Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

optimTE: an R package for searching the optimal design for two-phase experiment

Kevin C. Chang
University of Auckland,
New Zealand

Abstract

Two-phase experiments are used when the responses of experimental units to treatments cannot be measured directly in a single experiment. Subsequent processing (Phase 2) of the initial (Phase 1) experiment is necessary in order for the measurements to be made. In the case of proteomics experiments, the Phase 1 experiment involves the organisms that are to be perturbed by the experimental conditions of interest. Since the abundance of proteins cannot be measured directly from the organisms, the Phase 2 experiment uses multiplexing techniques such as iTRAQ peptide labelling, coupled with liquid chromatography-mass spectrometry (LC-MS), to measure the abundance of proteins in samples extracted from the organisms in the Phase 1 experiment.

For a given set of design parameters, there are often many ways to allocate the samples collected from the Phase 1 experiment to the iTRAQ labels of the Phase 2 experiment. The objective function can identify the best allocation in terms of allowing a formal test of treatment effects to be conducted with the highest average efficiency factor. The objective function is optimised using a simulated annealing algorithm (SA). This article defines the objective functions for finding the optimal two-phase design, where the Phase 1 experiment is arranged either in completely randomised design or randomised block design and the Phase 2 experiment is in a randomised block design. In addition, an improved version of the SA which is shown to be more efficient than the standard SA is presented.

Keywords: optimal design, two-phase experiment.

1. Introduction

This chapter discusses searching for optimal designs for two-phase proteomics experiments, where the Phase 1 experiment is arranged in a completely randomised design (CRD) with multiplexing technique such as iTRAQ in the Phase 2 experiment. The first phase (Phase 1) experiment involves the organisms that are to be perturbed by the experimental conditions of interest. Since the protein abundances cannot be measured directly from the organisms, the second phase (Phase 2) experiment involves the MudPIT-iTRAQ™ experiments for measuring the protein abundances of the experimental materials from the Phase 1 experiment.

The CRD of Phase 1 experiment consists of v treatments and r_b biological replicates. Hence, vr_b equals the total number of animals, denoted by n_A. To estimate the measurement errors coming from the Phase 2 experiment, r_t technical replicates are used. The Phase 2 experiment is arranged in a randomised block design (RBD) where the numbers of MudPIT runs, n_R, and iTRAQ tags, n_γ, correspond to the numbers of blocks and block size, respectively. The total number of observations, denoted by n, equals to both vr_br_t and n_Rn_γ. The numbers of treatments, biological replicates, technical replicates, runs and tags are collectively known as the *design parameters.*

There are many ways to allocate the samples from the Phase 1 experiment to the Phase 2 experiment. The number of ways that these samples can be assigned to each run is

$$\binom{n}{n_\gamma}.$$

For example, if $n = 10$, then there are 210 ways to assign these samples to a run in a four-plex experiment. An optimal design can be obtained from a specific allocation of these samples to each run, based on various optimality criteria. Some optimality criteria have been described in (**?**). This chapter focuses only on the MS- and A-optimality criteria which are further discussed in Section **??**.

The method of finding the optimal designs for different classes of design such as block, row-column and α_n designs has been previously discussed (**???**). These methods aim to find the designs with the most treatment information in the most bottom stratum. As for finding the optimal two-phase designs, the allocation of the block factors from the Phase 1 experiment to block factors of the Phase 2 experiment also needs to be monitored. This is because if there is confounding between the block effects of Phase 1 and 2 experiments, the design may not have a valid test for the treatment effects. Hence, the goal is not only to find two-phase optimal designs that has the most treatment information in the most bottom stratum, but also to find two-phase optimal designs where a formal test for the treatment effects can be conducted in that stratum. A suitable method of finding a such two-phase optimal designs has yet to be described; the main reason is that the optimality criterion has not been defined.

The methods of generating two-phase optimal designs by optimising the objective function uses the *simulated annealing algorithm* (SA). The optimality criterion is defined in the *objective function*, which is a mathematical expression describing the relationship between the variables. SA is a well-known heuristic method for finding the values of variables that result in a maximum or minimum of an objective function (**?**). In the case of finding optimal designs, the variables of the objective function correspond to the candidate designs. SA plays the role in continuously generating the new candidate design and comparing the values generated from the objective function between the candidate designs. The optimality criterion and the objective function are to be discussed in detail in Section **??**. In addition, this chapter presents an improved version of SA which shown to speed up the search for the optimal designs in Section **??**.

The article can be divided into three main sections: defining the objective function, constructing the initial design and searching the optimal design using SA.

