Conditional Independence Testing with Heteroskedastic Data and Applications to Causal Discovery

Wiebke Günther
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
wiebke.guenther@dlr.de

Urmi Ninad *
Technische Universität Berlin
10623 Berlin, Germany
urmi.ninad@tu-berlin.de

Jonas Wahl *
Technische Universität Berlin
10623 Berlin, Germany
wahl@tu-berlin.de

Jakob Runge
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
and
Technische Universität Berlin
10623 Berlin, Germany
jakob.runge@dlr.de

Abstract

Conditional independence (CI) testing is frequently used in data analysis and machine learning for various scientific fields and it forms the basis of constraint-based causal discovery. Oftentimes, CI testing relies on strong, rather unrealistic assumptions. One of these assumptions is homoskedasticity, in other words, a constant conditional variance is assumed. We frame heteroskedasticity in a structural causal model framework and present an adaptation of the partial correlation CI test that works well in the presence of heteroskedastic noise, given that expert knowledge about the heteroskedastic relationships is available. Further, we provide theoretical consistency results for the proposed CI test which carry over to causal discovery under certain assumptions. Numerical causal discovery experiments demonstrate that the adapted partial correlation CI test outperforms the standard test in the presence of heteroskedasticity and is on par for the homoskedastic case. Finally, we discuss the general challenges and limits as to how expert knowledge about heteroskedasticity can be accounted for in causal discovery.

1 Introduction

Conditional independence (CI) testing is a frequently used step across a wide range of machine learning tasks for various scientific fields. It is also very challenging. Discovering causal relationships from purely observational data is an even more challenging task and an important topic in sciences where real experiments are infeasible, e.g. in climate research [Ebert-Uphoff and Deng, 2012, Runge et al., 2019a]. One can distinguish several frameworks that address this problem: score-based approaches [Chickering, 2002], restricted structural causal models [Peters et al., 2017], and constraint-based methods [Spirtes et al., 2000] that rely on CI testing. A typical representative is the PC algorithm [Spirtes and Glymour, 1991] which can be combined with any conditional independence

*equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

test and is thus adaptable to a wide range of data distributions. It utilizes the Faithfulness assumption to conclude that no causal link can exist between two variables X and Y if a CI test suggests that they are independent given a set Z.

When a linear additive noise model may be assumed, a popular CI test is the partial correlation test [Lawrance, 1976] which has the advantages of fast computation time and that the null distribution is known analytically. As a disadvantage rather strict and often unrealistic assumptions have to be satisfied. One of these assumptions is homoskedasticity, meaning that the variance of the error term is constant. When this assumption is violated, i.e., in the case of heteroskedasticity, the variance can, for instance, depend on the sampling index or the value of one or multiple influencing variables.

In regression analysis one distinguishes between *impure* and *pure* heteroskedasticity. Impure heteroskedasticity stems from an insufficient model specification, e.g. if unobserved variables or confounders are present. Another reason might be that the model fails to capture the full relationships of the variables, e.g. the underlying model might be linear with multiplicative noise rather than additive. On the other hand, pure heteroskedasticity is non-constant noise variance that is present despite a correct model.

The sources for heteroskedasticity in real data are manifold. For example, in environmental sciences precipitation in different areas might exhibit different variances that are unaccounted for by other variables in the system, i.e. location-scaled noise. Such a problem could be introduced by aggregating data of different catchments and not adding a variable that is well enough correlated with catchment location [Merz et al., 2021]. An example for sampling index-dependent heteroskedasticity in the time series case are seasonal effects that are present in many climate variables [Proietti, 2004]. Finally, an example where the noise variance of a variable is dependent on an observed cause is the distance to sea influencing the variability of temperature. This is a special case of state-dependent noise.

One assumption of the PC algorithm is causal sufficiency, meaning that there are no unobserved confounders, formally excluding impure heteroskedasticity. However, in practice this assumption is often violated which can lead to heteroskedastic noise.

If unaccounted heteroskedasticity is present in the data, the estimator of the ordinary least squares (OLS) regression slope parameter is still unbiased. However, the estimator of the covariance matrix of the parameter estimates can be biased and inconsistent under heteroskedasticity [Long and Ervin, 2000]. This can skew subsequent partial correlation significance tests and affect the link detection rate of a causal discovery method. Furthermore, heteroskedasticity might even lead to the detection of wrong links (false positives). Moreover, the Gauss-Markov theorem assumes homoskedasticity and, hence, with heteroskedastic data the OLS slope estimator is no longer guaranteed to be the most efficient linear unbiased estimator, which can further harm power of the CI test.

There are multiple ways to treat heteroskedasticity, either already during the modeling step by addressing possible model misspecification, by pre-processing the data, e.g. by applying a $\log$-transform, or post-hoc using robust statistics for CI testing. Adapting the model, for example, by using CI tests allowing for multiplicative dependencies [Runge, 2018] might not always be feasible for limited sample sizes. Pre-processing in real-world problems also comes with drawbacks: The transformed variables are difficult to interpret and can introduce dependencies or spurious links.

In this work we propose an adapted weighted least-squares (WLS) partial correlation variant as a CI test for the PC algorithm that is able to deal with particular forms of heteroskedasticity. **Our contributions** are theoretical consistency results as well as numerical experiments demonstrating that this approach yields well-calibrated CI tests leading to controlled false positive rates and also improves upon detection power as compared to the standard partial correlation CI test. Our approach requires expert knowledge in that it needs to be known which of the variables the heteroskedasticity depends on, or if it depends on the sampling index.

2 Related Work

The effect of heteroskedasticity on the standard Pearson correlation test has been investigated, for example, by Wilcox and Muska [2001]. Remedies for heteroskedasticity also have been extensively studied. Hayes [2007] discuss and evaluate heteroskedasticity-consistent standard errors. Practical approaches to choose weights for WLS, consistency and asymptotic results have been obtained, e.g. by Neumann [1994], Fan and Yao [1998], Carroll [1982], Robinson [1987], Brown and Levine [2007].

Romano and Wolf [2017] propose a method that combines WLS with heteroskedasticity-consistent standard errors.

Recently, within the causal discovery framework of restricted structural causal models [Peters et al., 2017] several authors [Xu et al., 2022, Tagasovska et al., 2020] have relaxed the assumption of homoskedasticity. In these works the authors focus on identifying cause from effect only in the bivariate setting. Specifically, Xu et al. [2022] base their inference score on the log-likelihood of regression residuals and apply a binning-scheme on regions where variance is approximately constant.

More closely related to our work is the robust-PC method of Kalisch and Bühlmann [2008] who use an estimator in a recursive partial correlation formula to robustify the PC algorithm against outliers and otherwise contaminated data.

Non-constant conditional variance can also be regarded as a specific kind of distributional shift or context change. The effects of distributional shifts on causal discovery have been investigated, for instance, in Huang et al. [2020], Mooij et al. [2020] where the authors propose a framework that includes the environment into the structural causal model formulation using context variables.

3 Problem setting

3.1 Heteroskedasticity in causal models

In this work, we consider discovering causal relationships in linear models in the presence of non-constant error variance which is potentially dependent on the parents or on the sampling index. To translate this into a structural causal model (SCM), we represent heteroskedasticity as a scaling function of the noise variable. In this way, it can also be viewed as state-dependent or multiplicative noise.

Consider finitely many random variables $V = (X^1, \ldots, X^d)$ with joint distribution $\mathcal{P}_X$ over a domain $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. Then we are interested in n samples from the following SCM with assignments

$$X_t^i := f_i(Pa(X_t^i)) + h_i(H(X_t^i)) \cdot N_i, \qquad i = 1, \ldots, d \tag{1}$$

where f_i are linear functions, $t \in \mathcal{T}$ stands for the sample index, and we have the heteroskedasticity functions $h_i : \mathcal{X} \times \mathcal{T} \to \mathbf{R}_{\geq 0}$. The noise variables N_i are assumed independent standard Gaussian distributions. The parent set of the variable X_t^i is denoted by $Pa(X_t^i)$, and $H(X_t^i) \subset Pa(X_t^i) \cup \{t\}$ which can also be the empty set. Furthermore, we make the restriction that the causal relationships are stable over time, i.e. the parent sets $Pa(X_t^i)$ as well as the functions f_i are not time-dependent.

For $h_i \equiv const.$ the homoskedastic version of this SCM is obtained, which simply is a linear SCM with Gaussian noise. To learn such a causal model from data, a well-known and easy to implement method is to use a constraint-based causal discovery method with CI tests based on partial correlation.

