

# Multiple Hypotheses Testing in Pair-Matched Observational Studies: The R package CrossScreening

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania  
qyzhao@wharton.upenn.edu

April 20, 2017

## Abstract

This article describes the R package `CrossScreening` that provides useful functionality for multiple hypotheses testing in pair-matched observational studies, including performing sensitivity analyses with multiple signed score test (`sen` function), computing sensitivity value (`sen.value` function), and use planning sample to screen hypotheses to gain power (`cross.screen` function). These functions are demonstrated using four real datasets by reproducing results of several previous papers.

## 1 Introduction

In an observational study that tests many causal hypotheses, to be credible one must demonstrate that its conclusions are neither artifacts of multiple testing nor of small biases from nonrandom treatment assignment. Indeed, Young and Karr [2011] identified two main difficulties with observational studies: multiple testing/multiple modeling and bias due to unmeasured confounder. Existing R packages for multiple testing, such as the `p.adjust` function in the `stats` package [R Core Team, 2017] and the resampling based tests implemented in `multtest` package, only correct for the error due to simultaneous testing and ignore the systematic error due to confounding bias. Other R packages for sensitivity analysis such as `rbounds` [Keele, 2014] and `sensitivitymw` [Rosenbaum, 2015] consider how uncontrolled confounder may change the qualitative conclusion of a single causal hypothesis.

This article describes the new R package `CrossScreening` that provide useful functions to screen and test causal hypotheses. When hundreds or thousands of hypotheses are tested at the same time, the cross-screening method implemented in `cross.screen` can substantially improve power over directly applying multiple testing procedures on the  $p$ -values generated from `rbounds` or `sensitivitymw`. Intuitively, this is due to the conservativeness of the  $p$ -value of a sensitivity

analysis when the null hypothesis is correct. To avoid over-correcting for the conservative  $p$ -values, Heller et al. [2009] proposed to use a subsample of the data to screen the hypotheses before using rest of the data for sensitivity analysis. Zhao et al. [2017] further proposed to use both subsamples to screen the hypotheses and perform sensitivity analyses. The cross-screening procedure in Zhao et al. [2017] is usually more powerful and robust than the sample splitting procedure in Heller et al. [2009], and both procedures are implemented in this package.

To screen many bias-prone hypotheses, a useful function is `sen.value` in the package. The `sen.value` function returns the "sensitivity value"—the magnitude of departure from a randomized experiment needed to change the qualitative conclusions, a concept used in many existing observational studies and formalized in Zhao [2017]. The sensitivity value speaks to the assertion "it might be bias" in an observational study in much the same way as the  $p$ -value speaks to the assertion "it might be bad luck" in a randomized trial [Rosenbaum, 2015]. Just as the  $p$ -value in a randomized experiment summarizes the amount of bad luck needed for the association between treatment and outcome to be non-causal, the sensitivity value in an observational study summarizes the amount of bias needed for that association to be non-causal. Therefore, it is quite natural to use sensitivity values to screen hypotheses in an observational study.

The rest of this article is organized as follows. Section 2 introduces some notations for a pair-matched observational study and describes four datasets that will be used as examples in this article. Section 3 outlines sensitivity analysis of observational studies and function `sen` in the package. Section 4 describes using sensitivity value to screen hypotheses and the function `sen.value` in the package. Section 6 describes single-screening [Heller et al., 2009] and cross-screening [Zhao et al., 2017] and the corresponding function `cross.screen`. Section 7 concludes the article with a brief discussion.

## 2 Pair-matched observational studies

We first describe the basic setting of a pair-matched observational study. There are  $I$  independent matched pairs,  $i = 1, \dots, I$  and each pair has two subjects,  $j = 1, 2$ , one treated, denoted by  $Z_{ij} = 1$ , and one control, denoted by  $Z_{ij} = 0$ . Pairs are matched for observed covariates, but the investigator may be concerned that matching failed to control some unmeasured covariates  $u_{ij}$ . Let  $r_{Tij}$  be the potential outcome of the  $j$ -th subject in the  $i$ -th pair if  $ij$  receives treatment. Similarly,  $r_{Cij}$  is the potential outcome if  $ij$  receives control. The potential outcomes  $r_{Tij}$  and  $r_{Cij}$  can be a vector if multiple outcomes are observed. The observed outcome is  $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$  and the individual treatment effect  $r_{Tij} - r_{Cij}$  is not observed for any subject [Rubin, 1974]. Let  $D_i$  be the treatment-minus-control difference  $D_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$  for the  $i$ -th pair.

