Supplement 2

Modgo comparison

By Francisco Ojeda, George Koliopanos, Andreas Ziegler

Table of contents

1 Thyroid dataset

In our first example, we will use the Thyroid Disease from the UCI machine learning data repository (Dua and Graff 2017). The data set contains 22 variables and 8021 subjects. When all variables of this data set are used, the correlation matrix between variables is not positive definite. The package *sbgcop* can therefore not be run on the complete data set for estimating the correlation matrix. We therefore decided to select a subset of 11 variables, 5 of which are continuous, 5 are dichotomous, and one is categorical. Since *modgo* does not include an option for unordered categorical variables, the variable *clinic* is treated as ordinal.

Three different *modgo* runs are performed:
1) *modgo* with the selected variables.
2) *modgo* with the selected variables using the covariance matrix obtained from *sbgcop* (Hoff 2007). The *sigma* argument permits to import an external correlation matrix.
3) *modgo* with all variables by calculating nearest positive definite correlation matrix. Specifically, *modgo* relies on the nearPD package (Higham, Borsdorf, and Raydan, n.d.) for the calculation of the nearest positive definite correlation matrix.
All simulations produce 500 simulated data sets.

```r
bin_variables = c("sex","thyroid","pregnant","lithium","goitre","tumor","psych",
                  "TSH_measured","T3_measured","T4U_measured","FTI_measured",
                  "BG_measured","TT4_measured")

removed = c("TSH_measured","T3_measured","T4U_measured",
            "FTI_measured","BG_measured","lithium","goitre",
            "tumor","psych","y","pregnant")

categ_variable <- "clinic"

binies <- bin_variables[-which(bin_variables %in% removed)]

variables <- colnames(testThyDf)[-which(colnames(testThyDf)=="y")]
varis <- variables[-which(variables %in% removed)]

nrep = 500

sbgcop_approach <- sbgcop.mcmc(Y = as.matrix(testThyDf[,varis]), nsamp = 1,
                               verb = FALSE)
```

```r
Thy_modgo <- modgo(testThyDf[,varis],bin_variables = binies,
                         categ_variables = categ_variable,
                         nrep = nrep)


Thy_sbgcop_modgo <- modgo(testThyDf,variables = varis,
                         sigma = sbgcop_approach$C.psamp[,,1],
                         bin_variables = binies,
                         categ_variables = categ_variable,
                         nrep = nrep)

Thy_modgo_nearPD <- modgo(testThyDf,variables = variables,
                         bin_variables = bin_variables,
                         categ_variables = categ_variable,
                         nrep = nrep)
```

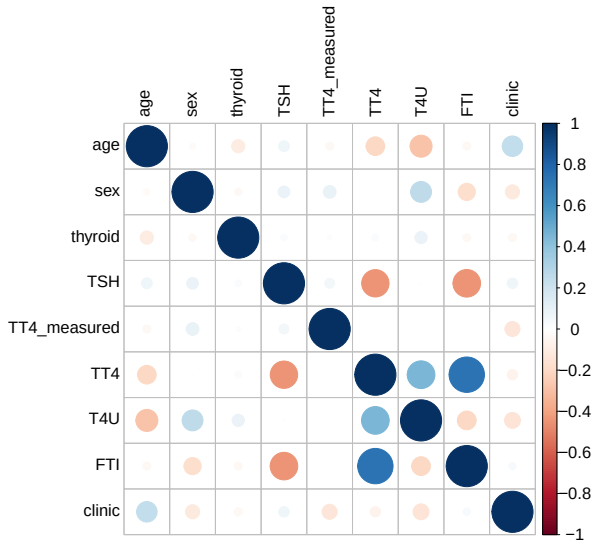
```
[1] "Covariance matrix is not positive definite."
[1] "It will be replaced with nearest positive definite matrix"
```

Figure 1 shows the original correlation matrix and the mean correlations for the three simulation approaches.

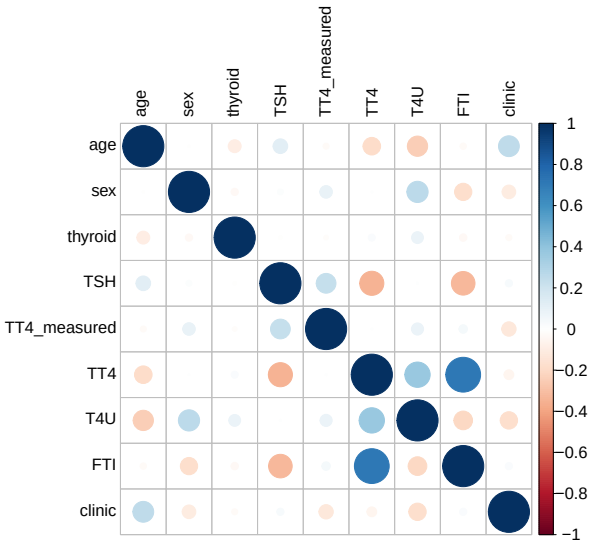
Figure 2 displays the differences between the original correlation matrix and the mean correlations for the three approaches. From the correlation plots, we can observe that the default *modgo* and *modgo* using nearest positive definite correlation matrix performs better than the combination of *modgo* with *sbgcop*.
Figure 3 shows the distribution for each selected variable for the original data and for one simulation for each of the three *modgo* runs. The figure shows that all three simulations mimic the original distribution well. This also holds true for the unordered categorical variable clinic.
Finally, Table 1 illustrates logistic regression results for the original data and the simulated data from the three simulation approaches. Thyroid disease was used as dependent variable, and age, TT4 (thyroxine test), FTU (Free Thyroxine Index), and sex were used as independent variables. Odds ratios are displayed for the original data. Median odds ratios and empirical 2.5% and 97.5% quantiles (in parenthesis) of the odds ratio estimates are displayed for each variable for the simulated data. While age, TT4, and FTI are close to the original values in the simulations, the variable sex shows a marked deviation when *sbgcop* is employed for estimating the correlation matrix. This deviation can also be noted in the correlation between thyroid and sex in Figure 2.

Figure 1: Correlation matrix plot for thyroid data. Displayed are the correlation matrices for the original data (top left), the mean correlation of *modgo* simulations (top right), the mean correlation of *modgo* simulations with selected variables when *sbgcop* was used for estimating the correlation matrix (bottom left) and the mean correlation of *modgo* simulations when all variables were used and the correlation matrix was calculated using the nearest positive definite matrix. The reader should note the light blue dots in the *sbgcop*/*modgo* plot for the variables sex and thyroid. Furthermore, no differences are visible between *modgo* when run on selected variables and *modgo* with nearest positive definite matrix when run on all variables.