To recap the partial correlation test and to illustrate our adaptations, consider the following simple example model for a time-series, or otherwise indexed, SCM (1). The general multivariate case is discussed in section 4.2.

$$\begin{aligned} X_t &= aZ_t + cE_t + h_X(Z_t, t) \cdot N_X \qquad & Z_t &= N_Z \\ Y_t &= bZ_t + cE_t + h_Y(Z_t, t) \cdot N_Y \qquad & E_t &= N_E \end{aligned} \tag{2}$$

for some constants a, b and c, and standard normal independent noise terms N_X, N_Y, N_Z, and N_E. The constant c is set to zero for (conditionally) independent X and Y. Note that here we are only interested in discovering the relationship between X and Y and not in learning the whole causal graph.

An intuitive way to define and understand partial correlation is in terms of the correlation between residuals. The partial correlation between X and Y given a controlling variable Z, or a set thereof, is the correlation between the residuals r_X and r_Y resulting from the linear regression of X on Z and of Y on Z, respectively. The linear regression that is used in the standard variant of the partial correlation test is ordinary least squares (OLS) regression, thus we refer to this CI test as *ParCorr-OLS*. The Pearson correlation coefficient ρ is then estimated by $\widehat{\rho}(X, Y|Z) = \frac{\widehat{Cov(r_X, r_Y)}}{\sqrt{\widehat{Var(r_X)}\widehat{Var(r_Y)}}}$. We use the hat operator to indicate estimators. For testing the null hypothesis $H_0 : \rho = 0$ versus the alternative

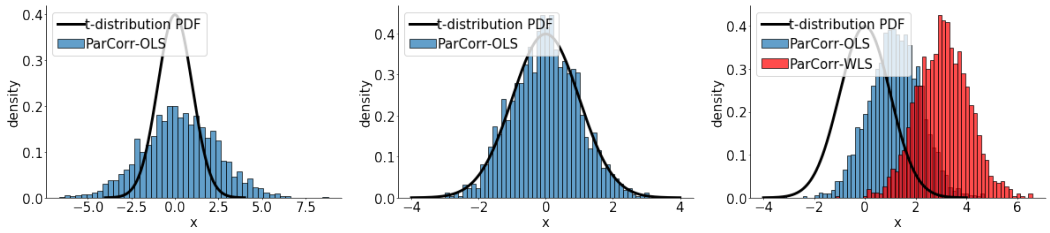

Figure 1: Data generated with SCM (2). (Left) Null distribution when X and Y are affected by the same kind of heteroskedasticity. (Middle) Null distribution when only X is affected by heteroskedasticity. (Right) Alternative distribution (assuming dependence with $c \neq 0$) when only X is affected by heteroskedasticity. In all plots the heteroskedasticity is a linear function of the value of Z and the solid black line depicts the t-distribution under the null hypothesis.

$H_1 : \rho \neq 0$, we use the studentized version of the partial correlation as a test statistic $T(\hat{\rho}) = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$. This statistic is $t(n-2-k)$-distributed, where k is the number of variables we are conditioning on, and n is the sample size.

However, for the partial correlation test to be correct, the assumptions of OLS, in particular homoskedasticity, have to be fulfilled. In the next section, we investigate what happens if this assumption is violated.

3.2 Effects of heteroskedasticity on partial correlation

Under heteroskedasticity, the estimator of slope regression parameters in OLS regression is still unbiased. However, the estimator of the covariance matrix of the parameter estimates can be biased and inconsistent under heteroskedasticity [Wooldridge, 2009, p.268]. This also affects the subsequent residual-based correlation test in ways that we summarize below. Proofs for and further discussions of the following statements are provided in the Supplement A.1.

Effect 1. *Under the null hypothesis, the studentized Pearson correlation coefficient is $t(n-2-k)$ distributed if X is independent of the variables inducing heteroskedasticity in Y.*

This means that if only one node is compromised by heteroskedasticity, or, more generally, the heteroskedasticity in X is independent of that in Y, the type-I error rate of the t-test will not be affected. Please refer to the middle plot in figure 1 for a visualization of the associated null distribution in comparison to the analytical one used by the partial correlation test. On the other hand, in the left plot the case where both X and Y are affected by the same kind of heteroskedasticity is shown, leading to a different null distribution.

Effect 2. *If X and Y are dependent and at least one is affected by heteroskedasticity, the detection power of the t-test might be degraded.*

In particular, if X and Y are dependent, X is affected by linear heteroskedasticity, the mean of the t-distribution is closer to zero compared to the distribution of weighted least squares based studentized partial correlation coefficient, which is introduced in the next section and essentially transforms the data to be homoskedastic, for fixed sample size. See the right plot in figure 1. This effect is an immediate consequence of the reduced efficiency of the OLS slope estimate. Intuitively, heteroskedasticity masks the relationship between X and Y.

4 Weighted least squares partial correlation test and causal discovery

We have seen that the standard partial correlation test is sensitive to heteroskedastic noise since it is based on an OLS regression step. Therefore, we propose to replace the OLS regression by the weighted least squares (WLS) approach which is known to be able to handle non constant error variance. We will refer to the resulting CI test as *ParCorr-WLS*.

4.1 Weighted least squares partial correlation test

The idea of WLS is to perform a re-weighting of each data point depending on how far it is from the true regression line. It is reasonable to assume data points where the error has low variance to be more informative than those with high error variance. Therefore, ideally the weights are chosen as the inverse variance of the associated error.

To formalize this idea, consider the linear model $y = X\beta + \varepsilon$ with $\mathrm{E}[\varepsilon \mid X] = 0,\ \mathrm{Cov}[\varepsilon \mid X] = \mathrm{diag}(\sigma_i^2)_{i=1,\dots,n}$ for n observations. The variance of the error term ε_i is denoted by σ_i^2. Note, as opposed to the assumptions of OLS, the entries of the conditional variance matrix are allowed to differ from each other. Denote the weight matrix by $W := \mathrm{diag}(\frac{1}{\sigma_i^2})_{i=1,\dots,n}$. The WLS method estimates β by solving the adjusted optimization problem

$$\hat{\beta} = \underset{b}{\operatorname{argmin}}(y - Xb)^{\mathsf{T}} W (y - Xb).$$

This objective is quadratic, thus we can write down the solution in closed form

$$\hat{\beta} = \left(X^{\mathsf{T}} W X\right)^{-1} X^{\mathsf{T}} W y.$$

If the true weights are known, WLS is equivalent to applying OLS to a linearly transformed, homoskedastic version of the data as the weights have the effect of standardizing the scale of the errors. Thus, the following lemma holds. Proofs for these statement can be found, for instance, in Greene [2003].

Lemma 1. *If the weights are chosen as the reciprocal of the error variance per sample, the WLS estimator is consistent, efficient, and asymptotically normal, as well as BLUE.*

Moreover, note that the weighted residuals $W(y - X\hat{\beta})$ are homoskedastic.

The goal now is to approximate the conditional variance function of both X and Y in SCM (2). Since it works analogously for both cases, we illustrate the approach for $X_t = aZ + s(Z,t) \cdot N_X$, i.e. we approximate $\sigma^2(z,t) = Var(X_t|Z = z)$.

For that, we use a residual-based non-parametric estimator for the conditional variance, similar to the approach of Robinson [1987]. Motivated by the identity $Var(X|Z) = \mathbb{E}[(X - \mathbb{E}[X|Z])^2|Z]$, and noting that this is the regression of $(X - aZ)^2$ on Z, the first step is using OLS regression to obtain the squared residuals $(X - \hat{a}Z)^2$. Afterwards, we use a non-parametric regression method to regress these residuals on Z and thereby predict the conditional mean by using a linear combination of the k residuals closest in Z value. For sampling index-dependent heteroskedasticity this turns into a windowing approach, which essentially smoothes the squared residuals.

Algorithm 1 details the proposed partial correlation CI test based on the feasible WLS approach that employs our weight approximation method. For the test to perform well, it is crucial to know the type of heteroskedasticity, more precisely, which of the predictors the variance depends on. In practice, this kind of expert knowledge could be obtained by performing a test for heteroskedasticity, e.g. as suggested in Wooldridge [2009, p.277] or by investigating plots of the residuals. Here we make the limiting assumption that the heteroskedasticity only depends on one of the predictors or on the sampling index. Further extensions are considered in the section 6.

Now, we formulate our main assumption under which it is possible to obtain a consistency result for the WLS method that uses our weight approximation method.

Assumption 1 (Heteroskedastic relationships)**.** *For each node X^i, $i = 1, \dots, d$ in SCM* (1)*, the skedasticity or noise scaling function h_i only depends on one of the predictors H or on the sampling index t, i.e. its domain is one dimensional, or it is constant. Furthermore, it is known what it depends on.*

We also have to impose a rather technical assumption on the functions h_i.

Assumption 2 (Weight approximation)**.** *Assumptions (3.3) - (3.5) from Robinson [1987].*

The proof of the following lemma can be found in Robinson [1987].

Lemma 2. *Under assumptions 1 and 2, the WLS estimator that uses the reciprocal of the approximated variance as weights is consistent and attains the correct covariance matrix asymptotically.*

Assumption 3 (Technical assumptions)**.** *See Supplement A.2.*

We now state the first main theorem.

Theorem 1. *Under the assumptions 1 - 3 ParCorr-WLS CI test with estimated weights (Algorithm 1) is consistent for testing the conditional independence between two potentially heteroskedastic variables X and Y conditioned on a set of variables Z.*

Algorithm 1: ParCorr-WLS

Data: Expert knowledge $\mathcal{E}$ as map with keys X and Y and values in {false, 'sampling index', 'heteroskedastic parent H'}, where $H \in V$, observational data with sample size n for nodes $X, Y, Z_1, \ldots, Z_k, H$, window length λ, significance level α

Result: boolean indicating whether there is a (conditional) dependence between X and Y given $Z_1, \ldots, Z_k$

$resid := []$;

for $node$ *in {X, Y}* **do**

  if $k == 0$ **then**

    $\tilde{r} = node$

  else

    obtain residuals $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)$ by regressing $node$ on $Z_1, \ldots, Z_k$ using OLS;

  end

  if $\mathcal{E}[node]$ == *'false'* **then**

    append $\tilde{r}$ to $resid$;

  else if $\mathcal{E}[node]$ == *'sampling index'* **then**

    compute weights $w_i = (\frac{1}{\lambda} \sum_{j=\max(1,i-\frac{\lambda}{2})}^{\min(n,i+\frac{\lambda}{2})} \tilde{r}_j^2)^{-1}$, for $i = 1, \ldots, n$;

    obtain residuals r by regressing $node$ on $Z_1, \ldots, Z_k$ using WLS with the weights w;

    append $w \cdot r$ to $resid$;

  else

    sort $\tilde{r}$ such that their corresponding values of H increase;

    compute weights $w_i = (\frac{1}{\lambda} \sum_{j=\max(1,i-\frac{\lambda}{2})}^{\min(n,i+\frac{\lambda}{2})} \tilde{r}_j^2)^{-1}$, for $i = 1, \ldots, n$;

    revert the sorting in the indices of w;

    obtain residuals r by regressing $node$ on $Z_1, \ldots, Z_k$ using WLS with the weights w;

    append $w \cdot r$ to $resid$;

end

calculate studentized Pearson correlation t between $r_X = resid[0]$ and $r_Y = resid[1]$;

perform t-test with (two-sided) significance level α, i.e. reject if $|t| > t(1 - \frac{\alpha}{2}, n - 2 - k)$

4.2 Extension to the PC algorithm

A well-known and widely used algorithm for discovering causal relationships in terms of the completed partially directed acyclic graph (CPDAG) from observational data is the PC algorithm as introduced in Spirtes and Glymour [1991]. It consists of two phases: The first one is concerned with learning the skeleton of adjacencies based on iterative CI testing. Subsequently, a set of rules is applied to determine the orientation of the found links.

To ensure consistency of this method, the following assumptions have to be fulfilled. Details can be found in Spirtes et al. [2000]. Let $G = (V, E)$ be a graph consisting of a set of vertices V and a set of edges $E \subset V \times V$. Let $\mathcal{P}$ denote the probability distribution of V.

Assumption 4 (PC algorithm)**.** *The Causal Markov condition, Causal Faithfulness, and Causal Sufficiency are fulfilled for the SCM* (1) *with graph G.*

Under these assumptions and if the utilized conditional independence test is consistent, it can be shown that the PC algorithm converges in probability to the correct causal structure, i.e.

$$\lim_{n \to \infty} \mathcal{P}(\hat{G}_n \neq G) = 0,$$

where G denotes the ground truth CPDAG and $\hat{G}_n$ is the finite sample output of the PC algorithm. See Kalisch and Bühlmann [2007] for a proof.

In the following, we need to further restrict Assumption 1 to prove consistency of the PC algorithm under heteroskedasticity.

Assumption 5 (Heteroskedastic relationships regarding PC algorithm)**.** *Assumption 1 is further limited to the case where there is only sampling index dependent heteroskedasticity or homoskedasticity.*

Given Assumption 5, we can apply our proposed method ParCorr-WLS in every CI test of the PC algorithm. Using Lemma 1, we thus can establish consistency of the PC algorithm with ParCorr-WLS if the true weights are known. Lemma 2 yields the same result for ParCorr-WLS with estimated weights.

Theorem 2. *Under Assumptions 2,3,4,5 the output of the PC algorithm, with ParCorr-WLS as a CI test, converges in probability to the correct causal graph.*

Note that we prove consistency of the PC algorithm with ParCorr-WLS only in the case of sampling index dependent heteroskedasticity. We discuss the challenges to extending this to more general types of heteroskedasticity in section 6.

5 Experiments

In the following, we conduct experiments evaluating our proposed CI test separately and in conjunction with the PC algorithm. Throughout the experiments, heteroskedasticity strength refers to the parameter s in the scaling functions h of linear and periodic type given by

$$\begin{aligned} h(x) &= 1 + se^T x \cdot \mathbb{1}_{x \geq 0} \qquad \text{(linear)} \\ h(x) &= 1 + se^T \sin(x) + s \qquad \text{(periodic)}\,. \end{aligned} \tag{3}$$

In other words, in case of linear heteroskedasticity strength is the slope of the variance function, and for periodic heteroskedasticity it refers to the amplitude of the variance function.

5.1 Conditional independence testing

We generate the data from the SCM (2) where we consider various types of heteroskedasticity, i.e. functions h, namely linear and periodic as given in Eq. (3). Plots of the simulated data are provided in the Supplement. Each of the types can either be Z- or sampling index-dependent. A visualization of the considered heteroskedasticity-types can be found in the Supplement A.5.

We use the Kolmogorov-Smirnov (KS) statistic to quantify how uniform the distribution of p-values is, and therefore as a metric for type-I errors, as in Runge [2018]. Type-II errors are measured by the area under the power curve (AUPC). The metrics were evaluated with a sample size of 500 from 100 realizations of the SCM. Error bars indicate the bootstrapped standard errors.

Figure 2 shows that ParCorr-WLS is well calibrated in the presence of heteroskedasticity, regardless if it affects only one or both of the variables X and Y. On the other hand, the ParCorr-OLS test becomes ill-calibrated in the case of heteroskedastic X and Y as expected by Effect 1. In particular, this means that if we can choose the weights reasonably well, we are able to overcome multiplicative confounding with our proposed CI test.

Regarding power as measured by AUPC, we observe a rather rapid decrease for ParCorr-OLS as the heteroskedasticity increases (compare to Effect 2). Our proposed method ParCorr-WLS has higher power in the heteroskedastic scenario, and even for homoskedastic noise (heteroskedasticity strength equal zero) the power is comparable to that of ParCorr-OLS. The drop in power for ParCorr-WLS can be explained by an overall increase in noisiness of the data as heteroskedasticity strength is growing. Refer to the Supplement A.5 for a plot of AUPC of ParCorr-OLS and ParCorr-WLS on data with homoskedastic but increasing noise.

Similar effects are present for all considered types of heteroskedasticity (see Supplement). The results remain very similar if we do not use the ground truth weights but estimate them using the window approach with a reasonable window length as detailed in Section 4.1.

5.2 Causal discovery

To test our proposed CI test in a more realistic setting, we apply it within the PC algorithm to recover the causal graph from simulated observational data. We build upon the PC-stable algorithm

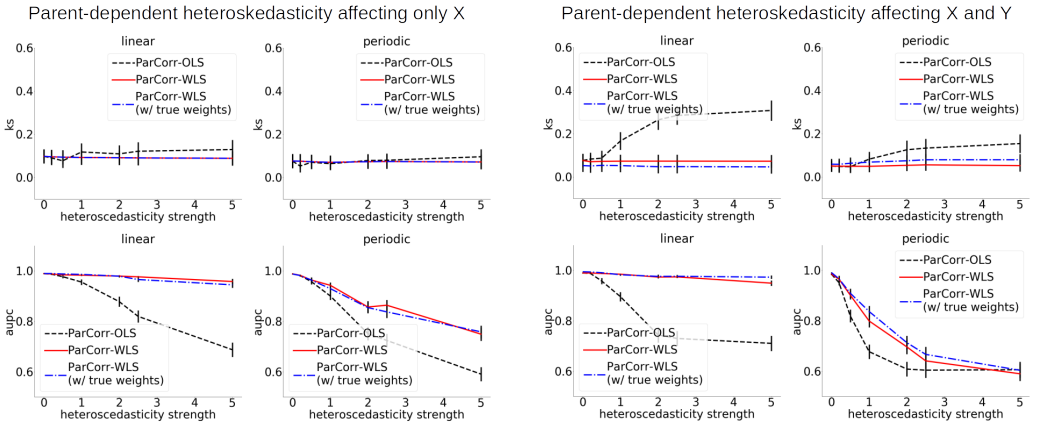

Figure 2: Performance of partial correlation CI tests for dependence ($c = 0.5$) and conditional independence ($c = 0$) between X and Y given Z. In all plots the heteroskedasticity is a function of the confounder Z and only affects X (left two columns) or both X and Y (right two columns). Shown are KS (top row) and AUPC (bottom row) for different strengths of heteroskedasticity for linear (left), periodic (right) noise scaling functions. The ground truth weights are used for WLS, or the weights are estimated using the window approach with window length 10 as detailed in section 4.1. A sample size of 500 is used and the experiments are repeated 100 times.

implementation within the Tigramite software package [Runge et al., 2019b] which is published under the GNU General Public License.

The data for these experiments was generated with SCM (1) in the following way. Given a random ground truth graph, we fix a percentage of heteroskedasticity-affected nodes which are then selected uniformly at random. Throughout the experiments this percentage is set to 0.3 to reduce the chance of parent and child being affected by the same kind of heteroskedasticity. For these affected nodes, we choose as a heteroskedasticity type linear or periodic with equal probability (Eq. (3)) and let the noise variance h either depend on one randomly selected parent or the sampling index. We also set a fixed strength s per experiment and investigated the effect of increasing the strength on the performance of our method compared to the PC algorithm with ParCorr-OLS. All linear dependencies have a coefficient $c = 0.5$. We tested for various small to medium sized causal graphs (see also Supplement).

In these experiments, we estimate the weights based on the expert knowledge as required by Assumption 1, namely which node is affected by heteroskedasticity, and whether the noise variance depends on the sampling index or another node. In the case of the heteroskedasticity depending on another node, we also need to know which node it is. We also compare with the PC algorithm which uses ParCorr-WLS based on the ground truth weights for all direct heteroskedastic relations. Note, however, that this does not take care of indirect heteroskedasticity due to heteroskedastic parents that are not part of the conditioning set. Note that this challenging setting is outside of the stricter Assumption 5 for which the PC algorithm is consistent.

Figure 3 shows, similar to the experiments of the previous section, that the PC algorithm with ParCorr-WLS continues to have a rather small false positive rate (FPR) even though the heteroskedasticity strength increases. In contrast, ParCorr-OLS shows an increase in FPR as the heteroskedasticity strength increases. Additionally, even in this rather complicated setting, we see that the true positive rate (TPR) can be improved by using our method. The PC algorithm with ParCorr-WLS has an average runtime of 0.29 seconds on homoskedastic data compared to 0.14 seconds for ParCorr-OLS evaluated on AMD 7763.

6 Discussion and Outlook

In this work we relaxed the common assumption of a constant variance by explicitly allowing for heteroskedasticity as a multiplicative scaling of noise in an otherwise linear SCM. Our proposed partial correlation test based on weighted least squares regression is a linear, computationally fast and

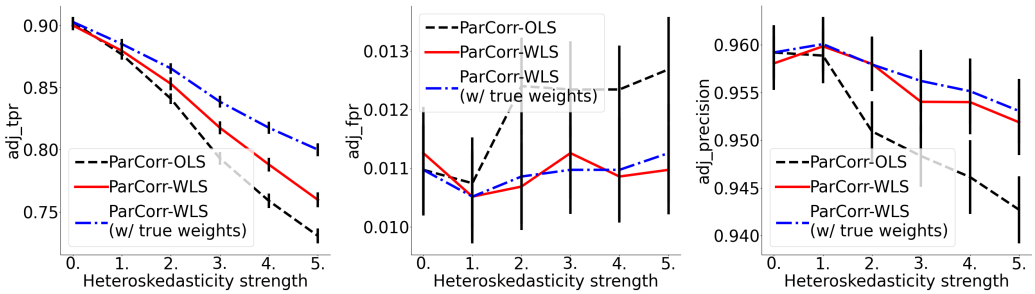

Figure 3: Results for the PC algorithm with the standard ParCorr-OLS CI test compared to that with the proposed ParCorr-WLS test. Shown are adjacency TPR (left) and FPR (middle), as well as the adjacency precision (right) for increasing strengths of heteroskedasticity. The graph has 10 nodes and 10 edges, a sample size of 500 is used. The significance level α is set to 0.05. The experiment is repeated 500 times. Errorbars show standard errors. Estimated weights with a window length of 5 or ground truth weights are used for WLS.

easy to implement method and constitutes a useful standalone method for various data science tasks, but our focus here is on its use in causal discovery. The main **strengths** of our approach are that it is able to produce more reliable results both as a CI test and in causal discovery in the presence of heteroskedasticity than the standard partial correlation variant. More reliable results here refers to controlled false positives as well as higher detection power. Furthermore, the suggested adaptations do not compromise calibratedness and power of the CI test on homoskedastic data.

The main **weakness** of our method is that the CI test requires substantial expert knowledge in Assumption 1 and for the consistency of the PC algorithm the even stricter Assumption 5. Assumption 1 requires that the heteroskedasticity only depends on one of the predictors or on the sampling index, i.e. its domain is one dimensional. Furthermore, one needs to know which of these types is the case. Assumption 5 only allows for sampling-index heteroskedasticity. The reason for this is that there can be indirect heteroskedasticity at the node X or Y that is not induced by a parent of X or Y, but by some other ancestor. This heteroskedasticity then propagates through the causal graph and essentially makes X or Y heteroskedastic whenever this path is not blocked by the conditioning set. In this case, the expert knowledge does not tell us about this heteroskedasticity and we would not be able to apply ParCorr-WLS to remove it. See also Section A.4 in the Supplement for a detailed treatment of cases in which the CI tests within the PC algorithm are consistent under more general heteroskedasticity forms. In the following, we discuss alternatives and further avenues of research.

In future work one may alter the required expert knowledge and weight approximation scheme to overcome the issues induced by indirect heteroskedasticity. Here, a possible remedy could be an iterative approach using information about causal relationships from earlier steps. An alternative would be to use the CMIknn CI test [Runge, 2018] that is able to fully treat heteroskedastic multiplicative confounding. However, this increased generality comes at the price of reduced detection power, also see Figure 10 in the Supplement.

Another open question is how to further improve the weight approximation method. For instance, by iteratively repeating the regression and smoothing steps. Another important consideration is the choice of the window length. Linear properties of the variance function allow us to use a larger window length. However, if the variance function shows a high variability, crucial information might be lost if the chosen window length is too large. Model selection criteria might be employed to alleviate this problem.

Furthermore, one can consider extensions of ParCorr-WLS to multiple heteroskedastic influencing factors, e.g., Spokoiny [2002] extend the residual-based conditional variance function estimation to dimensions larger than one. Another multidimensional method based on differences is discussed in Cai et al. [2009].

It would also be interesting to explore using generalized least squares to be able to account for additional correlation of the residuals. Potentially, the weighted least squares partial correlation

coefficient could also be combined with a permutation-based test to extend the method to problems with non-Gaussian noise.

Concluding, our proposed ParCorr-WLS CI test makes constraint-based causal discovery methods better applicable to real world problems where the assumption of homoskedasticity is violated, such as in climate research or neuroscience. Ethically, we believe that our rather fundamental work has a low potential for misuse.

Acknowledgments and Disclosure of Funding

WG was supported by the Helmholtz AI project *CausalFlood*. UN and JW were supported by grant no. 948112 *Causal Earth* of the European Research Council (ERC). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1083. We thank the anonymous reviewers for their helpful comments.

References

L. D. Brown and M. Levine. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232, 2007.

T. T. Cai, M. Levine, and L. Wang. Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, 100(1):126–136, 2009. ISSN 0047-259X.

R. J. Carroll. Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, 10(4):1224–1233, 1982.

D. M. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *J. Mach. Learn. Res.*, 2: 445–498, 2002. ISSN 1532-4435.

I. Ebert-Uphoff and Y. Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25:5648–5665, 2012.

J. Fan and Q. Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85 (3):645–660, 1998.

W. H. Greene. *Econometric Analysis*. Pearson Education, fifth edition, 2003. ISBN 0-13-066189-9.

P. Hall and R. J. Carroll. Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):3–14, 1989.

A. Hayes. Using heteroskedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior research methods*, 39:709–22, 12 2007.

B. Huang, K. Zhang, J. Zhang, J. D. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007.

M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789, 2008.

A. Lawrance. On conditional and partial correlation. *The American Statistician*, 30(3):146–149, 1976.

J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

B. Merz, G. Blöschl, S. Vorogushyn, F. Dottori, J. C. Aerts, P. Bates, M. Bertola, M. Kemter, H. Kreibich, U. Lall, et al. Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2(9):592–609, 2021.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

M. H. Neumann. Fully data-driven nonparametric variance estimators. *Statistics: A Journal of Theoretical and Applied Statistics*, 25(3):189–212, 1994.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.

T. Proietti. Seasonal specific structural time series. *Studies in Nonlinear Dynamics & Econometrics*, 8(2), 2004.

P. M. Robinson. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica: Journal of the Econometric Society*, pages 875–891, 1987.

J. P. Romano and M. Wolf. Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19, 2017.

J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In A. J. Storkey and F. Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947. PMLR, 2018.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019a.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019b.

P. Spirtes and C. Glymour. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Soc. Sci. Comput. Rev.*, 9 (1):62–72, 1991. ISSN 0894-4393.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

V. Spokoiny. Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82 (1):111–133, 2002. ISSN 0047-259X.

N. Tagasovska, V. Chavez-Demoulin, and T. Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9311–9323. PMLR, 2020.

R. Wilcox and J. Muska. Inferences about correlations when there is heteroscedasticity. *The British journal of mathematical and statistical psychology*, 54:39–47, 06 2001.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. ISE - International Student Edition. South-Western, 2009. ISBN 9780324581621.

S. Xu, A. Marx, O. Mian, and J. Vreeken. Causal inference with heteroscedastic noise models. In *Proceedings of the AAAI Workshop on Information Theoretic Causal Inference and Discovery (ITCI'22)*, Proceedings of the AAAI Workshop on Information Theoretic Causal Inference and Discovery, pages 9311–9323, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sections 1 and 6, as well as the assumptions, specifically Assumption 1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1-5.
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide references and intuitive justifications for the results in the main text and complete proofs in the Supplement.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions on how to execute it are provided in the supplemental material. The code includes a method for generating the synthetic data.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All hyperparameters are listed in Section 5.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figures 2 and 3.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.

(b) Did you mention the license of the assets? [Yes] See Section 5.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code is included in the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Supplement

Here we provide proofs of the statements made in the main text as well as further figures of numerical experiments and a more detailed discussion of heteroskedasticity effects regarding causal discovery.

A.1 Proof of Effects 1 and 2

A.1.1 Proof of Effect 1:

Proof. Let $(X_i, Y_i)_{i=1,\dots,n}$ be an independent sample with Pearson correlation coefficient ρ, and we assume the linear model $Y_i = X_i\beta + h(Z_i)\epsilon_i$, where Z_i and ϵ_i are independent and standard normal, and h is the noise scaling function.

Note that w.l.o.g. we assume only Y to be heteroskedastic w.r.t. Z.

Testing whether the Pearson correlation between X and Y is zero is equivalent to testing whether the slope parameter β is equal to zero.

Under the null hypothesis, calculating $\mathrm{Var}(h(Z_i)\epsilon_i|X_i)$ gives

$$\mathrm{Var}(h(Z_i)\epsilon_i|X_i) = \mathbb{E}[h^2(Z_i)\epsilon_i^2|X_i] = \mathbb{E}[h^2(Z_i)\epsilon_i^2] = \mathbb{E}[h^2(Z_i)] = \mathbb{E}[h^2(Z_1)]$$

using the independence of Z_i and X_i, the fact that ϵ_i is standard normal, and that the Z_i are identically distributed. Therefore, this is a homoskedastic problem.

□

A.1.2 Discussion of Effect 2:

We start by discussing the homoskedastic case to see where non-constant variance of noise leads to problems within the t-test. Let $(X_i, Y_i)_{i=1,\dots,n}$ be an independent sample from a bivariate normal distribution with Pearson correlation coefficient ρ, and we assume the linear model $Y_i = X_i\beta_0 + \epsilon_i$, where ϵ_i is standard normal, then

$$\beta_0 = \rho\sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}.$$

This also is true for the finite sample estimators of these entities, since

$$\hat{\rho} = \frac{\sum X_iY_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$$

and

$$\hat{\beta} = \beta_0 + \frac{\sum X_i\varepsilon_i}{\sum X_i^2} = \beta_0 + \frac{\sum X_i(Y_i - X_i\beta_0)}{\sum X_i^2} = \frac{\sum X_iY_i}{\sum X_i^2} = \hat{\rho}\sqrt{\frac{\sum Y_i^2}{\sum X_i^2}}.$$

Furthermore, the following relationship holds for the estimator of the studentized correlation coefficient and the estimator of the slope parameter β_0

$$\hat{\rho}\frac{\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} = \hat{\beta}\sqrt{\frac{\sum X_i^2}{\frac{1}{n}\sum(Y_i - \hat{\beta}X_i)^2}}. \tag{4}$$

For homoskedastic noise the second factor is an estimator of the standard error of $\hat{\beta}$, which we derive by using the mean of the squared residual as an estimator for the error variance. We know that

$$\hat{\beta} = \frac{\sum X_iY_i}{\sum X_i^2} = \frac{\sum X_i(\beta_0X_i + \epsilon_i)}{\sum X_i^2} = \beta_0 + \frac{\sum X_i\epsilon_i}{\sum X_i^2}$$

and thus

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\big(\frac{\sum X_i\epsilon_i}{\sum X_i^2}\big). \tag{5}$$

If we assume the design to be fixed and $\mathrm{Var}(\epsilon_i) = \sigma^2$ to be constant, this simplifies to

$$\mathrm{Var}(\hat{\beta}) = \frac{\sum X_i^2 \mathrm{Var}(\epsilon_i)}{(\sum X_i^2)^2} = \frac{\sigma^2 \sum X_i^2}{(\sum X_i^2)^2}.$$

Therefore we use the following estimator

$$\widehat{\mathrm{Var}(\hat{\beta})} = \frac{\frac{1}{n}\sum (Y_i - \hat{\beta} X_i)^2}{\sum X_i^2}. \tag{6}$$

We now assume that Y is heteroskedastic, i.e. the model takes the form

$$Y_i = X_i \beta_0 + h(Z_i)\epsilon_i,$$

We need to show

$$\mathbb{E}\left[\frac{\hat{\beta}^{\mathrm{OLS}}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}})}}}\right] \leq \mathbb{E}\left[\frac{\hat{\beta}^{\mathrm{WLS}}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}^{\mathrm{WLS}})}}}\right],$$

where $\hat{\beta}^{\mathrm{OLS}}$ is the OLS estimator for the slope parameter and $\hat{\beta}^{\mathrm{WLS}}$ is its WLS version.

Since OLS and WLS generally return similar values for the slope estimator (especially for large n) and both are unbiased estimators for β_0, we focus on the denominator.

We get $\mathbb{E}\left[\frac{\beta_0}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}})}}}\right] \leq \mathbb{E}\left[\frac{\beta_0}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}^{\mathrm{WLS}})}}}\right]$ from the reduced efficiency of the OLS regression on heteroskedastic data, i.e. from

$$\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}}) \geq \mathrm{Var}(\hat{\beta}^{\mathrm{WLS}}),$$

if we assume that the estimators (6) for $\mathrm{Var}(\hat{\beta}^{\mathrm{OLS}})$ and $\mathrm{Var}(\hat{\beta}^{\mathrm{WLS}})$ are unbiased. The unbiasedness is clear for the WLS-case, as well as for the case where the heteroskedasticity in X is independent from the one in Y, see Effect 1. In the case of heteroskedasticity in Y that is dependent on X, the estimator might be biased. If it is over-estimating $\mathrm{Var}(\beta)$, the power is also reduced. The situation of under-estimating the variance can, in fact, quasi-increase power but comes with the major drawback of increased probability of type I error.

A.2 Proof of Theorem 1

A.2.1 Assumptions

Consider random variables $X = (X_1, \ldots, X_n)^T, Y = (Y_1, \ldots, Y_n)^T$ and $Z^k = (Z_1^k, \ldots, Z_n^k)^T$ for $k = 1, \ldots, d$, where n is the number of samples. Denote the design matrix of the regressor variable by $Z = (Z^1, \ldots, Z^d) \in \mathbb{R}^{n \times d}$. We assume that all variables are centered Gaussians and generated according to the model

$$X = Z \cdot \alpha + \varepsilon^X$$
$$Y = Z \cdot \beta + \varepsilon^Y$$

with error variables $\varepsilon^X = (\varepsilon_1^X, \ldots, \varepsilon_n^X)^T, \varepsilon^Y = (\varepsilon_1^Y, \ldots, \varepsilon_n^Y)^T$. We make the following assumptions:

(A1) The error vectors $\varepsilon^X, \varepsilon^Y$ have zero mean and diagonal covariance matrices $\Sigma_X = \mathrm{diag}(\sigma_{X,1}^2, \ldots, \sigma_{X,n}^2)$, $\Sigma_Y = \mathrm{diag}(\sigma_{Y,1}^2, \ldots, \sigma_{Y,n}^2)$. Similarly, for any k, the regressor variables $Z^k = (Z_1^k, \ldots, Z_n^k)^T$ have diagonal covariance matrices.

(A2) The matrix $Z^T Z$ is almost surely positive definite, of full rank d, and we have that

$$\lim_{n \to \infty} \frac{Z^T Z}{n} = Q$$

converges entrywise in probability to a matrix Q. Note that the k, ℓ-entry of $Z^T Z$ is $\sum_{i=1}^n Z_i^k Z_i^\ell$.

(A3) We have well defined asymptotic average variances

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sigma_{X,i}^2 = \sigma_X^2$$
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sigma_{Y,i}^2 = \sigma_Y^2$$

and the limits

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (Z_i^k)^2 \sigma_{X,i}^2, \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (Z_i^k)^2 \sigma_{Y,i}^2 \qquad \text{for } k = 1, \dots, d$$

exist in probability.

(A4) The partial correlations

$$\mathrm{Cor}(X_i, Y_i | Z_i^1, \dots, Z_i^d) = \rho_{X,Y|Z}$$

are constant, i.e. independent of the sampling index i. Moreover, conditional covariances between different sample indices are zero.

(A5) The variances of each variable are bounded by a constant independent of the sampling index i.

(A6) For parent-dependent heteroskedasticity, let H be the heteroskedasticity-inducing parent. Denote by $\mathcal{S}_i$ the set of λ nearest neighbours of X_i (or Y_i respectively) in H-value. In the case of sampling index dependent heteroskedasticity, $\mathcal{S}_i$ denotes the set of nearest neighbours in sampling index value.
We assume

$$\lambda \to \infty, \quad \frac{\lambda}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

Furthermore, the limit of the variances averaged over the nearest neighbours exists and converges to the right value, i.e.

$$\lim_{n\to\infty} \frac{1}{\lambda} \sum_{r_j \in \mathcal{S}_i} \sigma_{X,j}^2 = \sigma_{X,i}^2, \qquad \lim_{n\to\infty} \frac{1}{\lambda} \sum_{r_j \in \mathcal{S}_i} \sigma_{Y,j}^2 = \sigma_{Y,i}^2.$$

A.2.2 Proof

We will now sketch the proof of consistency of the estimator for the partial correlation $\rho_{X,Y|Z}$ in the case where we regress on only one variable, i.e. $d = 1$. The result can however easily be generalized to higher dimensions although the computations become slightly more involved. Recall that in the first step of ParCorr-WLS, we compute the residuals

$$r_i^X = (\alpha - \hat{\alpha}) Z_i + \varepsilon_i^X, \quad r_i^Y = (\beta - \hat{\beta}) Z_i + \varepsilon_i^Y,$$

where $\hat{\alpha}, \hat{\beta}$ are OLS-estimates of α, β. More precisely, recall that $\hat{\alpha}$ is given by the formula

$$\hat{\alpha} = \alpha + \frac{\sum_{i=1}^n Z_i \varepsilon_i}{\sum_{i=1}^n Z_i^2}.$$

Hence, using Slutsky's theorem, we compute that

$$\mathbb{E}[(r_i^X)^2 | Z] = \frac{Z_i^2}{\sum_{j=1}^n Z_j^2} \left(\frac{\sum_{j=1}^n Z_j^2 \sigma_{X,j}^2}{\sum_{j=1}^n Z_j^2} + \sigma_{X,i}^2 \right) + \sigma_{X,i}^2$$

which converges in probability as $\frac{1}{n}(C \cdot Z_i^2 + \sigma_{X,i}^2) + \sigma_{X,i}^2$ where C is a constant. Recall the definition of the set of nearest neighbours $\mathcal{S}_i$ from assumption (A6). Following the results in Hall and Carroll [1989], we establish the convergence of $\hat{\sigma}_{X,i}^2 = \frac{1}{\lambda} \sum_{r_j \in \mathcal{S}_i} r_j^2$. Furthermore, we see that the smoothed

variances of the residuals behave asymptotically as

$$\begin{aligned}\hat{\sigma}^2_{X,i} = \frac{1}{\lambda}\sum_{r_j\in\mathcal{S}_i} r_j^2 &\approx \frac{1}{\lambda}\sum_{r_j\in\mathcal{S}_i}\mathbb{E}[r_j^2|Z]\\ &= \frac{1}{\lambda}\sum_{r_j\in\mathcal{S}_i}\left(\frac{C\cdot Z_j^2+\sigma^2_{X,j}}{n}+\sigma^2_{X,j}\right)\\ &\approx \sigma^2_{X,i}.\end{aligned}$$

After estimation of the variances, we determine the residuals of the WLS regression with weight matrix $\hat{W}_X$ which is equivalent to the OLS regression after scaling all variables by $\hat{W}_X$. In other words, we define

$$R_i^X = (\alpha-\tilde{\alpha})\hat{W}_X Z_i + \hat{W}_X\varepsilon_i,$$

where $\tilde{\alpha}$ is the OLS-estimator w.r.t. the regressor $\hat{W}_X Z$ and the error variable $\hat{W}_X\tilde{\varepsilon}$. Define $\hat{W}_Y$ and R_i^Y analogously.

The following asymptotic moment formulas can be computed from the definition of R_i^X, R_i^Y.

Lemma 3. *Assume (A1)-(A6). Then* $\mathbb{E}[R_i^X|Z]=0$ *and for large* n

$$\begin{aligned}\mathbb{E}[\left(R_i^X\right)^2|Z] &\approx 1+\frac{Z_i^2+2Z_i}{\sum_{j=1}^n Z_j^2}\\ \mathbb{E}[\left(R_i^X\right)^4|Z] &\approx 3+\frac{p(Z_i)}{\left(\sum_{j=1}^n Z_j^2\right)^4},\end{aligned}$$

where p *is a polynomial of degree* 4*. In particular, by (A2),(A5) and the fact that for Gaussians higher moments are functions of the variance,* $\mathrm{Var}(\left(R_i^X\right)^2)\le K$ *for some constant* $K>0$ *independent of* i*. The same statements hold for the residuals* R_i^Y*.*

Lemma 4. *Assume (A1)-(A6). Then, the estimator*

$$\hat{\rho}_{X,Y|Z} = \rho(R^X,R^Y) = \frac{\frac{1}{n}\sum_{i=1}^n R_i^X R_i^Y}{\sqrt{\frac{1}{n}\sum_{i=1}^n (R_i^X)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (R_i^Y)^2}}$$

consistently estimates $\rho_{X,Y|Z}$*.*

Proof. By Slutsky's theorem, it suffices to prove that the three sums in the definition of $\hat{\rho}_{X,Y|Z}$ converge in probability and to compute the limits. Convergence follows from the law of large numbers and the variance bounds $\mathrm{Var}(\left(R_i^X\right)^2)\le K, \mathrm{Var}(\left(R_i^Y\right)^2)\le K'$ of Lemma 3. The limit of the numerator is equal to $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E[R_i^X R_i^Y]$, so we compute this quantity first conditioned on Z. This yields for large n

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[R_i^X R_i^Y|Z] \approx \frac{1}{n}\sum_{i=1}^n \rho_{X,Y|Z}\left(1+\frac{3Z_i^2}{\sum_{j=1}^n Z_j^2}\right)\approx \rho_{X,Y|Z}.$$

Hence $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[R_i^X R_i^Y] = \rho_{X,Y|Z}$. To compute $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(R_i^X)^2]$ we use the first formula of Lemma 3 to see that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(R_i^X)^2|Z] \approx 1+\frac{1}{n}\sum_{i=1}^n\left(\frac{Z_i^2+2Z_i}{\sum_{j=1}^n Z_j^2}\right) = 1+\frac{1}{n}\left(1+\frac{\frac{2}{n}\sum_{i=1}^n Z_i}{\frac{1}{n}\sum_{j=1}^n Z_j^2}\right)\approx 1,$$

so that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(R_i^X)^2]=1$.

A similar computation yields $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(R_i^Y)^2]=1$, so that indeed $\hat{\rho}_{X,Y|Z}\to\rho_{X,Y|Z}$ as $n\to\infty$ in probability. □

A.3 Proof of Theorem 2

Proof. Assumption 5 allows us to apply ParCorr-WLS in every conditional independence test within the PC algorithm. Using Theorem 1, we get the result. □

A.4 Extension of Theorem 2 to different heteroskedasticity forms

As discussed in section 4.2, the PC algorithm with ParCorr-OLS is not consistent. If we use ParCorr-WLS with a consistent estimate for the variance of the residual in every conditional independence test, we can make the PC algorithm consistent.

The scope of the current work is a situation where heteroskedastic noise can be dependent on one of the parents of the node or on the sampling index. However, only for the stricter set of assumptions for Theorem 2 we can prove consistency, namely only for sample-index dependent heteroskedasticity.

In the following, we will discuss different situations of heteroskedasticity affecting variables involved in the CI tests within the PC algorithm. In some of these situations, we are able to apply our proposed ParCorr-WLS CI test, thus obtaining consistency. However, in other situations we are not able to use background knowledge about heteroskedasticity and therefore suffer the same problems as in the standard version of the PC algorithm with ParCorr-OLS.

Let us look at the following cases.

Case 1: The heteroskedasticity in X is independent of that in Y. In case of direct heteroskedasticity, we apply ParCorr-WLS. However, if we encounter a situation where the heteroskedasticity in either X or Y is caused indirectly through a parent, compare figure 4, we lack the background knowledge about it, and therefore apply ParCorr-OLS. Due to Effect 1, we get type I error control for both ParCorr-OLS and ParCorr-WLS. With regard to type II errors, by Effect 2 it is clear that the power of ParCorr-OLS is reduced.

In a setting where the sample size is large enough to get a reasonable weight approximation for WLS, we can only gain detection power by using ParCorr-WLS within the PC algorithm.

Case 2: There is dependence between the heteroskedasticity in X and that in Y. If X and Y are uncorrelated, i.e. $\rho_{X,Y|Z} = 0$, and the heteroskedasticity of at least one of them is directly induced by a parent or by the sampling index, then we apply WLS whereby we remove the heteroskedasticity and are in Case 1 or in the homoskedastic case again.

If the heteroskedasticity is indirect, meaning one of the parents of X and/or Y, that is not included in the conditioning set, is heteroskedastic, our expert knowledge doesn't include this information and we perform ParCorr-OLS. By Effect 1, we know that this suffers from inflated false positives. However, if the link is not removed in this step of the PC algorithm, the conditioning set will be increased in subsequent steps. This means at some point it will include the parent that caused the indirect heteroskedasticity, if the link did not get removed before. When this happens we again will be in Case 1 or in the homoskedastic case.

The main challenge, or in other words the reason why we cannot prove consistency for this more general form of heteroskedasticity, arises if X and Y are correlated. If at least one of them is affected by direct heteroskedasticity, then we again apply WLS whereby we remove the heteroskedasticity and are in Case 1 or in the homoskedastic case as above.

But, if the heteroskedasticity is indirect in both X and Y, then we apply ParCorr-OLS. Since the power for this test is very low in this situation, and we do not obtain a consistent estimator for the standard deviation of the regression parameter, we potentially remove the link between X and Y.

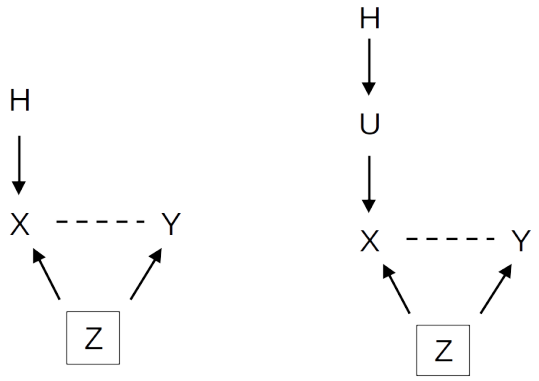

Figure 4: Illustration of Case 1. The dashed line indicates the relationship between X and Y which we are testing. On the left, only X is affected by direct heteroskedasticity introduced by H. In this situation, we would apply ParCorr-WLS since the information about this heteroskedasticity is included in the expert knowledge. On the other hand, in case of indirect heteroskedasticity (right), we do not know about the heteroskedasticity in X because we do not know whether U causes X before running the causal discovery algorithm. Therefore, we would apply ParCorr-OLS, which potentially leads to reduced power.

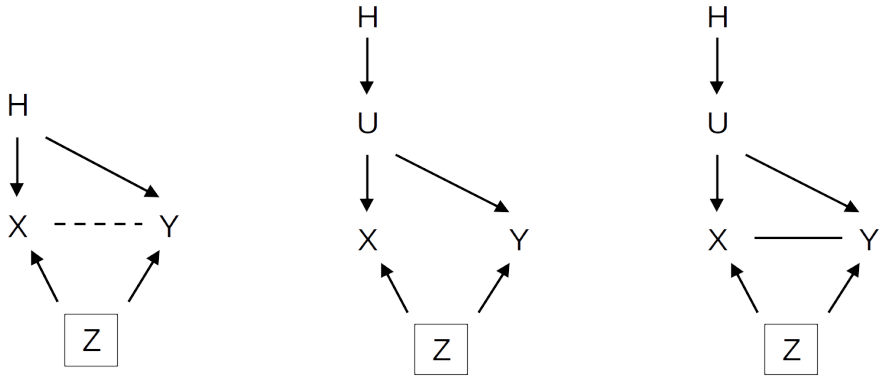

Figure 5: Illustration of Case 2. Here X and Y are both affected by the same kind of (indirect) heteroskedasticity. On the left, the heteroskedasticity is directly introduced by parent H. In this situation, we would apply ParCorr-WLS since the information about this heteroskedasticity is included in the expert knowledge. On the other hand, in case of indirect heteroskedasticity (middle and right), we do not know about the heteroskedasticity in X and Y because we do not know whether U causes X and/or Y beforehand. Therefore, we would apply ParCorr-OLS. If there is no ground truth link between X and Y (middle), this could lead to the wrong detection of a link. However, this means that we would enlarge the conditioning set by including U in a later PC algorithm step. Thereby, the heteroskedasticity is removed. If X and Y are dependent as indicated by the solid line (right), applying ParCorr-OLS might lead to missing this link, i.e. to inconsistency in the PC algorithm.

A.5 Additional Plots

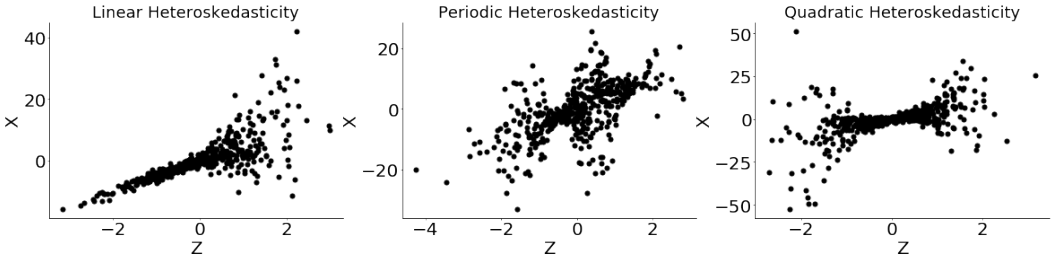

Figure 6: Visualizations of the skedasticity functions h that depend on the confounder Z for strength $s = 5$, compare equation 3.

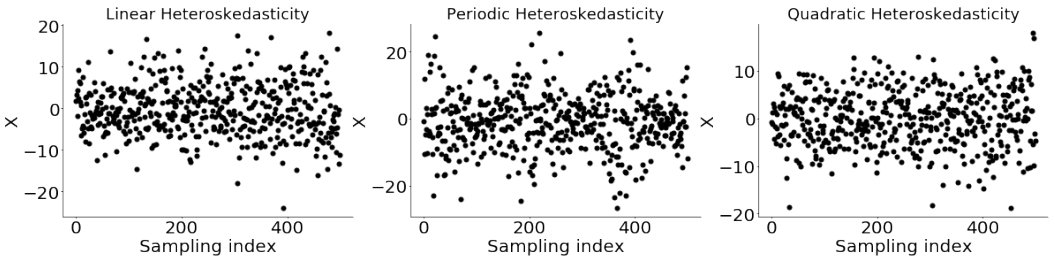

Figure 7: Visualizations of the skedasticity functions h that depend on the sampling index for strength $s = 5$, compare equation 3.

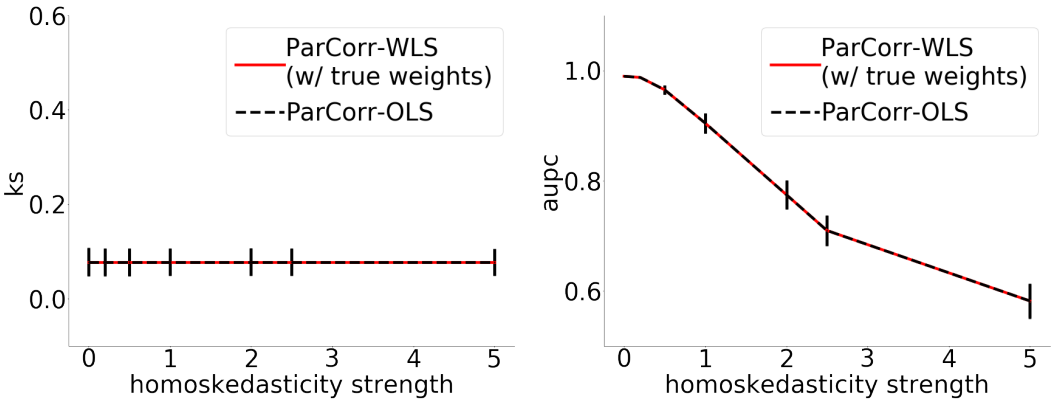

Figure 8: Performance of partial correlation CI tests for dependence ($c = 0.5$) and conditional independence ($c = 0$) between X and Y given Z. Shown are KS (left) and AUPC (right) for different strengths s of noise but the noise is homoskedastic. In other words, we consider the skedasticity function $h_X \equiv s$. This figure illustrates that the power of ParCorr-WLS and ParCorr-OLS reduces in the same way on increasingly noisy but homoskedastic data.

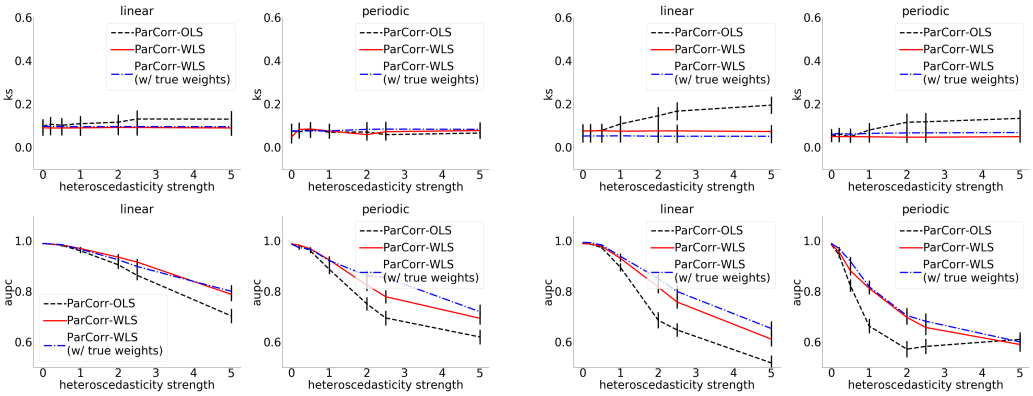

Figure 9: Performance of partial correlation CI tests for dependence ($c = 0.5$) and conditional independence ($c = 0$) between X and Y given Z. In all plots the heteroskedasticity is a function of the sampling index and only affects X (left two columns) or both X and Y (right two columns). Shown are KS (top row) and AUPC (bottom row) for different strengths of heteroskedasticity for linear (left), periodic (right) noise scaling functions. The ground truth weights are used for WLS, or the weights are estimated using the window approach with window length 10 as detailed in section 4.1. A sample size of 500 is used and the experiments are repeated 100 times.

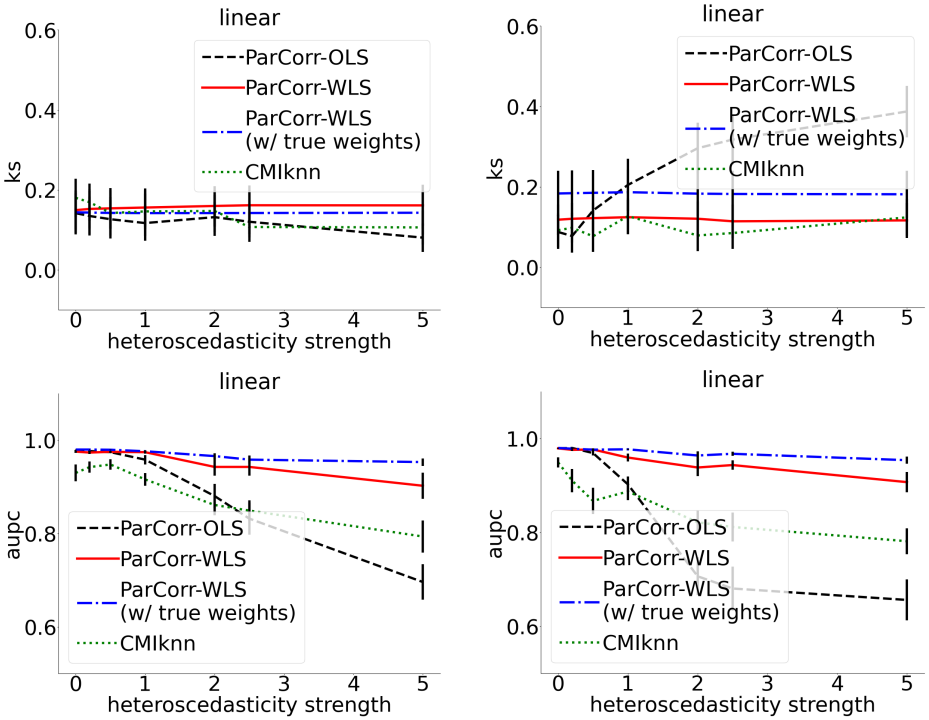

Figure 10: Performance of partial correlation CI tests for dependence ($c = 0.5$) and conditional independence ($c = 0$) between X and Y given Z. We compare our method to ParCorr-OLS and the non-parametric CI test CMIknn [Runge, 2018]. In all plots the heteroskedasticity is a function of Z and only affects X (left two columns) or both X and Y (right two columns). Shown are KS (top row) and AUPC (bottom row) for different strengths of heteroskedasticity for linear noise scaling functions. The ground truth weights are used for WLS, or the weights are estimated using the window approach with window length 10 as detailed in section 4.1. A sample size of 500 is used and the experiments are repeated 50 times. For CMIknn we use 0.1 as the number of nearest-neighbors around each sample point, otherwise the default values are used.

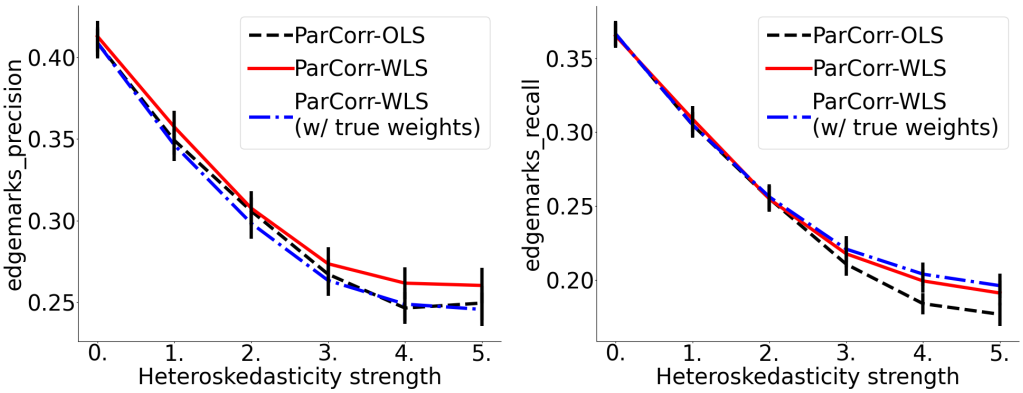

Figure 11: Results for the PC algorithm with the standard ParCorr-OLS CI test compared to that with the proposed ParCorr-WLS test. Shown are edgemark precision (right) and recall (left) for increasing strengths of heteroskedasticity. The graph has 10 nodes and 10 edges, a sample size of 500 is used. The significance level α is set to 0.05. The experiment is repeated 500 times. Errorbars show standard errors. Estimated weights with a window length of 5 or ground truth weights are used for WLS.