this can be achieved by projecting the raw data vector onto the grand mean space.

This grand mean space is spanned in the $1$-dimensional space, because $\mu$ is a vector of length $1$. The grand mean space can be represented by a $bt \times bt$ averaging matrix, $K$, with all elements equal to $1/bt$. The \emph{orthogonal projection} of the raw data vector $\bm{y}$ onto the grand mean space represented by matrix $K$ is derived by

\begin{equation}\label{eq:vectorProj}

K(K'K)^{-1}K' \bm{y} = P\_{K} \bm{y},

\end{equation}

where $P\_K$ denotes projection matrix of matrix $K$. Projection matrices are symmetric (i.e. $P' = P$), orthogonal (i.e. $P\_aP\_b = 0$) and idempotent (i.e. $P^2 = PP = P$) \citep{Hadi1996}. Note that the projection matrix of any averaging matrix is identical to the averaging matrix. Hence, the mean corrected observational vector, denoted by $(I-K)\bm{y}$, is derived from the raw data vector, $\bm{y}$, subtracting the grand mean vector, $K\bm{y}$. This newly mean corrected observational vector is spanned in $(bt - 1)$-dimensional space. Furthermore, the adjusted total SS is obtained from pre-multiplying the mean corrected observational vector by its transpose, i.e.

\begin{equation}\label{eq:adjustSS}

[(I - K)\bm{y}]'[(I - K)\bm{y}] = \bm{y}'(I-K)\bm{y}.

\end{equation}

The mean corrected observational vector is then projected onto the between blocks vector subspace. This between block vector subspace is spanned in a $b$-dimensional space and represented by the block design matrix, $Z$. The projection is applied by pre-multiplying the projection matrix of block design matrix denoted by $P\_Z$. This gives a new vector $(P\_Z - K)\bm{y}$ which corresponds to the between blocks effects. The orthogonal complement of $(P\_Z - K)\bm{y}$, denoted by $(I - P\_Z)\bm{y}$, corresponds to the within blocks effects. The SS of between and within blocks are calculated by Equation~\ref{eq:adjustSS} which gives $\bm{y}'(P\_Z-K)\bm{y}$ and $\bm{y}'(I-P\_Z)\bm{y}$, respectively.

The between and within blocks ESS are computed by applying a \emph{trace} operation, which is the sum of the diagonal elements of the square matrix \citep{Searle1982}. Subsequently, the EMS is calculated by dividing the corresponding DF, and the results are shown in Table~\ref{tab:infoDecomp}.

\begin{table}[ht]

\centering

\caption{Phase 1 ANOVA with the coefficients of variance components of the EMS.}

\begin{tabular}[t]{lllll}

\hline

\multicolumn{1}{l}{Source of Variation} & \multicolumn{1}{l}{DF} & \multicolumn{1}{l}{SS} & \multicolumn{1}{l}{ESS}& \multicolumn{1}{l}{EMS}\\

\hline

Between Blocks &$b-1$ & $\bm{y}'(P\_Z-K)\bm{y}$ & $(b-1)(\sigma^2 + t\sigma\_{B}^2)$ & $\sigma^2 + t\sigma\_{B}^2$\\

Within Blocks &$b(t - 1)$ & $\bm{y}'(I - P\_Z)\bm{y}$ & $b(t-1)\sigma^2$ & $\sigma^2$\\

\hline

Adjusted Total & $bt - 1$ & \\

\hline

\end{tabular}

\label{tab:infoDecomp}

\end{table}

If there is another block factor in the design, this decomposition process is repeated by projecting the between or within blocks vectors onto that additional block vector subspace and computing its orthogonal complements. Then, the ESS and EMS can be calculated as described.

**\subsection{Estimating the treatment effects}\label{subsec:estTrtRBD}**

The next step is to examine the presence of treatment information in the between and/or within blocks. This method has been described by \cite{Nelder1965B} and \cite{John1987}.

Another set of matrices are required to represent the treatment part of the experiment. These matrices are the square treatment model matrices \citep{Nelder1965B} or the contrast matrices \citep{John1987}. These matrices can be generated from a equation showing how the treatment effects can be partitioned into orthogonal components, namely \emph{yield identity}. Consider the example described in Section~\ref{subsec:decompRCD} with one treatment factor denoted by $\tau\_i$, $i= 1,\dots, t$, the yield identity is given by

\[\tau\_{i} = \overline{\tau\_{.}}+(\tau\_{i} -\overline{\tau\_{.}}),\]

where the dot in the subscript denotes the summation over the subscript it replaces and the over-line indicates the average over the terms associated with the nominal subscript. Thus, $\tau\_{.}$ denotes the mean of all treatment effects and the term in the parentheses, $\tau\_{i} -\overline{\tau\_{.}}$, is an estimate of the treatment effects. To convert the yield identity to matrix notation, the $\tau\_{i}$ is re-written as the identity matrix pre-multiply treatment parameter, $I\_t\bm{\tau}$, and $\overline{\tau\_{.}}$ becomes the averaging matrix pre-multiply treatment parameter, $K\_t\bm{\tau}$, hence, the contrast matrix, denoted by $C$, of the treatment effect is given by $(I\_t - K\_t)\bm{\tau} = C\bm{\tau}$.

The next step is to define the reduced normal equations for the treatment parameter obtained by eliminating the mean and block parameters from the full set of normal equations~\citep{John1987}. This can be shown as

\[

A\_i\tau = q\_i,

\]

where

\begin{eqnarray\*}

A\_i &=& CX' P\_i XC,\\

q\_i &=& CX' P\_i \bm{y},

\end{eqnarray\*}

as $A\_i$ denotes information matrix, vector $q\_i$ is the adjusted treatment totals, and $P\_i$ is the projection matrix of stratum $i$. Since we are estimating the treatment information, so the equation is rewritten in the respect of $\tau$

\[

\tau = A\_{i}^{-} q,

\]

where $A\_{i}^{-}$ is a generalised inverse of $A\_{i}$ satisfying $A\_{i} A\_{i}^{-} A\_{i} = A\_{i}$. The generalised inverse of matrix is applied, because the information matrix can be a singular matrix. Now consider the example in Section~\ref{subsec:decompRCD}, the treatment information in the within blocks can be written as

\[[CX' (I - P\_B) XC]^{-}CX'(I - P\_B)\bm{y},\]

and the SS

\begin{equation}\label{eq:genInvTrt}

\bm{y}'(I - P\_B)XC[CX' (I - P\_B) XC]^{-}CX'(I - P\_B)\bm{y}.

\end{equation}

The residual SS in the within block stratum is derived by subtraction and is given by

\[ \bm{y}'(I - P\_B)(I - XC[CX' (I - P\_B) XC]^{-}CX')(I - P\_B)\bm{y}. \]

The coefficients of the variance components in the treatment and residual ESS can be computed using the same procedure described in Section~\ref{subsec:decompRCD}.