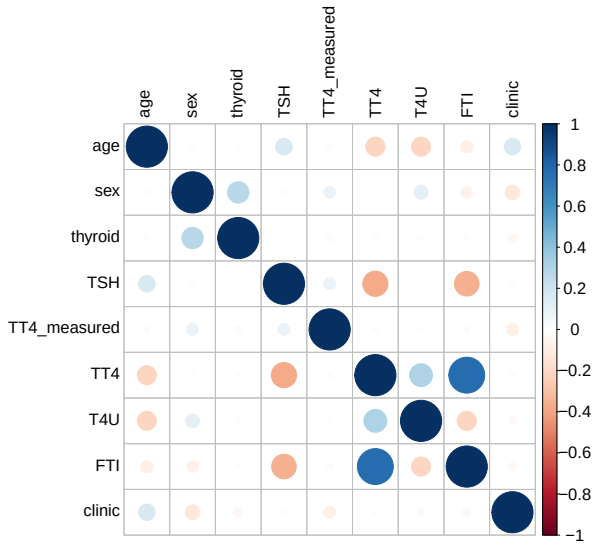
Original correlation



modgo mean correlation



modgo/sbgcop mean correlation



modgo/nearPD mean correlation

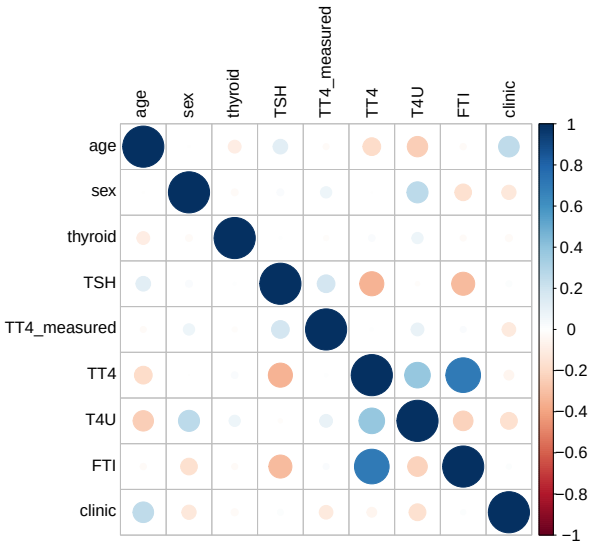


Figure 2: Differences of mean correlations between the correlation matrix of the original data and the three different *modgo* runs.

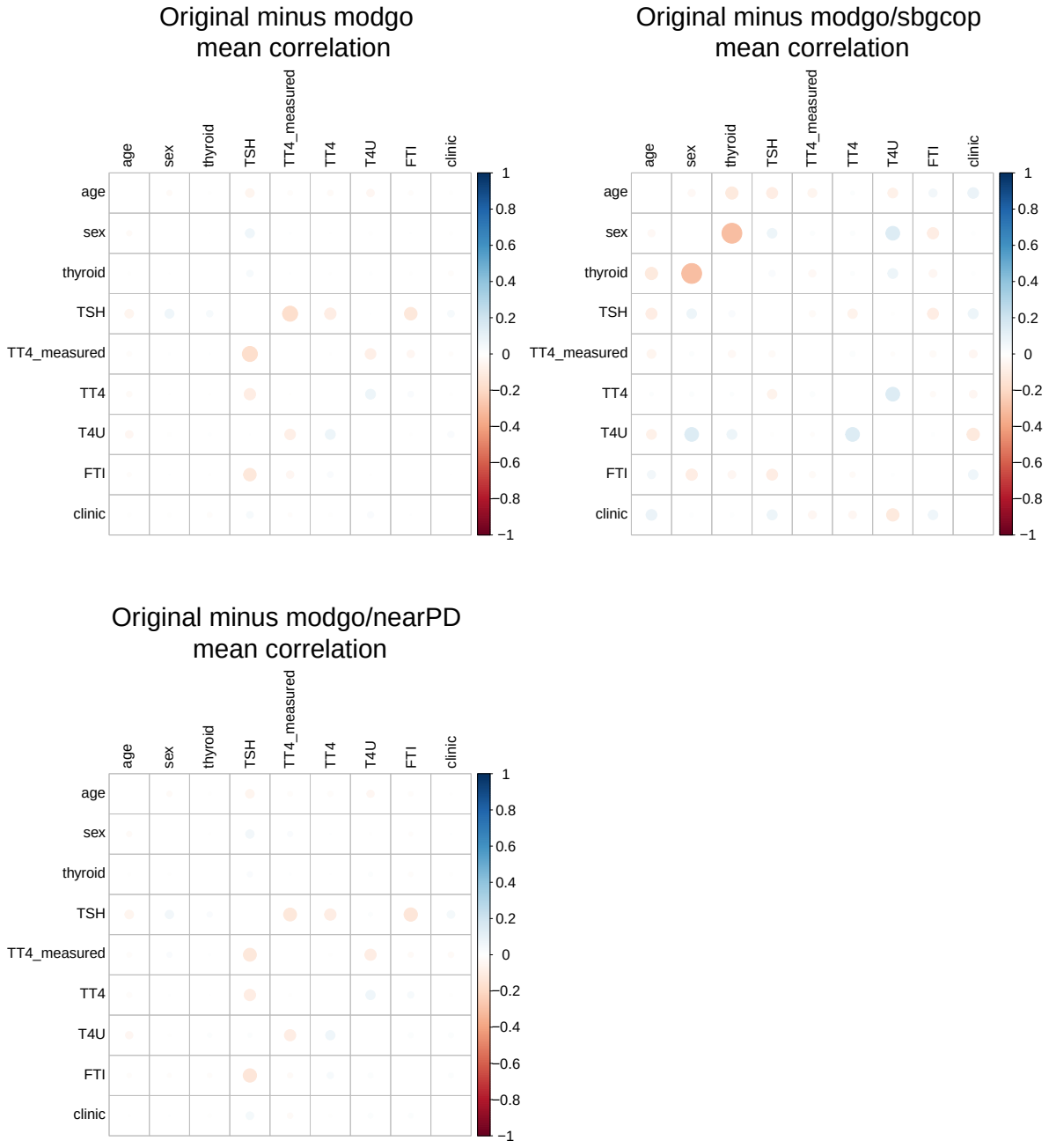


Figure 3: Distribution plot for each variable for the original data, simulated data from a default *modgo* run with the selected variables, a *modgo* run using the correlation matrix estimated by *sbgcop*, and a *modgo* run using all variables and the correlation matrix estiamted by nearest positive definite correlation. Box plots are used for continuous variables and bar plots for dichotomous and categorical variables.