2. Define model

Let y_{ijk} denote the abundance level of a given protein from animal ij with treatment i and j biological replicate with k technical replicates. The linear model of the Phase 1 experiments can be written as

$$y_{ijk} = \mu + \tau_i + A_{ij} \ (+\epsilon_{ijk}), \tag{1}$$

where μ denotes grand mean of all observations, τ_i denotes the fixed effect of treatment i, A_{ij} denotes the random effect from animal ij and ϵ_{ijk} denotes the dummy variable associated with the technical replicate k, $(i = 1, \ldots, v;\ j = 1, \ldots, r_b;\ k = 1, \ldots, r_t)$. Moreover, the animal effects, A_{ij}, are assumed $N(0, \sigma_A^2)$. The chapter focuses the experiments comparing between $v = 2, \ldots, 8$ treatments with $r_b = 2, \ldots, 10$ biological replicates and $r_t = 2$ technical replicates of the Phase 1 experiments.

In practice, the protein abundances cannot be obtained directly from the animals; hence, an additional experiment is required, i.e. Phase 2 experiment. Now let y_{ijk}^{lm} denote the abundance level of a same protein as in (3) from animal ij under treatment i with k technical replicate and is measured in run l with tag m. The superscript and subscript indexes of response, y_{ijk}^{lm}, correspond to the indexes of Phase 2 experiment and Phase 1 experiment, respectively. The linear model of the Phase 2 experiment can be written as

$$y_{ijk}^{lm} = \mu + \tau_i + A_{ij} + \gamma_m + R_l + \epsilon_{ijk}^{lm}, \tag{2}$$

where γ_m denotes the fixed effect of tag m, R_l denotes the run effect, and ϵ_{ijk}^{lm} denotes an experimental error, $(m = 1 \ldots n_\gamma;\ l = 1 \ldots n_R)$. The run effects, R_l, and measurement error, ϵ_{ijk}^{lm}, are assumed $N(0, \sigma_R^2)$ and $N(0, \sigma^2)$, respectively. The Phase 2 experiments consist of $n_\gamma = 4$ or 8 tags and $n_R = n/n_\gamma$ runs, where n denotes the total number of observations.

This article focuses on the two-phase experiments with $v = 2, \ldots, 8$ treatments, $r_b = 2, \ldots, 10$ biological replicates, $n_{A(B)}$ animals within cages, n_B cages, $r_t = 2$ technical replicates, , $n_\gamma = 4, 8$ tags and $n_R = n/n_\gamma$ runs. The animals can be either identified from the combinations of treatment and biological replicates or the combinations of Cage ID and Animal ID within each cage. The combination of Cage ID and Animal ID within each cage is used, because the relationship between the effects of cages and the animals within cages is important in the process of decomposition.

Let y_{ijkl} denote the abundance level of a given protein from animal jk in cage j under treatment i with l technical replicates. The linear model of the first phase experiments can be written as

$$y_{ijkl} = \mu + \tau_i + A_{jk} + B_j (+\epsilon_{ijkl}), \tag{3}$$

where μ denote grand mean of all observations, τ_i denotes the fixed effect of treatment i, A_{jk} denotes the random effect from animal jk and B_j denotes the random effects from cage k $(i = 1 \ldots v; j = 1 \ldots n_B;\ k = 1 \ldots n_{A(B)})$. The ϵ_{ijkl} is again denoting the dummy variable associated with the technical replicate k, $(l = 1, \ldots, r_t)$. The effects of Between Animals Within Cages and Between Cages are assumed to be $N(0, \sigma_A^2)$ and $N(0, \sigma_B^2)$, respectively.

Now let y_{ijkl}^{mn} denote the abundance level of a same protein as in (3) from animal jk in cage j under treatment i with l technical replicates and is measured in run m with tag n. The superscript and subscript indexes of response, y_{ijkl}^{mn}, correspond to the indexes of Phase 2 experiment and Phase 1 experiment, respectively. The linear model of the Phase 2 experiment can be written as

$$y_{ijkl}^{mn} = \mu + \tau_i + A_{jk} + B_j + \gamma_n + R_m + \epsilon_{ijkl}^{mn}, \tag{4}$$

where γ_n denotes the fixed effect of tag l, R_m denotes the run effect, and ϵ_{ijkl}^{mn} denotes an experimental error, $(n = 1 \ldots n_\gamma;\ m = 1 \ldots n_R)$.

3. Define objective function

Different objective functions are required for finding the optimal design with different Phase 1 experimental designs.

Here we present two objective functions for

completely randomised design and randomised block design in the Phase 1 experiment.

The phase 2 experiment is the MudPIT-iTRAQ experiment which is arranged with randomised block design.

