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

\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 between the phases.

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.

}

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

A primary objective of comparative experiments is to contrast measurements made on experimental units of material (e.g.\ humans, animals, plants, tissues, cells, etc.) in response to the interventions, or \emph{treatments}, applied to them. Many situations arise in practice which the response variable of interest cannot be measured directly from the experimental units in a single experiment (Phase 1). Instead, they must be further processed in a subsequent experiment (Phase 2) for the measurements to be made. Such \emph{two-phase experiments} were introduced by \cite{McIntyre1955} in the context of a study the effects of four light treatments on the synthesis of tobacco mosaic virus in the leaves of tobacco plants. Healthy tobacco plants were inoculated with the virus and subjected to different light treatments (Phase 1 experiment). To measure the severity of the disease, sap was first expressed from the experimental tobacco plants and then injected into the leaves of specific assay plants (Phase 2 experiment) on which lesions subsequently appeared and were counted. The relationship between the Phase 1 and 2 designs is important to be closely examined when designing a such experiment \citep{McIntyre1955}.

Efforts have been made towards developing a general theory for the design of two-phase experiments \citep{Brien1983, Wood1988, Brien1999, Jarrett2008}. Two-phase experiments consist of a two-step process: (1) the allocation of treatments to experimental units in the Phase 1 experiment and (2) the allocation of experimental units from the Phase 1 experiment to the experimental units in the Phase 2 experiment. Hence, the randomisation procedure generally needs to be preformed twice for each of these two allocations. \cite{Brien2006b} named the randomisation procedure for the two-phase experiments as \emph{multiple randomisation} and they compared and contrasted six different types multiple randomisation procedures for the different situations of the two-phase experiments. The procedure of separating the total variability into different sources of variation is know as \emph{information decomposition}. The decomposition of the data space for different multiple randomisation procedures are then discussed by \cite{Brien2009, Brien2010}. \cite{Brien2011} explained the theories to date in simpler terms with some fundamentals in designing two-phase experiments. In addition, \cite{Brien2011} provided a set of basic rules for deriving the EMS; but these rules applied manually and only for the balanced design \cite{Payne2003}.

Two-phase experiments commonly arise in studies using high-throughput biotechnologies for the identification and quantification of different molecular species (e.g. gene transcripts, proteins, metabolites, etc.) within the cell. \cite{Jarrett2008} conducted a detailed comparative study of the properties of two competing designs -- multiple dye-swap and the alternating loop \citep{Churchill2002} -- of the same size (i.e. same number of replicates of each treatment) for a two-colour microarray experiment (Phase 2), when the Phase 1 experiment was arranged in a completely randomised design (CRD). By constructing the theoretical analysis of variance (ANOVA) tables for these two designs, they demonstrated that the distribution of the treatment information across the strata of the ANOVA table was dependent on the Phase 2 design. In particular, while the multiple dye-swap design could be analysed using a simple ANOVA, whereas for the alternating loop design a more involved analysis was required in order to test for treatment effects. This is a consequence of the sources of variation introduced at Phase 2 interacting with those introduced at Phase 1. Thus, \cite{Jarrett2008} illustrated the importance of consider the sources of variation that are introduced at each phase when designing two-phase experiments and that a straightforward way of doing this is by constructing the relevant ANOVA tables.

Construction of ANOVA tables is a laborious manual task, even for small two-phase experiments. The commercial statistical software \proglang{Genstat} can be used to perform the decomposition of \emph{degrees of freedom} (DF) in the ANOVA but only for generally balanced designs which requires the use of pseudofactors \citep{Monod1992}. The \code{AMTIER} procedure within \proglang{Genstat} allows the fitting of three model formulae, two model formulae for the block factors of the Phase 1 and 2 experiments and one model formula for the treatment factors, to produce the ANOVA with the decomposition of DF \citep{Brien2006a}. The limitation of these approaches is that they do not generate the \emph{expected mean square} (EMS) of the treatment and residuals which are needed in order to assess the properties of competing designs. The \code{GLM} procedures in \proglang{JMP} and \proglang{SAS}, and the \code{ANOVA} command in \proglang{Minitab}, have the ability to compute the EMSs, but neither program has the ability to do this for two-phase experiments. While the \proglang{R} package \pkg{dae}, developed by \cite{Brien2011a}, has the ability to perform information decomposition of a given design. This package is only capable of generating a projection matrix for each factor and performing a decomposition within a single stratum. Therefore, a statistical software package that is capable of performing all of these tasks simultaneously and in a straightforward manner does not exist.

We introduce an \proglang{R} package called \pkg{infoDecompuTE} for the {\bf info}rmation {\bf decomp}osition of {\bf T}wo-phase {\bf E}xperiments. It is capable of quickly generating theoretical ANOVA tables, i.e.\ the ANOVA table consists of DF and EMS, for from the designs of any given single- or two-phase experiment, showing how the known sources of variation in the experiment are distributed across different strata and thereby enabling researchers to compare the properties of competing designs. As a direct consequence, one can quickly assess whether a particular design leads to a valid F-test.

In this article, we demonstrate the concepts and the methods underlying \pkg{infoDecompuTE} as well as its use. Section~\ref{sec:infoDecomp} explains the information decomposition for a single-phase experiment. Section~\ref{sec:infoiDecompTwoPase} discusses the information decomposition for the two-phase experiments and how it is different to the single-phase experiment. Section~\ref{sec:exampleTwoPase} shows an example of two-phase experiments with the theoretical ANOVA tables. Section~\ref{sec:package} demonstrates the usage of R package \pkg{infoDecompuTE}. Finally, section~\ref{sec:example} illustrates how the use of package with a published two-phase viticulture-sensory evaluation experiment by \cite{Brien1999}.

**\section{Information decomposition of designed experiments}**

**\label{sec:infoDecomp}**

The design of an experiment and the linear model which describes it are intimately linked. More specifically, the linear model describes the relationships between the experimental units, or \emph{block structure}, between the treatments, or \emph{treatment structure}, and the assignment of the treatments to the experimental units. In the following, an arbitrary design is considered involving both block and treatment factors whose relationships can be represented mathematically by a linear \emph{mixed-effects model}. We will show how the raw data from such an experiment is decomposed into its constituent components based on its block and treatment structures.

**\subsection{The linear mixed-effects model}**

**\label{subsec:matrixLMM}**

Consider an experiment involving $V$ treatments and $B$ block factors, where the $i$th treatment and $j$th block factors consist of $v\_i$ and $b\_j$ levels, respectively. The linear mixed-effects model for such an experiment can be written in matrix notation as

\begin{equation}\**label{eq:matrixLMM}**

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

\end{equation}

where $\bm{y}$ is an $n \times 1$ vector of responses, $\mathbf{1}$ is an $n \times 1$ vector with all elements unity, $\mu$ denotes the grand mean of the data, and $\bm{\epsilon}\sim \mathcal{N}(0,\sigma^2 \I)$ is a $n \times 1$ vector of unobserved random experimental errors, where $\I$ denotes identity matrix. The vector of treatment parameter is, therefore, defined as

\begin{equation}

\**label{eq:treatPar}**

\bm{\alpha} = (\alpha\_{11 \dots 1}, \alpha\_{11 \dots 2}, \dots, \alpha\_{1v\_2 \dots v\_V},\dots,\alpha\_{v\_1 v\_2 \dots v\_V}),

\end{equation}

where each $\alpha$ has $V$ number of indexes corresponding to $V$ treatment factors factor. Each $\alpha$ denotes the effects from a specific treatment combination. The treatment design matrix, $\X$, in (\ref{eq:matrixLMM}) describes the overall allocation of each treatment combination. The dimension of $\X$ consists of $n$ rows and the number of column is the length of vector $\bm{\alpha}$. As for the vector of block parameters, it is defined as

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

\bm{\beta} = (\bm{\beta}\_1, \bm{\beta}\_2, \ldots, \bm{\beta}\_B),

\end{equation}

where

\[

\bm{\beta}\_i = (\beta\_{i1}, \beta\_{i2}, \dots, \beta\_{i b\_i})

\]

and

$\beta\_{ij} ~ \sim \mathcal{N}(0, \sigma\_i^2)$ ($i=1,2,\dots,B; j=1,2,\dots, b\_i$). The block design matrix, $\Z$, in (\ref{eq:matrixLMM}) can be then expressed as

\begin{equation}\**label{eq:block1Mat}**

\Z = [\Z\_1 \vert \Z\_2 \vert \ldots \vert \Z\_B ],

\end{equation}

where $\Z\_i$ is the design matrix corresponding to $\bm{\beta}\_i$. Thus, the dimension of $\Z$ consists of $n$ rows and the number of column is the sum of the length of each vector in $\bm{\beta}$, i.e.\

\[

\sum^{B}\_{i = 1} b\_i.

\]

In summary, this subsection presents a general linear mixed-effect model for a designed experiment. Based on the treatment and block structures of an experiment, the treatment and block design matrices can be defined which are shown to be very different. The treatment design matrix describes the allocation of the treatment combinations corresponding to the elements in $\bm{\alpha}$. The block design matrix is made up of block design sub-matrices, where the $i$th sub-matrix corresponds to the $i$th vector of block parameter in $\bm{\beta}$. These two types of design matrix are the essential components in the method of decomposition which is described in the remaining of this section.

**\subsection{Null decomposition using projection matrices}**

**\label{subsec:strataDecompProj}**

The vector of response, $\bm{y}$, in (\ref{eq:matrixLMM}) spans an $n$-dimensional Euclidean space, commonly denoted by $\mathbb{R}^n$. A vector space, $\mathbb{V}$, is a \emph{subspace} of $\mathbb{R}^n$, i.e.\ $\mathbb{V} \subset \mathbb{R}^n$, if every vector in $\mathbb{V}$ is also in $\mathbb{R}^n$ \citep{Hadi1996}. The information decomposition of $\bm{y}$ is its separation from $\mathbb{R}^n$ space into its constituent vector subspace components. These vector subspaces can be referred to as the \emph{strata} of the ANOVA and can be mathematically expressed as

\begin{equation}

\**label{eq:vecSpace}**

\mathbb{R}^n = \mathbb{V}\_0 \oplus \mathbb{V}\_1 \oplus, \dots , \oplus \mathbb{V}\_{B},

\end{equation}

where $\oplus$ denotes addition operator of the vector spaces, $\mathbb{V}\_i$ denotes the stratum that corresponds to the $i$th block parameters and block design sub-matrices defined in (\ref{eq:block1Par}) and (\ref{eq:block1Mat}), respectively. For any designed experiment, the first element, $\mathbb{V}\_0$, in (\ref{eq:vecSpace}) always denotes the grand mean vector subspace.

Since the decomposition is the separation of the known variation in the data, the variance structure of the data, $\bm{y}$, can be expressed in a spectral form as

\begin{equation}

\**label{eq:strata}**

\operatorname{Var}(\bm{y}) = \sum\_{i=0}^{B} \xi\_i \Q\_i,

\end{equation}

where $\Q\_i$ is an $n \times n$ idempotent matrix and is, therefore, the \emph{orthogonal projector} of $\bm{y}$ onto the vector subspace $\mathbb{V}\_i$ (i.e.\ stratum $i$) and $\xi\_i$ is the $i$th stratum variance (i.e.\ $\xi\_i = \operatorname{Var}(\Q\_i \bm{y})$).

The matrix $\Q\_i$ can also be shown in the decomposition of the total SS, $\bm{y}'\bm{y}$, into $B$ components of the SS, i.e.\

\begin{equation}

\**label{eq:decomp}**

\bm{y}'\bm{y} = \sum\_{i=0}^{B}\bm{y}'\Q\_i\bm{y},

\end{equation}

where $\bm{y}'\Q\_i\bm{y}$ denotes the SS of the $i$th stratum. Hence, Equations~(\ref{eq:vecSpace}), (\ref{eq:strata}) and (\ref{eq:decomp}) give a basic illustration of decomposition ignoring the treatment, namely the \emph{null decomposition}. The process results from the sequential fitting of the block factors, computing the SS of each block factor and sweeping the computed SS from the total SS. The remainder of this section describes each step of null decomposition.

The initial step of null decomposition is to sweep the grand mean from raw data, $\bm{y}$. The $\mu$ is a vector of length $1$; thus, the grand mean vector spans in a $1$-dimensional grand mean vector subspace, denoted by $\mathbb{V}\_0$. To sweep the grand mean from $\bm{y}$, $\bm{y}$ is first projected onto the grand mean vector subspace, which is giving by $\Q\_{0}\bm{y}$. The grand mean vector subspace can be represented by a $n \times n$ averaging matrix, denoted by $\K\_n$ or $\K$, with all elements equal to ${n}^{-1}$. The orthogonal projector of $\mathbb{V}\_0$ is then derived by

\begin{equation}

\**label{eq:vectorProj}**

\Q\_{0} = \mP\_{\K} = {\K}({\K}'{\K})^{-1}{\K}',

\end{equation}

where $\mP\_K$ denotes the \emph{projection matrix} of matrix $\K$. The projection matrix of any averaging matrix can be shown to be identical as the averaging matrix, i.e.\

\[\Q\_{0} \bm{y} = \mP\_{\K} \bm{y} = {\K} \bm{y}.\]

Next, subtract $\bm{y}$ by the grand mean vector, $\K\bm{y}$, i.e.\

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

where $(\I-\K)\bm{y}$ denotes the \emph{mean corrected observational vector}. The $(\I-\K)\bm{y}$ spans in $\mathbb{V}^{\perp}\_0$ with the dimension of $(n - 1)$. The $\mathbb{V}^{\perp}\_0$ also stands for the \emph{orthogonal complement} of $\mathbb{V}\_0$. Furthermore, the adjusted total SS is obtained by pre-multiplying the $(\I-\K)\bm{y}$ 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 total adjusted SS can be calculated as

\[

\bm{y}'\bm{y} - \bm{y}'\K\bm{y} = \bm{y}'(\I-\K)\bm{y},

\]

which is based on the \emph{Pythagoras' theorem}.

The $(\I-\K)\bm{y}$ is then projected onto the next vector subspace, $\mathbb{V}\_1$, giving by $\Q\_{1}\bm{y}$, where the $\Q\_{1}$ is the orthogonal projector of the stratum for the first block parameter, $\bm{\beta}\_1$. The vector subspace $\mathbb{V}\_1$ is represented by the block design matrix, $\Z\_1$, and its projection matrix, $\mP\_{\Z\_1}$ can be computed from

\[\Z\_1(\Z\_1'\Z\_1)^{-1}\Z\_1'.\]

To project $(\I-\K)\bm{y}$ onto $\mathbb{V}\_1$, $(\I-\K)\bm{y}$ is pre-multiplying by $\mP\_{\Z\_1}$ and can be expressed as

\begin{equation}\**label{eq:projectB}**

\mP\_{\Z\_1}[(\I-\K)\bm{y}] = (\mP\_{\Z\_1} - \K)\bm{y} = \Q\_{1}\bm{y},

\end{equation}

where vector $(\mP\_{\Z\_1} - \K)\bm{y}$ estimates of the vector of block parameters in $\bm{\beta}\_1$. The orthogonal complement of $(\mP\_{\Z\_1} - \K)\bm{y}$ is derived by

\begin{equation}

\**label{eq:orthComp}**

(\I-\K)\bm{y}- (\mP\_{\Z\_1} - \K)\bm{y} = (\I -\mP\_{\Z\_1})\bm{y},

\end{equation}

which corresponds to the elimination on the effects of $\bm{\beta}\_1$.

The SS are derived by pre-multiplying the vectors in (\ref{eq:projectB}) and (\ref{eq:orthComp}) by its transpose as described in (\ref{eq:adjustSS}), i.e.\

\[

\bm{y}'(\I-\K)\bm{y}- \bm{y}'(\mP\_{\Z\_1} - \K)\bm{y} = \bm{y}'(\I -\mP\_{\Z\_1})\bm{y}.

\]

If the block structure contains additional block factors, e.g.\ plots and/or subplots, the vector $(I - P\_{Z\_1})\bm{y}$ is further projected onto the next vector subspace, $\mathbb{V}\_2$. To summarise, to project the raw data vector, $\bm{y}$ from $\mathbb{V}\_{i}$ onto $\mathbb{V}\_{i + 1}$ can be written as $P\_{Z\_{i+1}}Q\_{i}\bm{y}$. The orthogonal complement of $P\_{Z\_{i+1}}Q\_{i}\bm{y}$ can be derived by subtraction, i.e.\

\begin{equation}

\**label{eq:orthCompSummary}**

\Q\_{i}\bm{y}- \mP\_{\Z\_{i+1}}\Q\_{i}\bm{y} = (\I -\mP\_{\Z\_{i+1}})\Q\_{i}\bm{y} = \Q\_{i+ 1}\bm{y}.

\end{equation}

Thus, the SS are derived by pre-multiplying the vectors in (\ref{eq:orthCompSummary}) as

\[

\bm{y}'\Q\_{i}\bm{y}- \bm{y}'Q\_{i}\mP\_{\Z\_{i+1}}\Q\_{i}\bm{y} = \bm{y}'\Q\_{i}(\I -\mP\_{\Z\_{i+1}})\Q\_{i}\bm{y}=\bm{y}'\Q\_{i+ 1}\bm{y}.

\]

Providing the SS for each stratum is defined using the orthogonal projectors, the EMS can then be computed for the theoretical ANOVA table. From (\ref{eq:decomp}), the expected sum of squares (ESS) of the $i$th stratum without the treatment effects can be shown as

\begin{equation}

\**label{eq:ESSQuad}**

\operatorname{E}(\bm{y}'\Q\_i\bm{y})= \mathrm{tr}(\Q\_i)\operatorname{cov}(\bm{y}),

\end{equation}

where $\operatorname{cov}(\bm{y})$ is the variance covariance matrix and $\mathrm{tr}(\Q\_i)$ is the \emph{trace} of the matrix $\Q\_i$ \citep{Searle1982}.

Consider an experiment arranged in RCBD, the null decomposition of the total SS can be expressed as

\begin{equation}

\**label{eq:infoDecomp1}**

\bm{y}'\bm{y} = \bm{y}'\K\bm{y} + \bm{y}'(\mP\_{\Z\_1}-\K)\bm{y} + \bm{y}'(\I - \mP\_{\Z\_1})\bm{y},

\end{equation}

where $K$, $(\mP\_{\Z\_1}-\K)$ and $(\I - \mP\_{\Z\_1})$ denote the orthogonal projectors of grand mean, Between and Within Blocks strata, i.e.\ , $\mathbb{V}\_0$, $\mathbb{V}\_1$ and $\mathbb{V}\_2$, respectively.

Since the SS of Between Blocks stratum is $\bm{y}'(\mP\_{\Z\_1}-\K)\bm{y}$, the ESS of Between Blocks stratum can then be shown as

\begin{eqnarray}

\nonumber \operatorname{E}(\bm{y}'(\mP\_{\Z\_1}-\K)\bm{y}) &=& \mathrm{tr}(\mP\_{\Z\_1}-K)\operatorname{cov}(\bm{y})\\

\nonumber &=& \mathrm{tr}(\mP\_{\Z\_1}-\K)\operatorname{cov}(\bm{\epsilon}) + \mathrm{tr}[\Z'(\mP\_{\Z\_1}-\K)\Z]\operatorname{cov}(\bm{\beta})\\

\nonumber &=& (b\_1 - 1)\sigma^2 + (b\_1 - 1)v\sigma^2\_1\\

**\label{eq:computRBD}** &=& (b\_1 - 1)(\sigma^2 +v\sigma^2\_1),

\end{eqnarray}

where $b\_1$ and $v$ denotes the block number and block size, respectively. Subsequently, the EMS is calculated by dividing the ESS by the corresponding DF. Hence, the EMS of the Between Blocks stratum is $\sigma^2 +v\sigma^2\_1$. Similarly, the EMS of the Within Blocks stratum can be shown as $\sigma^2$. The theoretical ANOVA table without the treatment components for RCBD is shown in Table~\ref{tab:infoDecomp}.

\begin{table}[ht]

\centering

\caption{ANOVA with the coefficients of variance components of the EMS from an experiment arranged with RCBD.}

\begin{tabular}[t]{lllll}

\toprule

\multicolumn{1}{l}{\bf Source of Variation}& \multicolumn{1}{l}{\bf Vector (Sub)space} & \multicolumn{1}{l}{\bf DF} & \multicolumn{1}{l}{\bf SS} & \multicolumn{1}{l}{\bf EMS}\\

\midrule

Between Blocks & $\mathbb{V}\_1$ &$b\_1-1$ & $\bm{y}'(\mP\_{\Z\_1}-\K)\bm{y}$ & $\sigma^2 + v\sigma\_{1}^2$\\

Within Blocks & $\mathbb{V}^{\perp}\_1 = \mathbb{V}\_2$ &$b\_1(v - 1)$ & $\bm{y}'(\I - \mP\_{\Z\_1})\bm{y}$ & $\sigma^2$\\

\hline

Adjusted Total & $\mathbb{V}^{\perp}\_0$ & $n - 1$ & $\bm{y}'(\I - \K)\bm{y}$ \\

\hline

Grand Mean & $\mathbb{V}\_0$ & $1$ & $\bm{y}'\K\bm{y}$ \\

\midrule

Total & $\mathbb{R}^n$ & $n$ & $\bm{y}'\bm{y}$ \\

\bottomrule

\end{tabular}

**\label{tab:infoDecomp}**

\end{table}

**\subsection{Computing the treatment SS}**

**\label{subsec:estTrt}**

The previous section overview of the process to computing the SS and EMS from the null decomposition, the next step is to compute the treatment SS and EMS within each stratum. In order to compute the treatment SS, it is necessary to have the treatment design matrix, $X$, defined in (\ref{eq:matrixLMM}) and another set of matrices. This set of matrices are known as the treatment contrast matrices, denoted by $\C\_x$ \citep{John1987}. These matrices can be generated, based on the treatment structure and \emph{yield identity}, which describes how the treatment effects are partitioned into orthogonal components.

Consider an example where the treatment structure consists of a single treatment factor is denoted by $\alpha\_i$, $i= 1,\dots, v$, the yield identity is given by

\begin{equation}\**label{eq:yieldIdenity}**

\alpha\_{i} = \overline{\alpha\_{.}}+(\alpha\_{i} -\overline{\alpha\_{.}})

\end{equation}

The dot in the subscript denotes the summation over the subscript it replaces and the over-line, also known as \emph{bar}, indicates the average over the terms associated with the nominal subscript. Thus, $\overline{\alpha\_{.}}$ denotes the overall mean of the $\alpha\_i$. The term, $\alpha\_{i} -\overline{\alpha\_{.}}$, denotes the effect of treatment $i$ corrected from the mean. In matrix notation, (\ref{eq:yieldIdenity}) can be written as

\[

\bm{\alpha} = \C\_0 \bm{\alpha} + \C\_1 \bm{\alpha}

\]

where

\begin{eqnarray}

\nonumber \C\_0 &=& \K\_v\\

\nonumber \C\_1 &=& \I\_v - \K\_v

\end{eqnarray}

where $\I\_v$ is $v \times v$ identity matrix and $\K\_v$ is $v \times v$ averaging matrix. The $\C\_0 \bm{\alpha} $, corresponds to $\overline{\alpha\_{.}}$, can be seen as an operation which averages over the $\alpha\_i$. The $\C\_1\bm{\alpha}$, corresponds to $\alpha\_{i} - \overline{\alpha\_{.}}$, represents the effects of the treatment after the mean is swept.

Consider a factorial experiment with two factors $\tau$ and $\gamma$ at $v\_1$ and $v\_2$ levels, the yield identity of $\alpha\_{ij}$ can be written as,

\begin{equation}

\**label{eq:yieldIdentityFactorial}**

\alpha\_{ij} = \overline{\alpha\_{..}}+(\overline{\alpha\_{i.}} -\overline{\alpha\_{..}}) + (\overline{\alpha\_{.j}} -\overline{\alpha\_{..}}) +(\alpha\_{ij} + \overline{\alpha\_{i.}} + \overline{\alpha\_{.j}} -\overline{\alpha\_{..}}),

\end{equation}

where $\overline{\alpha\_{..}}$ denotes the overall mean of $\alpha\_{ij}$, $\overline{\alpha\_{i.}} -\overline{\alpha\_{..}}$ denotes the main effects of factor $\tau$, $\overline{\alpha\_{.l}} -\overline{\alpha\_{..}}$ denotes the main effects of factor $\gamma$ and $\alpha\_{ij} + \overline{\alpha\_{i.}} + \overline{\alpha\_{.j}} -\overline{\alpha\_{..}}$ is the interaction of two treatment factors. In the matrix notation, (\ref{eq:yieldIdentityFactorial})can be written as

\[

\bm{\alpha} = \C\_{00} \bm{\alpha} + \C\_{10} \bm{\alpha} + \C\_{01} \bm{\alpha} + \C\_{11} \bm{\alpha}

\]

where

\begin{eqnarray}

\nonumber \C\_{00} = & \K\_{v\_1} \otimes \K\_{v\_2}& = \K\_{v\_1 v\_2}\\

\nonumber \C\_{10} = & \I\_{v\_1} \otimes \K\_{v\_2} - \K\_{v\_1 v\_2}& = (\I\_{v\_1} - \K\_{v\_1}) \otimes \K\_{v\_2} \\

\nonumber \C\_{01} = &\K\_{v\_1} \otimes \I\_{v\_2} - \K\_{v\_1 v\_2}& = \K\_{v\_1} \otimes (\I\_{v\_2} - \K\_{v\_2})\\

\nonumber \C\_{11} = &\I\_{v\_1 v\_2}-\I\_{v\_1} \otimes \K\_{v\_2} - \K\_{v\_1} \otimes \I\_{v\_2} + \K\_{v\_1 v\_2} &= (\I\_{v\_1} - \K\_{v\_1}) \otimes (\I\_{v\_2} - \K\_{v\_2}),

\end{eqnarray}

where denotes $\otimes$ an operator for the Kronecker product. Notices the experiment with single treatment factor, the $C$ matrix is

\begin{equation}

\**label{eq:contrMat}**

\C\_{x\_i} =

\begin{cases}

\K\_{v\_i}, & x\_i = 0 \\

\I\_{v\_i} - \K\_{v\_i}, & x\_i = 1,

\end{cases}

\end{equation}

where $x\_i$ is a binary number. For the $v\_1 \times v\_2$ factorial experiment, the treatment contrast matrix is given by

\[

\C\_{x\_1 x\_2} = \C\_{x\_1} \otimes \C\_{x\_2}.

\]

Therefore, the treatment structure of an $V$-factor experiment is given by

\begin{equation}

\bm{\alpha} = \sum\_x{\C\_x\bm{\alpha}}

\end{equation}

where the $x$ is a set of binary numbers, $(x\_1 x\_2 \dots x\_{V})$, and

\[

\C\_x = \C\_{x\_1} \otimes \C\_{x\_2} \otimes \dots \otimes \C\_{x\_{V}} = \bigotimes^{V} \_{i = 1} \C\_{x\_i},

\]

where $\C\_{x\_i}$ is defined in (\ref{eq:contrMat}) \citep{John1987}.

The reduced normal equations for the treatment parameter can then be defined by eliminating the mean and block parameters from the full set of normal equations~\citep{John1987}. This can be shown as

\[

\A\_i\bm{\alpha} = \bm{q}\_i,

\]

where

\begin{eqnarray\*}

\A\_i &=& \mL\_{xi} \mL\_{xi}' ,\\

\bm{q}\_i &=& \mL\_{xi} \bm{y},

\end{eqnarray\*}

where $\mL\_{xi} = \C\_x \X' \Q\_i$, $\A\_i$ denotes the information matrix and $\bm{q}\_i$ is vector of the adjusted treatment totals. Since the main interest is in the treatment effects, the equation is rewritten with respect to $\bm{\alpha}$ as

\begin{equation}

\**label{eq:trtVec}**

\bm{\alpha} = \A\_{i}^{-} \bm{q}\_i = (\mL\_{xi}\mL\_{xi}')^{-}\mL\_{xi} \bm{y},

\end{equation}

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, which cannot be inverted using the normal inverse operation. Hence, $\A\_{i}^{-} \bm{q}\_i$ denotes the treatment vector in the $i$th vector subspace or stratum. The treatment SS for the $i$th stratum is given by

\begin{equation}

\**label{eq:trtSS}**

\bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i}.

\end{equation}

Providing the SS of $i$th stratum is $\bm{q}\_{i}'\bm{q}\_{i} = \bm{y}' \Q\_i\bm{y}$, the residual SS can be computed by subtraction as

\begin{equation}

\**label{eq:resSS}**

\bm{q}\_{i}'\bm{q}\_{i} - \bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i} = \bm{q}\_{i}' ( \I -\A\_{i}^{-}) \bm{q}\_{i},

\end{equation}

where $\bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i}$ and $\bm{q}\_{i}' ( \I -\A\_{i}^{-}) \bm{q}\_i$ can be re-written as $\bm{y}'\mL\_{xi}' \A\_{i}^{-}\mL\_{xi}\bm{y}$ and $\bm{y}\mL\_{xi}' ( \I - \A\_{i}^{-}) L\_{xi}\bm{y}$, respectively. Thus, $\mL\_{xi}' \A\_{i}^{-}\mL\_{xi}$ and $\mL\_{xi}' ( \I - \A\_{i}^{-}) \mL\_{xi}$ are the orthogonal projectors that decomposes the unadjusted SS, $\bm{y}' \bm{y}$, to the treatment SS and residual SS in stratum $i$, respectively. These two orthogonal projectors are also used to compute the coefficients of the variance components in the treatment and residual ESS as described in (\ref{eq:ESSQuad}).

The coefficients of the treatment parameters, commonly denoted by $\theta$, of the treatment EMS can be derived directly from the treatment replication, which is the total number of objection divided by the level of the corresponding treatment factor. If the design consists of a non-orthogonal treatment structure, the \emph{average efficiency factor}, which describes the amount of separation of the treatment information across different strata, is needed to be derived. The Equation~(\ref{eq:trtSS}) can be shown as

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

\bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i} = \dfrac{1}{r\_t} \sum^{v - 1}\_{j = 1} e\_j^{-1} (\bm{p}\_i' \bm{q}\_{i})^2

\end{equation}

where $r\_t$ denotes the number of treatment replication, $e\_i$ denotes the canonical efficiency factor and $\bm{p}\_i$ is $i$th eigenvector of the treatment information matrix. Given that $\lambda\_i$ and $\bm{p}\_i$ are $i$th eigenvalue and eigenvector of the treatment information matrix, the $i$th basic treatment contrast is given by $\bm{p}\_i'\bm{\alpha}$ and the $i$th \emph{canonical efficiency factors} is calculated by $\dfrac{\lambda\_i}{r\_t}$ \citep{John1987}. The harmonic mean of the canonical efficiency factors, i.e.\

\[

E = \dfrac{v-1}{\sum^{v - 1}\_{j = 1} e\_j^{-1}},

\]

gives the \emph{average efficiency factor}, denoted by $E$, which is the amount of the treatment information that is present in a stratum of the information matrix derived \citep{Yates1936}. Therefore, Equation~(\ref{eq:trtSS}) provides the estimation of the \emph{efficiency factor adjusted treatment SS}.

In summary, this section described the information decomposition of a single phase experiment, which consisted of three basic decomposition steps: adjusting for the grand mean, defining the stratum based on the block structures and computing the treatment SS based on the treatment structure in the block structure. The next section extends the decomposition method to the two-phase experiments.

**\section{Information decomposition of a two-phase experiment}**

**\label{sec:infoiDecompTwoPase}**

The decomposition of the two-phase experiment is more complicated than single phase experiment, because the two-phase experiment is made up of two sets of block structure from Phase 1 and 2 experiments and the overall treatment structure. The information decomposition of a two-phase experiment starts with the raw data vector being adjusted for the grand mean, the strata from the Phase 2 block structure are then defined. The Phase 2 strata are always defined before the Phase 1 strata, because the Phase 2 strata always contributes to the outer strata of the ANOVA table. The next step is to define the strata of the Phase 1 block structure from each stratum of the Phase 2 block structure. The final step is to estimate the treatment effect in block structure of the Phase 1 experiment. This section describes the information decomposition of the two-phase experiments.

**\subsection{The linear mixed-effects model}**

**\label{subsec:phase2Model}**

Consider a two-phase experiment with $V$ treatments, $B\_1$ block factors from Phase 1 experiment and $B\_2$ block factors from Phase 2 experiment, where the $i$th treatment, $j$th and $k$th block factors of Phase 1 and 2 experiments consist of $v\_i$, $b1\_j$ and $b2\_k$ levels, respectively. the general linear mixed-effect model in the matrix notation, can be 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 $\Z\_1$ and $\Z\_2$ are the block design matrices corresponding to the Phase 1 and Phase 2 experiments, respectively. The vectors block parameters for Phases 1 and 2 are $\bm{\beta}\_1$ and $\bm{\beta}\_2$, respectively. The $\bm{\beta}\_2$ and $\Z\_2$ have the same structure as the $\bm{\beta}$ and $\Z$ defined in (\ref{eq:matrixLMM}). Thus, the $\bm{\beta}\_2$ is consists of a set of vector from block parameters and $\Z\_2$ is made up of a set of sub-matrices. As for $\bm{\beta}\_1$ and $\Z\_1$, they are constructed with the same way as $\bm{\alpha}$ and $\X$ in (\ref{eq:matrixLMM}). This means the $\bm{\beta}\_1$ is consists of every combination of the block parameters from the Phase 1 experiments and the $\Z\_1$ is describing the allocation of the block combinations.

The block factors from the Phase 1 and 2 experiment are considered separately in the linear mixed-effect model is because the allocation of experimental units from the Phase 1 experiment to the experimental units in the Phase 2 experiment can be non-orthogonal; thus, the amount of block information from Phase 1 experiments may not stay intact in the strata of the Phase 2 experiments \cite{Wood1988}. Hence, the decomposition procedure of treatments factors to block factors described in Section~\ref{subsec:estTrt} is used by treating the Phase 1 block structure as the treatment structure. This can help us to adjust for the efficiency factors, under the situation of non-orthogonal block structure, while applying the projection from the Phase 1 block vectors onto Phase 2 block spaces as shown in (\ref{eq:projTrtBlock}).

**\subsection{Null decomposition of Phase 2 block structure}**

The same notation of $\Q\_i$ is used for the orthogonal projectors for the stratum $i$ for the block effects arise from the Phase 2 experiment. Since the same null decomposition is used to define the orthogonal projectors of each stratum, $\Q\_i$ is computed with the exact same way as described in Section~\ref{subsec:strataDecompProj}. Therefore, the $\bm{y}' \Q\_i \bm{y}$ denote SS decomposed from total SS, $\bm{y}'\bm{y}$, onto the stratum $i$ of the Phase 2 experiment.

**\subsection{Null decomposition of Phase 1 block structure}**

As mentioned in Section~\ref{subsec:phase2Model}, the allocation of experimental units from the Phase 1 experiment to the experimental units in the Phase 2 experiment can be non-orthogonal; so the Phase 1 block structure has to be treated as the treatment structure while performing the decomposition. This subsection uses a split-block design to describe the decomposition method which is similar to the method described in Section~\ref{subsec:estTrt}.

Consider an experiment arranged with a split-plot design with $b1\_1$ blocks and each block containing $b1\_2$ plots, the $\bm{\beta}\_1$ consists of the effects of plots within blocks as

\begin{equation}\**label{eq:blockParSplitBlk1}**

(P\_{11}, \dots, P\_{b1\_1 b1\_2}).

\end{equation}

where $P\_{ij}$ denotes the effect from plot $j$ within block $i$. The effect from block $i$ is denoted by $\overline{P\_{i.}}$. Thus, the block design matrix, $Z\_1$, is the design matrix of the plots, denoted by $\Z\_P$, which is as $n \times b1\_1 b1\_2$ matrix. The yield identity of $P\_{ij}$ can be written as,

\begin{equation}

\**label{eq:yieldIdentityNest}**

P\_{ij} = \overline{P\_{..}}+(\overline{P\_{i.}} -\overline{P\_{..}}) + (P\_{ij} - \overline{P\_{i.}}),

\end{equation}

where $\overline{P\_{..}}$ denotes the overall mean of $P\_{ij}$, $\overline{P\_{i.}} -\overline{P\_{..}}$ denotes the effect from block $i$ and $P\_{ij} - \overline{P\_{i.}}$ denotes the effect from plot $j$ within block $i$. In the matrix notation, (\ref{eq:yieldIdentityNest}) can be written as

\[

\bm{\beta}\_1 = \C\_{00} \bm{\beta}\_1 + \C\_{10}\bm{\beta}\_1 + \C\_{21}\bm{\beta}\_1

\]

where

\begin{eqnarray}

\nonumber \C\_{00} = & \K\_{b1\_1} \otimes \K\_{b1\_2}& = \K\_{b1\_1 b1\_2}\\

\nonumber \C\_{10} = & \I\_{b1\_1} \otimes \K\_{b1\_2} - \K\_{b1\_1 b1\_2}& = (\I\_{b1\_1} - \K\_{b1\_1}) \otimes \K\_{b1\_2} \\

\nonumber \C\_{21} = & \I\_{b1\_1 b1\_2} - \I\_{b1\_1} \otimes \K\_{b1\_2} & = \I\_{b1\_1} \otimes (\I\_{b1\_2} - K\_{b1\_2}).

\end{eqnarray}

Unlike the $x\_i$ in $\C\_{x\_i}$ matrix defined in (\ref{eq:contrMat}) which is a binary number, the block contrast matrix, denoted by $\C\_z$, can also considered the nested relationship between the block factors. Hence, an additional criterion is included, i.e.\

\begin{equation}

\**label{eq:contrMat1}**

\C\_{z\_i} =

\begin{cases}

\K\_{b1\_i}, & z\_i = 0 \\

\I\_{b1\_i} - \K\_{b\_i}, & z\_i = 1\\

\I\_{b1\_i}, & z\_i = 2,

\end{cases}

\end{equation}

where $z\_i$ is a \emph{ternary} number. Since the contrast matrix is expressed by the Kronecker product, not by the usual matrix multiplication, the identity matrix cannot be ignored.

Therefore, the Phase 1 block structure consists of $B1$-factor, the Phase 1 block parameter is then given by,

\begin{equation}

\bm{\beta}\_1 = \sum\_z{\C\_z\bm{\beta}\_1}

\end{equation}

where the $z$ is a set of ternary numbers, $(z\_1 z\_2 \dots z\_{B})$, and

\[

\C\_z = \C\_{z\_1} \otimes \C\_{z\_2} \otimes \dots \otimes \C\_{z\_{B}} = \bigotimes^{B} \_{i = 1} \C\_{z\_i},

\]

The $\C\_{z}$ matrix should also be applied to treatment contrast matrix $\C\_{x}$ defined in (\ref{eq:contrMat} to cover the situation when the relationship between the treatment factors is nested.

Providing the SS in stratum $i$ is $\bm{y}' \Q\_i\bm{y}$, SS stratum $j$ from the Phase 1 experiment in stratum $i$ from the Phase 2 experiment is given by

\[

\bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i},

\]

where

\begin{eqnarray\*}

\A\_i &=& \mL\_{zi}\mL\_{zi}',\\

\bm{q}\_i &=& \mL\_{zi} \bm{y},

\end{eqnarray\*}

where $\mL\_{xi} = \C\_z \Z\_1' \Q\_i$, $\A\_i$ denotes the block information matrix and $\bm{q}\_i$ is vector of the adjusted Phase 1 block effects in stratum $i$. The residual SS in the stratum $i$ is then derived by subtraction using the same equation in \ref{eq:resSS}, giving $\bm{q}\_{i}'(\I- \A\_{i}^{-}) \bm{q}\_{i}$.

The matrix $\mL\_{zi}' \A\_{i}^{-}\mL\_{zi}$ in $\bm{q}\_{i}' \A\_{i}^{-} \bm{q}\_{i}$ is re-expressed as $\Q\_{j(i)}$ which denotes the orthogonal projector of stratum $j$ of the Phase 1 experiment in stratum $i$ of the Phase 2 experiment. The orthogonal projector $\Q\_{j(i)}$ is then be used to compute the treatment SS in the following subsection.

**\subsection{Computing the treatment SS}**

**\label{subsec:trtDecomp}**

The next step is the information decomposition with respect to the treatment structure. Since the vector of treatment parameters, $\bm{\alpha}$ consists of the treatment parameters from both Phase 1 and Phase 2 experiments, the information decomposition procedure is identical to the single phase experiment as described earlier.

If the decomposition is in the stratum $j$ from the Phase 1 experiment within the stratum $i$ from the Phase 2 experiment, the orthogonal projector is expressed as $\Q\_{j(i)}$. The treatment SS can then be shown as

\[

\bm{q}\_{j(i)}' \A\_{j(i)}^{-} \bm{q}\_{j(i)},

\]

where

\begin{eqnarray\*}

\A\_{j(i)} &=& \mL\_{xj(i)}\mL\_{xj(i)}',\\

\bm{q}\_{j(i)} &=& \mL\_{xj(i)} \bm{y},

\end{eqnarray\*}

where $\mL\_{xj(i)} = \C\_x \X' \Q\_{j(i)}$, $\A\_{j(i)}$ denotes the information matrix and $\bm{q}\_{j(i)}$ is vector of the adjusted treatment totals.

The residual SS can then be computed by subtraction of the SS of the stratum $j$ ib the Phase 1 experiment within stratum $i$ in the Phase 2 experiment, denoted by $\bm{q}\_{j(i)}'\bm{q}\_{j(i)} = \bm{y}' \Q\_{j(i)} \bm{y}$, by the treatment SS, resulting

\[

\bm{q}\_{j(i)}' (\I - \A\_{j(i)}^{-}) \bm{q}\_{j(i)}.

\]

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

In summary, the information decomposition for the two-phase experiment is described in this section. The main difference to the single-phase experiment is an additional decomposition procedure of Phase 1 block factors to the Phase 2 factors. Furthermore, since the block structure can be non-orthogonal, this additional decomposition procedure has to performed with the same way as decomposing the treatment factors to block factors of the single-phase experiment. Thus, the vector of block parameters and matrix of block design for the Phase 1 experiment are constructed with the same way as treatment factors of the single-phase experiment.

**\section{Two-phase proteomics experiment}**

**\label{sec:exampleTwoPase}**

A \emph{quantitative proteomics experiment} is aiming to identify proteins that are differentially abundant between animals under different conditions of interest. This experiment can be treated as a two-phase experiment, where the animals are randomly perturbed to the conditions of interest in the Phase 1 experiment. Since the abundance of proteins cannot be measured directly from the animals, the Phase 2 experiment uses some biotechnologies to measure the abundances of proteins in the samples extracted from the organisms in the Phase 1 experiment.

One method to measure the protein abundances is \emph{Multi-dimensional Protein Identification Technology} (MudPIT), which is the series of instruments used to separate a protein mixture, based on their chemical properties, before measuring the abundance of each protein specie \citep{Florens2006a}. These series of separations allow a more accurate protein abundance measurement. However, the comparison of protein abundances between samples is difficult due to large variabilities 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 experiments \citep{Ross2004, Choe2007}. This article refers to each MudPIT experiment as a run and iTRAQ$^{\rm TM}$ tag as tag. The statistical aim is to find the best way to allocate the samples generated from the Phase 1 experiment to which runs and tags of Phase 2 experiment.

For the block and treatment structures of a most trivial two-phase proteomics experiment, the Phase 1 block structures is composed of Between Animals and Between Samples Within Animals. The Phase 2 block structure consists of Between and Within Runs. The treatment structure has the disease status and tag without the interaction, because the disease status and tag effects are assumed to not interact.

This section presents an example of two-phase proteomics experiment where the Phase 1 experiment consists of eight animals randomly assigned to either the healthy or diseased groups and the Phase 2 experiment consists of four runs and fours tags. Both Phase 1 and Phase 2 experimental designs and their theoretical ANOVA tables are described.

**\subsection{Phase 1 experiment}**

**\label{subsec:Phase1Exp}**

The experimental objective of the Phase 1 experiment is to compare treatment groups in the absence of nuisance sources of variation \citep{Mead1988}. Thus, the Phase 1 experiment is arranged in a CRD. As there are eight animals in this experiment, four animals can be randomly assigned to either the healthy or diseased group as shown in Table~\ref{tab:phase1Design}. The four animals in each treatment groups are used to assess the biological variation and known as \emph{biological replicates}. Note that Phase 2 experiment uses four runs and four tags; hence, a total of 16 samples are processed. Given there are eight animals from the Phase 1 experiment, each animal is further subdivided to two identical samples which is used for assessing the measurement error of the Phase 2 experiment, these identical samples are also known as \emph{technical replicates}.

\begin{table}[ht]

\centering

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

\begin{tabular}[t]{ccccc}

\hline

Healthy & A & C & E & G \\

Diseased & B & D & F & H \\

\hline

\end{tabular}

**\label{tab:phase1Design}**

\end{table}

Let $y\_{ij(k)}$ 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(k)}= \mu + A\_{i} + \tau\_{j} \; (+ \epsilon\_{ijk} ),

\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. The $\epsilon\_{ijk}$ denotes the dummy variable associated with the technical replicate from sample $k$ of rat $i$ under disease status $j$, ($k = 1, 2$). Despite the technical replicates is used to assessing the measurement error of the Phase 2 experiment, it is considered in the model of the Phase 1 experiment, because the combination of indexes from the Phase 1 experiment must be uniquely corresponded to the indexes of the Phase 2 experiment. For example, the first sample of rat A under the healthy group can be assigned to run 1 and tag 114. However, the term $\epsilon\_{ijk}$ does not contribute to the analysis, because the technical replicate is used to estimate the measurement error in the Phase 2 experiment; hence, there are treated as the dummy index and variable with parentheses as shown in the model (\ref{eq:phase1Model}).

Since the treatments are applied on the animals and animals are the smallest units on which a responses is measured, the animals are both the experimental and observational units \citep{Bailey2008}. Hence, the animals to animals variation is confounded with the measurement error. Furthermore, the animal effects are assumed to be normally distributed with mean zero and variance of $\sigma\_{A}^2$.

The theoretical ANOVA table for the Phase 1 experiment is shown in Table~\ref{tab:Phase1ANOVA}. The last column of this theoretical 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, $\theta\_{\tau}$. Given that DF$\_{\tau}$ denote the DF of disease status, $\theta\_{\tau}$ is defined as

\[\frac{\displaystyle \sum\_{j = healthy, diseased}(\tau\_{j} - \overline{\tau\_{.}})^2}{\mathrm{DF}\_{\tau}}=

\frac{(\tau\_{healthy} - \overline{\tau\_{.}})^2 + (\tau\_{diseased} - \overline{\tau\_{.}})^2}{2-1},\]

which is the variances between the treatment means. The coefficients of treatment components can be derived from the replication numbers of their corresponding treatment factors.

If the samples are considered, then there are total of 16 observations in the Phase 1 experiment. Since the differences between the samples does not contribute to the analysis, the total number of observation for the Phase 1 experiment is then 8, i.e.\ the number of animals. Hence, the DF associated with the adjusted total of Phase 1 experiment is 7. Since the animals are both experimental and observational units, the Between Animals stratum captured the overall analysis. Furthermore, the Between Animals stratum is further decomposed to 1 and 6 DF for the effects of disease status and residual, respectively. A valid test for disease status effects is present for this Phase 1 experiment, since the coefficients of the variance components of disease status and Residual EMS are identical. Hence, providing that the experimental data can be obtained, the F-ratio can be calculated from the 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}

\toprule

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

\midrule

Between Animals \\

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

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

Total & $7$ & \\

\bottomrule

\end{tabular}

**\label{tab:Phase1ANOVA}**

\end{table}

**\subsection{Phase 2 experiment}**

**\label{subsec:Phase2Exp}**

In practice, the protein abundances cannot be measured directly from the animals; so the Table~\ref{tab:Phase1ANOVA} cannot be used for conducting the test of treatment effects. Hence, it is necessary to harvest the tissues and extract the proteins from the animals for the subsequent Phase 2 MudPIT-iTRAQ$^{\rm TM}$ experiment. Hence, two additional components are introduced, i.e.\ the variation between runs, $\sigma\_{R}^2$, and the difference between tags, $\theta\_{\gamma}$. In addition, the variation between the samples, $\sigma^2$, is present for assessing the experimental error.