**\subsection{Information decomposition in infoDecompuTE}**

This laborious mathematical procedure is used, in \pkg{infoDecompuTE}, because equation~(\ref{eq:genInvTrt}) can be the shown to be the same as

\begin{equation}\label{eq:projTrtBlock}

\frac{1}{e}[ \bm{y}'(I - P\_B)P\_{\tau}(I - P\_B)\bm{y}],

\end{equation}

where $P\_{\tau}$ is the projection matrix derived from the treatment design matrix. Thus, $P\_{\tau}(I - P\_B)\bm{y}$ is the vector derived from the within blocks vector projected onto the between treatment vector subspace, so this vector is corresponding to the treatment effect in the within blocks stratum. However, there is an additional term $1/e$, where $e$ denotes \emph{efficiency factor}, which is the amount of treatment or block information that is present in the within blocks strata \citep{Yates1936}. Efficiency factors are typically used in non-orthogonal designs, for example \emph{balanced incomplete block designs} (BIBD). BIBD are where no treatments appear more than once in any block and every pair of treatment occurs together in exactly the same numbers of blocks. Therefore, Equation~(\ref{eq:genInvTrt}) can adjust the efficiency factor while computing the treatment SS.

\pkg{InfoDecompuTE} computes the \emph{canonical efficiency factors}, which are the efficiency factors of the basic treatment contrasts. Given that $\theta\_i$ and $p\_i$ are $i$th eigenvalue and eigenvector of the information matrix. The $i$th basic treatment contrast is given by $p\_i'\bm{\tau}$ and $i$th canonical efficiency factor is $\theta\_i/r$, where $r$ is the replication numbers of the corresponding treatment factors \citep{John1987}. The average efficiency factor can be calculated from the harmonic mean of the canonical efficiency factors \citep{John1987}.

Furthermore, \pkg{infoDecompuTE} obtains the fixed components' coefficients of the EMS from the total numbers of observations divided by the numbers of levels of a given treatment or treatment combination for interaction. If the design is non-orthogonal, the fixed components' coefficients of the EMS are multiplied by the efficiency factors to reflect the amount of separation of the treatment information across strata.

**\section[Information decomposition of two-phase experiment]{Information decomposition of a two-phase experiment} \label{sec:exampleTwoPase}**

In general, any single-phase experiment consists of three basic decomposition steps: adjusting for the grand mean, defining the block structures and estimating the treatment effect in the block structure. On the other hand, two-phase experiments are more complicated than single phase experiments, because they are made up of two block structures from Phase 1 and 2 experiments and the treatment structure. Hence, the decomposition starts with the raw data vector and adjusts for the grand mean, then the block structure of the Phase 2 experiment can be defined. The Phase 2 block structure is always defined before the Phase 1 block structure, because the Phase 2 block structure always contributes to the outer strata of the ANOVA table. The next step is to estimate the block effects of the Phase 1 experiment in the block structure of Phase 2 experiment, and finally the treatment effect in block structure Phase 1 experiment is estimated. However, the issue of non-orthogonal designs can occur when assigning the block factors of Phase 1 to the Phase 2 experiment as seen in \cite{Brien1999}. Hence, the equation described in (\ref{eq:genInvTrt}) is employed by treating the Phase 1 block structure as the treatment structure. This can help us to adjust for the efficiency factors while applying the projection from the Phase 2 block onto Phase 1 block spaces as shown in (\ref{eq:projTrtBlock}).

**\subsection{Example of two-phase experiments}**

\emph{Quantitative proteomics experiments}, measure and compare protein abundances between different treatment groups, are the good examples of two-phase experiments. The first phase consists of collecting cellular material from living organisms which have been experimentally perturbed, but, the protein abundances cannot be measured directly from these cellular materials. Hence, the second phase involves measuring the protein abundances on the materials using a specific biotechnology. Due to the chemical complexity of proteins, a series of instruments, collectively referred to as \emph{Multi-dimensional Protein Identification Technology} (MudPIT), are used to separate the protein mixture before measuring the abundance of each protein species. However, the comparison of protein abundances between samples is difficult due to large variability between different MudPIT experiments. This limitation has been resolved with the introduction of \emph{isobaric Tags for Relative and Absolute Quantitation} (iTRAQ$^{\rm TM}$) which enables the simultaneous analysis of up to eight distinct samples within a single MudPIT run \citep{Ross2004, Choe2007}.

The rest of this section will present an example experiment which aims to identify proteins that are differentially abundant between healthy and diseased animals. Consider there are eight animals available and MudPIT-iTRAQ$^{\rm TM}$ biotechnology was used to measure the protein abundances. This experiment can be treated as a two-phase experiment, where the animals are randomly perturbed to either the healthy or diseased treatment groups in the first phase experiment. Then, the tissues are harvested from the animals and the protein abundances are measured in the Phase 2 MudPIT-iTRAQ$^{\rm TM}$ experiment.

This section first describes both Phase 1 and Phase 2 experimental designs and the ANOVA table, then it explains the methods of information decomposition with the mathematical procedure of the two-phase experiments behind \pkg{infoDecompuTE}

**\subsection{Phase 1 experiment}\label{subsec:Phase1Exp}**

The experimental objective of the first phase is to compare treatment groups in the absence of nuisance sources of variation \citep{Mead1988}. Thus, the Phase 1 experiment is a completely randomised design. A total of eight animals are available for the current study, hence four animals can be randomly assigned to either the healthy or diseased treatment groups as shown Table~\ref{tab:phase1Design}.

\begin{table}[ht]

\centering

\caption{Phase 1 design shows the assignment of animals to disease status' groups where the letters denote the animal ID.}

\begin{tabular}[t]{ccccc}

\hline

Control & A & C & E & G \\

Diseased & B & D & F & H \\

\hline

\end{tabular}

\label{tab:phase1Design}

\end{table}

Let $y\_{ijk}$ denote the abundance of a given protein in sample $k$ of rat $i$ under disease status $j$. Then, the linear model of the Phase 1 design is given by

\begin{equation}\label{eq:phase1Model}

y\_{ij}= \mu + A\_{i} + \tau\_{j},

\end{equation}

where $\mu$ denotes grand mean from all observations, $\tau\_{j}$ denotes the fixed effect of disease status $j$, $j=$ healthy, diseased, and $A\_{i}$ denotes the effects from rat $i$, $i=$ A,$\dots$, H. In addition, $A\_i$ also denotes an observational error, because the animal is the smallest unit of the Phase 1 experiment. Terms with upper-case Roman letters are random effects, that have no fixed amount of fluctuation. Furthermore, these random effects are assumed to be normally distributed with mean zero and variance $\sigma\_{A}^2$.

