optimalAllocation: Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure

Jose Barrera-Gómez[a,b,c] jbarrera@creal.cat
Xavier Basagaña[a,b,c] xbasagana@creal.cat

June 13, 2013

[a]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
[b]IMIM (Hospital del Mar Research Institute), Barcelona, Spain.
[c]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

Contents

1 Introduction

This document is a guide for the R package optimalAllocation usage, which provides the optimal combination of number of participants and number of repeated measurements in observational longitudinal studies such that the power to detect the hypothesized effect is maximized without exceeding a fixed budget, or the cost of the study is minimized while achieving a certain target power. The response variable covariance structure is assumed

damped exponential, DEX(θ, σ, ρ), whose covariance matrix has diagonal elements σ^2 and off–diagonal $[j, j']$ elements, $\sigma^2 \rho^{|\frac{j'-j}{r}|^\theta}$ where ρ is the correlation between the first and the last response measurements and $\theta \in [0, 1]$ is the damping parameter ($\theta = 0$ corresponds to compound symmetry (CS); $\theta = 1$ corresponds to first order autoregressive), σ^2 is the residual variance and ρ is the correlation between the response at the first measurement and at the measurement at the end of follow-up. The exposure is assumed binary and can be time-varying. Two response patterns are considered under the alternative hypothesis, one assuming an acute and transient effect through a constant mean difference (CMD) between exposed and non-exposed, and the other a cumulative effect through a linearly divergent difference (LDD). Missing data due to dropout are allowed, considering a monotone dropout pattern, i.e., that losing one individual measurement implies losing all the subsequent measurements of that individual. No missing data at the first measurement is assumed. The methodology is described in the manuscript [1]. In this guide, the usage of the package is illustrated with some examples, including an application to a study relating cleanning tasks and respiratory health [2]. The optimalAllocation package can be downloaded at `http://www.creal.cat/xbasagana/software.html`.

2 Getting started

We can start the R session loading the package as follows:

```
> library(optimalAllocation)
```

2.1 The function `OA()`

We can get information about the function `OA()`, which performs the study design calculations:

```
> ?OA
```

The input parameters for the function `OA()` are:

- `target`: `"maxPower"` for maximizing the power or `"minCost"` for minimizing the total cost of the study.
- `pattern`: Response pattern under the alternative hypothesis. `"CMD"` assumes an acute and transient exposure effect while `"LDD"` assumes a cumulative exposure effect, i.e., an exposure-time interaction.
- `rMax`: Maximum value of the number of repeated measurements avaluated. Note that the value of `rMax` exponentially increases the computational time.
- `theta`: Damping parameter of the damped exponential covariance structure of the response. The structure is compound symmetry if `theta` $= 0$ and first order autoregressive if `theta` $= 1$.

- `rho`: Correlation between the response at the first measurement and at the measurement at the end of follow-up.

- `sigma2`: The response residual variance, σ^2.

- `rhoe`: Intraclass correlation of the exposure:

$$\rho_e = \frac{\text{sum}(\mathbf{\Sigma}_E) - \text{Tr}(\mathbf{\Sigma}_E)}{r\text{Tr}(\mathbf{\Sigma}_E)},$$

  where sum() and Tr() denote the sum of the elements and the trace of a matrix respectively [3]. ρ_e can be interpreted as a measure of within-participant variation of exposure. When ρ_e takes its maximum value of one, each of the participants are either exposed or non exposed for the entire follow-up (i.e., the exposure is time-invariant). Conversely, when ρ_e takes its minimum value, the within-participant variation of exposure is greatest [3]. The upper bound of ρ_e is lower than 1 when the exposure prevalence is time-varying [4] and 1 otherwise. For binary variables, as here, the lower bound of ρ_e is

$$-\frac{1}{r} + \frac{\text{frac}((r+1)\bar{p}_e)\,[1 - \text{frac}((r+1)\bar{p}_e)]}{r(r+1)\bar{p}_e(1-\bar{p}_e)}$$

  where frac(x) denotes the fractional (non-integer) part of x and $\bar{p}_e$ is the mean exposure prevalence [6]. When the exposure covariance structure, $\mathbf{\Sigma}_E$, is CS and the exposure prevalence is constant, ρ_e becomes the common off-diagonal term of the exposure correlation matrix. As a tool for deciding an appropriate value for ρ_e at the study design phase, it can be useful to explore the distribution of the number of exposed periods per participant, once the values of $\rho_e, \bar{p}_e$, and r have been fixed and $\mathbf{\Sigma}_E$ is assumed to follow CS. For this purpose, the package includes the functions `plotExposedPeriods()` and its interactive version, `plotExposedPeriodsInt()` (Section 2.2).

- `pe0`: Exposure prevalence at the first measurement.

- `per`: Exposure prevalence at the end of follow-up. Exposure prevalence can linearly vary from `pe0` to `per`. If `per` is equal to `pe0`, the exposure prevalence is assumed to be constant.

- `piM`: Fraction of individuals lost at the end of the study. The dropout pattern is assumed to be monotone, i.e. losing one individual measurement implies losing all the subsequent measurements of that individual. No missing data at the first measurement is assumed.

- `kappa`: Ratio between the cost of the first measurement (including recruitment) and each of the subsequent ones.

- `budget`: Total budget for the study if `target = "maxPower"`.

- `c1`: Cost of the first measurement (including recruitment).
- `reqPower`: Required power if `target = "minCost"`.
- `beta`: Expected effect under the alternative hypothesis. If `pattern = "CMD"`, `beta` can be interpreted as the expected difference in the mean of the response variable, at any time point, between exposed and non-exposed. If the `pattern = "LDD"`, `beta` can be interpreted as the expected difference in the mean of the response variable between the worst exposure pattern (i.e., exposed at all measurements) and non exposed, at the end of follow-up.
- `alpha`: Significance level.

2.2 The functions `plotExposedPeriods()` and `plotExposedPeriodsInt()`

The function `plotExposedPeriods()` plots the distribution of the number of exposed periods per participant, once the values of the exposure intraclass correlation, ρ_e, the constant exposure prevalence, p_e, and the number of repeated measurements, r, have been fixed and the exposure covariance structure is assumed to follow CS [5]. This is equivalent to the distribution of the sum of $r+1$ non-independent Bernoulli(p_e) variables with correlation ρ_e for each possible pair of them. The plot obtained is a tool to decide the exposure intraclass correlation value.

We can get information about the function `plotExposedPeriods()`:

```
> ?plotExposedPeriods
```

The input parameters for the function `plotExposedPeriods()` are:

- `r`: Number of repeated measurements, i.e., the total number of measurements is `r` + 1.
- `pe`: Exposure prevalence, assumed constant.
- `rhoe`: Exposure intraclass correlation, defined in Section 2.1.
- `eps`: Precision in the results (relative error). Default value is 0.001.
- `maxIter`: Maximum number of iterations for the computation of the distribution. Default value is 1000.

For example, fixing the number of repeated measurements at 3 and assuming a constant exposure prevalence of 0.2, we can explore the distribution of the number of exposed periods for several values of the exposure intraclass correlation with the code:

```
> rhoes <- c(-0.2, 0, 0.5, 0.9)
> par(las = 1, mfrow = c(2, 2))
> for (i in 1:4)
+  plotExposedPeriods(r = 3, pe = 0.2, rhoe = rhoes[i])
```

which provides the Figure 1. Thus, the value of ρ_e can be fixed at a value that provides a reasonable distribution for the number of exposed periods per participant in the study population.

The function `plotExposedPeriodsInt()` is an interactive version of the function `plotExposedPeriods()` which dinamically updates the distribution of the number of exposed periods when the user changes the value of ρ_e using a scroll bar.

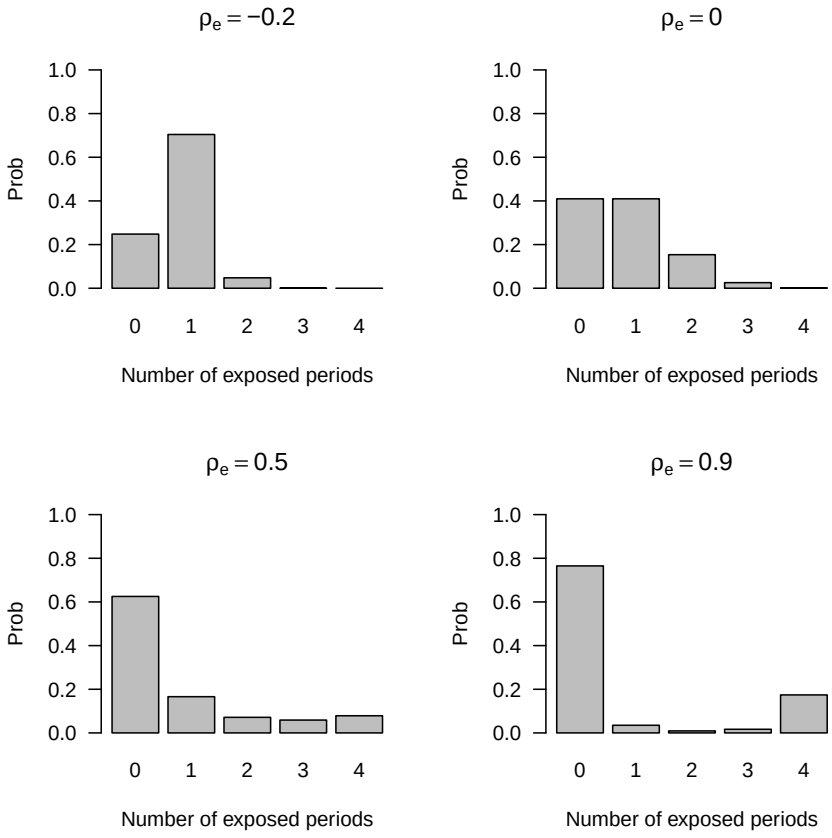

Figure 1: Distribution of the number of exposed periods per participant for several values of the exposure intraclass correlation. The number of repeated measurements was fixed at 3 and a constant exposure prevalence of 0.2 was assumed. The exposure covariance structure was assumed to follow CS.

3 Longitudinal study design examples

3.1 Study 1. Maximizing power

Suppose we are interested in maximizing the power of a longitudinal study assuming the CMD response pattern without exceeding a budget of 40 monetary units, where the monetary unit is the cost of the first measurement. The cost of the first measurement is $\kappa = 3$ times the cost of the subsequent ones. The response covariance structure is DEX($\sigma = 1, \rho = 0.7, \theta = 0.5$). The exposure intraclass correlation is $\rho_e = 0.2$. The expected proportion of dropout at the end of the study is $\pi_M = 0.2$. The exposure prevalence is assumed to increase linearly from $p_{e0} = 0.2$ at the first measurement to $p_{er} = 0.3$ at the last measurement. The effect size to be detected is $\beta = -0.3$ and the significance level is fixed at $\alpha = 0.05$. The maximum number of repeated measurements allowed is $r_{\max} = 20$. Thus, we can perform the study calculations and store the results in the object `study1`:

```
> study1 <- OA(target = "maxPower", pattern = "CMD", rMax = 20,
+              theta = 0.5, rho = 0.7, sigma2 = 1, rhoe = 0.2, pe0 = 0.2,
+              per = 0.3, piM = 0.2, kappa = 3, budget = 40, c1 = 1,
+              beta = -0.3, alpha = 0.05)
> study1

  Results subject to r not greater than 20:
-----------------------------------------------------
Optimal total number of measurements (r+1): 20
Optimal number of participants (N)        : 6
Maximized power                           : 0.9670238
```

Thus, the optimal is to perform a longitudinal study with $N_{\text{opt}} = 6$ participants and taking $r_{\text{opt}} + 1 = 20$ measurements. The maximized power of such study is 0.97.

A graphical representation can be obtained with the function `plot()`. For instance,

```
> plot(study1)
```

generates Figure 2, which shows that the optimal strategy is to take as many measurements as possible, as well as departures from the optimal design when varying the number of repeated measurements.

Further results, including the estimated standard error of β, can be obtained with the function `summary()`:

```
> summary(study1)

$roptreal
[1] 20

$ropt
```

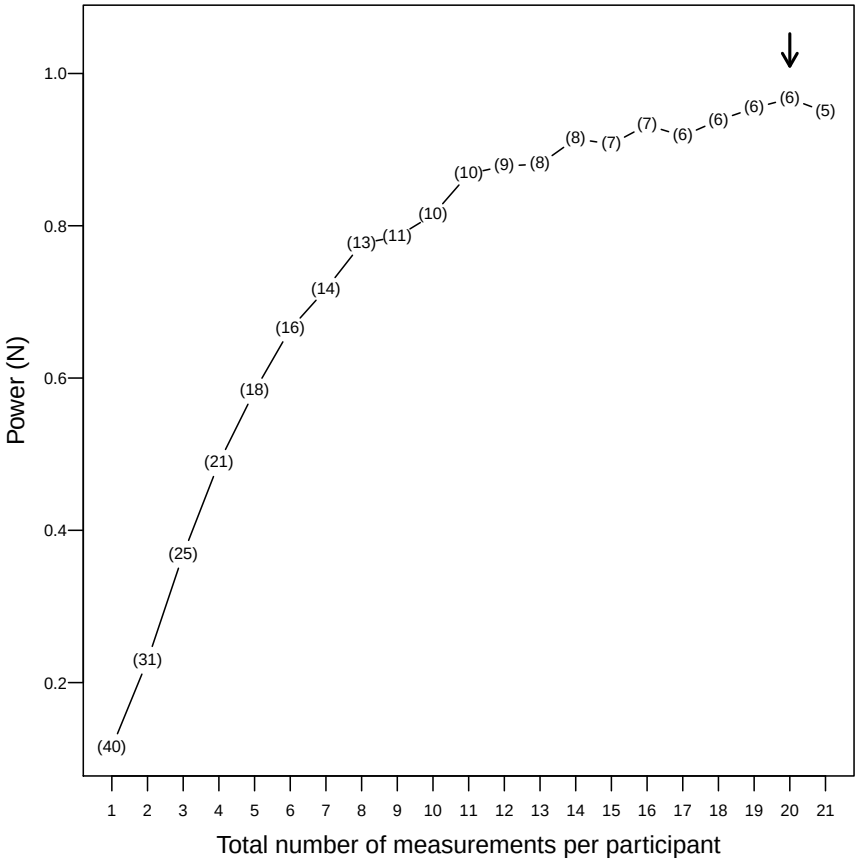

Figure 2: Maximized power and number of participants (in brackets) as a function of the total number of measurements per participant. The arrow points to the optimal allocation.

```
[1] 19

$Nopt
[1] 6

$maxPower
[1] 0.9670238

$cost
[1] 39.85918

$sdBeta
```

```
[1] 0.07897415

$parameters
    target pattern rMax theta rho sigma2 rhoe pe0 per piM kappa budget c1 beta alpha
1 maxPower     CMD   20   0.5 0.7      1  0.2 0.2 0.3 0.2     3     40  1 -0.3  0.05

$f
    r  N     power
1   0 40 0.1148722
2   1 31 0.2291079
3   2 25 0.3686537
4   3 21 0.4887750
5   4 18 0.5829804
6   5 16 0.6643834
7   6 14 0.7166546
8   7 13 0.7766321
9   8 11 0.7857411
10  9 10 0.8143079
11 10 10 0.8682019
12 11  9 0.8784733
13 12  8 0.8809511
14 13  8 0.9142015
15 14  7 0.9072734
16 15  7 0.9322918
17 16  6 0.9178524
18 17  6 0.9384925
19 18  6 0.9546310
20 19  6 0.9670238
21 20  5 0.9495415
```

3.2 Study 2. Minimizing cost

Suppose now we are interested in minimizing the cost of a longitudinal study assuming the LDD response pattern and achieving a power of at least 0.8. The cost of the first measurement is $c_1 = 50$ monetary units, which is $\kappa = 3$ times the cost of the subsequent ones. The response covariance structure is CS($\sigma = 1, \rho = 0.6$). The exposure intraclass correlation is $\rho_e = 0.6$. The expected proportion of dropout at the end of the study is $\pi_M = 0.2$. The exposure prevalence is assumed to be constant and equal to 0.2. The effect size to be detected is $\beta = 0.8$ and the significance level is fixed at $\alpha = 0.05$. The maximum number of repeated measurements allowed is $r_{\max} = 20$. Thus, we can perform the study calculations and store the results in the object `study2`:

```
> study2 <- OA(target = "minCost", pattern = "LDD", rMax = 20,
+              theta = 0, rho = 0.6, sigma2 = 1, rhoe = 0.6, pe0 = 0.2,
```

```
+              per = 0.2, piM = 0.2, kappa = 3, reqPower = 0.8, c1 = 50,
+              beta = 0.8, alpha = 0.05)
> study2

  Results subject to r not greater than 20:
-----------------------------------------------------
Optimal total number of measurements (r+1): 2
Optimal number of participants (N)        : 66
Minimized cost                            : 4180
```

Thus, the optimal is to perform a longitudinal study with $N_{\text{opt}} = 66$ participants and taking $r_{\text{opt}} + 1 = 2$ measurements. The minimized cost of such study is 4180 monetary units. Figure 3 has been obtained using the function `plot()` as in the previous section.

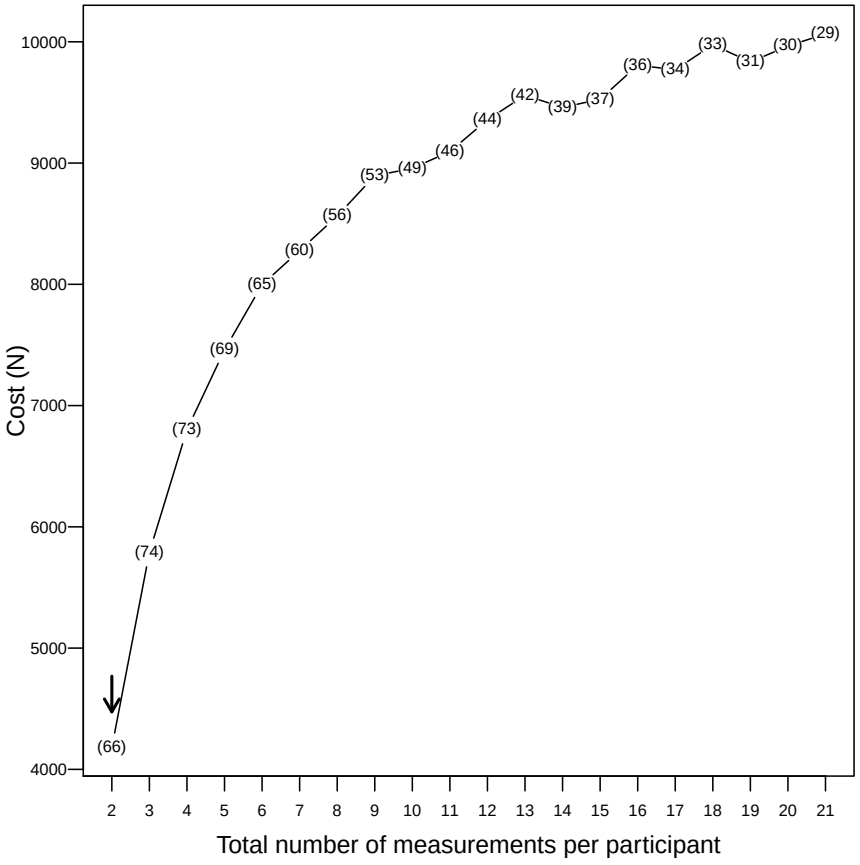

Figure 3: Minimized cost and number of participants (in brackets) as a function of the total number of measurements per participant. The arrow points to the optimal allocation.

4 Particular case: cross-sectional study design

The function `OA()` can also be used for a cross-sectional study design. In this case, we should fix `pattern = "CMD"` and `rMax = 0`, as shown in the two following examples.

4.1 Study 3. Cost of a cross-sectional study

Suppose we are interested in finding the cost of a cross-sectional study achieving a power of at least 0.9 to detect an effect size $\beta = -0.3$ with a significance level $\alpha = 0.05$. The cost of the unique measurement per participant is $c_1 = 25$ monetary units. The proportion of exposed is assumed to be 0.3 and the residual variance is estimated in $\sigma = 1$. Thus, the study calculations are:

```
> study3 <- OA(target = "minCost", pattern = "CMD", rMax = 0, sigma2 = 1,
+              pe0 = 0.3, reqPower = 0.9, c1 = 25, beta = -0.3,
+              alpha = 0.05)
> study3

  Results subject to a cross-sectional design:
-----------------------------------------------------
Number of participants (N): 556
Cost                      : 13900
```

Thus, the required number of participants is $N = 556$ and the total cost is 13900 monetary units.

4.2 Study 4. Power of a cross-sectional study

Suppose we are interested in finding the power of a cross-sectional study to detect an effect size $\beta = -0.3$ with a significance level $\alpha = 0.05$. The total budget for the study is 10000 monetary units and the cost of the unique measurement per participant is $c_1 = 25$ monetary units. The proportion of exposed is assumed to be 0.2 and the residual variance is estimated in $\sigma = 1$. Thus, the study calculations are:

```
> study4 <- OA(target = "maxPower", pattern = "CMD", rMax = 0, sigma2 = 1,
+              pe0 = 0.2, budget = 10000, c1 = 25, beta = -0.3,
+              alpha = 0.05)
> study4

  Results subject to a cross-sectional design:
-----------------------------------------------------
Number of participants (N): 400
Power                     : 0.6700445
```

Thus, the required number of participants is $N = 400$ and the achieved power is 0.67.

5 Illustrative example

In this section, we used data from a study on cleaners and respiratory health to provide an optimal design for a new hypothetical study on the same topic. Briefly, Medina-Ramón *et al.* [2] followed a group of $N = 43$ female domestic cleaners during $r + 1 = 15$ days. Each day, they provided measures of pulmonary function and annotated in a diary whether they performed certain cleaning tasks or used certain cleaning products. The study was observational and therefore the exposures were not assigned by design, rather, the cleaners performed the tasks and used the products that their work day required. All exposures showed day-to-day variations within-subjects. Here, we focus on the two exposures that had the highest and lowest value of ρ_e, namely vacuum cleaning and using air freshener sprays. The first one had $\rho_e = 0.13$ and an average prevalence of $\bar{p}_e = 0.37$ while the second had $\rho_e = 0.60$ and $\bar{p}_e = 0.17$. As expected, the prevalence of the exposures showed no trend so we assumed a constant prevalence of the exposures. Thirty-one participants in the original study provided complete data, so we set $\pi_M = 0.28$. The residual variance and the response covariance damping parameter were taken from the study and set to $\sigma^2 = 0.43$ and $\theta = 0.12$, respectively. We used low (0.3) and high (0.7) values for ρ. Regarding the hypothesized effect, we fixed it at a difference of 10% in the expected mean value of the response between exposed and non exposed assuming the CMD response pattern. This results in $\tilde{\beta} = -0.39$. The objective was to minimize the total cost of the study fixing a minimum required power of 0.9. The first measurement was assumed to be 2 times more expensive than each of the subsequent ones (i.e., $\kappa = 2$). We constrained the maximum number of repeated measurements to 20. All calculations were performed fixing a significance level $\alpha = 0.05$.

Then, all calculations for the study design in each scenario can be performed with the following code:

```
> # Creating scenarios:
>
> res <- expand.grid(Exposure = c("Vacuuming", "Air freshener sprays"),
+                    rho = c(0.3, 0.7))
> res$pe0 <- 0.37
> res$pe0[res$Exposure == "Air freshener sprays"] <- 0.17
> res$per <- res$pe0
> res$rhoe <- 0.13
> res$rhoe[res$Exposure == "Air freshener sprays"] <- 0.60
> res$TimeVaryingExposure <- TRUE
> aux <- res
> aux$TimeVaryingExposure <- FALSE
> aux$rhoe <- 1
> res <- rbind(res, aux)
> res$r <- NA
> res$N <- NA
> res$cost <- NA
```

```
> # Sorting the table:
> ord <- order(res$Exposure, res$rho, 1 - res$TimeVaryingExposure)
> res <- res[ord, ]
> rownames(res) <- NULL
> res
```

	Exposure	rho	pe0	per	rhoe	TimeVaryingExposure	r	N	cost
1	Vacuuming	0.3	0.37	0.37	0.13	TRUE	NA	NA	NA
2	Vacuuming	0.3	0.37	0.37	1.00	FALSE	NA	NA	NA
3	Vacuuming	0.7	0.37	0.37	0.13	TRUE	NA	NA	NA
4	Vacuuming	0.7	0.37	0.37	1.00	FALSE	NA	NA	NA
5	Air freshener sprays	0.3	0.17	0.17	0.60	TRUE	NA	NA	NA
6	Air freshener sprays	0.3	0.17	0.17	1.00	FALSE	NA	NA	NA
7	Air freshener sprays	0.7	0.17	0.17	0.60	TRUE	NA	NA	NA
8	Air freshener sprays	0.7	0.17	0.17	1.00	FALSE	NA	NA	NA

```
> # Optimal allocation calculations
> # for all scenarios:
>
> studies <- list()
> for (i in 1:nrow(res))
+  {
+   studies[[i]] <- OA(target = "minCost", pattern = "CMD", rMax = 20,
+                      theta = 0.12, rho = res$rho[i], sigma2 = 0.43,
+                      rhoe = res$rhoe[i], pe0 = res$pe0[i], per = res$per[i],
+                      piM = 0.28, kappa = 2, reqPower = 0.9, c1 = 1,
+                      beta = -0.39, alpha = 0.05)
+   res$r[i] <- studies[[i]]$ropt
+   res$N[i] <- studies[[i]]$Nopt
+   res$cost[i] <- round(studies[[i]]$minCost, 1)
+  }
> # Results:
>
> studyDesigns <- res[, -c(3:5)]
> studyDesigns
```

	Exposure	rho	TimeVaryingExposure	r	N	cost
1	Vacuuming	0.3	TRUE	18	6	51.6
2	Vacuuming	0.3	FALSE	1	92	125.1
3	Vacuuming	0.7	TRUE	15	3	22.0
4	Vacuuming	0.7	FALSE	0	128	128.0
5	Air freshener sprays	0.3	TRUE	20	17	160.7
6	Air freshener sprays	0.3	FALSE	1	152	206.7
7	Air freshener sprays	0.7	TRUE	19	8	72.2
8	Air freshener sprays	0.7	FALSE	0	211	211.0

Results reveal, in some scenarios, a notable discrepancy in both the optimal number of repeated measurements and optimal number of participants between the assumptions of a time-invariant exposure (i.e., $\rho_e = 1$) and a time-varying exposure (using the observed value of ρ_e). In the scenarios with $\rho = 0.7$, when using the observed value of ρ_e, the optimal design was to take a high number of measurements (15 for vacuum cleaning and 19 for using air freshener sprays) while, assuming $\rho_e = 1$, the optimal was to perform a cross-sectional study. Thus, incorrectly using the design formulas for $\rho_e = 1$ when the exposure is actually time-varying can lead to discrepancies not only in r_{opt} and N_{opt}, but also in the final cost of the study. For example, in the scenarios with $\rho = 0.7$, using the time-invariant exposure formulas leads to designs with an increase in cost of 192% for using air freshener sprays, and of 482% for vacuum cleaning.

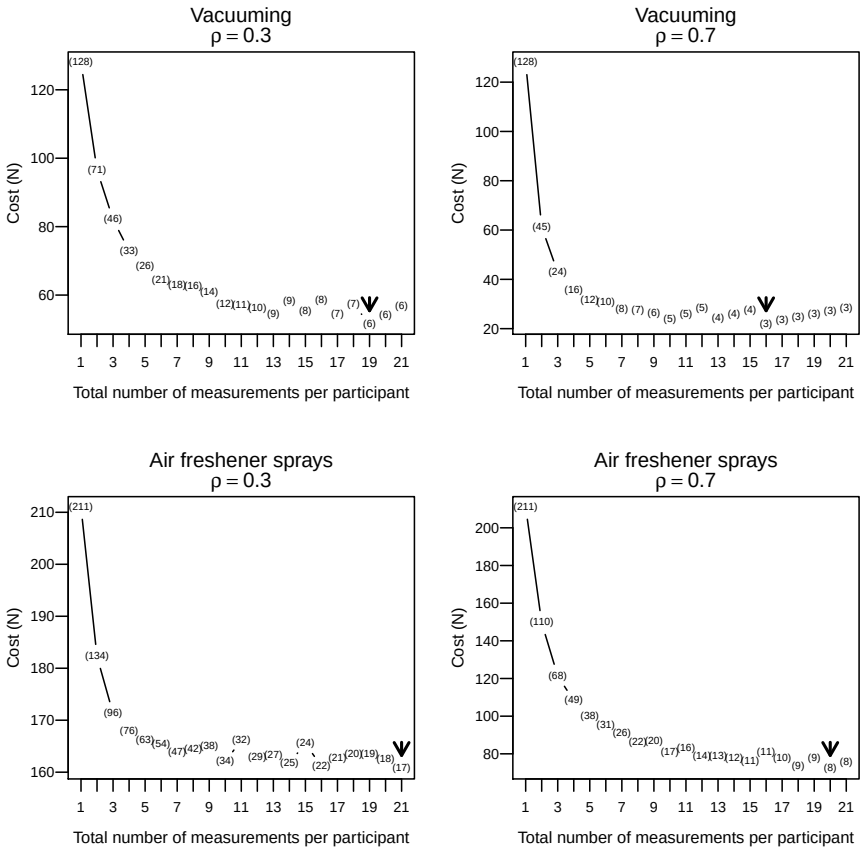

Figure 4: Minimized cost and number of participants (in brackets) as a function of the total number of measurements per participant. The arrow points to the optimal allocation.

In some cases, the slope of the optimal cost as a function of r attenuates as r increases

and thus, the investigator could be interested in increasing the number of participants in exchange for reducing the number of repeated measurements without a significant increase of the cost. In order to explore that, we can create the Figure 4 using the following code:

```
> for (i in c(1,3,5,7))
+  {
+   plot(studies[[i]], Ncex = 0.5)
+   mtext(text = res$Exposure[i], side = 3, line = 1, cex = 0.8)
+   mtext(text = bquote(paste(rho, sep = "") ==. (res$rho[i])),
+         side = 3, line = 0, cex = 0.8)
+  }
```

In fact, Figure 4 shows how, for large values of r (in scenarios where $\rho = 0.7$), the investigator can increase the number of participants in exchange for reducing the number of repeated measurements without a significant increase of the cost.

Acknowledgments

The authors wish to thank Dr Juan Ramón González and Dr Alejandro Cáceres for their help with the R package compilation as well as Dr Medina-Ramón, Dr Antó and Dr Zock for letting us use the EPIASLI data in our example.

References

[1] Barrera-Gómez J, Spiegelman D, Basagaña X. Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure. *Statistics in Medicine* [Epub ahead of print].

[2] Medina-Ramón M, Zock JP, Kogevinas M, Sunyer J, Basagaña X, Schwartz J, Burge PS, Moore V, Antó JM. Short-term respiratory effects of cleaning exposures in female domestic cleaners. *European Respiratory Journal* 2006; **27**(6):1196–1203. DOI:10.1183/09031936.06.00085405.

[3] Kistner EO, Muller KE. Exact distributions of intraclass correlation and Cronbach's alpha with gaussian data and general covariance. *Psychometrika* 2004; **69**(3):459–474. DOI:10.1007/BF02295646.

[4] Basagaña X, Spiegelman D. Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure. *Statistics in Medicine* 2010; **29**(2):181–192. DOI: 10.1002/sim.3772.

[5] Gange SJ. Generating Multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 1995; **49**(2): 134–138. DOI: 10.1080/00031305.1995.10476130.

[6] Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; **55**(1):137–148. DOI:10.1111/j.0006-341X.1999.00137.x.