3.1. CRD for the Phase 1 experiment

The optimal design can be obtained using the objective function can be written as

$$\frac{3}{4}E_A + \frac{1}{4}\left(\frac{DF_2 + E_\tau}{v}\right). \tag{5}$$

where E_τ and E_A denote the average efficiency factors of treatments and animals, and DF_2 denotes the DF associated with the treatment effects in the Between Animals Within Runs stratum of the Phase 2 experiment.

3.2. RBD for the Phase 1 experiment

The objective function for finding the optimal design where the Phase 1 experiment is arranged in RBD is more complicated. The single objective function is to split into two objective functions. The process is to optimised one objective function after another. The average efficiency factors associated with the animal within cages effect in the Within Runs stratum has shown to be consistent at 100%. In addition, the treatment allocation to runs has to connected, i.e. there should be $v1$ DF associated with treatment effects in the Between Animals Within Cages Within Runs stratum. These two components becomes two checks within the objective function. If either of two checks failed, the objective function will give a very low value, i.e. zero, so there is still a very low chance of the design been accepted for the SA search.

On the other hand, if a design found that satisfy these two checks, then DF associated residual in the Between Animals Within Cages Within Runs stratum is calculated for the first objective function. The SA search is aiming to find the design that passed two criteria while maximising the DF associated residual in the Between Animals Within Cages Within Runs stratum. The optimisation of the first objective function is shown in Pseudocode 1.

Once the design with maximum DF associated residual in the Between Animals Within Cages Within Runs stratum is founded, then this design is used as the initial design for the second objective function. The structure of second objective function also ensure the two checks are satisfied and DF associated residual stays at its maximum while computing the treatment

Pseudocode 1 First objective function

```
Let $D_i$ be design from $i$th iteration of SA
Let $E_A$ be the average efficiency factor associated with the animals within cages effects and compute from $D_i$
Let $DF_1$ be the DF associated with treatment effects from Phase 1 experiments
Let $DF_2$ be the DF associated with treatment effects from Phase 2 experiments and compute from $D_i$
Let $RDF_2$ be the DF associated with residual MS in Between Animals Within Cages Within Runs stratum and compute from $D_i$
Let $Max(RDF_2)$ be the maximum DF associated with residual MS in Between Animals Within Cages Within Runs stratum
Let $D_O$ be optimal design from the search.
if $E_A \neq 1$ then
    return 0
end if
if $DF_2 \neq DF_1$ then
    return 0
end if
Find a $D_O$ where $RDF_2$ is maximised
return ($D_O$, $Max(RDF_2)$)
```

average efficiency factor. The SA search is aiming to find the design that passed two criteria with DF associated residual at its maximum while maximising the treatment average efficiency factor in the Between Animals Within Cages Within Runs stratum. The optimisation of the second objective function is shown in Pseudocode 2.

The advantage of having two objective functions is allowing in optimising all four components at the same time, but without deciding which component is more important than the other components. One major drawback of this method is that it requires to perform SA twice for optimising two objective functions. Hence, this method can be slower than using a single objective function.

4. Construct the initial design

The construction of the initial design for the CRD and RBD are slightly different.

the initial design where cage is confounded more with run or tag has shown can affect in the residual DF of the ANOVA tables. The most noticeable differences are when using the initial design where cage is confounded more with tag is that the DF associated with the tag effects can be pushed from the Between Animals Within Cages Within Runs stratum into the Between Cages Within Runs stratum. Thus, the DF associated with residual MS in the Between Animals Within Cages Within Runs stratum will then increase.

Since the number of tags can be used are either four or eight, this situation occurs when the cage number is even to have some tag contrasts completely confounded with cage. Based on the first case, if the treatment number is odd and cage number is even, the initial allocation of cages should be confounded more with tag. For example, if the design parameter consists of $v = 3, r_b = 4, n_B = 4, n_{A(B)} = 3, r_t = 2, n_R = 6$ and $n_\gamma = 4$, the initial design where

Pseudocode 2 Second objective function.

```
Let $D_i$ be design from $i$th iteration of SA
Let $E_A$ be the average efficiency factor associated with the animals within cages effects and compute from $D_i$
Let $E_\tau$ be the treatment average efficiency factor and compute from $D_i$
Let $DF_1$ be the DF associated with treatment effects from Phase 1 experiments
Let $DF_2$ be the DF associated with treatment effects from Phase 2 experiments and compute from $D_i$
Let $RDF_2$ be the DF associated with residual MS in Between Animals Within Cages Within Runs stratum and compute from $D_i$
Let $Max(RDF_2)$ be the maximum DF associated with residual MS in Between Animals Within Cages Within Runs stratum and found from Pseudocode 1
Let $D_O$ be optimal design from the search.
if $E_A \neq 1$ then
    return 0
end if
if $DF_2 \neq DF_1$ then
    return 0
end if
if $Res_2 \neq Max(RDF_2)$ then
    return 0
end if
Find a $D_O$ where $E_\tau$ is maximised.
return $D_O$
```

cage is confounded more with tag in Table 2 is better than the initial design where cage is confounded more with run in Table 3. This is because the initial design in Table 2 comprised of Tag 114 and 115 contain Cage 1 and 2, and Tag 116 and 117 contain Cage 3 and 4. As for the initial design in Table 3, there is no any obvious grouping of cages with respect to runs and tags.