The ANOVA table for the Phase 1 experiment is shown in Table~\ref{tab:Phase1ANOVA}. Since there are total of 8 observations, the Phase 1 experiment consists of 7 DF. These 7 DF are decomposed to 1 and 6 DF for the disease status and residual, respectively. The last column of this ANOVA table is the EMS which is the linear combination of the variance components. These EMS are made up of the variation between animals, $\sigma\_{A}^2$, and the treatment component for disease statuses, $\theta\_{\tau}$. This $\theta\_{\tau}$ is defined as $\frac{(\overline{\tau\_{healthy}} - \overline{\tau\_{diseased}})^2}{2-1}$, which is the variance between the treatment means. The coefficients of these components can be derived from the replication numbers of their corresponding factors. Hence, a valid test for disease status is obtained from the F-ratio of disease status and the residual mean squares.

\begin{table}[ht]

\centering

\caption{Phase 1 ANOVA with the coefficients of variance components of the EMS.}

\begin{tabular}[t]{lrl}

\hline

\multicolumn{1}{l}{Source of Variation} & \multicolumn{1}{l}{DF} & \multicolumn{1}{l}{EMS}\\

\hline

Between Animals \\

\hspace{3mm}Disease status & $1$ & $\sigma\_{A}^2 + 4\theta\_{\tau}$\\

\hspace{3mm}Residual & $6$ & $\sigma\_{A}^2$\\

\hline

Total & $7$ & \\

\hline

\end{tabular}

\label{tab:Phase1ANOVA}

\end{table}

\subsection{Phase 2 experiment}\label{subsec:Phase2Exp}

Since the protein abundances cannot be measured directly from the animals, the tissues are harvested and proteins are extracted from the animals for the subsequent Phase 2 MudPIT-iTRAQ$^{\rm TM}$ experiment. Hence, two additional variance components are introduced, the variation between runs, $\sigma\_{R}^2$, and the difference between tags, $\theta\_{\gamma}$.

This example uses four MudPIT runs with a four-plex iTRAQ$^{\rm TM}$ labelling system to measure protein abundances, hence a total of 16 samples are processed. This allows us to obtain two identical samples from each of the eight animals of the Phase 1 experiment. The samples are used for assessing the measurement error of the MudPIT-iTRAQ$^{\rm TM}$ experiment, namely \emph{technical replicates}.

The 16 samples generated from Phase 1 can be directly allocated to MudPIT-iTRAQ$^{\rm TM}$ using a four-by-four grid. The goal of the design is to estimate the differences in protein abundance between the treatment groups as precisely as possible. This is achieved by ensuring that the fixed effects can be estimated independently of the random effects, i.e., they are not \emph{confounded} with the random effects.

The assignment of disease status is done using a Latin square design. However, there are eight rats available in this experiment and only four MudPIT runs of four-plex iTRAQ$^{\rm TM}$ labelling system are used, so any assignment of animals is unavoidably confounded with runs and tags. Hence, now the aim is to minimise the amount of confounding. A tabular representation of the Phase 2 design is shown in Table~\ref{tab:phase2Design}. The confounding of animals with runs occurs when comparing between Run 1, 3 and Run 2, 4, hence, one DF of animal is confounded with run. Likewise, the confounding of animals with tags occurs when comparing between Tag 114, 116 and Tag 115, 117, so one DF for tag is confounded with animal. The structure of the ANOVA with the DF is presented in Table~\ref{tab:Phase2ANOVA}.

\begin{table}[ht]

\centering

\caption{Phase 2 design showing the animal assignment to runs and tags. The letters denote animal ID.}

\begin{tabular}[t]{c|cccc}

& \multicolumn{4}{c}{{\bf Tag}} \\

{\bf Run} & 114 & 115 & 116 & 117 \\

\hline

1 & A & B & C & D \\

2 & F & E & H & G \\

3 & C & D & A & B \\

4 & H & G & F & E \\

\end{tabular}

\label{tab:phase2Design}

\end{table}

Let $y\_{ijklm}$ denote the abundances of a given protein in sample $k$ of rat $i$ under disease status $j$ and measured from $l$th MudPIT run with iTRAQ$^{\rm TM}$ tag $m$. Then, the linear model of the Phase 2 design is given by

\begin{equation}\label{eq:phase2Model}

y\_{ijklm}= \mu + R\_{k} + A\_{i}+ \tau\_{j} + \gamma\_{l} + \epsilon\_{ijklm},

\end{equation}

where $R\_{k}$ denotes the random effects from run $k$, $k=1,\dots, 4$, $\gamma\_{l}$ denotes the fixed effects of tag $l$, $l = 114,\dots, 117$, $\epsilon\_{ijklm}$ denotes the effects from sample $m$, $m = 1, 2$, in animal $i$ from run $k$ under disease status $j$ and tag $l$. In addition, $\epsilon\_{ijklm}$ also denotes an experimental error or measurement error, because the sample is the smallest unit of the Phase 2 experiment. The remaining terms are defined as in~(\ref{eq:phase1Model}). The disease status and tag effects are assumed not to interact.

This two-phase experiment is used to demonstrate each step of the decomposition from the raw data vector to each source of variation and the computation of the EMS of the ANOVA table.

\subsection{Design and projection matrices of the block structures}\label{subsec:desAndProj}

The EMS are computed by the matrix algebra, hence model~(\ref{eq:phase2Model}), defined in the matrix notation, is written as

\begin{equation}\label{eq:matrixTwoPhase}

\bm{y} = \bm{1}\mu + X\bm{\alpha} + Z\_1\bm{\beta}\_1 + Z\_2\bm{\beta}\_2 + \bm{\epsilon},

\end{equation}

where $\bm{y}$ is a $16 \times 1$ vector of responses, $\bm{1}$ is also a $16 \times 1$ vector with all elements unity, and $\bm{\epsilon}$ is a $16 \times 1$ vector of error terms. The vectors of Phase 1 and 2 block parameters are

\begin{equation}\label{eq:block1Par}

\bm{\beta}\_1= (\bm{\beta\_A}) = (\beta\_{A\_i}) = (\beta\_{A\_A}, \dots, \beta\_{A\_H})

\end{equation}

and

\begin{equation}\label{eq:block2Par}

\bm{\beta}\_2= (\bm{\beta\_R}) = (\beta\_{R\_k}) = (\beta\_{R\_1}, \dots, \beta\_{R\_4}),

\end{equation}

respectively. Note the Phase 1 parameter does not contain the samples, because it is in the error terms, $\bm{\epsilon}$.

Since two treatment factors are involved, the treatment parameter consists of every treatment combination, i.e.

\begin{equation}\label{eq:treatPar}

\bm{\alpha} = (\bm{\alpha}\_{\bm{\tau} \bm{\gamma}}) = (\alpha\_{\tau\_{diseased}:\gamma\_{114}}, \alpha\_{\tau\_{control}:\gamma\_{114}}, \alpha\_{\tau\_{diseased}:\gamma\_{115}}, \dots, \alpha\_{\tau\_{control}:\gamma\_{117}}).

\end{equation}

The treatment parameter is constructed this way because it has to match the columns of treatment design matrix $X$ in model~(\ref{eq:matrix}), where its column denotes every combination of all the treatment factors.