This example uses four MudPIT runs with a four-plex iTRAQ$^{\rm TM}$ labelling system to measure the protein abundances; hence, a total of 16 samples are processed. 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, the aim of the assignment 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 $y\_{ijk}^{lm}$ denote the abundances of a given protein in sample $k$ of rat $i$ under disease status $j$ and measured from the $l$th MudPIT run with iTRAQ$^{\rm TM}$ tag $m$. The superscript and subscript indexes of response, $y\_{ijk}^{lm}$, correspond to the indexes from the Phase 2 experiment and Phase 1 experiment, respectively. The linear model of the Phase 2 experiment is then given by

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

y\_{ijk}^{lm}= \mu + R\_{l} + A\_{i}+ \tau\_{j} + \gamma\_{m} + \epsilon\_{ijk}^{lm},

\end{equation}

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

Using the information decomposition method described in Section~\ref{sec:infoiDecompTwoPase}, the theoretical ANOVA table of the Phase 2 experiment can be generated and shown in Table~\ref{tab:Phase2ANOVA}. Since there are total of 16 observations, the total number of DF is 15 DF. These 15 DF are decomposed to 3 and 12 DF for Between and Within Runs strata, respectively. The Between Runs stratum is further decomposed to the effects of 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 the effects of disease status (1 DF), tag (1 DF) and residual (4 DF). Note that all of the disease status information stay intact in the Between Animals Within Runs stratum. Finally, the Between Sample Within Animals Within runs is decomposed to the effects of tag (2 DF) and residual (4 DF).

Comparing this ANOVA table (Table~\ref{tab:Phase2ANOVA}) to the ANOVA table of Phase 1 experiment (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 confounded with the tag effects. A valid test for the disease status effect can still be conducted in the Between Animals Within Runs stratum, but the DF of the residual is reduced from 6 to 4, which reduces the statistical power in detecting for a significant differences between the treatment groups. However, this is out of the cope of this paper.

\begin{table}[ht]

\centering

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

\begin{tabular}[t]{lrl}

\toprule

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

\midrule

Between Runs \\

\quad Between Animals & $1$ & $\sigma^2 + 2\sigma\_{A}^2 + 4\sigma\_{R}^2$\\

\quad Residual & $2$ & $\sigma^2 + 4\sigma\_{R}^2$\\

\hline

Within runs \\

\quad Between Animals \\

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

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

\quad\quad Residual & $4$ & $\sigma^2 + 2\sigma\_{A}^2$\\\hline

\quad Between Samples Within Animals &\\

\quad\quad Tag & $2$ & $\sigma^2 + 4\theta\_{\gamma}$\\

\quad\quad Residual & $4$ & $\sigma^2$\\

\hline

Total & $15$ & \\

\bottomrule

\end{tabular}

**\label{tab:Phase2ANOVA}**

\end{table}

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

\pkg{InfoDecompuTE} is written in the well known \proglang{R} programming language. This package automatically performs the decomposition procedure, described in Section~\ref{sec:exampleTwoPase}, and produce the theoretical ANOVA table similar to Table~\ref{tab:Phase2ANOVA}. The main purpose of this package is to allow the researchers to study the properties of the theoretical ANOVA of various experimental designs rapidly. This package is made up of series of functions used to construct the theoretical ANOVA tables. This article focuses on two main functions: \code{summaryAovOnePhase} and \code{summaryAovTwoPhase} for single and two-phase experiments, respectively.

This section first explains the installation procedure of this package and then the arguments that are needed for the two functions: \code{summaryAovOnePhase} and \code{summaryAovTwoPhase}.

**\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{R2012}. The system requirements for this package depends on the number of factors and observations in the experimental design that the user plans to analyse. This is because the number of factors and observations affects the dimensions of matrices for computation. \citeauthor{Brien1999}'s \citeyearpar{Brien1999} two-phase experiment can be analysed in under a minute using two-gigabytes of RAM with a Duo Core 3GHz machine running Microsoft Windows 7 (Section~\ref{sec:example}).

Given that the user has an internet connection, \pkg{infoDecompuTE} can be installed and initiated by typing the following two commands in a 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{Functions}**

This section explains the arguments for the two main functions in the \pkg{infoDecompuTE} package, these two functions and their arguments are:

\begin{CodeChunk}

\begin{CodeInput}

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

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

table.legend = FALSE, response = NA, latex = FALSE,

fixed.names = NA)

summaryAovTwoPhase(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,

response = NA, latex = FALSE, fixed.names = NA)

\end{CodeInput}

\end{CodeChunk}

The rest of this section exlains each of these arguments using the two-phase experiment example described in Section~\ref{sec:exampleTwoPase}.

**\subsection{Design data frame}**

The first argument, \code{design.df}, consists of the experimental design in a data frame format. The class of each vector in the data frame should be factor. The single-phase and two-phase experimental design in Table~\ref{tab:phase1Design} and \ref{tab:phase2Design} can be shown in two data frames: \code{design1} and \code{design2}. The contents of these two data frames are

\begin{CodeChunk}

\end{CodeChunk}

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

**\subsection{Block and treatment structures}**

The arguments \code{blk.str} and \code{trt.str} allows the user to input the block and treatment structures, respectively. The \citeauthor{Wilkinson1973}' syntax is used to present these two structures, this presentation is also known as \emph{structure formula} \citep{Wilkinson1973}. The user can also refer to the \code{formula} function in \proglang{R} for further information on the structure formula. (This will be in the intro of my thesis.)

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. For the first phase experiment (Section~\ref{subsec:Phase1Exp}), the structure formula for the block factors is \code{Ani/Sam} which denotes the two technical replicated samples nested from the animals.

The structure formula for the treatment factor contains a single term for the disease status, \code{Trt}. The output from \code{summaryAovOnePhase} is

\begin{CodeChunk}

\end{CodeChunk}

The output is a list of two tables. The first table, denoted by \code{ANOVA}, is the random effects table which gives the DF with the coefficients of the variance components for each source of variation. The \code{Between Ani} denotes Between Animals stratum. The second table, denoted by \code{EF}, is the fixed effects table provides the coefficients of the fixed components with the average efficiency factors. The efficiency factor is denoted by \code{eff.Trt}. This example indicates all the disease status information is in the Between Animals stratum.

In order to analyse two-phase experiments, the function \code{summaryAovTwoPhase} is used. Recall the two-phase experiment from Section~\ref{subsec:Phase2Exp}, the Phase 1 block structure consists of animals and Phase 2 block structure is MudPIT run. The Phase 1 and 2 block structures are represented by the arguments \code{blk.str1} and \code{blk.str2}, and their structure formulae can be written as \code{Ani} and \code{Run}, respectively.

The treatment structure is defined in the argument \code{trt.str}. Since the diseases status and tag do not interact, the structure formula can be written as \code{Tag + Trt}. The tag effects should always be fitted to the ANOVA model before the disease status effects or any other treatment effects of interest. This is because if there is confounding between the tag effects and diseases status effects and the diseases status is fitted before the tags, then some of the tag information will be present in the diseases status MS and test for the disease status effects will not be as accurate.

The output from the \code{summaryAovTwoPhase} of the two-phase experiment from Section~\ref{subsec:Phase2Exp} is

\begin{CodeChunk}

\end{CodeChunk}

The structure of the output is identical to the output from \code{summaryAovOnePhase} and is similar to the ANOVA Table~\ref{tab:Phase2ANOVA}. The second column name of the ANOVA table, \code{e}, denote the variance component of experimental error. Note the \code{Within Run} and \code{Within Ani} denote the Within Runs and Within Animals strata, respectively.

A more informative way to construct the block structure of the Phase 1 experiment for the current example is to include the sample factor. The Phase 1 block structure is then samples nested from the animals, i.e.\ \code{Ani/Sam} and the output from the \code{summaryAovTwoPhase} becomes

\begin{CodeChunk}

\end{CodeChunk}

The name in the source of variation and variance component for \code{Ani:Sam} denotes samples within animals. Since samples are the observational units, the samples within animals variance component is identical to the variance component for the experimental error. Thus, the column of the ANOVA table with \code{e} is omitted.

