6. Supplementary Material/ Appendix for paper: Necessary and sufficient conditions for causal feature selection in time series with latent common causes

6.1. Definitions

Here we briefly mention some fundamental definitions such as Causal Markov Condition and Granger Causality, which we use in our paper to present and prove our methodology. For a thorough study see (Pearl, 2009; Peters et al., 2017; Spirtes et al., 1993).

Definition 1 (Bivariate Granger Causality). *Under the assumption of Causal Sufficiency, X influences Y whenever the past values of X help in predicting Y from its own past. Formally, we write X Granger-causes $Y \iff Y_t \not\perp\!\!\!\perp X_{past(t)} \mid Y_{past(t)}$*

Definition 2 (Multivariate Granger Causality). *X_j Granger causes X_k if $X_t^k \not\perp\!\!\!\perp X^j_{past(t)} \mid \mathbf{X}^{-j}_{past(t)}$ Granger emphasized that proper use of Granger causality would actually require to condition on all relevant variables in the world. Nevertheless, Granger causality is often used in its bivariate version or in situations, in which clearly important variables are unobserved. Such a use can yield misleading statements when interpreting the results causally. (Peters et al., 2017)*

Definition 3 (d-separation (Pearl, 2009)). *In a directed acyclic graph (DAG) G, a path between nodes I_1 and I_m is blocked by a set $\mathcal{S}$ (with neither I_1 nor I_m in $\mathcal{S}$) whenever there is a node $I_k, k = 2, ..., m-1$, such that one of the following two possibilities holds:*

(i) $I_k \in \mathcal{S}$ and $I_{k-1} \to I_k \to I_{k+1}$ or $I_{k-1} \leftarrow I_k \leftarrow I_{k+1}$ or $I_{k-1} \leftarrow I_k \to I_{k+1}$

(ii) Neither I_k nor any of its descendants is in $\mathcal{S}$ and $I_{k-1} \to I_k \leftarrow I_{k+1}$.

In a DAG G, we say that two nodes A and B are d-separated by a third node C if every path between nodes A and B is blocked by C. We then write $A \perp\!\!\!\perp_G B \mid C$.

Definition 4 (Causal Markov Condition (Spirtes et al., 1993)). *Let G be a causal graph with vertex set $\mathcal{V}$ and P be a probability distribution over the vertices in V generated by the causal structure represented by G. G and P satisfy the Causal Markov Condition if and only if for every W in $\mathcal{V}$, W is independent of $\mathcal{V} \setminus (Descendants(W) \cup Parents(W))$ given $Parents(W)$.*

Here we use the global version of Markov condition, which reads: if $\mathcal{X} \perp\!\!\!\perp_G \mathcal{Y} \mid \mathcal{Z} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ for all disjoint vertex sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ (where $\perp\!\!\!\perp_G$ denotes d-separation, as defined above)

Definition 5 (Causal Faithfulness). *A distribution P is faithful to a directed acyclic graph (DAG) G if no conditional independence relations other than the ones entailed by the Markov property are present.*

6.2. Proof of Theorem 1b

Proof. **(Proof by contradiction)**
We need to show that if $X_t^i \not\dashrightarrow Y_{t+w_i}$ then at least one of the conditions 1 and 2 is violated.

Assume that there is no directed path between X_t^i and Y_{t+w_i}: $X_t^i \not\dashrightarrow Y_{t+w_i}$. Then, there is a confounding path $X_t^i \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$ without any colliders. (Colliders cannot exist in the path by the definition of the lag T8.) In that case we will show that either condition 1 or 2 is violated. If all the existing confounding paths $X_t^i \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$ contain an observed confounder $Q^j_{t'} \equiv X^j_{t'} \in \{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ (there can be only one confounder since in this case there are no colliders in the path), then condition 1 is violated, because we condition on $X^j_{t'}$ which d-separates X_t^i and Y_{t+w_i}. If in all the existing confounding paths the confounder node $Q^j_{t'} \notin \{\boldsymbol{S^i}, Y_{t+w_i-1}\}, t' \leq t$ but some observed non-collider node is in the path and this node belongs to $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$, then condition 1 is violated, because we condition on $\boldsymbol{S^i}$ which d-separates X_t^i and Y_{t+w_i}. If there is at least one confounding path and its confounder node does no belong in $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ and no other observed (non-collider or descendant of collider) node which is in the path belongs in $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ then condition 2 is violated for the following reasons: Let's name $p1 : X_t^i \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$. We know the existence of the path $p2 : X^i_{t-1} \to X^i_t$, due to assumption A7.

(1I) If $p1$ and $p2$ have X_t^i in common, then X_t^i is a collider. Therefore, adding X_t^i in the conditioning set would unblock the path between X^i_{t-1} and Y_{t+w_i}.

(1II) If $p1$ and $p2$ have X^i_{t-1} in common, that means X^i_{t-1} lies on $p1$. In this case X_t^i is not in the path from X^i_{t-1} to Y_{t+w_i} and hence adding X_t^i to the conditioning set could not d-separate X^i_{t-1} and Y_{t+w_i}.

In both cases condition 2 is violated. Therefore we showed that if conditions 1 and 2 hold, then $X_t^i \dashrightarrow Y_{t+w_i}$. □

6.3. Proof of lemmas 1, 2

Lemma 1. *If the paths between X^j and Y are directed then the minimum lag w_j as defined in T8 coincides with the minimum non-negative integer w'_j for which $X^j_t \not\perp\!\!\!\perp Y_{t+w'_j} \mid X^j_{past(t)}$. The only case where $w'_j \not\equiv w_j$ is when there is a confounding path between X^j and Y that contains a node from a third time series with memory. In this case $w'_j = 0$.*

Proof. This is obvious by the fact that in the first two cases

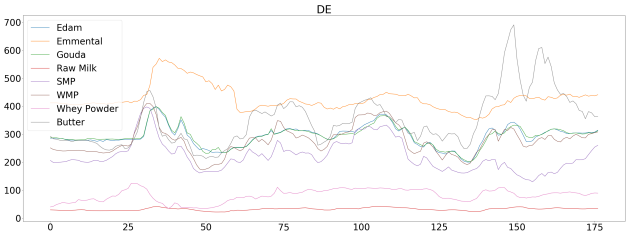
and when a memoryless confounder exists in the path X_t^j- - -$Y_{t+w'_j}$, the path does not contain horizontal arrows of the type $Q_s^r \to Q_{s+1}^r$. □

Lemma 2. *Theorems 1a/1b and 2 are valid if the minimum lag w_j as defined in T8 is replaced with w'_j obtained in lemma 1.*

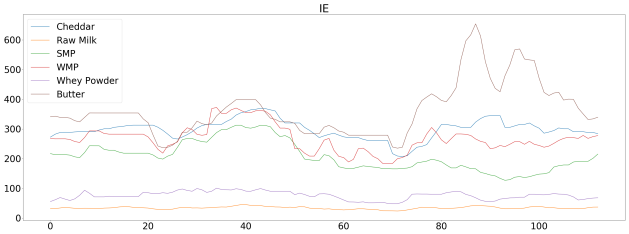
Proof. Claims of theorem 1a/1b remain unaffected because the conditions of theorem 1a/1b hold for any lag according to remark 1. According to lemma 1 the only occasion that the minimum non-negative integer w'_j identified by its simple condition, does not coincide with the minimum lag w_j of the definition in T8 is when there exist confounding paths X_t^j- - -Y_{t+w_j} in which the confounder or any intermediate node in the path has memory. In this case w'_j will always be 0. If the confounder in the paths is hidden then due to assumption A9 it will be memoryless. In this case the w'_j will coincide with the minimum lag and therefore according to the proof of theorem 2 the appropriate node of X^j will be in the conditioning set and no cause will be rejected. Therefore it is enough to show that theorem 2 is valid using $w'_j = 0$ when there is an observed confounder in the path.

Assume that condition 2 is violated. Then this will mean that the set $\{X_t^i, \boldsymbol{S^i}, Y_{t+w_i-1}\}$ does not d-separate X_{t-1}^i and Y_{t+w_i}. This would mean that there is a path X_{t-1}^i - - - Y_{t+w_i} in which one of the elements of this set is a collider or descendent of collider and there is no non-collider node in the conditioning set. The proof for the cases (a1), (a2), (a4) and (b) remain the same for the proof of theorem 2. Assume that $X_{t+w_{ij'}-1}^j$ is a collider and no non-collider node in the path belong to the conditioning set. However the observed common causes of $X_{t+w_{ij'}-1}^j$ and Y_{t+w_i-1} are always in the path. Because all these observed common causes are connected via a directed path with Y_{t+w_i-1}, their minimum lag will be correctly identified and so by construction they will be added in the conditioning set. This contradicts the statement "and there is no non-collider node in the path that belongs in the conditioning set". Therefore we showed that condition 2, thus theorem 2 is not violated. □

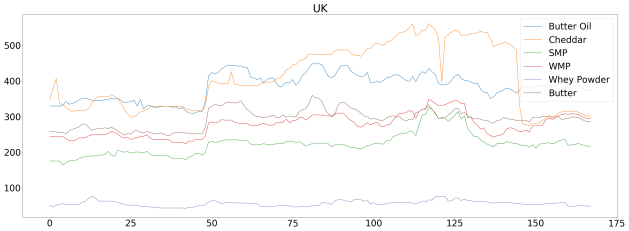
6.4. Experiments on real datasets: Dairy product prices for DE, IE and UK



(a) Dairy product prices for Germany. Raw Milk prices provided in the dataset.



(b) Dairy product prices for Ireland. Raw Milk prices provided in the dataset.



(c) Dairy product prices for Germany. Notice that Raw Milk prices are no provided in the dataset. This dataset was on purpose selected, as it would represent a realistic case of a hidden confounder between Butter and the rest dairy products.

Figure 6. Dairy product prices provided by the EU for a span of 8.5-14.5 years (one price recording per month). Here we provide the ground truth time-series from the dairy product prices for each of the three countries 'DE', 'IE' and 'UK'. For UK Raw Milk prices are not provided.

6.5. Additional results for simulated full-time graphs for various number of observed time-series, noise levels and hidden variables

Here we provide results for all the different hyperparameters that were tested during the simulations. In practice for our simulations where our models are linear with weights < 1 we assume that a shorter indirect edge will have a stronger indirect effect compared to a longer indirect edge. Therefore, we assume that the minimum integer that corresponds to the shortest lag between X^i and Y will also correspond to the maximum coefficient given by the lasso regression.

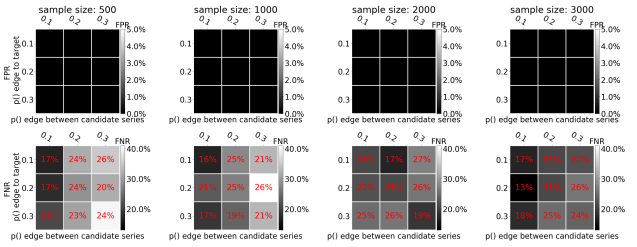
6.5.1. FPR AND FNR FOR VARIOUS DENSITIES

The x-axis on the heatmaps shows the probability of existence of an edge from the candidates to the target. The y-axis shows the probability of existence of an edge among the candidate time series. The first row of results always refers to the FPR and the bottom row to the FNR. The four different columns correspond to different sample sizes. Black color cell corresponds to 0 and white to 100%. The red-colored percentages incide the cells of the FNR correspond to the missed *dirrect* causes. We focus only on the FNR that correspond to the missed dirrect causes, because our conditions (Theorem 2) are necessary only for direct.

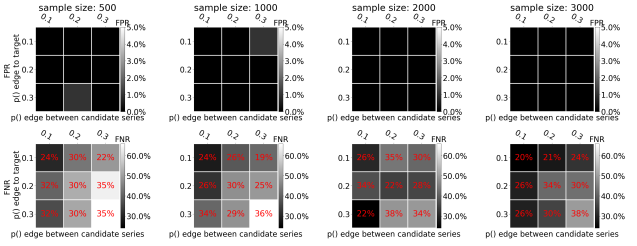
Here we provide heatmaps for all the noise variance levels and for all various number of observed time series that were simulated, for one hidden time series. The false positive and false negative rates are calculated over 100 random graphs created for each combination tested here.

Overall, the noise in the data does not seem to affect the results for sample sizes $\geq$ 1000. The false positive rate (FPR) is constantly close to zero for sample size > 500, and is not affected by the density and the size of the graph. The total false negative rate that refers to both direct and indirect missed causes (FNR) seem to gradually increase with the size and the density of the graph. On the other hand, the FNR that refers to the direct causes, for which we proved that our method is complete and sound, does not increase above 50% in very dense and large simulated graphs.

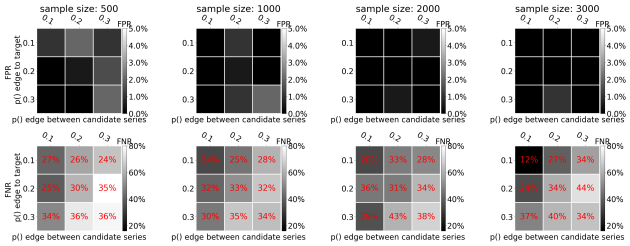
Results for low noise (0.1 noise variance):



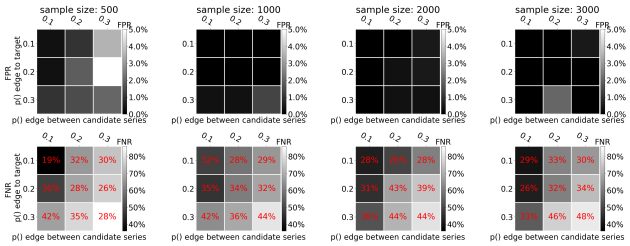
(a) 1 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).



(b) 2 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).

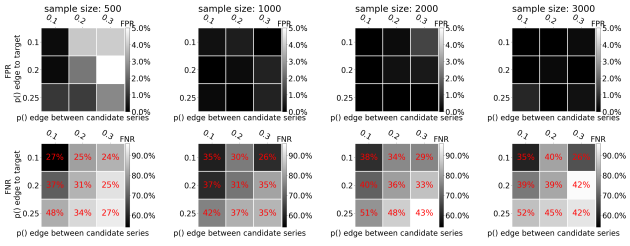


(c) 3 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).

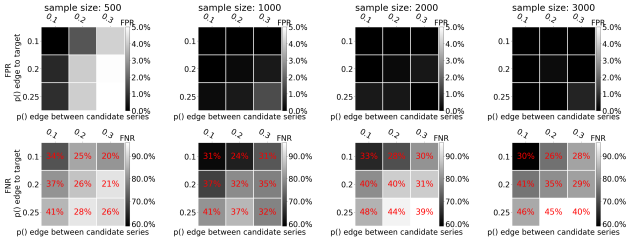


(d) 4 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).

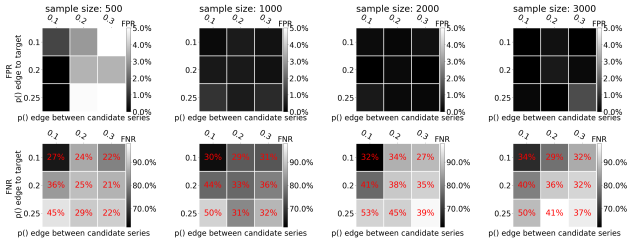
(e) 5 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).



(f) 6 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).



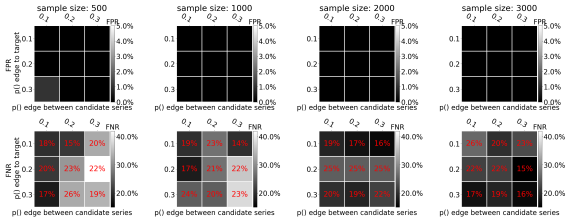
(g) 7 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).



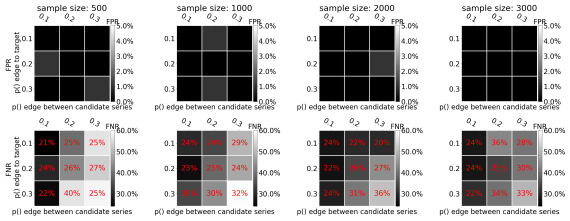
(h) 8 observed, 1 hidden and 1 target time-series, for low noise (variance 0.1).

Figure 7. FPR and FNR for low noise, various observed, 1 additional hidden and 1 additional target time-series, for different sample size (columns) and sparsity of edges among the candidate causes (x-axis) and between the candidate causes and the target (y-axis). The total FNR (for indirect and direct causes) is depicted by the heatmap color. The FNR that refers to the direct causes (for which our method is proved to be complete) is depicted with red in the middle of each cell. Overall we see tat for sample size above 500 the false positives are very low and they keep decreasing as the number of examples increase. False negatives for both direct and indirect causes increases with the number of nodes and the density of the graph, however the FNR that refers only to the direct causes for which our method provides necessary conditions (red coloured numbers) ranges just from 12% up to 52% for dense large graphs.

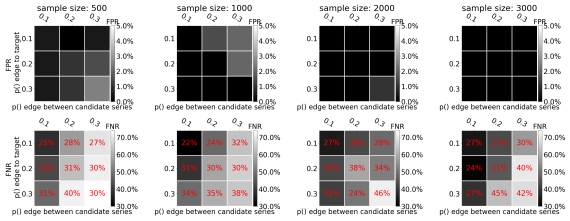
Results for medium noise (0.2 noise vairance):



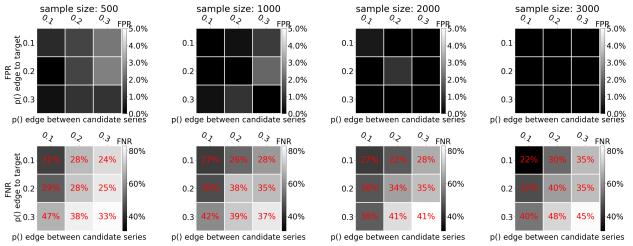
(a) 1 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



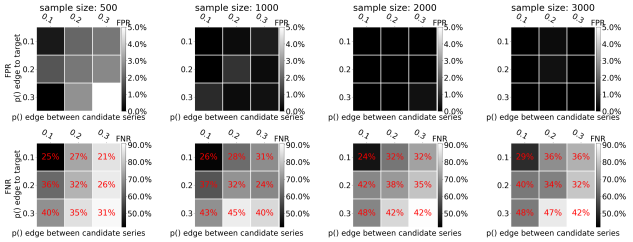
(b) 2 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



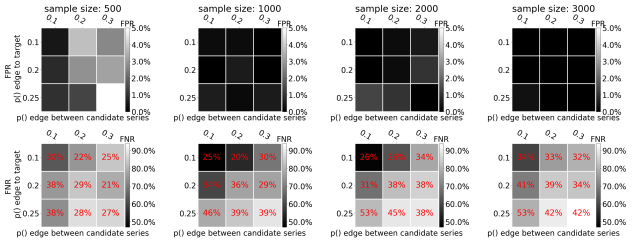
(c) 3 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



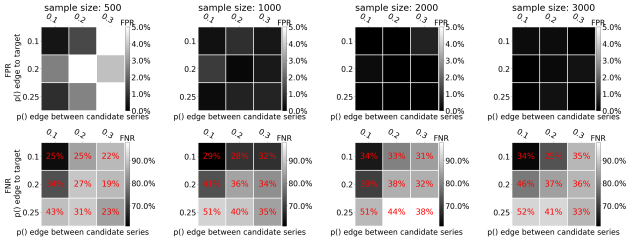
(d) 4 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



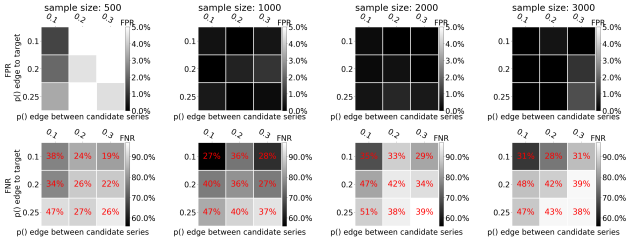
(e) 5 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



(f) 6 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



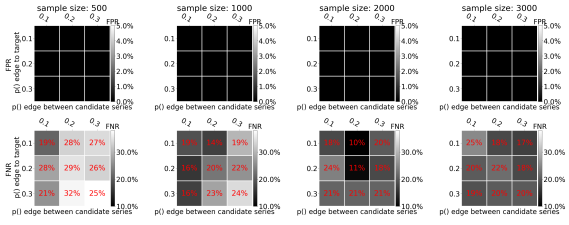
(g) 7 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).



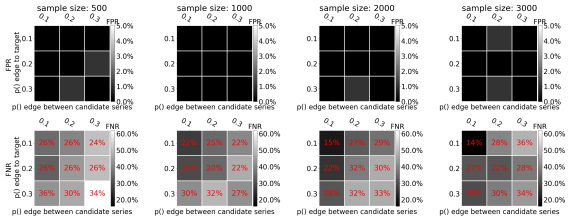
(h) 8 observed, 1 hidden and 1 target time-series, for medium noise (variance 0.2).

Figure 8. FPR and FNR for medium noise, various observed, 1 additional hidden and 1 additional target time-series, for different sample size (columns) and sparsity of edges among the candidate causes (x-axis) and between the candidate causes and the target (y-axis). Similar to the rest of the noise levels, the total FNR (for indirect and direct causes) is depicted by the heatmap color. The FNR that refers to the direct causes (for which our method is proved to be complete) is depicted with red in the middle of each cell. Overall we see tat for sample size above 500 the false positives are very low and they keep decreasing as the number of examples increase. False negatives for both direct and indirect causes increases with the number of nodes and the density of the graph, however the FNR that refers only to the direct causes for which our method provides necessary conditions (red coloured numbers) ranges just from 15% up to 52% for dense large graphs.

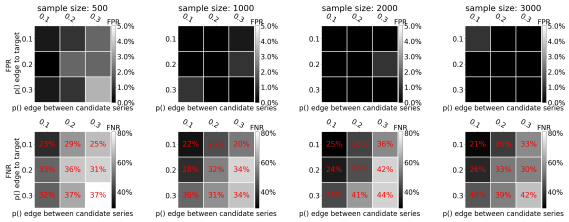
Results for high noise (0.3 noise vairance):



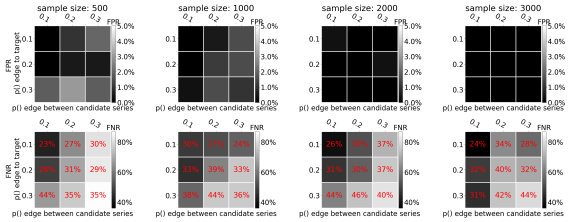
(a) 1 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).



(b) 2 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).

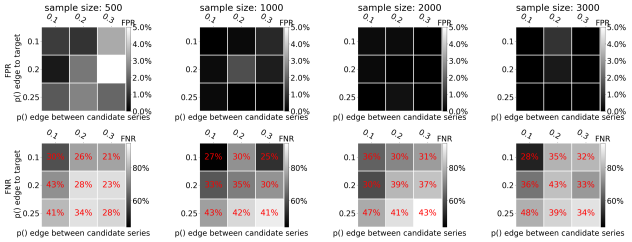


(c) 3 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).



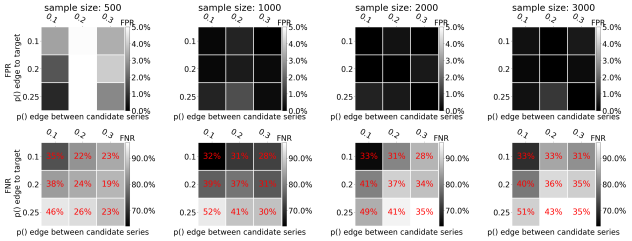
(d) 4 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).

(e) 5 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).



(f) 6 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).

(g) 7 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).



(h) 8 observed, 1 hidden and 1 target time-series, for high noise (variance 0.3).

Figure 9. FPR and FNR for high noise, various observed, 1 additional hidden and 1 additional target time-series, for different sample size (columns) and sparsity of edges among the candidate causes (x-axis) and between the candidate causes and the target (y-axis). Similar to the rest of the noise levels, the total FNR (for indirect and direct causes) is depicted by the heatmap color. The FNR that refers to the direct causes (for which our method is proved to be complete) is depicted with red in the middle of each cell. Overall we see tat for sample size above 500 the false positives are very low and they keep decreasing as the number of examples increase. False negatives for both direct and indirect causes increases with the number of nodes and the density of the graph, however the FNR that refers only to the direct causes for which our method provides necessary conditions (red coloured numbers) ranges just from 11% up to 51% for dense large graphs.

6.5.2. FPR AND FNR WITH VARIOUS HIDDEN VARIABLE FOR VARIOUS DENSITIES

In the presence of zero hidden variables our method has practically 0 false positives, which remains below 1% for large noise.

6.5.3. ROC CURVES OF SYPI AND LASSO GRANGER

To build the ROC curves, for Lasso-Granger, we varied the λ parameter across $\{0.00001, 0.0001, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For our method (SyPI), we varied the two thresholds *threshold1* and *threshold2*, keeping their ratio equal to 1, using values in $\{0.01, 0.02, \ldots, 0.12\}$. Figure 10 shows the ROC curve for the performance of SyPI and Lasso-Granger for the same graphs. Since our method functions with two conditions and two p-values, we did not manage to find logical pairs of thresholds that increase further the FPR. We see that at all operating points our method outperforms Lasso-Granger, with SyPI's ROC curve being above the Lasso-Granger one.

6.5.4. FPR AND FNR FOR "MULTIPLE-LAG DEPENDENCIES"

Although our theory is complete against false negatives only for single-lag dependencies, we wanted to test the performance of our method even in the presence of multiple lags. Therefore we examined the performance for 5 observed, 1 additional hidden and 1 target time series, for 2, 3 and 4 co-existing lag direct effects. We decide for the existence of a lag sampling from a Bernoulli distribution with $p = 0.5$. For Bernoulli probability of existence of an edge between the time series ($p1 = \{0.1, 0.2, 0.3, 0.4\}$) and the time series and the target ($p2 = \{0.1, 0.2, 0.3, 0.4\}$) we calculate FPR and FNR for different number of *lags* that can exist between the time series.

We fix the number of samples at 2000, and the noise variance at 20%. As depicted in Figure 11, our method seems to perform very well in terms of FPR, independent of the number of co-existing lags between the time series. As our method is complete only for single-lag dependencies, we don't expect low FNR.

6.6. Future work on multiple lags

The reason why our conditions are not necessary for "multiple-lag dependencies" is the difficulty in identifying just one lag from each time series to look at and to add in the conditioning set. If we did not put a lot of weight on keeping the conditioning set to a minimum size for assuring a decent statistical strength, we could still construct a conditioning set with as many nodes per time series as the multiple lags and have necessary conditions. In single-lag effects we describe why a single node from each time-series is necessary and sufficient and we show why this single lag can be the minimum lag as defined in 1. Without making any strong claims about the multi-lag case as it is out of the scope of this paper, we found the following: If we use the following condition, instead of the one defined in lemma 1, as $\max(v) \neq \inf$ s.t. $A_t \not\perp\!\!\!\perp B_{t+v} \mid \{A_{\text{past}(t)}, A_{\text{future}(t)}, B_{\text{past}(t+v)}, B_{\text{future}(t+v)}\}$, we managed to find only a very constrained bivariate case where it is enough to use the *maximum* integer v (max-int) as the w_i in the theorems and still have necessary and sufficient conditions. The case we managed to find that this is possible for multiple-lags is the following: Only in a bivariate (2 observed series) full-time graph with one candidate time series and one target time series, where hidden confounders are memoryless and with unique lag, given the above condition, max-int could be enough for differentiating between the time series causing the target with multiple lags and the time-series being confounded. If a node X_t^i has a direct edge both to Y_{t+1} and Y_{t+2}, then the "maximum " v would be equal to 2. If we used this as w_i in our conditions then, $Y_{t+w_i} \equiv Y_{t+2}$. Then conditioning on X_t^i, which is w_i steps back, and on $Y_{t+1} \equiv Y_{t+w_i-1}$, would render X_{t-1}^i and Y_{t+w_i} independent, so the two conditions of our theorems would hold.

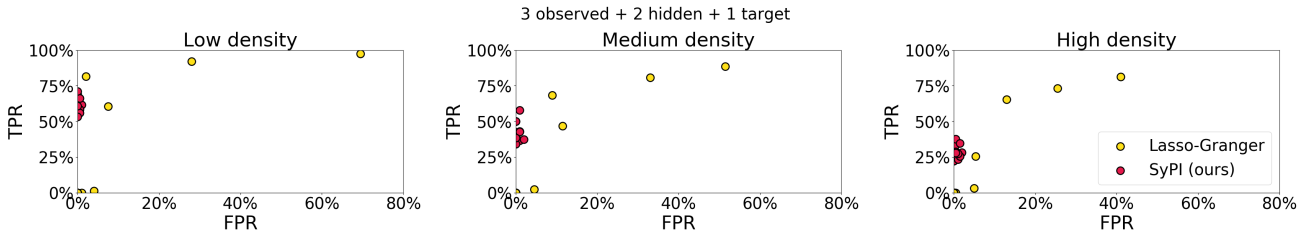


Figure 10. Yellow: ROC curve of Lasso-Granger for different values of the λ parameter. Red: ROC curve of our method for different values of *threshold1* and *threshold2* with fixed ratio of 1. The ROC curves were calculated over 100 random graphs, for different density of edges (three columns) and a moderate number of observed series with additional two hidden ones. Our method's ROC curve is always above the Granger's ROC.

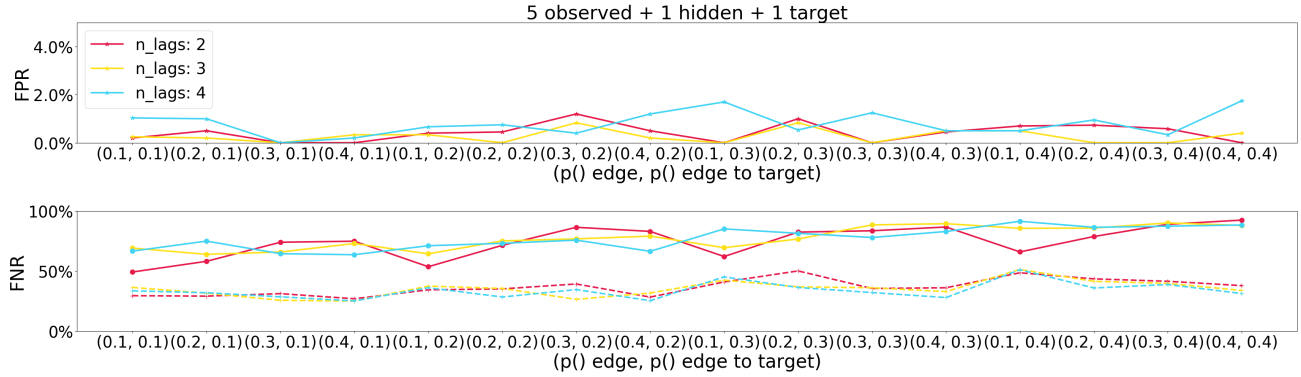


Figure 11. FPR and FNR for different number of coexisitng lags. Notice that the FPR is very low as expecteed by Theorem 1a/1b. Since our method is complete only for single-lag dependencies, we notice large FNR both for direct causes (dashed lines) and for indirect.

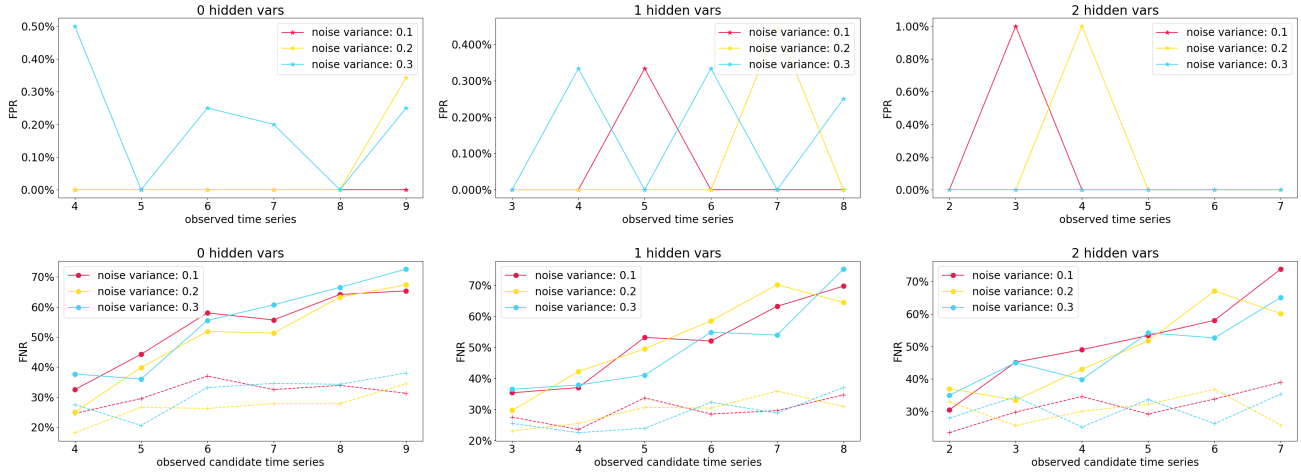


Figure 12. FPR and FNR for various number of hidden and observed series, noise variance and sample size 2000, for sparse edges among the $\mathbf{X}$ and Y (0.1, 0.1). As we can see, FPR is very low (max 1%) for any number of hidden series. Although the total FNR is gradually increasing with the graph size, notice that the FNR that corresponds to direct causes (dashed lines, for which our method is complete) does not exceed 35%.

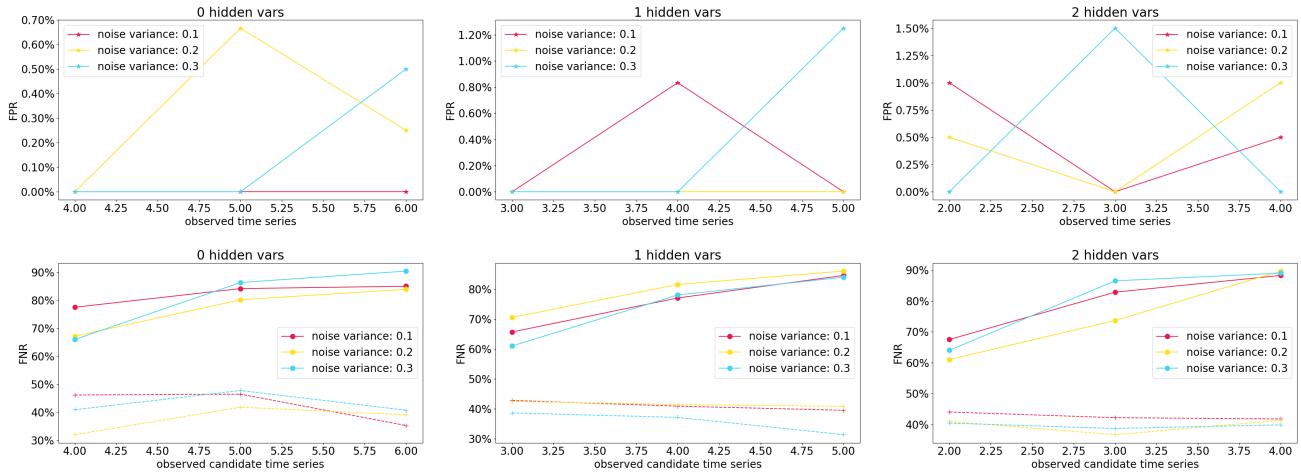


Figure 13. FPR and FNR for various number of hidden and observed series, noise variance and sample size 2000, for dense edges among the $\mathbf{X}$ and Y (0.3, 0.3). As we can see, FPR remains very low (max 1.5% for high noise) for any number of hidden series. Although the total FNR is gradually increasing with the graph size, notice that the FNR that corresponds to direct causes (dashed lines, for which our method is complete) does not exceed 45%.