The $Z\_1$ and $Z\_2$ in model~(\ref{eq:matrix}) are a block design matrix for the Phase 1 and Phase 2 experiments, respectively. Note that, if the Phase 1 experiment consists of more than one block factor, then the block parameter and design matrix will be constructed the same way as treatment as described in the beginning Section~\ref{sec:exampleTwoPase}.

The block design matrix of the Phase 2 and Phase 1 experiments are $Z\_2 = (Z\_R)$ and $Z\_1 = (Z\_A)$, respectively, where $Z\_R$ are block design matrices for the MudPIT runs and $Z\_A$ are block design matrices for the animals.

% and can be written as

%\[Z\_R = I\_4 \otimes 1\_4 \]

%where $I\_n$ is the identity matrix of order $n$ and $\otimes$ denotes the \emph{Kronecker product}.

%and can be written as,

%\[Z\_A =

%\begin{pmatrix}

%I\_4& 0\\

%0& \Gamma\_{4(2)}^{\parallel}\\

%\Gamma\_{4(2)}& 0\\

%0& I\_{4}^{\parallel}

%\end{pmatrix}\]

%where $\Gamma\_{s(h)}$ with the order of $s$ denotes a number one in column $h+1$ and all other elements are zero in the first row. The subsequent rows of $\Gamma\_{s(h)}$ are obtained by shifting the elements in the previous row right one position with a wrap-around at the end. The $\Gamma\_{s(h)}$ also known as \emph{square circulant matrix} \citep{Davis1969}. Another symbol $\parallel$ as the superscript of the matrix notation denotes a left-right mirror-transformations of the matrices \citep{Li2003}.

From Equation~(\ref{eq:vectorProj}), the projection matrices for runs and animals are,

\begin{eqnarray\*}

P\_{R} &=& Z\_{R}(Z\_{R}'Z\_{R})^{-1}Z\_{R}',\\

P\_{A} &=& Z\_{A}(Z\_{A}'Z\_{A})^{-1}Z\_{A}'.

\end{eqnarray\*}

The rest of this section will demonstrate each step of the decomposition of this two-phase experiment. The raw data vector $\bm{y}$ is \emph{projected} onto different vector subspaces to calculate the coefficients of the random effects' variance components for every source of variation as explained in Section~\ref{sec:infoDecomp}.

\subsection{Phase 2 block structure decomposition}

The Phase 2 block structure corresponds to MudPIT runs. After sweeping the grand mean from the raw data vector, the mean corrected observational vector, $(I-K)\bm{y}$, is projected onto the between runs vector subspace. This process of projection is obtained by pre-multiplying the mean corrected observational vector by the projection matrix of run design matrix as described in Equation~(\ref{eq:vectorProj}). The orthogonal complement from the projection, i.e. within runs, can be deduced by subtraction as shown in Table~\ref{tab:block2Projection}.

\begin{table}[ht]

\centering

\caption{Phase 2 ignoring phase 1 block structure ANOVA with projection.}

\begin{tabular}[t]{lr}

\hline

\multicolumn{1}{l}{Source of Variation} & \multicolumn{1}{l}{Projection of $\bm{y}$}\\

\hline

Between runs &$(P\_R - K)\bm{y}$ \\

Within runs &$(I-P\_R)\bm{y}$ \\

\hline

Adjusted total &$(I-K)\bm{y}$ \\

\hline

\end{tabular}

\label{tab:block2Projection}

\end{table}

\subsection{Phase 1 block structure decomposition}

The Phase 1 block structure consists of the between animals and between samples within animals spaces. The vectors of between runs, $(P\_R - K)\bm{y}$, and within runs, $(I-P\_R)\bm{y}$, are further decomposed by subsequent projection onto the between animals vector subspace. The method described in Section~\ref{subsec:estTrtRBD} is applied by estimating the Phase 1 block information in the Phase 2 block structure. Since there is only one factor,the treatment incidence matrix is identical to the block design matrix design. Furthermore, the contrast matrix is constructed the same way as in Section~\ref{subsec:estTrtRBD} giving

\[C\_A = I\_8 - K\_8.\]

Hence, the SS of the animals between runs and within runs are

\begin{equation}\label{eq:ssAniBeRun}

\bm{y}'(P\_R-K)Z\_{A}'C\_{A}[C\_{A}Z\_{A}'(P\_R - K)Z\_{A} C\_{A}]^{-} C\_{A}Z\_{A}'(P\_R- K)\bm{y}

\end{equation}

and

\begin{equation}\label{eq:ssAniWiRun}

\bm{y}'(I-P\_R)Z\_{A}'C\_{A}[C\_{A}Z\_{A}'(I-P\_R)Z\_{A} C\_{A}]^{-} C\_{A}Z\_{A}'(I-P\_R)\bm{y},

\end{equation}

respectively. The SS Equation~(\ref{eq:ssAniBeRun}) and (\ref{eq:ssAniWiRun}) can be shown to be the same as $\bm{y}'(P\_R - K)P\_{A}(P\_R - K)\bm{y}$ and $\bm{y}'(I-P\_R)P\_{A}(I-P\_R)\bm{y},$

respectively. This indicates SS Equation~(\ref{eq:ssAniBeRun}) and (\ref{eq:ssAniWiRun}) are the SS of the vectors resulting from the projection onto the between animal vector subspace. The orthogonal complement of the projection, i.e. between samples within animals, can then be deduced by subtraction. The results from this decomposition step is shown in Table~\ref{tab:block1Projection}. The coefficients of variance components are computed using trace operation as explained in Section~\ref{subsec:decompRCD}.

\begin{table}[ht]

\centering

\caption{Two-phase block structure ANOVA with projection}

\begin{tabular}[t]{lr}

\hline

\multicolumn{1}{l}{Source of Variation} & \multicolumn{1}{r}{Projection of $\bm{y}$}\\

\hline

Between runs & \\

\hspace{3mm}Between animals & $P\_{A}(P\_R - K)\bm{y}$ \\

\hspace{3mm}Between samples within animals & $(I - P\_{A})(P\_R - K)\bm{y}$ \\

\hline

Within runs & \\

\hspace{3mm}Between animals & $P\_{A}(I-P\_R)\bm{y}$\\

\hspace{3mm}Between samples within animals &$(I - P\_{A})(I-P\_R)\bm{y}$ \\

\hline

Adjusted total &$(I-K)\bm{y}$ \\

\hline

\end{tabular}

\label{tab:block1Projection}

\end{table}

\subsection{Treatment structure decomposition}{\label{subsec:trtDecomp}}

The treatment factors for this two-phase experiment are tag and disease status. This next step examines whether there is treatment information in the between animals and within animals in either between runs or within runs strata.

The contrast matrices are again constructed from the yield identities of the treatment effects of the experiment. Using the treatment parameters in~(\ref{eq:treatPar}) and the suffixes in model~(\ref{eq:phase2Model}), the yield identities of the treatment effects, $\alpha\_{jl}$, $j=$ healthy, diseased; $l = 114,\dots, 117$, can be written as,

\[

\alpha\_{jl} = \overline{\alpha\_{..}}+(\overline{\alpha\_{j.}} -\overline{\alpha\_{..}}) + (\overline{\alpha\_{.l}} -\overline{\alpha\_{..}}) +(\alpha\_{jl} + \overline{\alpha\_{j.}} + \overline{\alpha\_{.l}} -\overline{\alpha\_{..}}),

\]

where $\overline{\alpha\_{..}}$ denotes the means of tag and disease status effects, $\overline{\alpha\_{j.}} -\overline{\alpha\_{..}}$ denotes the disease status effect, $\overline{\alpha\_{.l}} -\overline{\alpha\_{..}}$ denotes the tag effect and $\alpha\_{jl} + \overline{\alpha\_{j.}} + \overline{\alpha\_{.l}} -\overline{\alpha\_{..}}$ is the interaction. However, the assumption is that there is no interaction between tag and disease status, so the last term can be ignored.

The yield identity for the disease status effect is $\overline{\alpha\_{j.}} -\overline{\alpha\_{..}}$. The term $\overline{\alpha\_{j.}}$ is the average over tags, thus, it can be written in matrix as $(I\_2 \otimes K\_4)\bm{\alpha}$ and $\overline{\alpha\_{..}}$ is $K\_8\bm{\alpha}$, where $\otimes$ denotes the \emph{Kronecker product}. Hence, the contrast matrix for the disease status effect is

\[ C\_{\tau} = I\_2 \otimes K\_4 - K\_8 = (I\_2 - K\_2) \otimes K\_4.\]

Similarly, the contrast matrix for the tag effect is given by

\[ C\_{\gamma} = K\_2 \otimes I\_4 - K\_8 = K\_2 \otimes (I\_4 - K\_4).\]

Having the treatment incidence matrix and the contrast matrices, we can then use the method described in Section~\ref{subsec:estTrtRBD} to estimate for the treatment information in each stratum in Table~\ref{tab:block1Projection}. To estimate the disease status information in the between animals within runs stratum, the between animal within runs vector is given by $P\_A(I-P\_R)\bm{y}$. Hence, the SS of the disease status in the within runs is

\[ \bm{y}'(I-P\_R)P\_AX'C\_{\tau}[C\_{\tau}X'P\_A(I-P\_R)X C\_{\tau}]^{-1} C\_{\tau}X'P\_A(I-P\_R)\bm{y}.\]

The coefficients of the variance and fixed components are then calculated as explained in Section~\ref{sec:infoDecomp}.

\subsection{Phase 2 ANOVA}

The ANOVA table of the Phase 2 experiment is shown in Table~\ref{tab:Phase2ANOVA}. Since there are total of 16 observations, this give us a total of 15 DF. These 15 DF are decomposed to 3 and 12 DF for between runs and within runs strata, respectively. The between runs stratum is further decomposed to between animals and residual with 1 and 2 DF, respectively. The within runs stratum is decomposed to 6 and 6 DF for between animals and between samples within animals strata, respectively. The between animals within runs is further decomposed to disease status (1 DF), tag (1 DF) and residual (4 DF). Finally, the between sample within animals within runs is decomposed to tag (2 DF) and residual (4 DF).

Comparing this ANOVA table to the ANOVA table of Phase 1 experiment in Table~\ref{tab:Phase1ANOVA}, the animals originally have 7 DF, however, 1 DF of the animals is now in the between runs stratum and another 1 DF is with tags. This is because of the confounding between the animals and runs and tags introduced from the Phase 2 design. The valid test for the disease status effect is still obtainable from the between animals within the within runs stratum, but the DF of residual is reduced from 6 to 4, which may affect the outcome of the test.

\begin{table}[ht]

\centering

\caption{Phase 2 ANOVA with the coefficients of the variance components of EMS.}

\begin{tabular}[t]{lrl}

\hline

\multicolumn{1}{l}{Source of Variation} & \multicolumn{1}{l}{DF} & \multicolumn{1}{l}{EMS}\\

\hline

Between runs \\

\hspace{3mm}Between animals & $1$ & $\sigma^2 + 2\sigma\_{A}^2 + 4\sigma\_{R}^2$\\

\hspace{3mm}Residual & $2$ & $\sigma^2 + 4\sigma\_{R}^2$\\

\hline

Within runs \\

\hspace{3mm}Between animals \\

\hspace{6mm}Disease status & $1$ & $\sigma^2 + 2\sigma\_{A}^2 + 8\theta\_{\tau}$\\

\hspace{6mm}Tag & $1$ & $\sigma^2 + 2\sigma\_{A}^2 + 4\theta\_{\gamma}$\\

\hspace{6mm}Residual & $4$ & $\sigma^2 + 2\sigma\_{A}^2$\\

\hspace{3mm}Between samples within animals &\\

\hspace{6mm}Tag & $2$ & $\sigma^2 + 4\theta\_{\gamma}$\\

\hspace{6mm}Residual & $4$ & $\sigma^2$\\

\hline

Total & $15$ & \\

\hline

\end{tabular}

\label{tab:Phase2ANOVA}

\end{table}

\section[InfoDecompuTE]{An \proglang{R} package: \pkg{InfoDecompuTE}} \label{sec:package}

\pkg{InfoDecompuTE}, applied the methods addressed in Section~\ref{sec:infoDecomp}, and is written in well known \proglang{R} programming language. The purpose of this package is allowing the researchers to automatically perform the decomposition described in Section~\ref{sec:exampleTwoPase} and produce the ANOVA table similar to Table~\ref{tab:Phase2ANOVA}. Since this package is open-source, most statistical researchers can study and adjust any part of these functions. This package is made up of two functions: \code{getVCs.onePhase} and \code{getVCs.twoPhase} for single and two-phase experiments, respectively. This section will first explain the installation procedure of this package and then the arguments that are needed for the two functions, \code{getVCs.onePhase} and \code{getVCs.twoPhase}, of package \pkg{InfoDecompuTE}.

\subsection{Installation instructions}

\pkg{InfoDecompuTE} requires a recent version of the \proglang{R} statistical programming environment which is available from the Comprehensive R Archive Network at \url{http://CRAN.R-project.org/} \citep{R2010}. The system requirements for this package depend on the number of factors and observations in the experimental design that the users attempt to analyse. This is because the number of factors and observations reflects on the dimensions of matrices for computation. Two-gigabytes of RAM with a Duo Core 3GHz machine is sufficient for \pkg{InfoDecompuTE} to analyse \citeauthor{Brien1999}'s two-phase experiment under a minute (Section~\ref{sec:example}).

Given the users has an internet connection, \pkg{InfoDecompuTE} can be installed and initiated by typing the following two command lines at new \proglang{R} session:

\begin{CodeChunk}

\begin{CodeInput}

> install.packages("infoDecompuTE")

> library("infoDecompuTE")

\end{CodeInput}

\end{CodeChunk}

The package can also be downloaded from \url{http://cran.r-project.org/web/packages/infoDecompuTE/index.html}.