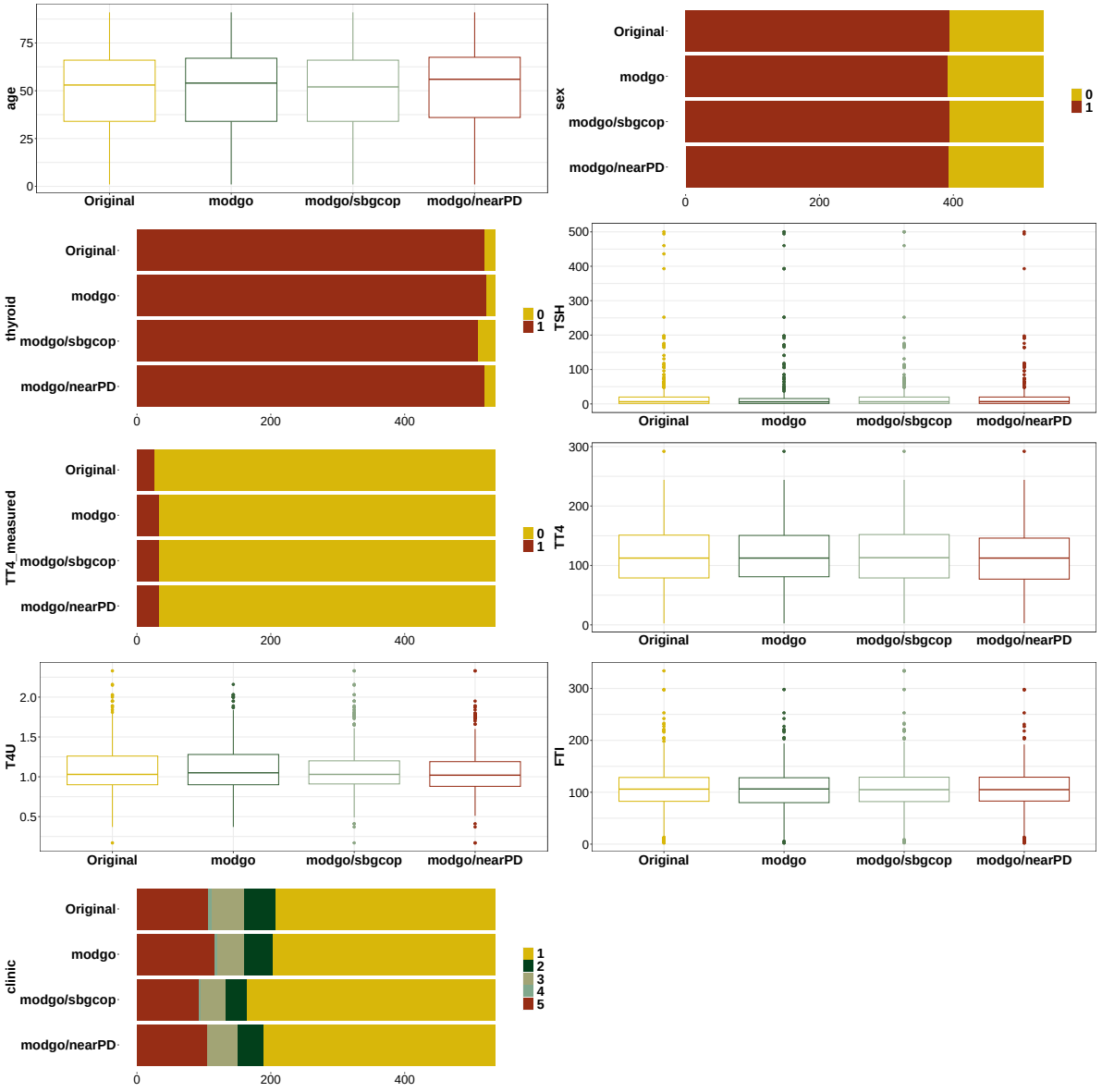


Table 1: Odds ratio estimates for the thyroid data. Displayed are estimates from the original data (original), the default *modgo* run (default), the combination *modgo* with *sbgcop*, and the *modgo* using nearest positive definite correlation matrix. Odds ratios are displayed for the original data, median odds ratios and empirical 2.5% and 97.5% quantiles for the odds ratio estimates are displayed in parenthesis for each variable included in the logistic regression. Results are shown for 500 simulated data sets for each *modgo* run.

	Original	modgo-default	modgo/sbgcop	modgo-nearPD
age	0.97	0.97 (0.93 - 1)	1 (0.97 - 1.04)	0.97 (0.93 - 1)
TT4	1.01	1.01 (0.99 - 1.03)	1 (0.98 - 1.02)	1.01 (0.99 - 1.02)
FTI	0.98	0.99 (0.97 - 1)	1.01 (0.99 - 1.02)	0.99 (0.98 - 1.01)
sex	0.49	0.46 (0 - 1.62)	70.87 (13.91 - 466651031.89)	0.51 (0 - 1.8)

2 Cleveland dataset

The Cleveland Clinic Heart Disease Data set from the UCI machine learning data repository served as second data set for illustration (Dua and Graff 2017). We selected 11 variables, 5 of which are continuous, four are dichotomous, and two ordinal categorical variables.

Three methods were used for simulations on the Cleveland data set:
1) *modgo* default run,
2) *modgo* with *sbgcop*'s correlation as an intermediate covariance matrix,
3) *SimMultiCorrData* package.

To simulate data sets that can mimic the original data using the package *SimMultiCorrData*, the user has to provide:
1) a vector with the means of all continuous variables,
2) a vector with the variances of all continuous variables,
3) a vector with the skewness values of all continuous variables,
4) a vector with the kurtosis values of all continuous variables,
5) a list with the cumulative probabilities for each categorical variable (categorical and dichotomous),
6) the original correlation matrix.