As for the cases where both treatment and cage numbers are even, these are covered by second and third cases. Even though the result from the second case shows using the initial design where the cage is confounded more with run is better; the later section will present the contrary. Thus, the recommendation is to compare the optimal design founded from both initial designs.

5. Simulated annealing algorithm

There are two elements to be discussed while searching for the optimal design using simulated annealing algorithm. These are temperature control and swapping method or generating the new design.

Table 1: Initial animal and cage allocation to runs and tags with $v = 3, r_b = 4, r_t = 2, n_R = 6$ and $n_\gamma = 4$ where cage is confounded more with tag.

	Tag			
Run	114	115	116	117
1	A	B	C	F
2	B	A	F	C
3	G	H	I	J
4	H	G	J	I
5	K	L	M	N
6	L	K	N	M

Table 2: Initial animal and cage allocation to runs and tags with $v = 3, r_b = 4, n_B = 4, n_{A(B)} = 3, r_t = 2, n_R = 6$ and $n_\gamma = 4$ where cage is confounded more with tag.

	Tag			
Run	114	115	116	117
1	1A	1B	3A	3B
2	1B	1A	3B	3A
3	1C	2A	3C	4A
4	2A	1C	4A	3C
5	2B	2C	4B	4C
6	2C	2B	4C	4B

5.1. Temperature control

5.2. Swapping method

Swapping with respect to the same technical replicates

Three-stage swapping for this case

This depends on the design of the second phase experiment ignoring the first phase.

6. Presents the package with examples

The method described in this chapter has implement in R function `optCRD` for finding the optimal design. The arguments of this function are shown as follows,

optCRD(nTrt, bRep, tRep, nPlot, iter)

where the argument `nTrt` denotes the number of treatments, `bRep` denotes the number of biological replicates, `tRep` denotes the number of technical replicates, `nPlot` denoted number of tags and `iter` denotes the number of iterations for each stage within each level. Note that Section **??** stated 30000 iterations should be used for each stage within each level, from the experience, 10000 iterations should be enough to cope all examples mentioned in this chapter.

The R function `optRBD` or `optBIBD` is used to find the optimal design where the Phase 1 experiment is arrange in RBD or BIBD, respectively. The arguments of these two functions are shown as follows,

Table 3: Initial animal and cage allocation to runs and tags with $v = 3, r_b = 4, n_B = 4, n_{A(B)} = 3, r_t = 2, n_R = 6$ and $n_\gamma = 4$ where cage is confounded more with run.

	Tag			
Run	114	115	116	117
1	1A	1B	1C	2A
2	1B	1A	2A	1C
3	2B	2C	3A	3B
4	2C	2B	3B	3A
5	3C	4A	4B	4C
6	4A	3C	4C	4B

```
optRBD(nTrt, bRep, nCag, tRep, nPlot, resDF = NA, confoundCag = FALSE,
upperValue = 1,  iter = 10000)

optBIBD(nTrt, bRep, nCag, tRep, nPlot, resDF = NA, confoundCag = FALSE,
upperValue = 1,  iter = 10000)
```

where the argument `nTrt` denotes the number of treatments, `bRep` denotes the number of biological replicates, `nCag` denotes the number of cages, `tRep` denotes the number of technical replicates, `nPlot` denotes number of tags. If the user already knew the residual DF, they can optimise the second objective function directly by inputting the know residual DF to `resDF`. The default setting is `NA`, as both first at second objective functions are optimised. The argument `confoundCag` allows initial design where the cage to be either confounded more with runs or tag, the default setting of `FALSE` gives the initial design where the cage is confounded more with runs. The argument `upperValue` allows the used to set upper-bound of the treatment average efficiency factor while performing the SA search and `iter` denotes the number of iterations for each stage within each level, 10000 iterations is recommended.

Affiliation:

Kevin C. Chang
Bioinformatics Institute
School of Biological Sciences
The University of Auckland
New Zealand
E-mail: kcha193@aucklanduni.ac.nz
September 6, 2013

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd