**Supplemental Material for**

**“Covariate Selection in Causal Learning under Non-Gaussianity”**

**Additional Simulation Results**

In this supplement, we present results of additional Monte-Carlo simulation experiments that evaluated the performance of the nGFS algorithm. Specifically, 1) we provide a more fine-grained analysis of the effect the skewness of model errors has on the accuracy of nGFS through considering a wider range of levels of asymmetry, 2) we evaluate the impact of the magnitude of covariate asymmetry on nGFS model selection, 3) we describe the behavior of nGFS for various nominal significance levels of Wald, Gaussianity, and independence tests, and 4) we illustrate potential adverse effects of ignoring the significance status of a considered covariate in the multiple linear regression model (i.e., performing model search without considering OLS Wald tests to evaluate the significance of the considered covariate).

**Performance of nGFS as a Function of Error Asymmetry**

In the following section, we report results of additional simulation experiments that considered a wider and more fine-grained range of asymmetry of the model errors. While, in the main experiment, the skewness of errors was varied from 0 to 2.25 in increments of 0.75, in the present simulation experiment, the error skewnesses were set to 0, 0.5, 1, 1.5, 2, 2.5, and 3 for selected samples sizes of *N* = 100 and 250 (reflecting small to moderate sample sizes). Similar to the Monte-Carlo simulation study in the main text, we considered four types of continuous covariates (i.e., confounder: *x* ← *z* → *y*, control variable: *z* → *y* and no *x* relation, collider: *x* → *z* ← *y*, and uncorrelated variable: no *x* relation and *y* relation). Confounders, control variables, and uncorrelated variables were randomly drawn from gamma-distributed populations exhibiting a skewness of 0.75. The OLS Wald test was applied to evaluate the significance of a covariate, the Lilliefors test was applied to evaluate the non-Gaussianity assumption, and HSIC and non-linear correlation tests were used to test the independence of model residuals. All significance tests were performed using a nominal significance level of 5%.

Figure A1 summarizes the accuracy of nGFS to correctly classify confounders, colliders, control variables, and uncorrelated variables as a function of the skewness of *rx*, the skewness of *ry*, and sample size. In line with the results obtained in our main simulation experiment, the power to correctly identify confounders and control variables increases with the skewness of *rx*. Here, for *N* = 100 cases, a skewness of = 1.5 is required to achieve acceptable accuracy rates. For *N* = 250 observations, a skewness of = 1 is sufficient to guarantee acceptable rates of correct decisions. In contrast, rates of correctly classifying collider variables decrease with the skewness of *rx* and increase with the skewness of *ry*. Here, non-linear correlation tests tend to outperform the HSIC when the skewness of *ry* is small. In the latter case, for *N* = 250, power curves of the HSIC tend to follow a U-shaped pattern indicating that highly skewed residuals *rx* can mitigate the loss of power. Overall, we can conclude that, for small to moderate sample sizes, larger deviations from error symmetry are required to ensure adequate performance.

**A picture containing text, diagram, line, plot

Description automatically generated**

*Figure A1.* Performance of nGFS for different types of continuous covariates with various levels of asymmetry of *rx* and *ry*.

**Impact of Covariate Asymmetry**

In the following section, we focus on the effect the distributional shapes of confounders, control variables, and uncorrelated variables have on nGFS model selection. In the main simulation experiment, confounders, control variables, and uncorrelated variables were drawn from gamma distributions with a skewness of 0.75 (i.e., cases of small covariate skewness). In the present simulation experiment, we investigate whether the distributional shape of covariates has an impact of nGFS’s accuracy. For this purpose, we repeated the main simulation study (focusing on continuous covariates) with heavily skewed covariates. That is, confounders, control variables, and uncorrelated variables were generated from a gamma-distributed population exhibiting a skewness of 2. To ensure comparability, remaining simulation factors were the same as in the main experiment, that is, sample sizes were *N* = 50, 100, 250, and 500, and residual (*rx* and *ry*) skewnesses were set to 0, 0.75, 1.5, and 2.25. OLS Wald tests were applied to evaluate the significance status of a considered covariate, the Lilliefors test was used to test the non-Gaussianity assumption, and HSIC and non-linear correlation tests were used to test the independence of model residuals. All significance tests were performed using a nominal significance level of 5%.

nGFS accuracy rates (based on 1000 resamples per simulation condition) of correctly classifying the four considered types of covariates are summarized in Figure A2. Overall, results suggest that selection rates for confounders, control variables, and uncorrelated variables are not affected by the distributional shape of the variables. However, accuracy rates for collider detection dramatically improve, in particular, when the skewness of *ry* is low. The reason for this is that the distribution of a collider is indirectly affected by the distribution of confounder and control variables, that is, larger magnitudes of asymmetry of confounders and control variables lead to more skewed *x* and *y* variables which, in turn, result in a more skewed collider variable. Thus, we can conclude that the presence of highly skewed causes of *x* and *y* can improve the performance of nGFS to detect collider biases.

**A picture containing text, diagram, line, map

Description automatically generated**

*Figure A2.* Performance of nGFS for heavily skewed confounders, control variables, and uncorrelated variables ( = 2).

**Impact of Nominal Significance Thresholds**

The nGFS algorithm uses three significance tests to determine the status of a covariate: The OLS Wald test to determine the significance of the covariate, the Gaussianity test to check the non-Gaussianity assumption, and an independence procedure to test error independence. So far, the default nominal significance threshold for the three tests was set to 5%. In the present section, we evaluate the influence of a priori selecting the nominal significance level for the three tests. For this purpose, three separate Monte-Carlo simulation experiments were conducted in which we systematically varied the significance threshold for one test (i.e., either using 1%, 5%, or 10%) while holding the significance thresholds for the remaining two significance procedures at 5%. Sample sizes were fixed at either *N* = 100 or 250, reflecting small to moderate sample sizes. Skewnesses of error terms were set to 0, 0.75, 1.5, and 2.25, and continuous covariates (i.e., confounders, control variables, and uncorrelated variables) were sampled from a gamma distribution with a skewness of 0.75. Two types of independence tests were applied (HSIC and non-linear correlations).

Figure A3 summarizes the performance of nGFS when varying the significance threshold for the Gaussianity test. In general, when using a lower significance threshold (i.e., 1% instead of 5%), the power of the Gaussianity tests to confirm sufficient non-Gaussianity of *rx* decreases, as expected. As a consequence, compared to a nominal threshold of 5%, selection rates for confounders and control variables tend to be lower when *rx* is slightly skewed. However, for moderate to strong deviations from Gaussianity, accuracy rates are not affected by the lower significance threshold. In contrast, exclusion rates for colliders tend to increase with a lower significance threshold for the Gaussianity test. This can be explained by the fact that colliders tend to be excluded due to a higher chance of retaining the null hypothesis of error Gaussianity.

When increasing the nominal significance level of the Gaussianity test to 10%, rates of confounder and control variable detection tend to be higher than the ones observed for a significance threshold of 5%. The reason for this is that the Gaussianity test accepts less skewed confounders and control variables which, however, are still sufficiently non-Gaussian to detect systematic error dependence. Collider detection tends to be unaffected by the higher significance threshold.

**A picture containing text, diagram, line, map

Description automatically generated**

*Figure A3.* Performance of nGFS when performing Gaussianity tests with different significance thresholds.

Next, we focus on the performance of nGFS when changing the significance threshold of the independence procedure to evaluate the independence of the error terms (see Figure A4). Recall that, in the present context, one is interested in retaining the null hypothesis (indicating that the causal estimate of interest is consistent). For a nominal significance level 1% (instead of 5%) one, therefore, observes higher rates of correctly classifying confounders and control variables than in the case of a nominal significance level of 5%. In contrast, collider detection rates tend to be lower than in the main analysis due to the lower power of independence tests.

A reversed picture can be observed when increasing the significance threshold of the independence tests to 10%. Here, compared to a significance level of 5%, selection rates for confounders and control variables tend to be lower and rates of detecting colliders tends to increase. The reason for this is that one now has a higher chance to reject the null hypothesis of error independence and it is, therefore, more likely to detect a collider bias and less likely to retain the null hypothesis of independence (which is required to detect confounders and control variables).

**A picture containing text, diagram, line, map

Description automatically generated**

*Figure A4*. Performance of nGFS when performing independence tests with different significance thresholds.

Finally, we turn to the effects of varying the significance threshold for testing the significance of a covariate using the OLS Wald test (see Figure A5). Here, changing the significance threshold does not have an impact on detecting confounders, colliders, and control variables. However, the choice of the significance threshold can have small effects on the decision rates for uncorrelated variables. Here, accuracy rates slightly drop when the nominal significance level is set to 10%.

Overall, we can conclude that the choice of the significance thresholds of the three tests can impact the accuracy of nGFS to distinguish confounders, control variables, colliders, and uncorrelated variables to some degrees. While a global significance level of 5% constitutes a reasonable choice for the majority of considered data conditions, the performance of nGFS may further be improved, depending on the goal of the analysis. If, for example, one prioritizes the detection of confounders and control variables, a larger nominal significance threshold for the Gaussianity test (e.g., 10% instead of 5%) and a lower threshold for the independence test (e.g., 1% instead of 5%) are preferable. In contrast, when one wishes to be sensitive to potential collider biases, a lower significance threshold for the Gaussianity test and a higher threshold for testing independence should be preferred. Note, however, that the accuracy of nGFS tends to be unaffected by the choice of the significance threshold when sample sizes are large and data strongly deviate from Gaussianity.

**A picture containing text, diagram, line, parallel

Description automatically generated**

*Figure A5*. Performance of nGFS when performing OLS Wald tests with different significance thresholds.

**Potential Adverse Effects of Omitting OLS Significance Testing**

One of the proposed extensions of Entner et al.’s (2012) original approach concerns the inclusion of OLS Wald tests to determine the significance status of a considered covariate. To illustrate the importance of testing the significance status of a covariate using OLS Wald tests in the phase of model building, we repeated the Monte-Carlo simulation experiment reported in the main text using a simplified version of the nGFS algorithm in which covariates are not evaluated with respect to their status of significance. In other words, in this simulation study, nGFS model selection solely relied on Gaussianity and independence tests. To ensure comparability, we used the same simulation design as in the main experiment, that is, sample sizes were set to *N* = 50, 100, 250, and 500, and the skewnesses of *rx* and *ry* were 0.75, 1.5, and 2.25. Covariates were generated from a gamma distribution with a skewness of 0.75. The simulation factors were fully crossed, and one thousand samples were generated for each simulation condition. The Lilliefors test was applied to evaluate the non-Gaussianity assumption, and HSIC and non-linear correlation tests were used to test the independence of model residuals. All significance tests were performed using a nominal significance level of 5%.

Figure A6 gives the accuracy rates of nGFS to correctly classify confounders (*x* ← *z* → *y*), control variables (*z* → *y* and no *x* relation), colliders (*x* → *z* ← *y*), and uncorrelated variables (no *x* relation and *y* relation). Overall, nGFS’s performance to correctly include confounders and control variables and correctly exclude colliders is unaffected by omitting OLS significance testing. However, rates of correctly excluding the uncorrelated variable systematically decrease with the skewness of *rx* and approach zero, in particular, for larger sample sizes. In other words, as sample size and the magnitude of non-Gaussianity increase, the uncorrelated variable fulfills nGFS’s non-Gaussianity and independence requirements and is, therefore, falsely retained in the set of admissible covariates. In contrast, accuracy rates for correctly excluding uncorrelated variables remain close to 80% for slightly asymmetric residuals *rx* and small samples (e.g., *N* = 50). The reason for this is that, under these data conditions, the Gaussianity test tends to retain the null hypothesis of Gaussianity and, therefore, excludes the variable from further consideration. In other words, these reflect cases in which the uncorrelated variable is correctly classified, but for the wrong reason. Overall, these simulation results emphasize the necessity to extend Entner et al.’s (2012) original procedure to incorporating the significance status of a covariate during model building.

A picture containing text, diagram, line, parallel

Description automatically generated

*Figure A6.* Performance of nGFS for different types of continuous covariates without the significance test of regression coefficients.