Figure 4 shows the original correlation and the mean correlations for the three simulation approaches. Figure 5 presents the difference between the original correlations and the mean correlations estimated for each simulation approach. Table 2 displays summary statistics for these differences. In Figure 6, distribution plots are provided for each variable for the original data and for one simulated data set for each simulation approach. A visual inspection of the distributions plots indicates that the three simulation approaches produce comparable simulated data. However, by examining the correlation plots and the table of summary statistics, it becomes apparent that *modgo* and *SimMultiCorrData* performed better than the combination of *modgo*/*sbgcop*.
Table 3 provides odds ratio estimates from logistic regression with coronary artery disease (CAD) as dependent variable for the original data and the three different simulation approaches. In this illustration, exercise-induced angina, age, max heart rate, and ST depression were used as independent variables. This results show that default *modgo* and *SimMultiCorrData* matched the original odds ratios well for all three independent variables, with *modgo* being slightly closer to the original odds ratio estimate. In contrast, *modgo*/*sbgcop* showed a substantial difference in odds ratios for exercise-induced angina.

```
binary_variables = c("Sex","HighFastBloodSugar","CAD","ExInducedAngina")
categorical_variables = c("Chestpaintype","RestingECG")
continuous_variables = setdiff(colnames(Cleveland),c(categorical_variables,
                                                     binary_variables))
```

```r
# SimMultiCorrData needs to order variables in the following way.
# We keep this order for all other methods
Cleveland <- Cleveland[,c(categorical_variables,binary_variables,
                          continuous_variables)]
sbgcop_approach_Cleve <- sbgcop.mcmc(Y = as.matrix(Cleveland), nsamp = 1,
                            verb = FALSE)
```

```r
Cleve_modgo <- modgo(Cleveland,bin_variables = binary_variables,
                     categ_variables = categorical_variables,nrep = nrep)


Cleve_sbgcop_modgo <- modgo(Cleveland,sigma = sbgcop_approach_Cleve$C.psamp[,,1],
                          bin_variables = binary_variables,
                          categ_variables = categorical_variables,nrep = nrep)
```

Figure 4: Correlation matrix plots for the Cleveland Clinic data. Displayed are the correlation matrices as heat maps for the original data (top left), the mean correlation of *modgo* simulations (top right), the mean correlation matrix of *modgo* simulations when *sbgcop* was used for estimating the correlation matrix (bottom left) and the mean correlation from simulations with *SimMultiCorrData* (bottom right).

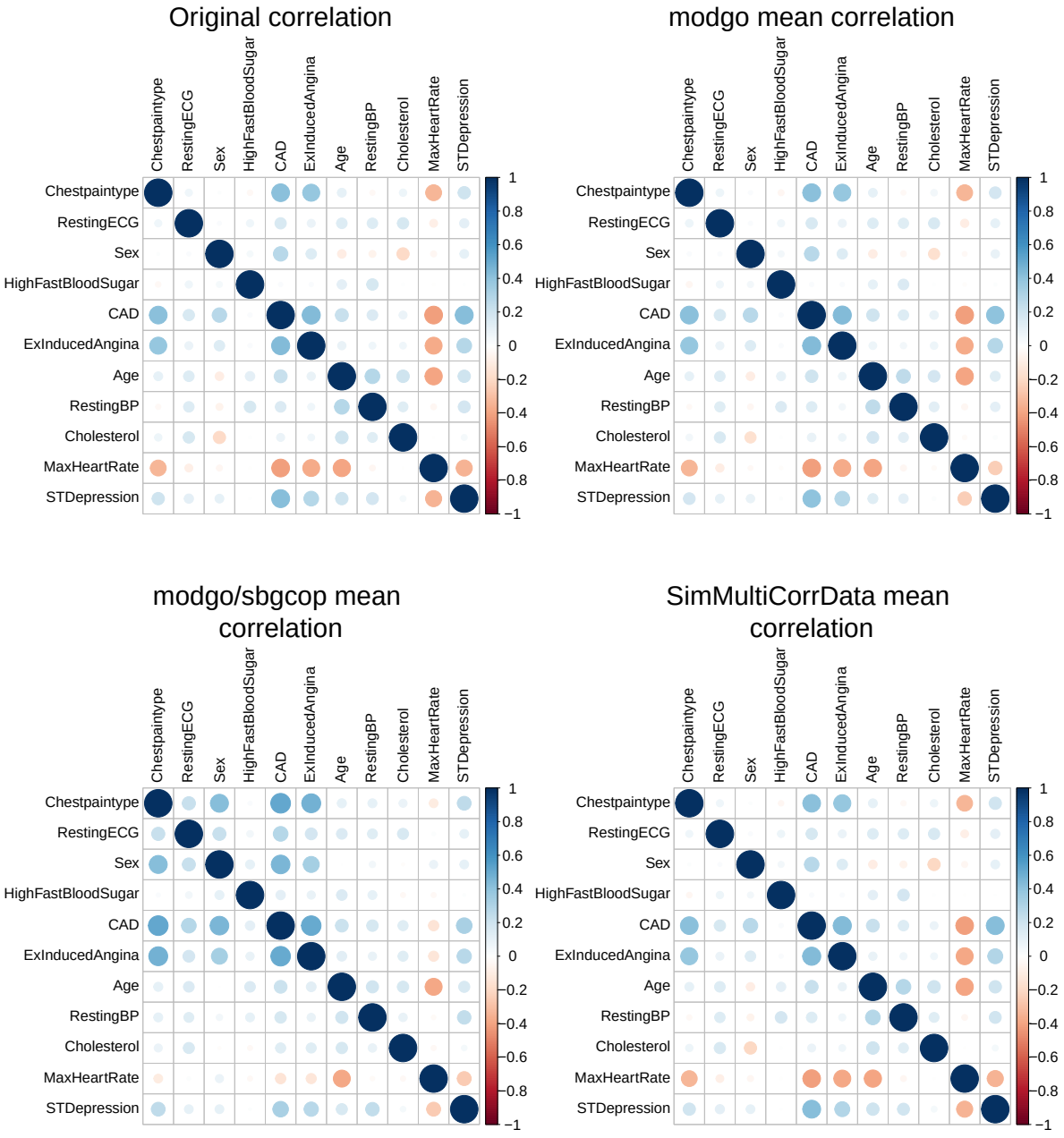

Original correlation
modgo mean correlation
modgo/sbgcop mean correlation
SimMultiCorrData mean correlation
Chestpaintype
RestingECG
Sex
HighFastBloodSugar
CAD
ExInducedAngina
Age
RestingBP
Cholesterol
MaxHeartRate
STDepression
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 2: Differences of mean correlations between the correlation matrix of the original data and the two different *modgo* runs and the *SimMultiCorrData* simulation runs.

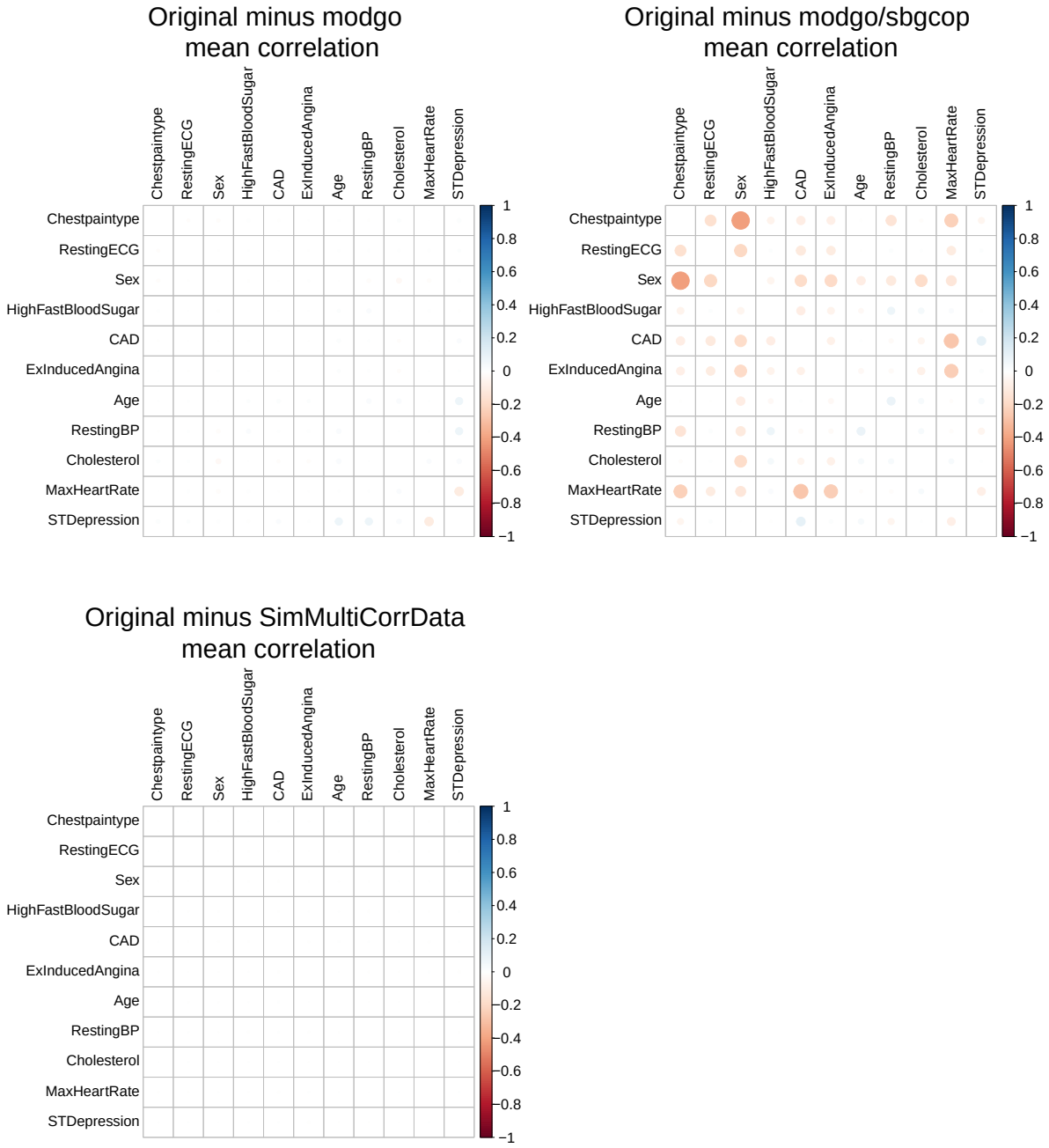


Table 2: Descriptive statistics for the correlations of the three simulation approaches when applied to the Cleveland Clinic data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
modgo	-0.0071874	0.0012243	0.0025551	0.0037233	0.0078056	0.0162565
modgo/sbgcop	-0.1451253	-0.0869803	-0.0583012	-0.0558512	-0.0151673	-0.0012993
SimMultiCorrData	-0.0012614	-0.0004757	-0.0002910	-0.0002469	0.0000504	0.0005090

Figure 6: Distribution plot for each variable for the original data, simulated data from a default *modgo* run with the selected variables, a *modgo* run using the correlation matrix estimated by *sbgcop*, and a *SimMultiCorrData* run. Box plots are used for continuous variables and bar plots for dichotomous and categorical variables.

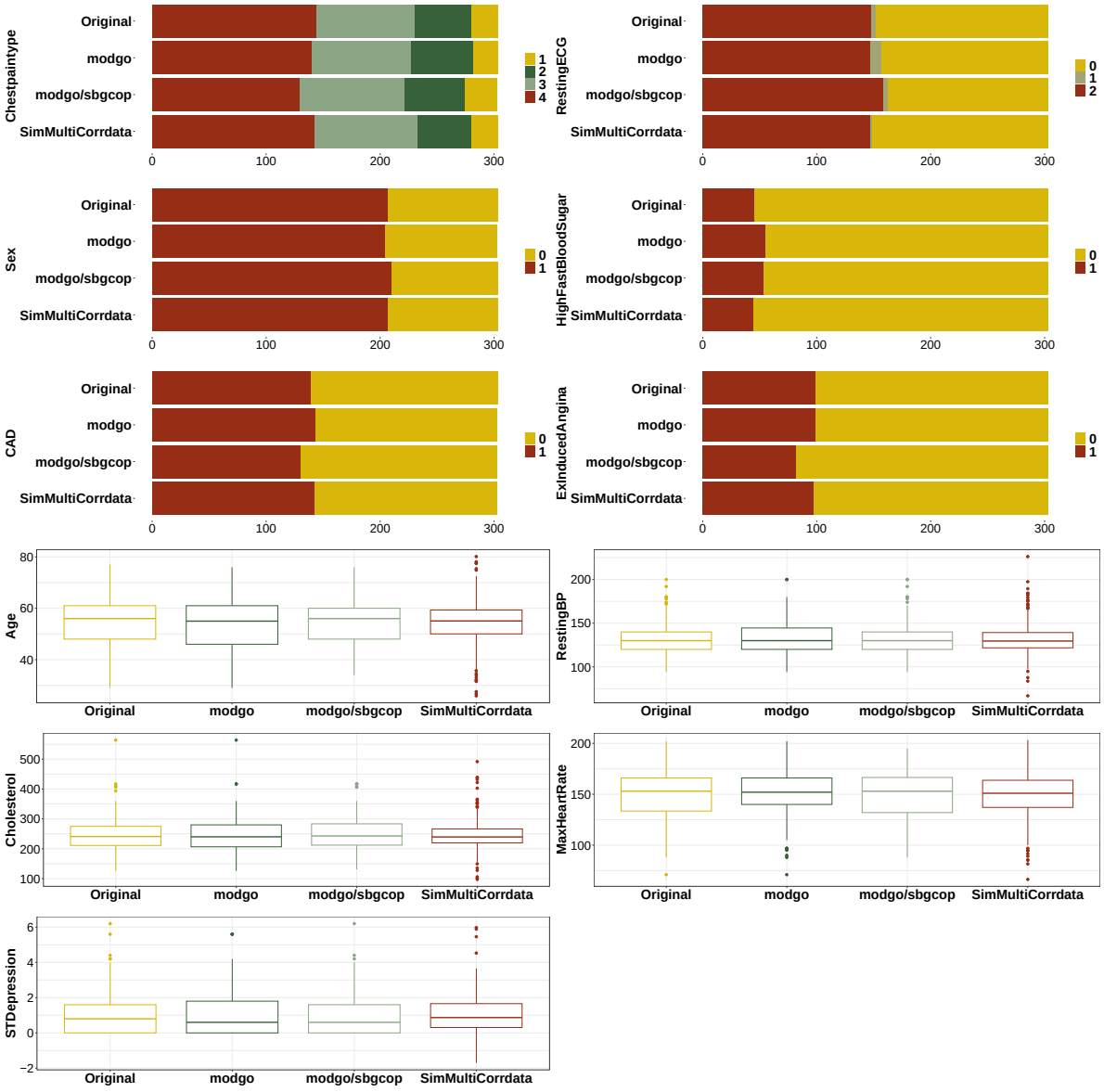


Table 3 Odds ratio estimates for the Cleveland Clinic Data. Displayed are estimates from the original data (original), the default *modgo* run (default), the combination of *modgo* and *sbgcop* and the use of *SimMultiCorrData*. Odds ratios are displayed for the original data, median odds ratios and empirical 2.5% and 97.5% quantiles for the odds ratio estimates are displayed in parenthesis for each variable included in the logistic regression. Results are shown for 500 simulated data sets for each *modgo* run.

	Original	modgo-default	modgo/sbgcop	SimMultiCorrData
ExInducedAngina	4.43	4.45 (2.38 - 8.17)	10.16 (5.52 - 19.26)	4.42 (2.58 - 8.61)
Age	1.02	1.02 (0.98 - 1.05)	1.05 (1.01 - 1.08)	1.02 (1 - 1.05)
MaxHeartRate	0.98	0.97 (0.95 - 0.99)	1 (0.99 - 1.02)	0.97 (0.96 - 0.99)
STDepression	1.92	2.07 (1.52 - 2.86)	1.6 (1.26 - 2.23)	2.2 (1.69 - 2.97)

3 Heart failure dataset

The final example for illustration is based on the Heart Failure data set from the UCI machine learning data repository (Dua and Graff 2017). It consists of 299 subjects and 13 variables, 6 of which are dichotomous, and 7 are continuous. Three simulation methods were used in this example:
1) *modgo*,
2) *modgo* with the correlation matrix estimated by *sbgcop*,
3) *BinNor* (Demirtas, Amatya, and Doganay 2014).
Each method was run to generated 500 simulated data sets.

```
heartFail_data <- read.table(heartFail_link, sep = ",",header = TRUE)
binary_variables = c("anaemia","diabetes","high_blood_pressure","sex"
                     ,"smoking","DEATH_EVENT")

#BinNor needs to order variables in the following way.
#We keep this order for all other methods
ordered_variable <- c("anaemia","diabetes","high_blood_pressure","sex","smoking"
                      , "DEATH_EVENT","age","creatinine_phosphokinase",
                        "ejection_fraction","platelets","serum_creatinine",
                        "serum_sodium","time")
heartFail_data <- heartFail_data[,ordered_variable]
sbgcop_approach_heartFail <- sbgcop.mcmc(Y = as.matrix(heartFail_data), nsamp = 1,
                             verb = FALSE)
```

```
# modgo default run
heartFail_modgo <- modgo(heartFail_data,bin_variables = binary_variables,nrep=nrep)

# modgo/sbgcop default run
heartFail_sbgcop_modgo <- modgo(heartFail_data,
                                sigma = sbgcop_approach_heartFail$C.psamp[,,1],
                                bin_variables = binary_variables,nrep=nrep)
```

To simulate datasets with the package *BinNor*, the user has to provide
1) a vector with the means of all continuous variables,
2) a vector with the variances of all continuous variables,
3) a list with the cumulative probabilities for each binary variable,
4) the original correlation matrix.

Figure 7 shows the original correlation and the mean correlations of the three simulation approaches. Figure 8 presents the difference of mean correlation for each approach from the original correlation. Summary statistics for these differences are displayed in Table 4. Figure

9 provides the distribution plots for each variable for the original data and for one data set for each simulation approach. Although all three simulation methods seems to be close to the original data set when the distributions plots are visually inspected, both the correlation plots and the summary statistics show that *modgo* and *BinNor* outperformed the combination of *modgo*/*sbgcop* combination. Table 5 shows the odds ratios for logistic regressions with death (yes/no) as dependent variable for the original data and for the three simulation approaches. Hypertension (high blood pressure), smoking, serum creatinine and time were used as independent variables. The *modgo*/*sbgcop* odds ratios substantially differed from the original odds ratios. For serum creatinine, default *modgo* showed the largest variability and the largest deviation from the original odds ratio estimate, and *BinNor* had the smallest variability and the smallest difference.

Figure 7: Correlation matrix plots for the Heart Failure data. Displayed are the correlation matrices as heat maps for the original data (top left), the mean correlation of *modgo* simulations (top right), the mean correlation matrix of *modgo* simulations when *sbgcop* was used for estimating the correlation matrix (bottom left) and the mean correlation from simulations with *BinNor* (bottom right).

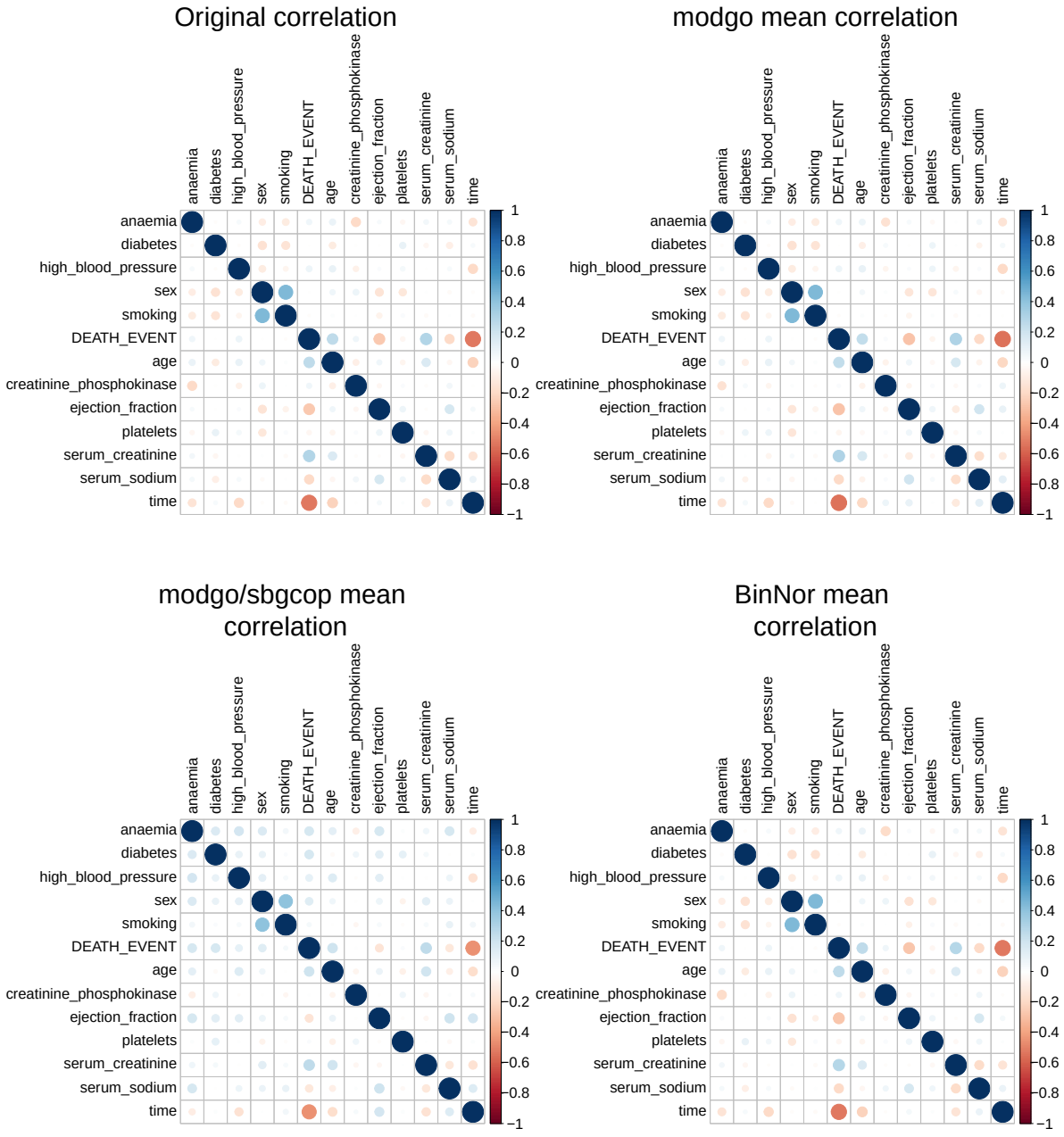

Original correlation
modgo mean correlation
modgo/sbgcop mean correlation
BinNor mean correlation
anaemia
diabetes
high_blood_pressure
sex
smoking
DEATH_EVENT
age
creatinine_phosphokinase
ejection_fraction
platelets
serum_creatinine
serum_sodium
time
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 8: Differences of mean correlations between the correlation matrix of the original data and the two different *modgo* runs and the *BinNor* simulation runs.

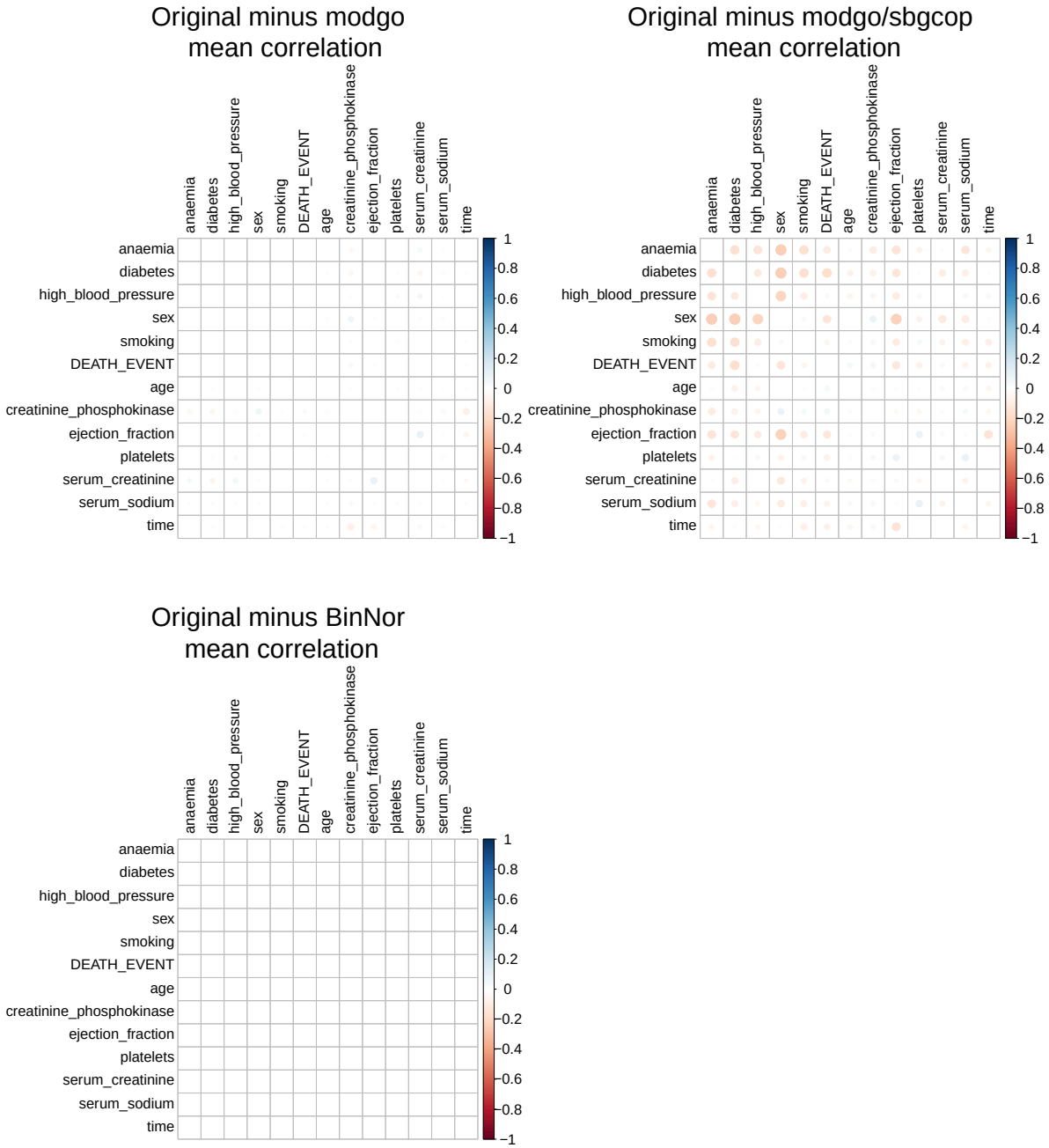


Table 4: Descriptive statistics for the correlations of the three simulation approaches when applied to the Heart Failure data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
modgo	-0.0166395	-0.0027437	0.0011207	-0.0006726	0.0025698	0.0059735
modgo/sbgcop	-0.1051556	-0.0725674	-0.0513956	-0.0516293	-0.0243134	-0.0017070
BinNor	-0.0011186	-0.0003370	0.0001492	0.0000481	0.0007070	0.0011571

Figure 9: Distribution plot for each variable for the original data of the Heart Failure data, simulated data from a default *modgo* run with the selected variables, a *modgo* run using the correlation matrix estimated by *sbgcop*, and a *BinNor* run. Box plots are used for continuous variables and bar plots for dichotomous and categorical variables.

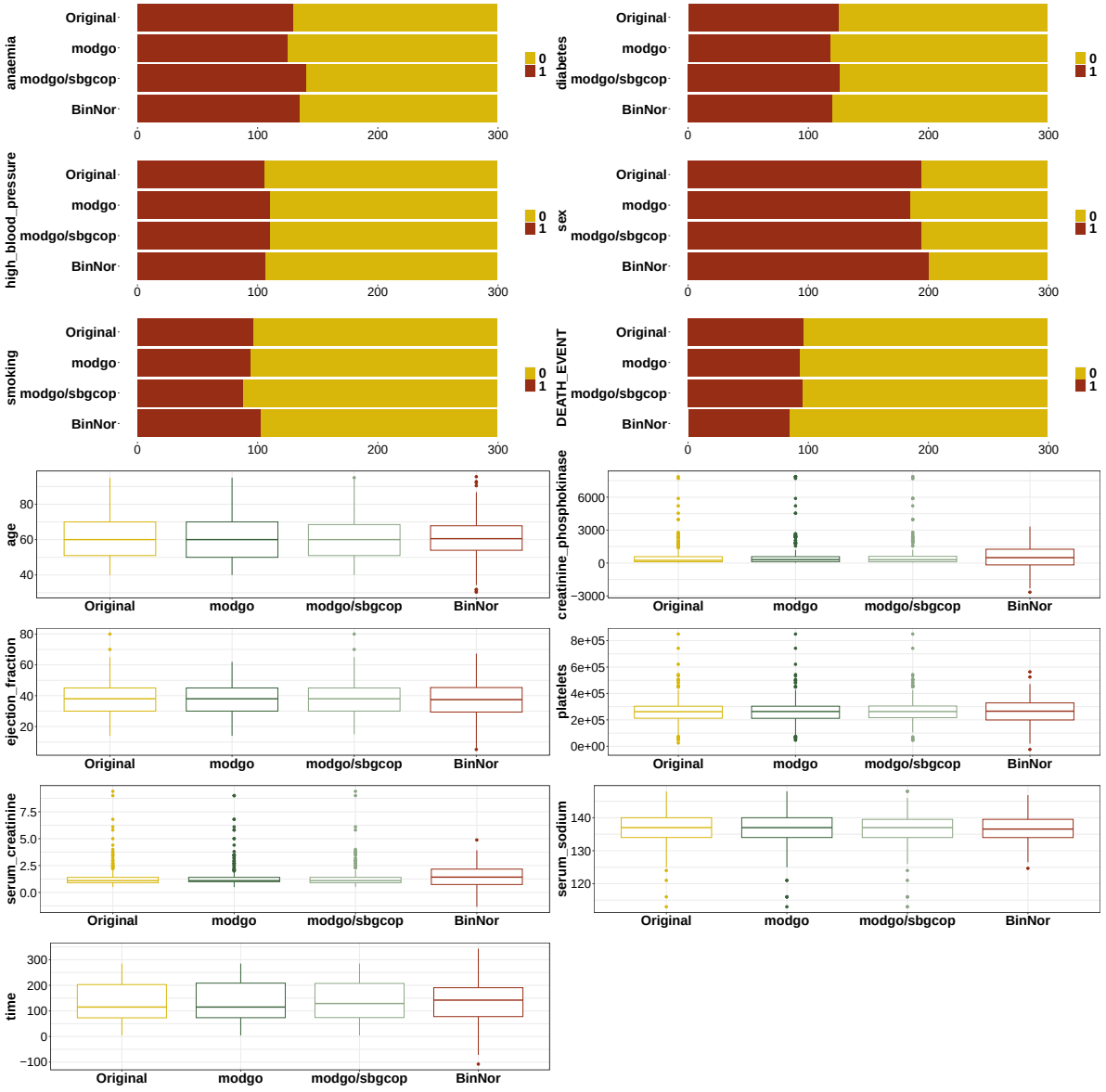


Table 5: Odds ratio estimates for the Heart Fail Data. Displayed are estimates from the original data (original), the default *modgo* run (default), the combination *modgo* with *sbgcop*, and the *BinNor*. Odds ratios are displayed for the original data, median odds ratios and empirical 2.5% and 97.5% quantiles for the odds ratio estimates are displayed in parenthesis for each variable included in the logistic regression. Results are shown for 500 simulated data sets for each *modgo* run.

	Original	modgo-default	modgo/sbgcop	BinCorrData
high_blood_pressure	0.97	0.93 (0.45 - 1.69)	1.29 (0.68 - 2.27)	0.87 (0.44 - 1.75)
smoking	0.99	0.92 (0.44 - 1.87)	1.45 (0.68 - 2.71)	0.84 (0.44 - 1.61)
serum_creatinine	2.15	2.88 (1.66 - 7.39)	1.76 (1.21 - 3.4)	2.05 (1.51 - 2.85)
time	0.98	0.98 (0.97 - 0.98)	0.98 (0.98 - 0.99)	0.98 (0.97 - 0.98)

References

Demirtas, Hakan, Anup Amatya, and Beyza Doganay. 2014. "BinNor: An r Package for Concurrent Generation of Binary and Normal Data." *Communications in Statistics - Simulation and Computation* 43 (3): 569–79. https://doi.org/10.1080/03610918.2012.707725.

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Higham, Nick, Rüdiger Borsdorf, and Marcos Raydan. n.d. "Computing a Nearest Correlation Matrix with Factor Structure."

Hoff, Peter D. 2007. "Extending the Rank Likelihood for Semiparametric Copula Estimation." *The Annals of Applied Statistics* 1 (1): 265–83.