The `CrossScreening` package includes three datasets of observational studies:

**lead** Morton et al. [1982] compared the blood lead levels of 33 children whose father worked in a factory that used lead in manufacturing batteries to 33 control children of the same age from the same neighborhood.

**methotrexate** Deng et al. [2005] compared the genetic damage of 21 workers from a plant producing methotrexate to 21 controls matched according to age, gender and smoking. Genetic damage of the workers are studied using four assays (four outcomes).

**nhanes.fish** Using the 2013–2014 National Health and Nutrition Examination Survey (NHANES), Zhao et al. [2017] compared 46 laboratory outcomes of 234 adults with high fish consumption (more than 12 servings of fish or shellfish in the previous month) with 234 adults with low fish consumption (0 or 1 servings of fish).

**gender** In a microarray experiment, Vawter et al. [2004] investigates genes differentially expressed in human brain with respect to gender. Using their dataset, Zhao [2017] obtained 41 matched pairs of males and females by the laboratory and microarray platform that analyzed the sample. There are in total 12,600 genes in this dataset.

The first three datasets can be load into R by

```
library(CrossScreening)
data(lead, methotrexate, nhanes.fish, nhanes.fish.match)
```

The next code chunk obtains the treat-minus-control differences (a matrix for methotrexate and `nhanes.fish`). Note that function `nhanes.log2diff` in the package computes the  $\log_2$  differences of laboratory variables in the `nhanes.fish` dataset.

```
d.lead <- lead$exposed[-21] - lead$control[-21] # the 21st control outcome is NA
d.methotrexate <- methotrexate[, 1:4] - methotrexate[, 6:9]
d.nhanes <- nhanes.log2diff()
```

The gender dataset is too large to be put inside a R package. It can be obtained by

```
url <- "https://raw.githubusercontent.com/qingyuanzhao/CrossScreening/master/"
d.gender <- read.csv(paste0(url, "data/d.gender.csv"))
```

### 3 Sensitivity analysis

The sharp null hypothesis of no treatment effect assumes that  $H_0 : r_{Tij} = r_{Cij}, \forall i, j$ . If  $H_0$  is true and the treatments are randomly assigned, then conditioning on the potential outcomes and observed and unobserved covariates,  $D_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$  attaches equal probabilities to  $\pm|r_{Ci1} - r_{Ci2}|$ . When there is no concern of bias due to unmeasured confounders, a randomization test can be used to test  $H_0$ . One popular choice is Wilcoxon's signed rank test which uses the ranks of the absolute differences  $|D_i|, i = 1, \dots, n$ . This can be done using the `sen` function by setting the sensitivity parameter  $\Gamma$  to 1.

```
sen(d.lead, gamma = 1)$p.value
```

```
##          1  
## 5.062e-06
```

It is easy to check that the  $p$ -value computed by `sen` is very close to the `wilcox.test` function in the `stats` package. They are not exactly equal because `d.lead` has tied values and also `sen` always uses a normal approximation while `wilcox.test` computes an exact  $p$ -value when sample size is less than 50.

```
wilcox.test(d.lead, alternative = "greater")$p.value
```

```
## [1] 5.774e-06
```

In a sensitivity analysis, the user specifies the sensitivity parameter  $\Gamma \geq 1$ , the upper bound of the odds ratio of treatment for two matched people.  $\Gamma = 1$  means the odds ratio can only be 1, so the matched observational study mimics a randomized experiment. The larger the parameter  $\Gamma$ , the more bias we allow in the study. When  $\Gamma > 1$ ,  $p$ -value is no longer a single value, but rather an interval of possible  $p$ -values. Typically the largest possible (worst case)  $p$ -value is reported. For more technical detail about sensitivity analysis, we refer the reader to Rosenbaum [2002, Chapter 4].

To run a sensitivity analysis, simply call `sen` with a vector of sensitivity parameters  $\Gamma$ :

```
gamma <- c(1, 4, 4.5, 5, 5.5, 5.8)  
round(sen(d.lead, gamma = gamma)$p.value, 3)
```

```
##      1      4      4.5      5      5.5      5.8  
## 0.000 0.039 0.055 0.074 0.094 0.107
```

This reproduces the first column of Table 2(a) in Rosenbaum [2011]. Notice that the `p.value` field in the returned list of `sen` contains the upper bound(s) of one-sided  $p$ -values (default `alternative` is greater than 0). The field `p.value2` contains the upper bound(s) of two-sided  $p$ -values.

Rosenbaum [2011] proposed a new class of signed score tests for sensitivity analysis in observational studies. By choosing an appropriate non-linear transform (indexed by three numbers,  $(m, \underline{m}, \bar{m})$ ) to the ranks, the tests are usually less sensitive to unmeasured bias than Wilcoxon's signed rank test. The `sen` function implements this class of tests and supports multiple test statistics by inputting a matrix `mm` with 3 rows. (By default, `mm = NULL` is Wilcoxon's test.) The next code chunk reproduces Table 2(b) in Rosenbaum [2011].

```

mm <- matrix(c(2, 2, 2, 5, 4, 5, 8, 7, 8, 8, 6, 8, 8, 5, 8, 8, 6, 7), nrow = 3)
gamma <- c(1, 1.3, 1.4, 2, 2.5)
round(sen(d.methotrexate$wmtm, mm, gamma, score.method = "exact")$p.value, 4)

##      (2,2,2) (5,4,5) (8,7,8) (8,6,8) (8,5,8) (8,6,7)
## 1      0.0151 0.0149 0.0546 0.0157 0.0096 0.0025
## 1.3    0.0475 0.0399 0.1024 0.0392 0.0288 0.0076
## 1.4    0.0624 0.0507 0.1194 0.0489 0.0375 0.0100
## 2      0.1810 0.1307 0.2222 0.1189 0.1073 0.0291
## 2.5    0.2954 0.2062 0.3017 0.1831 0.1776 0.0490

```

Additionally, under the assumption that the treatment effect is an additive constant  $r_{Tij} - r_{Cij} \equiv \tau$ , one can obtain confidence interval of the treatment effect  $\tau$  allowing for a bias of  $\Gamma$  in treatment assignment by inverting sensitivity analysis. The confidence interval is implemented in the `sen.ci` function which returns two intervals, `point.estimate` and `ci`. Since we allow bias in treatment assignment up to  $\Gamma$ , the Hodges-Lehmann point estimate of  $\tau$  is indeed a range of values when  $\Gamma > 1$ . We refer the reader to Rosenbaum [2011, Section 5] for more detail. Next we reproduce the results for the lead example in Rosenbaum [2011, Section 5].

```

sen.ci(d.lead, gamma = c(1, 2), alpha.up = 0, alpha.low = 0.05)

## $point.estimate
##      low  up
## 1 15.0 15.0
## 2 10.5 19.5
##
## $ci
##      low  up
## 1 10.5 Inf
## 2  5.5 Inf

```

By default, Wilcoxon's test is used. At  $\Gamma = 2$ , the one-sided confidence interval of  $\tau$  is  $[5.5, \text{inf})$  and the range of point estimate is  $[10.5, 19.5]$ .

## 4 Using the sensitivity value to screen hypotheses

Since sensitivity analysis gives an upper bound of possible  $p$ -values when  $\Gamma > 1$ , the null hypotheses will typically have very conservative  $p$ -value upper bounds (stochastically larger than the uniform distribution on  $[0, 1]$ ). In fact, in absence of bias, it is extremely unlikely that random chance alone can create an association insensitive to moderate amount of bias. To see this, we run two-sided sensitivity analysis using Wilcoxon's test on the first 8 outcomes in the NHANES fish dataset:

```

gamma <- c(1, 1.25, 1.5)
round(apply(d.nhanes[, 1:8], 2, function(d) sen(d, gamma = gamma)$p.value2), 3)
##      o.LBXSAL o.LBXSBU o.LBXSCL o.LBXSCH o.LBXSCK o.LBXSCL o.LBXSGB o.LBXSGL
## 1      0.529   0.126   0.412   0.583   0.829   0.475   0.245   0.781
## 1.25   1.000   0.942   1.000   1.000   1.000   1.000   1.000   1.000
## 1.5    1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000

```

The  $p$ -value bounds for  $\Gamma = 1.25$  and  $1.5$  quickly become very close to 1. In contrast, a true causal effect may fend off a large amount of bias. In the NHANES fish dataset, `o.LBXTHG` is the total blood mercury of the surveyee and it remains significant

```

mm <- matrix(c(2, 2, 2, 8, 5, 8), nrow = 3)
sen(d.nhanes$o.LBXTHG, mm, gamma = c(1, 5, 11, 14))$p.value2
##      (2,2,2) (8,5,8)
## 1  0.000e+00 0.000e+00
## 5  1.226e-06 1.298e-06
## 11 1.099e-02 3.780e-03
## 14 5.340e-02 1.621e-02

```

The (2,2,2) test closely resembles Wilcoxon's test and is more sensitive to bias than the (8,5,8) test.

Based on the observation above, Heller et al. [2009] proposed a sample splitting method that uses part of the data to screen the hypotheses and uses the other part for sensitivity analysis. What is a reasonable way to screen out the hypotheses that are sensitive to a small amount of bias? One possibility is to keep the hypotheses whose  $p$ -value upper bound at some  $\Gamma$  is small. A more natural measure of the "sensitivity" of a hypothesis, is the sensitivity value—a concept formalized in Zhao [2017]. Briefly speaking, sensitivity value is the critical parameter  $\Gamma$  where the  $p$ -value upper bound just becomes insignificant. For example, if we zoom in to  $13.6 \leq \Gamma \leq 14$  in the fish-mercury example, sensitivity analysis outputs

```

sen(d.nhanes$o.LBXTHG, gamma = seq(13.6, 14, 0.05))$p.value2
##      13.6  13.65  13.7  13.75  13.8  13.85  13.9  13.95  14
## 0.04512 0.04611 0.04711 0.04812 0.04915 0.05019 0.05124 0.05231 0.05340

```

If the significance level is  $\alpha = 0.05$ , the sensitivity value in this case is between 13.8 and 13.85. This can be computed via the `sen.value` function by setting `alpha` to `0.05/2` (divided by 2 because `sen.value` is one-sided by nature):

```
kappa2gamma(sen.value(d.nhanes$o.LBXTHG, alpha = 0.05, alternative = "two.sided"))
## (2,2,2)
## 13.84
```

The function `sen.value` outputs the sensitivity value in the  $\kappa = \Gamma / (1 - \Gamma)$  scale, and `kappa2gamma` transforms the value to the familiar  $\Gamma$  scale. Note that rather than searching over a range of  $\Gamma$ , `sen.value` directly computes the sensitivity value and is much faster than searching [Zhao, 2017]. The function `sen.value` also supports matrix input of the differences and test statistics. For example,

```
kappa2gamma(sen.value(d.nhanes[, c(1:5, 18, 21, 23)],
alpha = 0.05, mm = mm, alternative = "two.sided"))
##          o.LBXSAL o.LBXSBU o.LBXSCA o.LBXSCH o.LBXSCK o.LBXTHG o.LBXIHG o.LBxBGM
## (2,2,2)  0.7824   0.9154   0.8274   0.8088   0.7695   13.84   1.200   14.33
## (8,5,8)  0.7676   0.9309   0.8144   0.7999   0.7511   17.70   1.972   20.27
```

When the sensitivity value  $\Gamma^*$  is less than 1, this means the usual hypothesis test at  $\Gamma = 1$  is not significant and  $1/\Gamma^*$  is the critical value that the *lower* bound of  $p$ -values becomes significant.

We demonstrate the use of sensitivity value in screening hypotheses by reproducing the results in Zhao [2017, Section 4]. Gagnon-Bartsch and Speed [2012] and Wang et al. [2016] studied the gender microarray dataset and found that it is very likely that the associations between genes and gender have unmeasured confounding. Sensitivity values provide a brief summary of which associations are less sensitive to unmeasured bias.

```
gamma.star <- kappa2gamma(sen.value(d.gender, alpha = 0.05, alternative = "two.sided"))

library(ggplot2)
ggplot() + aes(sample = gamma.star) + stat_qq() +
xlab("Normal quantile") + ylab("Sensitivity value quantile")
```

Figure 1 shows the quantile-quantile plot of the two-sided sensitivity values of the 12,600 genes using Wilcoxon's test. It is clear from the plot that two genes have large sensitivity values and thus are more robust to confounding bias.

## 5 Approximating the power of a sensitivity analysis

To facilitate the design of observational studies, The `power.sen` function in the package computes approximate power of a sensitivity analysis using asymptotic formulas in Zhao [2017]. We assume the data are in favorable situation that random treatment assignment is satisfied after matching

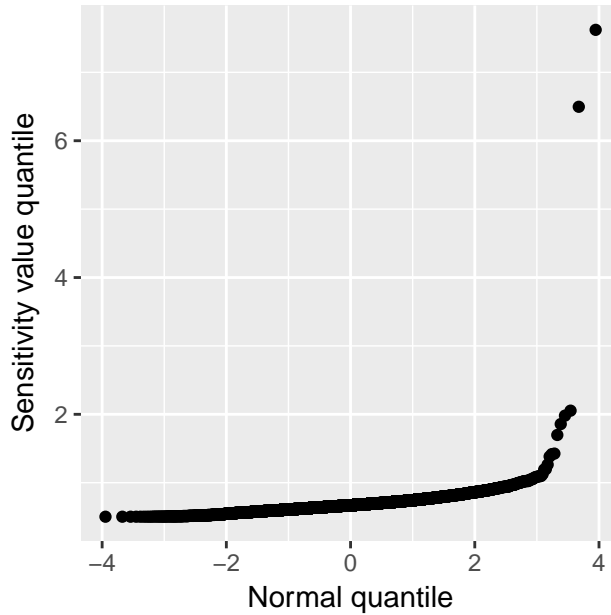


Figure 1: Quantile-quantile plot of the sensitivity values in the gender dataset

and there is a positive treatment effect. The function `power.sen` first estimates the large-sample mean and variance of the signed score statistics, then use the estimates to compute the power of a subsequent sensitivity analysis using a different dataset. Alternatively, the user can specify the large-sample mean and variance of the test statistic.

For example, suppose we observe 100 planning samples whose treatment-minus-control differences are distributed as  $N(0.3, 1)$ . The following code gives the power of a subsequent sensitivity analysis using 200 and 500 samples at sensitivity level  $\Gamma = 2$  and significance level  $\alpha = 0.05$ . We also consider two test statistics,  $(2, 2, 2)$  which closely approximates Wilcoxon's test and  $(8, 5, 8)$ .

```
d <- rnorm(100) + 0.3
mm <- matrix(c(2, 2, 2, 8, 5, 8), nrow = 3)
power.sen(d = d, mm = mm, I = 200, gamma = 2)

## $power
## [1] 0.01860 0.03563
##
## $mu.F
## [1] 0.6499 0.6602
##
## $sigma.F
## [1] 0.5503 0.6052

power.sen(d = d, mm = mm, I = 500, gamma = 2)$power
```



```
## [1] 0.01014 0.03020
```

## 6 Using cross-screening to improve the power of multiple testing

Sample splitting was introduced by Heller et al. [2009] to reduce the number of hypotheses and gain power in sensitivity analysis. We shall first explain the single screening procedure in Heller et al. [2009] before moving to the cross-screening procedure in Zhao et al. [2017]. Both procedures are implemented in the function `cross.screen` in the package.

In single screening, the entire data are randomly split into two parts, the first part ( $0 < \zeta < 1$  proportion of the sample) is used to screen the hypotheses and the second part ( $1 - \zeta$  proportion of the sample) is used for sensitivity analysis. Heller et al. [2009] used the  $p$ -value upper bounds at fixed  $\Gamma = 2$  to screen the hypotheses (implemented in `cross.screen.fg` by specifying `gamma.screen = 2`). The user can use the option `screen.method` to decide which hypotheses are kept. Option “`threshold`” means the hypotheses with  $p$ -value upper bounds less than `alpha.screen` are kept, and option “`least.sensitive`” means the `least.sensitive`-smallest  $p$ -value upper bounds are kept. As explained earlier, sensitivity value is a more natural metric of the hypotheses’ sensitivity to unmeasured bias (implemented in `cross.screen` with the same options). After screening, a sensitivity analysis is then performed for each selected hypothesis as explained in Section 3.

Single-screening can sometimes discard the wrong hypotheses in the screening step. A closely related but more robust procedure is cross-screening, in which both splits are used for screening and testing. Since the role of the two parts are now the same, it is recommended to choose  $\zeta = 1/2$ . Cross-screening rejects the union of the rejected hypotheses in both halves, but since a false positive can occur in both halves, hence a Bonferroni correction is needed to control the familywise error rate. See Zhao et al. [2017] for more technical detail.

We illustrate the usage of `cross.screen` using the `nhanes.fish` dataset (see also the example of `cross.screen.fg` which reproduces Table 1 in Zhao et al. [2017]). First, randomly split the sample into two halves

```
set.seed(11)
split <- sample(1:nrow(d.nhanes), nrow(d.nhanes) / 2, replace = FALSE)
d1 <- d.nhanes[split, ]
d2 <- d.nhanes[-split, ]
```

Then, run `cross.screen` with the desired options. In the returned list, field `p` is the adjusted  $p$ -value that can be directly used to control FWER. For example, if we reject all the hypotheses whose corresponding `p` are less than  $\alpha$ , then the FWER is controlled at level  $\alpha$ .

```

cross.screen(d1, d2, gamma = 9, gamma.screen = 1.25, mm = c(2, 2, 2),
screen.method = "least.sensitive", least.sensitive = 3)$p
## [1] NA NA NA NA NA NA NA NA NA
## [10] NA NA NA NA NA 1.00000 NA NA 0.03882
## [19] NA NA 1.00000 NA 0.04295 NA NA NA NA
## [28] NA NA NA NA NA NA NA NA NA NA
## [37] NA NA NA NA NA NA NA NA NA NA
## [46] NA

```

Therefore, the 18th (total blood mercury) and 23rd (methyl mercury) outcomes can be rejected at  $\Gamma = 9$  and  $\alpha = 0.05$ .

To use single screening, simply use an appropriate split (usually the screening portion  $\zeta < 1/2$ ) and use the p1 field of the returned list of `cross.screen`. Notice that the p1 field is not adjusted for multiple testing.

```

split <- sample(1:nrow(d.nhanes), nrow(d.nhanes) * 0.3, replace = FALSE)
d1 <- d.nhanes[split, ]
d2 <- d.nhanes[- split, ]
p.single <- cross.screen(d1, d2, gamma = 9, gamma.screen = 1.25, mm = c(2, 2, 2),
screen.method = "least.sensitive", least.sensitive = 3)$p1
p.adjust(p.single, "bonferroni")
## [1] NA NA NA NA NA NA NA NA NA
## [10] NA 1.00000 NA NA NA NA NA NA 0.03673
## [19] NA NA NA NA 0.03673 NA NA NA NA
## [28] NA NA NA NA NA NA NA NA NA NA
## [37] NA NA NA NA NA NA NA NA NA NA
## [46] NA

```

By default, `cross.screen` tests for the two-sided alternative (`two.sided = TRUE`) and uses the planning sample can also be used to adaptively select the alternative direction. `cross.screen` can also adaptively select test statistic using the planning sample if the input `mm` is a matrix (each column correspond to a signed score statistic). In the three examples, three candidate statistics are considered ((8,5,8) is robust against most alternative distributions, (8,7,8) is more powerful for light-tailed distributions and (8,6,7) is more powerful for heavy-tailed distributions, see Rosenbaum [2011]).

```

mm <- matrix(c(8, 5, 8, 8, 7, 8, 8, 6, 7), nrow = 3)
cross.screen(d1, d2, gamma = 9, gamma.screen = 1.25, mm = mm,
screen.method = "least.sensitive", least.sensitive = 3)$p

```

```
## [1] NA NA NA NA NA NA NA NA NA
## [10] NA NA NA NA NA NA NA NA NA 0.03476
## [19] NA NA 1.00000 NA 0.02063 NA NA NA NA
## [28] NA NA NA NA NA NA NA NA NA NA
## [37] NA NA NA NA NA NA NA NA NA NA
## [46] NA
```

In addition to screening out some hypotheses that are sensitive to bias, we can also use the planning step to order the hypotheses (`s1.order` and `s2.order` in the returned list of `cross.screen`). In the next example, we set `gamma = 1.5` and `screen.method = "threshold"`.

```
output <- cross.screen(d1, d2, gamma = 1.5, gamma.screen = 1.25, mm = mm,
screen.method = "threshold")
attach(output)
s1.order

## [1] 23 18 21 10 17 11 16 27

s2.order

## [1] 23 18 21 19 15 42 44 46
```

Given the ordered hypotheses, we can use a sequential test that controls FWER. The `fallback.test` function implements the fallback procedure of Wiens [2003] that spreads the significance equally into the first few hypotheses. Note that to control FWER at  $\alpha$ , the sequential tests should be performed at significance level  $\alpha/2$ . In the above example, the rejected outcomes are the union of

```
s1.order[fallback.test(p1[s1.order], alpha = 0.05/2)]

## [1] 23 18 21

s2.order[fallback.test(p2[s2.order], alpha = 0.05/2)]

## [1] 23 18
```

If we ignore the ordering and just use Bonferroni's correction, the rejected outcomes are

```
which(p <= 0.05)

## [1] 18 23
```

In the above example, ordering the hypotheses helps to reject one more outcome at  $\Gamma = 1.5$ .

## 7 Discussion

The `CrossScreening` package provides many useful R functions to test multiple hypotheses in a pair matched observational study. We have demonstrated the usage and options of these functions through reproducing the numerical results in some previous papers, including Heller et al. [2009], Rosenbaum [2011], Zhao [2017], Zhao et al. [2017].

Cross-screening is a very flexible method and this package does not implement all the possible extensions. For instance, other sequential testing methods (e.g. the recycling tests in Burman et al. [2009]) can be used after the hypotheses are ordered. The Bonferroni adjusted  $p$ -values (`p` in the returned list of `cross.screen`) provide a quick test that controls FWER, but to gain more power we recommend the investigator to choose a sequential test based on her subject knowledge. The idea of cross-screening can also be used in problems that do not use sensitivity analysis. However, it is attractive only if there are dozens of tests and many of them are conservative.

So far we have only used cross-screening with random splits of the data. In some problems, we may want to test the hypotheses for different subgroups defined by a binary covariate. In this case, we can call `cross.screen` with the appropriate `d1` and `d2`. We refer the reader to Zhao et al. [2017, Section 6] for more discussion on nonrandom cross-screening.

## References

- C-F Burman, C Sonesson, and O Guilbaud. A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in Medicine*, 28(5):739–761, 2009.
- Hongping Deng, Meibian Zhang, Jiliang He, Wei Wu, Lifen Jin, Wei Zheng, Jianlin Lou, and Baohong Wang. Investigating genetic damage in workers occupationally exposed to methotrexate using three genetic end-points. *Mutagenesis*, 20(5):351–357, 2005.
- Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.
- Ruth Heller, Paul R Rosenbaum, and Dylan S Small. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101, 2009.
- Luke J Keele. *rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data*, 2014. URL <https://CRAN.R-project.org/package=rbounds>. R package version 2.1.
- David E Morton, Alfred J Saah, Stanley L Silberg, Willis L Owens, MARK A ROBERTS, and Marylou D Saah. Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, 115(4):549–555, 1982.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>.
- Paul R Rosenbaum. *Observational Studies*. Springer, 2002.

- Paul R Rosenbaum. A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics*, 67(3):1017–1027, 2011.
- Paul R Rosenbaum. Two r packages for sensitivity analysis in observational studies. *Observational Studies*, 1:1–17, 2015.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- Marquis P Vawter, Simon Evans, Prabhakara Choudary, Hiroaki Tomita, Jim Meador-Woodruff, Margherita Molnar, Jun Li, Juan F Lopez, Rick Myers, David Cox, et al. Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology*, 29(2):373–384, 2004.
- Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B. Owen. Confounder adjustment in multiple hypothesis testing. *to appear in Annals of Statistics*, 2016.
- Brian L Wiens. A fixed sequence bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2(3):211–215, 2003.
- S Stanley Young and Alan Karr. Deming, data and observational studies. *Significance*, 8(3):116–120, 2011.
- Qingyuan Zhao. On sensitivity value of pair-matched observational studies. *arXiv preprint arXiv:1702.03442*, 2017.
- Qingyuan Zhao, Dylan S Small, and Paul R Rosenbaum. Cross-screening in observational studies that test many hypotheses. *arXiv preprint arXiv:1703.02078*, 2017.