\subsection{Arguments of the functions}

This section explains the arguments for the two functions in the package \pkg{InfoDecompuTE}: \code{getVCs.onePhase} and \code{getVCs.twoPhase}. The two-phase experiment example described in Section~\ref{sec:exampleTwoPase} is used repeatedly to aid explanation in this section.

The two functions and their arguments are:

\begin{CodeChunk}

\begin{CodeInput}

getVCs.onePhase(design.df, blk.str, trt.str, var.comp = NA,

trt.contr = NA, contr.matrix = all(is.na(trt.contr)),

table.legend = FALSE)

getVCs.twoPhase(design.df, blk.str1, blk.str2, trt.str, var.comp = NA,

blk.contr = NA, trt.contr = NA,

contr.matrix = all(is.na(trt.contr)), table.legend = FALSE)

\end{CodeInput}

\end{CodeChunk}

The first argument, \code{design.df}, consists of the experimental design in a data frame format. The classes of each vector in the data frame should be factors. The single-phase and two-phase experimental design in Table~\ref{tab:phase1Design} and \ref{tab:phase2Design} are

\begin{CodeChunk}

\begin{CodeInput}

> design1

\end{CodeInput}

\begin{CodeOutput}

Ani Trt

1 A a

2 B a

3 C a

4 D a

5 E b

6 F b

7 G b

8 H b

\end{CodeOutput}

\begin{CodeInput}

> design2

\end{CodeInput}

\begin{CodeOutput}

Run Ani Sam Tag Trt

1 1 A 1 114 a

2 1 B 1 115 b

3 1 C 1 116 a

4 1 D 1 117 b

5 2 E 1 114 b

6 2 F 1 115 a

7 2 G 1 116 b

8 2 H 1 117 a

9 3 C 2 114 a

10 3 D 2 115 b

11 3 A 2 116 a

12 3 B 2 117 b

13 4 G 2 114 b

14 4 H 2 115 a

15 4 E 2 116 b

16 4 F 2 117 a

\end{CodeOutput}

\end{CodeChunk}

where \code{Run} denotes MudPIT runs, \code{Ani} denotes rat ID, \code{Sam} denotes samples, \code{Tag} denotes iTRAQ$^{\rm TM}$ tags and \code{Trt} denotes disease status.

The relationship between block and treatment factors can be shown using the \citeauthor{Wilkinson1973}' syntax and written as the structure formula \citep{Wilkinson1973}. The user can also refer to the \code{formula} function in \proglang{R} for further information on the structure formula. For a single-phase experiment, the relationships between block and treatment factors are represented by two arguments: \code{blk.str} for block factors and \code{trt.str} for treatment factors. The first phase experiment in Section~\ref{subsec:Phase1Exp}, the structure formula for the block factors is \code{Ani} which denotes the animals. The structure formula for the treatment factor contains a single term for the disease status, \code{Trt}. The output from \code{getVCs.onePhase} is

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.onePhase(design1, blk.str = "Ani", trt.str = "Trt")

\end{CodeInput}

\begin{CodeOutput}

1. Preparing the block structure.

2. Preparing the treatment structure.

3. Start calculating the variance components.

4. Pre- and post-multiply NTginvATN by block projection matrices.

5. Get coefficients of each source of variation for the random effects,

and for fixed effects.

################################################################################

$random

DF Ani

Between Ani

Trt 1 1

Residual 6 1

$fixed

Trt eff.Trt

Between Ani

Trt 4 1

\end{CodeOutput}

\end{CodeChunk}

Initially, the function prints out a progress report which can be used for debugging. Note that this progress report will not be printed for the rest of this paper. The output consists of two tables for random and fixed components; both tables have a similar structure and also similar to the ANOVA Table~\ref{tab:Phase1ANOVA}. The random component's table gives the DF with the coefficients of the variance components for each source of variation. The fixed component's table provides coefficients of the fixed components with the efficiency factors, here denoted by \code{eff.Trt}. This example indicates all the disease status information is in the between animals stratum.

The two-phase experiment is represented by three arguments: these are the block factors of the phase two experiment, the block factors of the phase one experiment and the treatment factors of the overall experiment. The arguments are \code{blk.str2}, \code{blk.str1} and \code{trt.str}, respectively. In the two-phase experiment from Section~\ref{subsec:Phase2Exp}, the structure formula for the phase one block factor has samples introduced to the animals giving \code{Ani/Sam}, which denotes the two technical replicated samples nested from the animals. The phase 2 block structure contains a single term for the MudPIT run, \code{Run}. The structure formula for the treatment factors consists of two terms, one for the iTRAQ$^{\rm TM}$ tag and one for disease status, \code{Tag + Trt}. The output from the \code{getVCs.twoPhase} is

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.twoPhase(design2, blk.str1 = "Ani/Sam", blk.str2 = "Run",

trt.str = "Trt + Tag")

\end{CodeInput}

\begin{CodeOutput}

$random

DF Ani:Sam Ani Run

Between Run

Between Ani 1 1 2 4

Between Ani:Sam 2 1 0 4

Within

Between Ani

Trt 1 1 2 0

Tag 1 1 2 0

Residual 4 1 2 0

Between Ani:Sam

Tag 2 1 0 0

Residual 4 1 0 0

$fixed

Trt Tag eff.Trt eff.Tag

Between Run

Between Ani

Between Ani:Sam

Within

Between Ani

Trt 8 1

Tag 4 1

Between Ani:Sam

Tag 4 1

\end{CodeOutput}

\end{CodeChunk}

The structure of the output is identical to the output from \code{getVCs.onePhase} and is similar to the ANOVA Table~\ref{tab:Phase2ANOVA}.

The \code{var.comp} argument allows the researchers to have artificial strata to facilitate decomposition. For the example in Section~\ref{subsec:Phase2Exp}, four of eight animals can be grouped as an animal set, denoted by \code{AniSet}. This new vector is generated as,

\begin{CodeChunk}

\begin{CodeInput}

> AniSet

\end{CodeInput}

\begin{CodeOutput}

[1] 1 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2

Levels: 1 2

\end{CodeOutput}

\end{CodeChunk}

MudPIT runs are now nested within the animal set as the Phase 2 block structure, \code{AniSet/Run}. Since, the purpose of having animal set is for an artificial stratum, then there should not be any variance components for the animal set. The \code{var.comp} argument is a vector of characters indicating which variance components to appear in the output table of the random effects. In addition, the user can also use this argument to set of the order of the variance components to appear in the output table, i.e.,

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.twoPhase(design2, blk.str1 = "Ani/Sam",blk.str2 = "AniSet/Run",

trt.str = "Trt + Tag", var.comp = c("Ani:Sam", "Ani", "Run"))

\end{CodeInput}

\begin{CodeOutput}

$random

DF Ani:Sam Ani Run

Between AniSet

Between Ani 1 1 2 4

Between AniSet:Run

Between Ani:Sam 2 1 0 4

Within

Between Ani

Trt 1 1 2 0

Tag 1 1 2 0

Residual 4 1 2 0

Between Ani:Sam

Tag 2 1 0 0

Residual 4 1 0 0

$fixed

Trt Tag eff.Trt eff.Tag

Between AniSet

Between Ani

Between AniSet:Run

Between Ani:Sam

Within

Between Ani

Trt 8 1

Tag 4 1

Between Ani:Sam

Tag 4 1

\end{CodeOutput}

\end{CodeChunk}

The statistical researchers can also specify the contrasts for the block and treatment factors using the argument \code{blk.contr} and \code{trt.contr}, respectively. The argument \code{blk.contr} and \code{trt.contr} and a list of contrast vectors for each block factor in Phase 1 experiment and treatment factor. The order factors for these contrasts needs to be identical to the order of the block and treatment factors in the arguments \code{blk.str1} and \code{trt.str}. The example here will only focus on the treatment contrasts, four iTRAQ$^{\rm TM}$ tags can be represented by three orthogonal contrasts using the contrasts from a classical 2$^k$ design. Thus, the four iTRAQ$^{\rm TM}$ tags can be represented by three contrast vectors representing each orthogonal contrast as shown below

\begin{CodeChunk}

\begin{CodeInput}

> Tag = list(Tag1 = Tag1, Tag2 = Tag2, Tag3 = Tag3)

> Tag

\end{CodeInput}

\begin{CodeOutput}

$Tag1

[1] 1 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 1 -1 -1

$Tag2

[1] 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1

$Tag3

[1] 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1

\end{CodeOutput}

\end{CodeChunk}

Note that it is essential to have names for each of the contrasts for identification in the output table. The contrasts for disease status can be written as

\begin{CodeChunk}

\begin{CodeInput}

> Trt

\end{CodeInput}

\begin{CodeOutput}

[1] -0.5 0.5 -0.5 0.5 0.5 -0.5 0.5 -0.5 -0.5 0.5 -0.5 0.5 0.5 -0.5 0.5

[16] -0.5

\end{CodeOutput}

\end{CodeChunk}

Using the design described in Section~\ref{subsec:Phase2Exp}, the treatment contrasts can be input into the function \code{getVCs.twoPhase}. The argument \code{contr.matrix} is \code{TRUE}, because this case is about the contrast matrices. The output is shown as follows

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.twoPhase(design2, blk.str1 = "Ani/Sam", blk.str2 = "Run",

+ trt.str = "Trt + Tag",

+ trt.contr = list(Trt = Trt, Tag = list(Tag1 = Tag1, Tag2 = Tag2, Tag3 = Tag3))

+ contr.matrix = TRUE)

\end{CodeInput}

\begin{CodeOutput}

$random

DF Ani:Sam Ani Run

Between Run

Between Ani 1 1 2 4

Between Ani:Sam 2 1 0 4

Within

Between Ani

Trt 1 1 2 0

Tag 1 1 2 0

Residual 4 1 2 0

Between Ani:Sam

Tag 2 1 0 0

Residual 4 1 0 0

$fixed

Trt Tag eff.Trt eff.Tag

Between Run

Between Ani

Between Ani:Sam

Within

Between Ani

Trt 8 1

Tag 4 1

Between Ani:Sam

Tag 4 1

\end{CodeOutput}

\end{CodeChunk}

In addition, the researchers can break down the treatment factors into multiple orthogonal contrasts to study how these contrasts contribute to each source of variation. This is achieved by changing the argument \code{contr.matrix} to \code{FALSE}. Using the design described in Section~\ref{subsec:Phase2Exp}, the treatment contrasts can be input into the function \code{getVCs.twoPhase}, and the output is shown as follows,

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.twoPhase(design2, blk.str1 = "Ani/Sam", blk.str2 = "Run",

+ trt.str = "Trt + Tag",

+ trt.contr = list(Trt = Trt, Tag = list(Tag1 = Tag1, Tag2 = Tag2, Tag3 = Tag3)),

+ contr.matrix = FALSE, table.legend = TRUE)

\end{CodeInput}

\begin{CodeOutput}

$random

$random$VC

DF a b c

Between Run

Between Ani 1 1 2 4

Between Ani:Sam 2 1 0 4

Within

Between Ani

Trt 1 1 2 0

Tag.Tag2 1 1 2 0

Residual 4 1 2 0

Between Ani:Sam

Tag.Tag1 1 1 0 0

Tag.Tag3 1 1 0 0

Residual 4 1 0 0

$random$Legend

[1] "a = Ani:Sam" "b = Ani" "c = Run"

$fixed

$fixed$trt

a b c d e f g h

Between Run

Between Ani

Between Ani:Sam

Within

Between Ani

Trt 8 1

Tag.Tag2 4 1

Between Ani:Sam

Tag.Tag1 4 1

Tag.Tag3 4 1

$fixed$Legend

[1] "a = Trt" "b = Tag.TagA" "c = Tag.TagB" "d = Tag.TagC"

[5] "e = eff.Trt" "f = eff.Tag.TagA" "g = eff.Tag.TagB" "h = eff.Tag.TagC"

\end{CodeOutput}

\end{CodeChunk}

Note that, having broken down the tag contrasts, the table of the random components shows the tag contrast $2$ is in the between animals within runs stratum, and tag contrasts $1$ and $3$ are in the between samples within animals within runs stratum.

The argument \code{table.legend} in \code{getVCs.twoPhase} was set to \code{TRUE}, because once the contrasts are broken down, the number of columns in the table of fixed components will increase which can be difficult to read. Furthermore, a complex experimental design may comprise many block or treatment factors. The argument \code{table.legend} allows the researcher to use the legend which is shown in lower case for the variance components table.

\section[Example]{Two-phase viticultural experiment using \pkg{InfoDecompuTE}}\label{sec:example}

This section will show how the function \code{getVCs.twoPhase} is used to generate the table of EMS for the viticultural experiment described by \cite{Brien1999}. \cite{Brien1999} used the structure formulae to represent the block and treatment structures in their two-phase experiment. The first phase was the viticultural experiment comparing four different types of trellising and two pruning methods. The second phase involved the evaluation of the wines made from the viticultural experiment.

%\cite{Brien1999} described a two-phase experiment in which the first phase involved a viticultural experiment comparing four different types of trellising and two pruning methods. The second phase involved the sensory evaluation of wines made from the grapes that were grown in the viticultural experiment.

The first phase viticultural experiment was arranged into two adjacent squares, each with three rows and four column blocks. The four trellising methods were assigned to the row blocks as a randomised complete block design and to the column blocks as a BIBD. Furthermore, each plot was halved, and one of two different pruning methods was randomly assigned to each half-plot. Hence, 48 observations were made from the viticultural experiment.

The second phase experiment consisted of six judges evaluating the wines made from the grapes grown in the viticultural experiment. The wines were evaluated on two separate occasions, with wines made from grapes grown within the same square at the first phase being evaluated on the same occasion at the second phase. Each occasion was divided into three intervals, with four sittings per interval. At each sitting, each judge was presented with four glasses of duplicate wines from each of two half-plots from the same main plots in the first phase. \cite{Brien1999} referred to these glasses as the positions. They used the row and column numbers in the viticultural experiment to assign the plots in the evaluation experiment. These row and column numbers are important because they refer to the block and treatment factors of the viticultural experiment are in-cooperated into the Phase 2 wine evaluation experiment. The two-phase experiment yield a total of 576 measurements.

The structure formulae of Phase 2 and 1 block and the treatment factors are

\begin{eqnarray}

\label{eq:stru1}&&\mathrm{((Occasions/Intervals/Sittings)\*Judges)/Positions,}\\

\label{eq:stru2}&&\mathrm{(Rows\*(Squares/Columns))/Halfplots}

\end{eqnarray}

and

\begin{equation}\label{eq:stru3}

\mathrm{Trellis\*Method.}

\end{equation}

The block structure in~(\ref{eq:stru1}) indicates that Sittings are nested within Intervals which are nested within Occasions. However, since all Judges are present at every Sitting, Judges is crossed with Sittings within Intervals within Occasions. Finally, positions are nested within Judges and Sittings because four glasses of wine were evaluated by each Judge at each Sitting. The block structure defined in~(\ref{eq:stru2}) for the Phase 1 experiment indicates that the main plots, to which Trellising methods are assigned, are defined as the Rows crossed with the Columns nested within squares, with Half-plots being nested within plots. The treatment structure defined in~(\ref{eq:stru3}) is a 2-2 factorial experiment, thus, Trellising and pruning Methods are crossed.

These three structure formulae and the design are input into the function \code{getVCs.twoPhase} and the output is as follows:

\begin{CodeChunk}

\begin{CodeInput}

> getVCs.twoPhase(design, blk.str1 = "(Row\*(Squ/Col))/Hal",

+ blk.str2 = "((Oc/In/St)\*Ju)/Pos", trt.str = "Tre\*Met",

+ table.legend = TRUE)

\end{CodeInput}

\begin{CodeOutput}

$random

$random$VC

DF a b c d e f g h i j k l m n

Between Oc

Between Squ 1 12 24 96 72 288 0 1 4 16 48 0 24 96 288

Between Oc:In

Between Residual 4 0 0 0 0 0 0 1 4 16 0 0 24 96 0

Between Oc:In:St

Between Squ:Col

Tre 3 4 8 0 24 0 0 1 4 0 0 0 24 0 0

Residual 3 4 8 0 24 0 0 1 4 0 0 0 24 0 0

Residual 12 0 0 0 0 0 0 1 4 0 0 0 24 0 0

Between Ju

Between Residual 5 0 0 0 0 0 0 1 4 16 48 96 0 0 0

Between Oc:Ju

Between Residual 5 0 0 0 0 0 0 1 4 16 48 0 0 0 0

Between Oc:In:Ju

Between Row 2 12 24 96 0 0 192 1 4 16 0 0 0 0 0

Between Row:Squ 2 12 24 96 0 0 0 1 4 16 0 0 0 0 0

Residual 16 0 0 0 0 0 0 1 4 16 0 0 0 0 0

Between Oc:In:St:Ju

Between Squ:Col

Tre 3 8 16 0 48 0 0 1 4 0 0 0 0 0 0

Residual 3 8 16 0 48 0 0 1 4 0 0 0 0 0 0

Between Row:Squ:Col

Tre 3 12 24 0 0 0 0 1 4 0 0 0 0 0 0

Residual 9 12 24 0 0 0 0 1 4 0 0 0 0 0 0

Residual 72 0 0 0 0 0 0 1 4 0 0 0 0 0 0

Between Oc:In:St:Ju:Pos

Between Row:Squ:Col:Hal

Met 1 12 0 0 0 0 0 1 0 0 0 0 0 0 0

Tre:Met 3 12 0 0 0 0 0 1 0 0 0 0 0 0 0

Residual 20 12 0 0 0 0 0 1 0 0 0 0 0 0 0

Residual 408 0 0 0 0 0 0 1 0 0 0 0 0 0 0

$random$Legend

[1] "a = Row:Squ:Col:Hal" "b = Row:Squ:Col" "c = Row:Squ"

[4] "d = Squ:Col" "e = Squ" "f = Row"

[7] "g = Oc:In:St:Ju:Pos" "h = Oc:In:St:Ju" "i = Oc:In:Ju"

[10] "j = Oc:Ju" "k = Ju" "l = Oc:In:St"

[13] "m = Oc:In" "n = Oc"

$fixed

$fixed$trt

a b c d e f

Between Oc

Between Squ

Between Oc:In

Residual

Between Oc:In:St

Between Squ:Col

Tre 16/3 1/27

Residual

Between Ju

Residual

Between Oc:Ju

Residual

Between Oc:In:Ju

Between Row

Between Row:Squ

Residual

Between Oc:In:St:Ju

Between Squ:Col

Tre 32/3 2/27

Between Row:Squ:Col

Tre 128 8/9

Residual

Between Oc:In:St:Ju:Pos

Between Row:Squ:Col:Hal

Met 288 1

Tre:Met 72 1

Residual

$fixed$Legend

[1] "a = Tre" "b = Met" "c = Tre:Met" "d = eff.Tre"

[5] "e = eff.Met" "f = eff.Tre:Met"

\end{CodeOutput}

\end{CodeChunk}

\section[Conclusion]{Conclusion}

\pkg{InfoDecompuTE}, a freely available \proglang{R} package, allows statistical researchers to enter any complex single or two-phase experimental design with the structure formulae of the block and treatment factors. The package then generates the structure of the ANOVA table with the coefficients of variance components of the EMS as shown in Section~\ref{sec:infoDecomp} and \ref{sec:example}. This package will not only allow the researchers to check for valid statistical test, it will also allow them to study how the raw data is decomposed across different strata and different sources of variation of random and fixed effects.

This package can analyse non-orhtogonal designs and produce the efficiency factors for each fixed effect as shown in Section~\ref{sec:example}. The users also can represent the treatment factors using treatment contrasts allowing more flexibility in the analysis as shown in Section~\ref{sec:infoDecomp}.

However, this package has some limitations. Currently it can only analyse the single and two-phase experiments. If another phase was added, it would increase the computation time from $n^2$ to $n^3$. This is due to an additional for-loop being required to define the block structure of the additional phase. The best solution would be to re-implement the matrix calculation in another programming language such as \proglang{C} to speed up the computation time.

In addition, users need to have some understanding on how to build the model using the structure formulae for block and treatment structures of the two-phase experiments. Nonetheless, \pkg{infoDecompuTE} gives statistical researchers an additional tool for better understanding experimental designs and hence construct better experiments in the future.

\section\*{Acknowledgement}

Emma Mark, Steven Wu and Chew-Seng Chee for proofreading.

\bibliography{Reference/ref}

\end{document}