**\subsection{Crossed or nested}**

Typically, we cannot distinguish the output of \code{terms} function that whether the relationship of two block or treatment factors is nested or crossed. For example, if the sample is assumed to be nested from the animals, then the output from the \code{terms} function gives

\begin{CodeChunk}

\begin{CodeInput}

> attr(terms(~Ani / Sam), "term.labels")

\end{CodeInput}

\begin{CodeOutput}

[1] "Ani" "Ani:Sam"

\end{CodeOutput}

\end{CodeChunk}

where \code{Ani} denotes the effect of between different animals and \code{Ani:Sam} denotes the effect of between different samples within animals. If the sample is assumed to be crossed from the animals, then the output from the \code{terms} function gives

\begin{CodeChunk}

\begin{CodeInput}

> attr(terms(~Ani \* Sam), "term.labels")

\end{CodeInput}

\begin{CodeOutput}

[1] "Ani" "Sam" "Ani:Sam"

\end{CodeOutput}

\end{CodeChunk}

where \code{Ani} and \code{Sam} denotes the effects of between different animals and samples, respectively. However, \code{Ani:Sam}, which is identical from the previous output, denotes the interaction effect between the animals and samples. Thus, to distinguish between the nesting and crossing relationship, the interaction effect between the animals and samples is denoted by \code{Ani\*Sam}.

If the animals and samples are assumed to interact, the Phase 1 block structure become \code{Ani\*Sam}. The output of \code{summaryAovTwoPhase} can be expressed as

\begin{CodeChunk}

\end{CodeChunk}

Notices the name of sources of variation and the variance component for the interaction effects of animals and sample is \code{Ani\*Sam}. This procedure is also applied on the treatment effects.

**\subsection{Artificial strata}**

The \code{var.comp} argument allows the researchers to use the artificial strata to facilitate decomposition. For the example in Section~\ref{subsec:Phase2Exp}, four of eight animals can be grouped as an animal set. Animals in first and second sets are denoted by 1 and 2, respectively. This new vector of animal set, denoted by \code{AniSet}, can be generated as

\begin{CodeChunk}

\end{CodeChunk}

Phase 2 block structure is then written as \code{AniSet/Run}, which means MudPIT runs are now nested within the animal sets. Since the purpose of having animal set is to create the artificial strata, there should not be any variation between the animal sets and the variance components associated with the animal set should be zero. To remove the animal set from the variance component estimate, the \code{var.comp} argument is used which contains a vector of characters indicating which variance components should appear in the output table of the random effects. In addition, the user can also use this argument to define the order of the variance components to appear in the output table, i.e.,

\begin{CodeChunk}

\end{CodeChunk}

Note the \code{Within AniSet.Run} denotes Within Runs and Animal Sets where these two factors are connected by a dot.

**\subsection{Manually defined contrasts}**

The contrasts for the Phase 1 block and treatment factors can also be specified using the argument \code{blk.contr} and \code{trt.contr}, respectively. The contrasts are made up of lists of numeric vectors. The order of factors 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 illustrates the treatment contrasts, where four iTRAQ$^{\rm TM}$ tags can be represented by three orthogonal 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

\begin{CodeChunk}

\end{CodeChunk}

Note that it is essential to assign names, i.e. \code{Tag1}, \code{Tag2} and \code{Tag3}, for each of the contrast vector for identification in the output table.

Using the design described in Section~\ref{subsec:Phase2Exp}, the treatment contrasts can be input into the function \code{summaryAovTwoPhase}. The argument \code{contr.matrix} is set to \code{TRUE}, as it indicates that we are working with contrast matrices. The output is shown as follows

\begin{CodeChunk}

\end{CodeChunk}

In addition, it is possible to break down the treatment factors into multiple orthogonal contrasts to study how these contrasts contribute to each source of variation. This is achieved by setting 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{summaryAovTwoPhase}, and the output is shown as follows,

\begin{CodeChunk}

\end{CodeChunk}

Note that, having broken down the tag contrasts, the random effects table shows that 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} allows the users to use letters to represent column names, and then insert a legend at the bottom of the table. In this example, \code{summaryAovTwoPhase} is set to \code{TRUE}, because once the treatment contrasts are fitted separately, the number of columns in the table of fixed components increases making it difficult to read.

\**subsection{Mean squares computation}**

The argument \code{response} allows the computation of the MS from the experimental data. Using the example of Section~\ref{subsec:Phase2Exp}, \code{response} argument can be input with 16 random numbers corresponding to the 16 observations are drawn from the normal distribution, i.e.\ \code{rnorm(16)}. The output is shown as,

\begin{CodeChunk}

\end{CodeChunk}

Since the fixed effects table is identical to the table without inserting the experimental data, the fixed effects table is omitted from the output shown. The random effects table, however, has an extra column for the mean squares.

**\subsection{Latex output}**

The output from \proglang{R} is not always easy to read on the screen. The argument \code{latex} allows the user to transform the \proglang{R} output into {\LaTeX} script. Using the example of Section~\ref{subsec:Phase2Exp}, the \code{latex} of the \code{summaryAovTwoPhase} function is then set to \code{TRUE}, i.e.\

\begin{CodeChunk}

\end{CodeChunk}

The output from the compiling the {\LaTeX} script is shown in Table~\ref{tab:ouputFromR}. Two {\LaTeX} packages \code{bm} and \code{booktabs} are required to compile the {\LaTeX} script.

When the {\LaTeX} script is to be produced, the Greek letters are required to define the fixed effects. Since different experiments can have different set of Greek letters, the users can choose their own set of Greek letters using the argument \code{fixed.names}. Note that this table does not only contain the DF and EMS, it also has the average efficiency factors for all the treatment effects in the last two columns. However, some further modifications may still required to this table, such as adjusting the names in the source of variation, e.g.\ from "Between Ani" to "Between Animals". However, this additional functionality allows a quick way to generate theoretical ANOVA table from the {\LaTeX} scripts.

\begin{table}[ht]

\end{table}

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

This section show how the function \code{summaryAovTwoPhase} is used to generate the table of EMS for the viticultural experiment described by \cite{Brien1999}. \citeauthor{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.

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. The two-phase experiment yields 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~(\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-by-2 factorial experiment, thus, Trellising and pruning Methods are crossed.

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

\begin{CodeChunk}

\begin{CodeInput}

> summaryAovTwoPhase(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}

Note: Complete confounding between Squ and Oc!

\end{CodeChunk}

Three additional features, which were not discussed in Section~\ref{sec:exampleTwoPase}, can be observed from this output. Since the Occasion factor of the Phase 2 experiment is completely confounded with Square factor of the Phase 1 experiment, the function will first output, \code{Note: Complete confounding between Squ and Oc!}, then the \code{Squ} is changed to \code{SquCCW}. The \code{CCW} stands for "completely confounded with". Furthermore, since \code{:} and \code{\*} are used to denote the nesting and crossing between the treatment or block factors in the names in sources of variation and variance component, the parentheses in the structure formula are integrated into these names to further clarify the relationship between the treatment or block factors. Lastly, the lowermost stratum or the Within stratum always denote by all block factors, from either Phase 1 or Phase 2 experiment, concatenated with dots.

\**section[Conclusion]{Conclusion}**

\pkg{InfoDecompuTE}, a freely available \proglang{R} package, allows researchers to study any complex single or two-phase experimental design by generating the structure of the theoretical ANOVA table with the coefficients of variance components of the EMS as shown in Section~\ref{sec:example}. This package not only allow the researchers to check for valid statistical test can be conducted, it also allows them to study how the raw data is decomposed across different strata and different sources of variation.

This package can also analyse designs with non-orthogonal block or treatment structure and produce the average efficiency factors for each fixed effect as shown in Section~\ref{sec:example}. Hence, the researcher will know the amount of the treatment information remains when conducting the test for the treatment effects. In addition, the users can also fit the each block or treatment contrast separately allowing more flexibility in the analysis and really dissect how the block or treatment information is split across different strata.

However, this package still 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 as described by \cite{Wilkinson1973} 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 ability for constructing better experiments.