

Causal discovery and inference of data from production flows

Jonas Bruun Hubrechts



Kongens Lyngby 2024

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

The goal of the thesis is to ...

Summary (Danish)

Målet for denne afhandling er at ...

Preface

This thesis was prepared at DTU Compute in fulfilment of the requirements for acquiring an M.Sc. in Engineering.

The thesis deals with ...

The thesis consists of ...

Lyngby, 31-July-2024

Not Real

Jonas Bruun Hubrechts

Acknowledgements

I would like to thank my....

Contents

Summary (English)	i
Summary (Danish)	iii
Preface	v
Acknowledgements	vii
1 Problemformulering / Introduktion	1
2 Data	3
2.1 Basic statistics	4
2.2 Incompleteness on trailing batches	5
2.3 Production phases	7
2.3.1 Correlations	10
2.4 Cleaning operations	15
3 Method	23
3.1 Causal discovery	23
3.1.1 Setup and assumptions	24
3.2 Information measures and computation	28
3.2.1 Copula	28
3.2.2 Mutual information and Copula entropy	30
3.2.3 Entropy and mutual information in the limit	33
3.2.4 Correlation	35
3.3 Copula based network discovery	36
3.3.1 Network deconvolution	38
3.3.2 Ensuring convergence and the effect of β	40
3.3.3 Robustness to noise	41

3.4	Estimation of mutual information	45
3.4.1	B-splines	45
3.4.2	M-splines	46
3.4.3	A better spline	47
3.4.4	Naïve KDE	49
3.4.5	Boundary corrected KDE	53
4	Results	61
4.1	Gaussian chains	62
4.1.1	Gaussian chain deconvolution using correlation	64
4.1.2	Gaussian chain deconvolution using mutual information .	70
4.2	Directed acyclic Gaussian graphs	75
4.3	CE computation	82
4.3.1	Exponentiated multivariate Gaussian	82
4.4	sammenligning af metoder for at finde MI	91
4.4.1	10D gaussian example	91
5	Conclusion	93
6	Appendix	95
6.1	Suicide data	95
6.2	Confidence interval for absolute correlation in bivariate Gaussian	96
6.3	Gaussian chain deconvolution	98
6.4	Gaussian network deconvolution	99
	Bibliography	101

CHAPTER 1

Problemformulering / Introduktion

In many production facilities, planning is a big part of maximizing some index. Whether this is production throughput over some time period and thus often also the economic surplus or some other key index, it is of great importance to have an underlying model to describe the observed variation. In particular in operational research, the schedules may drift in suboptimal ways if the variation is not considered.

Furthermore, from a salesman point of view, expected production and time intervals can be of great use when planning and also building production facilities. Namely, one might find that increasing the volume or efficiency of some part of the facility would increase the production throughput and profitability. This is also known as bottleneck analysis and require some understanding of the underlying mechanics and a stochastic model of this could improve the strength of such results.

Therefore, the primary objective of this paper/thesis is to investigate and model the yield and time of a production flow with focus on the pharmaceutical and chemical production industry. More precisely, we will be building a statistical model for a single process, with the purpose of being able to describe the variation in the yield of the production cycle and production times. This will then be used to analyze potential bottlenecks.

Furthermore, it will be interesting to construct a network of such processes as is typically the case in industry. We shall see how much can be said about such a network and what obstacles one may encounter when trying to analyze such networks which is this thesis will initially be treated as networks of queues.

CHAPTER 2

Data

In this section, we will describe and analyze the data used in this thesis. The data stems from a simulation study on a chemical batch production system and comprises six time series from which the time stamps (in hours) and level in a tank is of the most interest. The goal of this section is to describe how different phases of the production covary. The data is chosen as it resembles an actual production data set but is also implemented with what at first glance seem to be fairly realistic variation and noise in measurements. Overall, the point of this example data is to exemplify what one may encounter in a real batch production system and from this try to build a model in order to predict or quantify the behavior of the system or learn hidden (causal) structure important for optimization etc.

We define a set of phases/units \mathcal{U} that each batch consists of. In this case, units are identified with IDs 1 through 10 (with subunits such as 3.1) described further in Table 2.2. Then, for each unit, we define the random variable X_u to be the total time a batch spends in unit u i.e. the sojourn time related to unit u . We also define the random variables X_u^P and X_u^D to be the duration of the actual *processing* and *delay* respectively such that $X_u = X_u^P + X_u^D$. It is also important to keep track of the level in the tank after each unit is finished, and we thus define variables M_u and M_u^D to be the level in the tank after unit u and its associated delay respectively. As the units are executed in sequence, as is the case in this data, a simple representation of the variables can easily be

visualized and is shown in Figure 2.1 below.

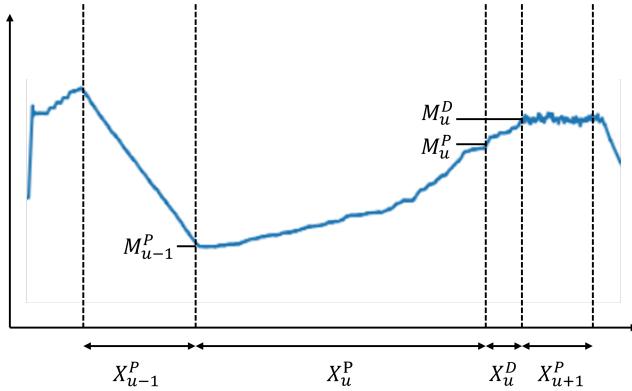


Figure 2.1: Exemplification of the variables X_u^P , X_u^D , M_u^P , M_u^D .

We note that during the delay for unit u , there is a change in mass while also, delays are will not always occur between operations. A further investigation of the data set shows that it is common in some units that delays effect the mass while for others it does not.

2.1 Basic statistics

Initially, we present some basic statistics in Table 2.1 for batches and proceed to discuss the nature of the batches, removing outliers and faulty observations etc.

Cycle	#batch	$\hat{\mu}$	s^2	s	$s/\hat{\mu}$
A	66	14.776	3.641	1.908	0.1291
B	64	15.644	3.915	1.979	0.1265
C	61	17.714	2.330	1.526	0.08617
D	60	18.069	6.922	2.631	0.1456
E	60	18.088	9.613	3.100	0.1714
F	63	17.227	7.766	2.787	0.1618
all	374	16.876	7.218	2.687	0.1592

Table 2.1: Per cycle batch duration statistics i.e. $\sum_u X_u$ per batch and for all batches combined

Each batch comprises several states. These include adding materials (IDs 1 through 4), centrifugation (ID 5), product transfer (the precipitate generated

from the centrifugation, ID 6), chemical reaction (ID 7), a post operation state (**Probably to let it cool down to a point where it is ready for further processing**, ID 8), Cooling of the product (ID 9), material transfer (transfer the gained product before cleaning of the reaction vessel and/or prepare for the next reaction batch, ID 10). Notice that there is a total of 374 batches throughout the 6 observed cycles.

2.2 Incompleteness on trailing batches

As it may be of interest to investigate the correlation structure of different metrics and variables later on, it is important to understand how each of the batches across the cycles behave. Initially, when looking through the dataset, we observe a few negative phase IDs which will need investigation. However, before we do so, we check that each of the batches actually go through all the states mentioned in [1]. Thus, we take the absolute value of the negative phase IDs to ease the analysis prior to the analysis of the negative phase IDs. After this is done, we observe that not all batches go through all the phases and that some seem to have extra phases not described by [1]. Namely, from Table 2.2, we see that IDs 3 and 4 (which are not described in [1]) have significantly fewer batches going through this phase. But perhaps even more interesting is the final 4 phases where almost all batches goes through these phases.

ID	Count	Description
1.0	374	Addition of liquid raw material Educt1
2.0	374	Addition of liquid raw material Educt2
3.0	181	
3.1	374	Addition of liquid raw material Educt3
3.2	374	Agitation
4.0	163	
4.1	374	Waiting for field operation
4.2	374	Addition of solids
4.3	374	Waiting for control operator
5.0	374	centrifugation
6.0	374	Product transfer
7.0	370	Reaction
8.0	369	Post reaction
9.0	369	Cooling
10.0	368	Material transfer

Table 2.2: The number of batches across all cycles that contains at least one observation for each different absolute phase ID.

Investigating when these apparent inconsistencies occur reveals that they are the last batch from each of the cycles. For example, the final batch from cycle A only goes to phase 6 (the product transfer). This can however be explained from the fact that simulation only last for 1100 hours for each cycle and is thus simply cut-off here. As we do not know if these final operations were done at the time the simulations were cut off (which is likely not the case), the final phase for each of the final batches should be disregarded. The cut-off can also be observed in Figure 2.2. Furthermore, throwing away 6 incomplete batches out of the total 374 will likely not harm the analysis and is thus thrown away as this will make the analysis much simpler later on.

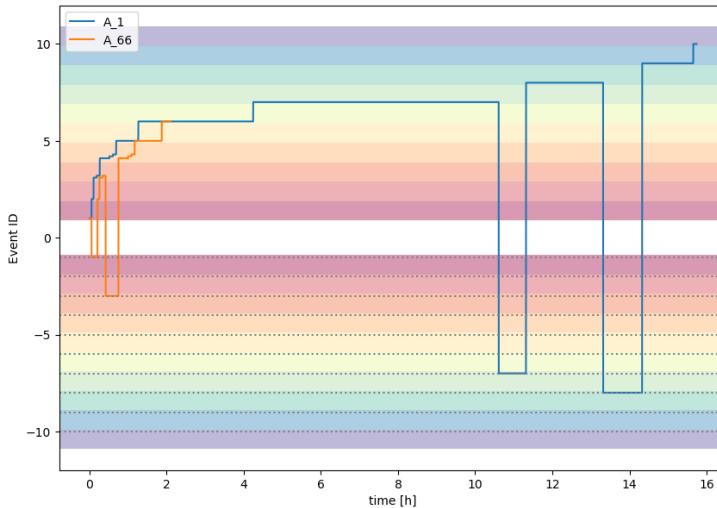


Figure 2.2: The first and last batch from cycle A. It is clear that the final batch is cut-off even before the current phase it finished.

After cutting of the final 6 batches, we have a total of 368 batches of which each goes through all the phases. We thus proceed to discuss the negative phase IDs in the following section, where we also discuss the first four phases.

2.3 Production phases

This part of the process corresponds to events tagged with ID 1 through 10 but will initially concern itself with ID 1 through 4 as much can be learned from the data set here. In Figure 2.3 an example of how the process evolves over time through the different phases is shown. Immediately, we observe something weird, namely the negative event IDs.

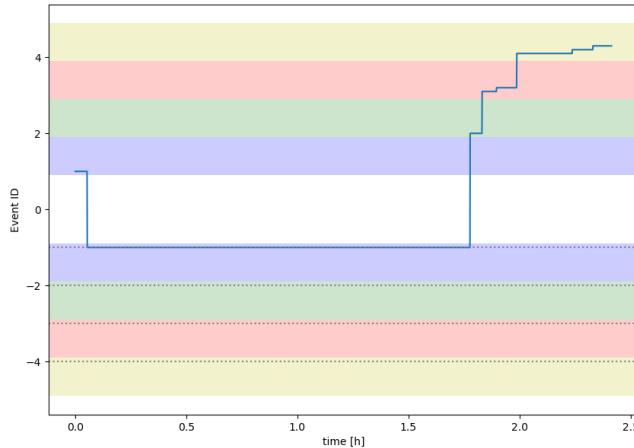


Figure 2.3

To see what is going on here, from data we can see that negative values occur throughout all the six cycles. More specifically, for each negative phase observed, we log in which cycles this occurs. The result is shown in Table 2.3. Notice that -4.1, -4.2 and -4.3 only show up in cycle F, which from [1] is known to be the only one with wrongly labelled phases. We thus suspect that this is indeed the case for these labels and might just have supposed to be the original 4.1, 4.2 and 4.3. To see if nothing funny goes on with these values, these batches are plotted as in Figure 2.3 in Figure 2.4.

Figure 2.4 shows that nothing weird is going on except for the negation of the sub phase's ID. The same can be said for the remaining of the cases where phase ID -4.1, -4.2 and/or -4.3 is used. We thus conclude that these may simply be wrongly labelled thus we convert every such instance to its absolute value and continue with this modified data set from this point on.

Event \ Cycle	A	B	C	D	E	F
Event						
-1						
-2						
-3						
-4						
-4.1						
-4.2						
-4.3						
-5						
-6						
-7						
-8						
-9						
-10						

Table 2.3: Occurrences of negative phases IDs. It is observed that sub phases 4.1, 4.2, 4.3 only occur in cycle F which is known to be the only cycle with wrongly labelled phases.

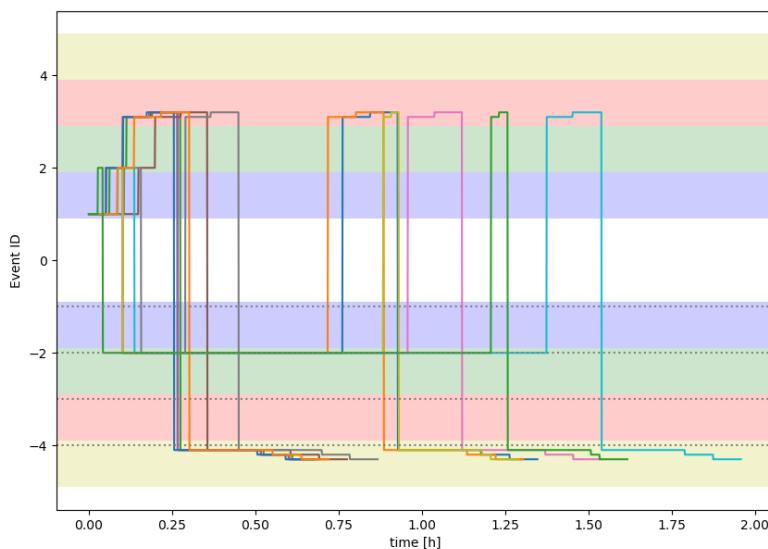


Figure 2.4: 13 of the 48 batches with at least one of the sub phases 4.1, 4.2, 4.3 negative.

Having converted the above sub phase IDs we summarize the current situation in regard to negative phase IDs in the following table, Table 2.4. Now all the remaining occurrences of negative phase IDs does not seem to exhibit any structure from looking at Table 2.4. We thus proceed to understand what is going on with the remaining negative phase IDs.

Event \ Cycle	A	B	C	D	E	F
Event						
-1						
-2						
-3						
-4						
-5						
-6						
-7						
-8						
-9						
-10						

Table 2.4: Occurrences of negative phases IDs. It is observed that sub phases 4.1, 4.2, 4.3 only occur in cycle F which is known to be the only cycle with wrongly labelled phases.

When plotting different batches, it is clear that the negative phase IDs only occur at the end of a phase e.g. -1 only happens after 1 and so on. This together with the fact that -3 and -4 also only happen after 3.1, 3.2 and 4.1, 4.2, 4.3 respectively (and we never see a phase labelled 3 or 4) indicate that the negative phase IDs could very well correspond to delays at the end of a phase, which both makes sense from a production point of view but also from [1] where they note that all simulated cycles have been implemented with delays.

At this point, it would seem that the labels of the processes are understood for phases 1 through 10 corresponding to the actual production in each batch. Thus, we proceed by searching relationships and otherwise quantifying the durations of each phase, both delays and duration for each of the phases. As a beginning, histograms for each of the phases and delays are plotted in Figure 2.5. Notice that phases 4.1, 4.3 and 8 are not shown, this is because they always last 15 min, 5 min and 2 hours respectively with the only derivation being in machine precision either when loaded or during calculations. Furthermore, notice that for the negative IDs i.e. the delays, the orange bar. This bar represents the cases where no delay was observed which is thus modelled as an atom at 1.

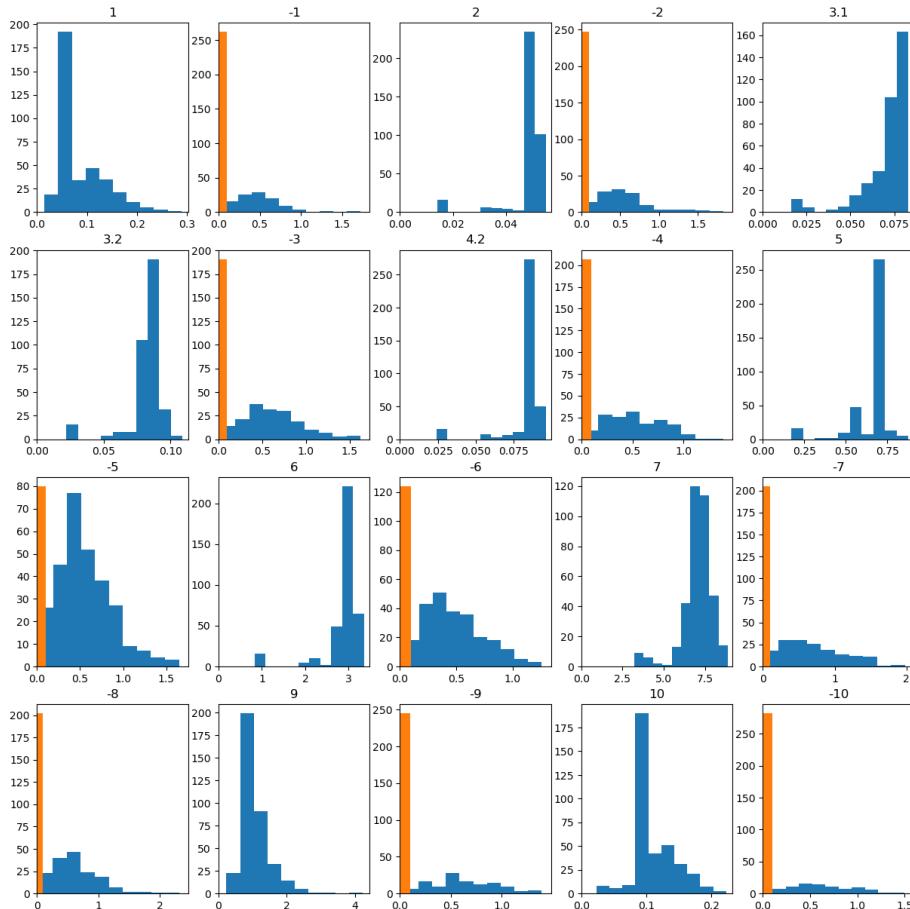


Figure 2.5: Histograms of all phases and delays which are non-constant.

Apart from the above comments, not much catches the eye when looking at Figure 2.5, and we thus proceed by checking if any correlation is immediately present.

2.3.1 Correlations

Lige en korrelationsmatrix

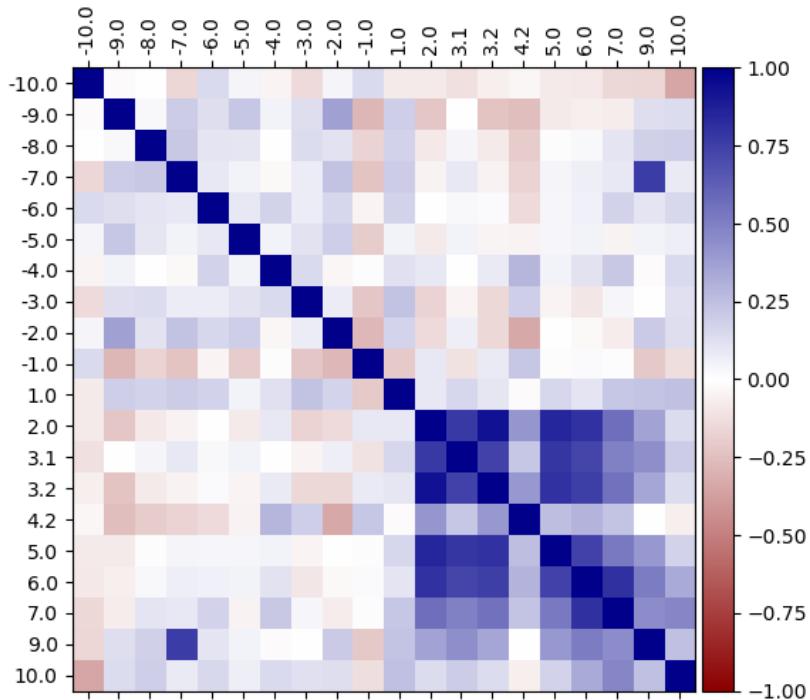


Figure 2.6: Correlation matrix for all phases with non-constant duration.

9 og 10 er ikke specielt korreleret med noget (afkøling og materiale overførsel). Ellers er 2 frem til og med reaktionen alle korrelerede med hinanden. Eftersom rent fysisk det udvikles i tid, må handlingen i 2 påvirke de næste osv.

Umiddelbart lidt spøjst hvis delay på 7 (reaktion) skulle have noget med tiden for afkøling at gøre, især at den skulle være positiv (ville man ikke tro delay efter produktion ville afkøle mere og dermed reducere behov for afkøling, medmindre varmt steam bliver tilføjet også under delay på 7)

Herunder er samme korrelationsmatrix, dog hvor delay og phasens varighed lagt sammen (også med sub phases såsom 3.1, 3.2 og -3 tilsammen bliver 3)

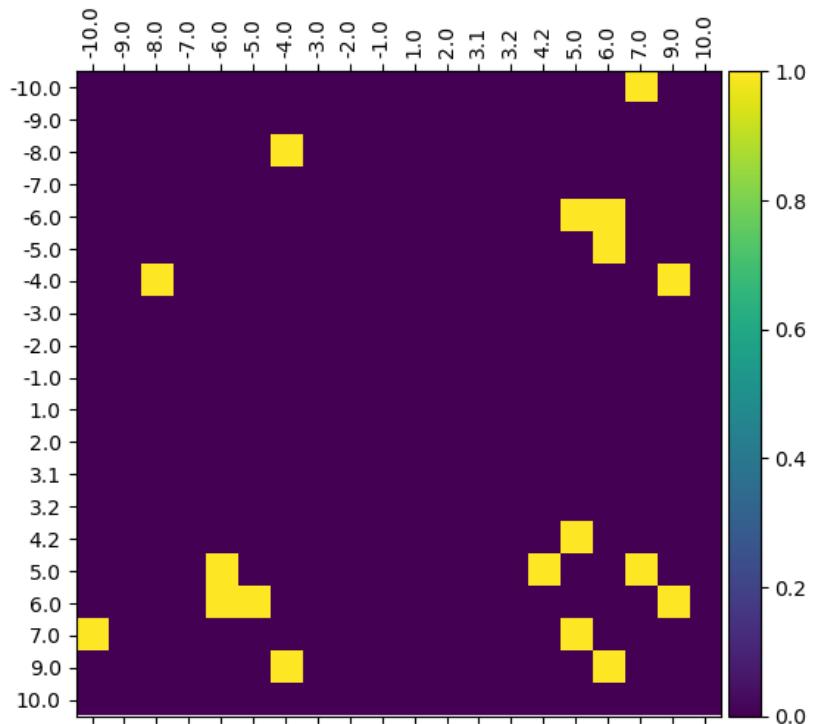


Figure 2.7: Permuteringstest med $\alpha = 0.05$. Also run with less simulations but same result at 1 mil and 10k sims. The Benjamin-Hochberg procedure on the upper (or lower) triangle reveals that none of the correlations are significant.

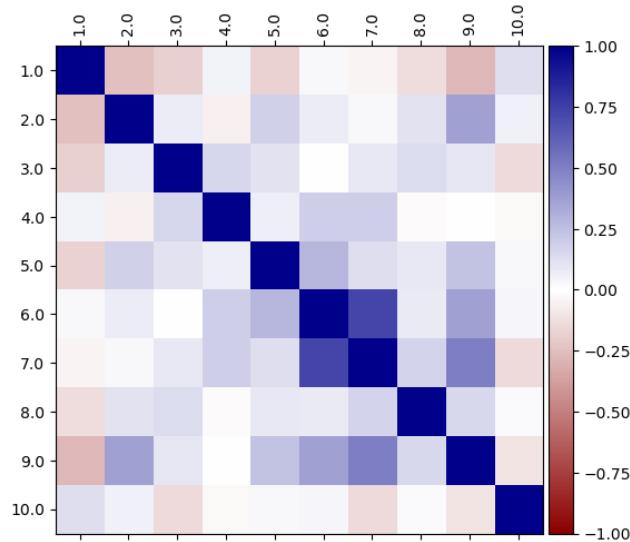


Figure 2.8: Correlation matrix for all phases collapsed

Ligeledes scatter plots for superdiagonalen i ovenstående matrix. Altså phase 1 overfor phase 2, phase 2 overfor phase 3 osv. Er farvelagt efter hvilken cycle de kommer fra. Table 2.4 forklarer hvorfor nogle af de horisontale fremkommer sammen med Figure 2.5 (selve produktionstiden er ret kort sammenlignet med delay.)

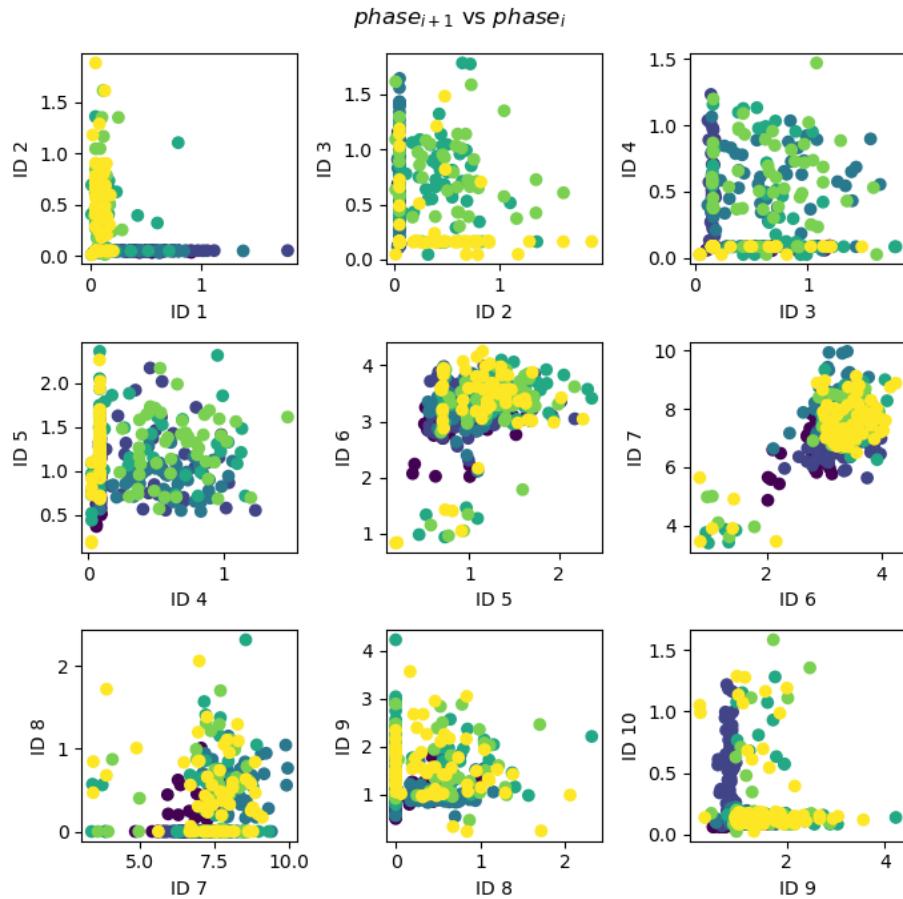


Figure 2.9: Phases vs their next phase when collecting everything regarding a single phase into a total time duration

2.4 Cleaning operations

Sometimes, the vessel is cleansed. This is however not every time after a batch so might be interesting to investigate further. Initially, per cycle, the cleanings are summarized in the following table with basic statistics. As can be seen, there is quite some differences.

The most notifiable differences per batch are the number of cleanses especially when comparing to Table 2.1. For the first two cycles, the cleanses seem to be in between every batch, which is indeed also the while the later four are only sometimes. Furthermore, although the cleanses are between every batch for cycles A and B, the variances are extremely different. For the last four cycles, they seem to be grouped further, E and F are very alike while cleanses in C and D are generally longer although D has a substantially smaller variance than C.

Cycle	#ops	min	max	μ	σ^2	σ	σ/μ
A	65	1.113	3.067	1.917	0.269	0.518	0.270
B	63	1.324	1.751	1.566	0.00883	0.0939	0.0600
C	9	1.544	3.306	2.153	0.277	0.526	0.245
D	10	1.474	2.009	1.581	0.0212	0.146	0.0922
E	10	0.827	1.584	1.465	0.0462	0.215	0.147
F	10	0.748	1.610	1.466	0.0595	0.244	0.166

Table 2.5: Per cycle cleansing statistics

Cycle	A	B	C	D	E	F
$\mathbb{E}[\sum X_u]$	13.993	13.898	15.343	14.471	14.589	14.418
$\text{Var}(\sum X_u)$	0.95636	0.46587	0.76111	4.9589	4.2678	5.3545
$\sum \text{Var}(X_u)$	0.50590	0.31182	0.36667	1.8322	1.5788	1.9696
$\mathbb{E}[\sum X_u^D]$	0.96398	1.9402	2.4503	3.6050	3.7390	3.0041
$\text{Var}(\sum X_u^D)$	0.31843	0.39117	0.90187	1.2468	1.2787	1.0462
$\sum \text{Var}(X_u^D)$	0.34921	0.53198	0.74914	1.4357	1.2454	1.3099
$\mathbb{E}[\sum X_u X_u^D]$	1.9321	1.5001	4.8191	6.4225	6.0405	6.3343
$\text{Var}(\sum X_u X_u^D)$	3.7798	0.89920	16.870	22.133	12.660	16.194

Table 2.6: Each of the time related variables X_i and D_i and variance description.

To verify these observations and potentially discovering more important facts of their probability distributions, histograms are plotted in the following Figure 2.10. We indeed again observe the likeliness between the cycles A and B, C

and D, E and F respectively. Also, for the first two cycles and more so cycle B, the cleaning times are somewhat normally distributed although cycle A has a very heavy right tail in that case. The later four cycles only have 10 observations but the mode (i.e. peak) seem to be about the same.

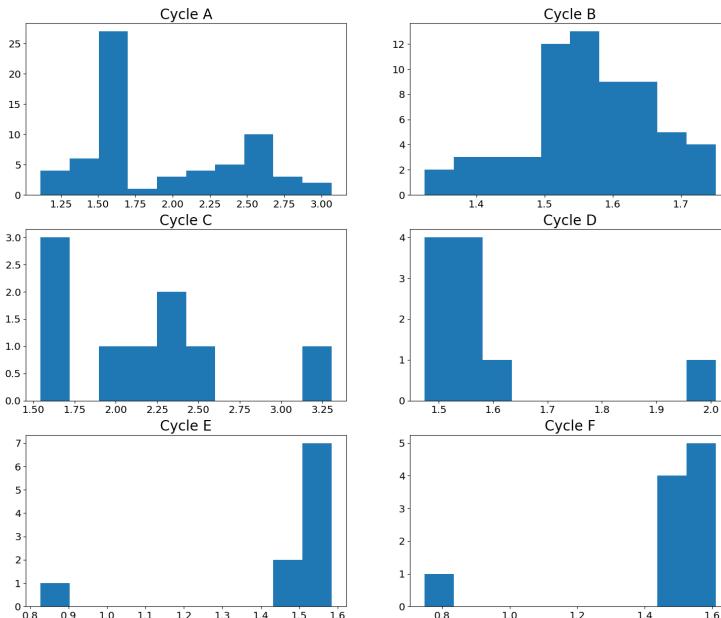


Figure 2.10: Each of the 6 cycles, cleaning operations histograms.

From the above observation of like modes one may want to observe the histogram of the combined set of cleaning times. In particular, under the hypothesis that the durations are actually from the same probability distributions and realized independently within each cycle a histogram of all the observations are of interest and is shown in Figure 2.11 below.

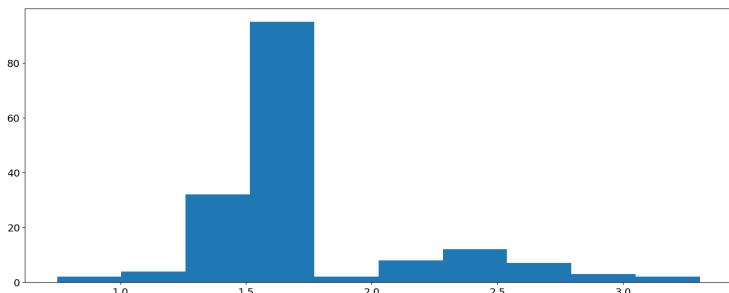


Figure 2.11: Combined cleaning operations histograms.

Finally, to get a better overview of the irregularities in the number of cleaning periods (mostly concerning cycles C through F), each cleaning operation is shown in the following Figure 2.12. The vertical shaded rectangles signify the period in which a cleaning operation is taking place. Furthermore, the event IDs are shown but to get a clearer view on what is going on, a single rectangle (zoomed in) is shown in Figure 2.13.

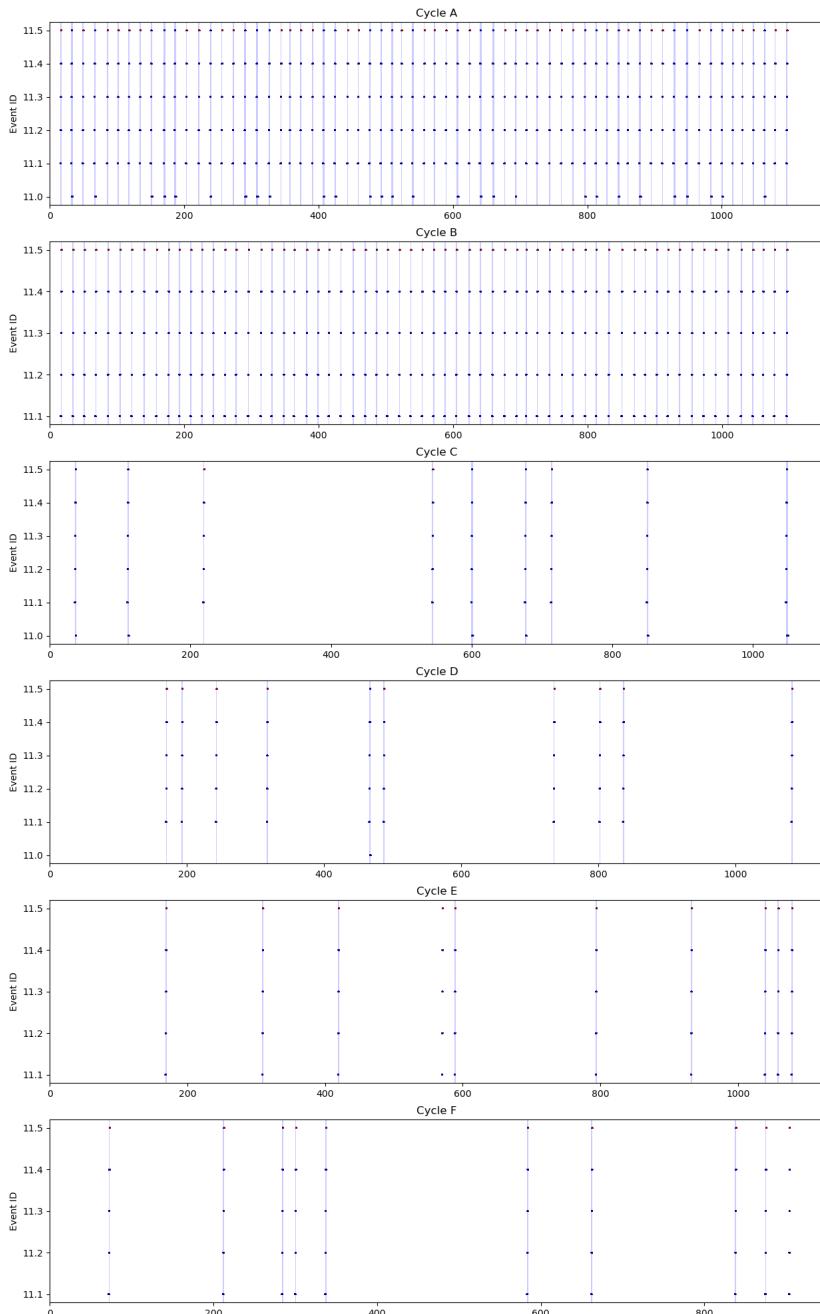


Figure 2.12: Each of the 6 cycles, cleaning (corresponding to $\text{BatchID} = 0$). Each (Cleaning Procedure), CIP, is highlighted with an opaque interval (the blue rectangles). The dots marked with red (only ID 11.5, but not all of these are red), is if the Cleaning ID is 0.

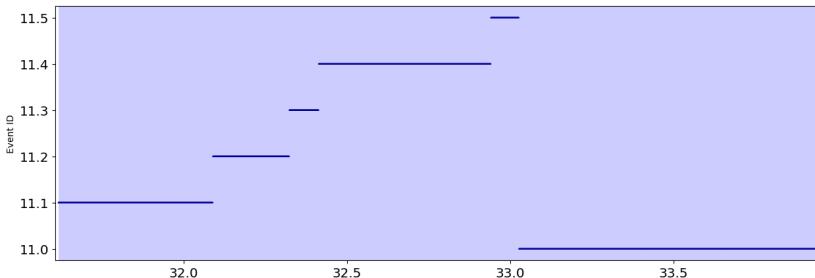


Figure 2.13: A single blue rectangle zoomed in

It is observed that the observations marked with red in figure 2.12 occur exactly when that specific cleaning operation does not go to the state 11.0 after the flush of the tank (event ID 11.5) and vice versa. It is hard to conclude what this may mean, but the cleaning being in state 11.0 may indicate that the system is idle before continuing the next batch like what is observed from the other steps of the process flow. Also, it is noted that while the red dots occur nothing else is happening according to the dataset.

From a modelling point of view, the cycles C through F can be thought of as the cleansing operation having a probability of not happening or equivalently as having a duration of 0. It is thus of interest to observe what the probability of cleaning after an operation is. From Table 2.1 and Table 2.5, we see that indeed for cycles A and B, the probability is 100 % when disregarding the possibility of cleaning after the final batch. Hence, we see that for the remaining cycles, the probabilities of cleaning the tank after an operation are as in Table 2.7

Cycle	% cleaning
A	100.00
B	100.00
C	15.00
D	16.95
E	16.95
F	16.13

Table 2.7: Per cycle probability of cleaning

Furthermore, let C_i denote whether the i th batch is followed by a cleaning of the tank or not. It is then of interest if the next batch is followed by a cleaning given whether the current batch is followed by a cleaning. In particular, we count for each of the cycles the transitions which are shown in the following tables. Notice that the number of observations is two less than the total number of batches within each specific cycle. This is due to the last batch is never followed by

a cleaning (nor is the first batch superseded by a cleaning procedure) which results in one less observation and also due to the fact that we are logging transitions and hence lose another observation. To test for randomness, a Chi-squared test is carried out on each of the cycles to check for independence. It is observed all the cycles exhibit independence between the groups i.e. there is no statistical evidence for information is gained about if the next batch is followed by a cleaning operation given whether the current batch is followed by a cleaning operation.

C_i	C_{i+1}	No	Yes
C_i	No	41	9
	Yes	9	0

(a) C, $p = 0.3293$

C_i	C_{i+1}	No	Yes
C_i	No	41	7
	Yes	7	3

(c) E, $p = 0.3532$

C_i	C_{i+1}	No	Yes
C_i	No	41	8
	Yes	7	2

(b) D, $p = 0.6456$

C_i	C_{i+1}	No	Yes
C_i	No	41	9
	Yes	9	1

(d) F, $p = 1.0000$ **Table 2.8:** Contingency table for Cycle C-F

Thus collecting the observations from all the last four cycles, we may want to model the atom of the cleaning procedure independently of the previous batch and with a probability of 0.8375 corresponding to the cleaning procedure only being carried out 16, 25% of cases.

Finally, we show the autocorrelation function for each the four cycles C-F in Figure 2.14 and note that all the ACF stay within the 95% confidence interval.

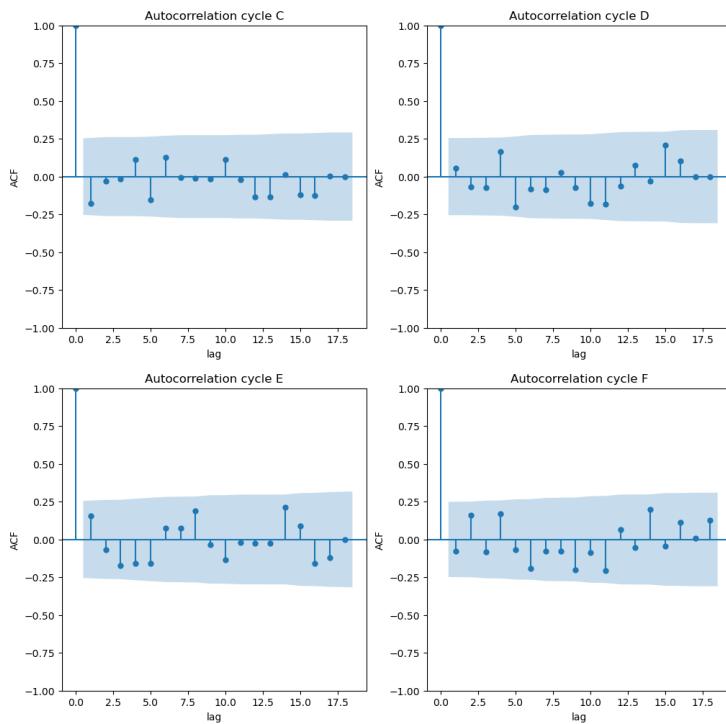


Figure 2.14: Autocorrelation function for each of the final 4 cycles. As can also be seen from this there seem to be no information to be gained of C_i from C_{i-1} .

CHAPTER 3

Method

General introduction to metodeafsnittet og causalitet

3.1 Causal discovery

In this section, we shall discuss the method for network deconvolution, originally proposed by [2]. The underlying problem is inferring direct effects and dependencies. From this, using prior information on the production setup, we shall be able to infer causal dependencies by directing the resulting edges from the network deconvolution (ND) algorithm. Particularly, the framework and general algorithm proposed by Feizi et al. stems from a graph-theoretic approach to the problem of inferring direct dependencies. **Namely, suppose that observations are made of some properties such as levels and sojourn times of in this case a chemical process.** We shall represent these properties as vertices (nodes) V and dependencies between properties as edges. Initially, when observing the vertices, we observe both direct and indirect effects. Particularly, a vertex v_1 might influence some other vertex v_3 through another vertex v_2 if v_2 depends on v_1 and v_3 or v_2 . In this case, we will observe that v_1 influences v_3 , but actually it is v_2 that has a direct influence on v_3 . In graph-theoretical terms, we thus observe the transitive closure of the information that flows between vertices but want to infer the underlying network structure.

An important note on the algorithm to come is that we only use vertices that we have observed. Namely, the underlying structure might be as in Figure 3.1a with an unobserved node/variable (named U in this case). However, without any more assumptions or modelling choices we would (ideally) infer the network structure depicted in Figure 3.1b. With these initial comments, we proceed

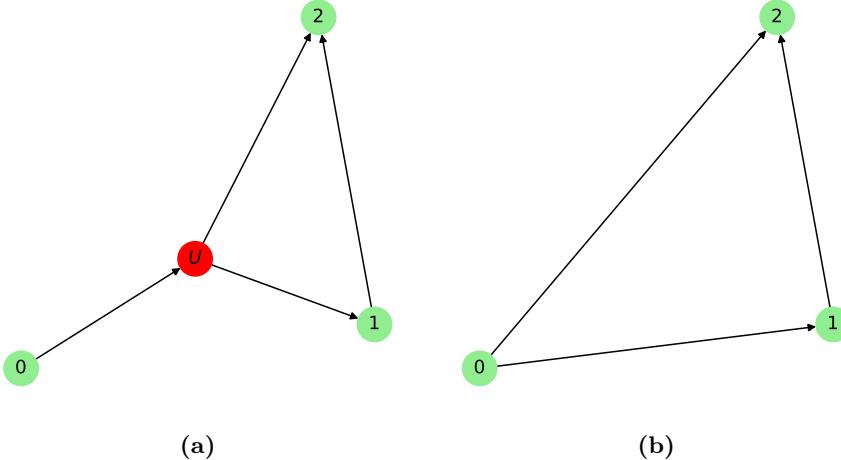


Figure 3.1: (a) An example of a causal structure depicted as a graph. When observing the network, only nodes 0, 1 and 2 are observed/recorded. (b) The resulting inferred graph from observational data. Although this is not a complete picture of the true underlying dynamics of the system, if only the observed variables are of interest, this will be an equally proper representation of the system. Furthermore, in practice this means no further assumptions are made which can and can not be of desire. Namely, if prior information is accessible one might introduce new nodes in the inferred network.

with the general setup and assumptions for network deconvolution based on observations.

3.1.1 Setup and assumptions

Suppose a set of N random variables (X_i) is given. The method presented in this section aims to discover direct relationships between pairs X_i and X_j for $i \neq j$. These relationships will be presented by a directed graph as in the previous section or an undirected graph in case the causal direction is either unknown or such an assumption on direction is not plausible. In particular, we shall let each random variable X_i be represented by a vertex in a graph. We

will later discuss a way of directing edges such that a causal network may be discovered i.e. a directed acyclic graph that may be used for inference.

The method proposed by [2] then works as follows. Given an observed matrix $G_{obs} \in \mathbb{R}^{N \times N}$ of similarities between each pair of variables, we shall deduce a matrix $G_{dir} \in \mathbb{R}^{N \times N}$ of direct similarities between each pair of random variables X_i and X_j . In particular, we wish to filter out indirect effects which we will denote by G_{dir} defined as effects between pairs of variables that is the result of effects propagating through other variables. The measure of similarity, can in practice be any desired measure such as correlation or mutual information which we will focus on in this thesis. See Section 3.2 for a further discussion on these two measures and Section 3.3 for how to obtain such a matrix. Note that the algorithm presented will in theory work for non-symmetric measures as well such as *Interaction information*, *Directed information* and *Normalized information*.

Illustration af hvordan algoritmen fungerer/formål

The (direct) network is then presented by the discovered G_{dir} containing only the direct effects i.e. interaction between pairs of variables which can be viewed as weights on the edges of the complete graph with nodes representing the random variables. As we shall see in Subsection 3.3.3, the algorithm is somewhat robust to noise in the sense that we can ensure accuracy depending on the level of noise observed present in G_{obs} and on the norm chosen (from a certain, although rather general, set of norms). Namely, if G_{obs} is subject to noise, we find a bound on how different the inferred directed effects can be to the true direct effects using different matrix norms to measure this difference. This hints to that a threshold on the inferred weights on the edges of the network might be a good idea to remove small inferred effects. This is further supported by the facts that often only the most influential variables are of importance when trying to control the process.

The first assumption is that the observed matrix of co-dependence G_{obs} may be expressed as

$$G_{obs} = G_{dir} + G_{indir} \quad (3.1)$$

Namely, that the direct and indirect effects can be added together to get the total and thus observed interdependence between each pair of variables. Often, this is not the case as we shall see later on. However, the error made from this assumption and the ones to be presented seem to be small enough that the discovered network accurately resemble the true underlying network.

The second and final assumption is that the indirect effects G_{indir} can be com-

puted in terms of G_{dir} . Namely, that

$$G_{indir} = G_{dir}^2 + G_{dir}^3 + \dots \quad (3.2)$$

i.e. that the observed *information* exchanged on an edge e_{ij} between nodes X_i and X_j is the sum of the second, third etc. order effects, each given by the information on the n -path (where n is the order of the (diminishing) indirect effect) again assumed to be a sum of products. In other terms, the second order indirect effect between X_i and X_j (given as the (i, j) element of G_{dir}^2) is the sum of products on edges e_{ik} and e_{kj} for all k

$$[G_{dir}^2]_{ij} = \sum_{k=1}^N e_{ik} e_{kj}$$

where e_{ij} is the (i, j) element of G_{dir} . This is of course not true in general, but through the error analysis in Subsection 3.3.3 and the examples in chapter 4 this is not necessarily a problem. Immediately, we observe that e_{ii} is of interest in terms of its physical meaning. The co-dependence between a random variable and itself might be somewhat ambiguous or even undefined depending on the measure. Thus, the notion of (non-existing) edges e_{ii} will be of interest later on when using the method on controlled cases. We note that in G_{obs} we shall in general set these elements to 0.

Thus, from the above assumptions, it follows that we can express G_{obs} as

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir}^3 + \dots = G_{dir} + G_{dir} G_{obs} \quad (3.3)$$

Clearly, such a G_{dir} must have spectral radius at most 1 as otherwise, the above sum diverges and thus G_{obs} will not exist. I.e. $\rho(G_{dir}) < 1$. Thus, assuming convergence we can rewrite the infinite series as

$$G_{obs} = G_{dir} (I - G_{dir})^{-1} \quad (3.4)$$

It immediately follows that G_{dir} is given by (can be proved by directly inserting the above expression for G_{obs})

$$G_{dir} = G_{obs} (I + G_{obs})^{-1} \quad (3.5)$$

Furthermore, if the measure of dependence between pairs of variables is symmetric, then so is G_{obs} and hence diagonalizable by some orthogonal matrix U and diagonal matrix Λ_{obs} such that $G_{obs} = U \Lambda_{obs} U^T$ (with the columns of U being right eigenvectors of G_{obs}). It follows that G_{dir} can be expressed in a simple (and later computationally efficient) way

$$G_{dir} = U \Lambda_{dir} U^T$$

where $\Lambda_{dir} = \Lambda_{obs} (I + \Lambda_{obs})^{-1}$ which is also a diagonal matrix.

We note that from the above one needs $(I + G_{obs})^{-1}$ to be well-defined which is equivalent to $-1 \notin \sigma_{G_{obs}}$ i.e. -1 is not an eigenvalue of G_{obs} . To see that this is indeed the case whenever $\rho(G_{dir}) < 1$, and that $I + G_{obs}$ is thus invertible we use Equation 3.4 and simplify

$$\begin{aligned} I + G_{obs} &= I + G_{dir} (I - G_{dir})^{-1} \\ &= (I - G_{dir}) (I - G_{dir})^{-1} + G_{dir} (I - G_{dir})^{-1} \\ &= (I - G_{dir})^{-1} \end{aligned}$$

which is clearly invertible. Furthermore, we note that under the assumption $\rho(G_{dir}) < 1$ we can not place any bound on the spectral radius of G_{obs} . Namely, if v is a eigenvector of G_{dir} with eigenvalue λ such that $|\lambda| < 1$, then v is also an eigenvector of G_{obs} as

$$G_{obs}v = \sum_{k=1}^{\infty} G_{dir}^k v = \sum_{k=1}^{\infty} \lambda^k v = \frac{\lambda}{1-\lambda} v$$

i.e. $\left(\frac{\lambda}{1-\lambda}, v\right)$ is an eigenpair of G_{obs} and since $\frac{\lambda}{1-\lambda} \in (-1/2, \infty)$ for $\lambda \in (-1, 1)$ we can in general not bound the spectral radius of G_{obs} , although we should never observe an eigenvalue equal to or below $-1/2$ (which again proves that -1 is not an eigenvalue of G_{obs}).

As we shall later use some assumptions regarding causality i.e. directing the edges in the graph, we shall investigate the effect of letting G_{obs} be a triangular matrix which corresponds to a directed acyclic graph. Namely, in the following, we show that given the existence of G_{dir} (with necessary and sufficient conditions on G_{obs} as given above), G_{obs} is triangular if and only if G_{dir} is triangular. Thus, by directing the observed similarity (by removing half the edge weights in G_{obs}), we also infer a directed graph G_{dir} .

Clearly, if G_{dir} is triangular, so are the powers G_{dir}^i for all $i \in \mathbb{N}$ and hence if the infinite sum $\sum_{i=1}^{\infty} G_{dir}^i$ converges, G_{obs} is triangular as well. To show the other way, assume that G_{obs} is triangular and is the result of a G_{dir} with spectral radius smaller than 1. By Equation 3.5, G_{dir} is triangular if the inverse of $I + G_{obs}$ is triangular (upper triangular if G_{obs} is also upper triangular and similarly for lower triangular). This is indeed the case as in general, the inverse of a triangular matrix is also triangular provided that the diagonal elements are non-zero which is true as $I + G_{obs}$ has only ones in the diagonal as we will later assume in Equation 3.8. A simple proof of this is as follows. Without loss of generality, we assume that a matrix T is upper triangular. Let D be the diagonal elements of T and T_u be the remaining strictly upper triangular

part of T such that $T = D + T_u$. Then, assuming that D has non-zero diagonal elements, $T = D(I + D^{-1}T_u)$ and hence, we have that

$$\begin{aligned} T^{-1} &= (I + D^{-1}T_u)^{-1} D^{-1} \\ &= \sum_{i=0}^{\infty} (-D^{-1}T_u)^i D^{-1} \end{aligned}$$

which is clearly also upper triangular. Thus, we conclude that G_{obs} is triangular if and only if G_{dir} is (under the assumption G_{dir} exists).

Finally, before discussing the implementation and analyzing the algorithm both analytically and through examples, we will take a closer look at the similarity measures that are to be used with this method and that in the end will make up the matrix G_{obs} . Namely, *mutual information* and *correlation*.

3.2 Information measures and computation

In this section we discuss two measures that can be used to construct the matrices of codependency from the previous section. Namely, we shall touch on correlation and discuss what one might choose to call Copula-based entropy. However, before discussing Copula entropy (CE) we first need to define what a copula is.

3.2.1 Copula

Given a set of N random variables X_1, \dots, X_d , a copula is loosely speaking a distribution function with support $[0, 1]^d$ incorporating the dependence structure between the random variables. Given a joint distribution function F and (invertible) marginals F_1, \dots, F_N we define a copula C as

$$\begin{aligned} F(x_1, \dots, x_N) &= \mathbb{P}(X_1 \leq x_1, \dots, X_N \leq x_N) \\ &= \mathbb{P}(F_1(X_1) \leq F_1(x_1), \dots, F_d(X_d) \leq F_d(x_d)) \\ &= C(F_1(x_1), \dots, F_N(x_N)) \end{aligned}$$

Letting $u_i = F_i(x_i) \in [0, 1]$ it is clear that C is a distribution function as described above [3]. Furthermore, it follows that the marginals of C are uniform as $F_i(X_i)$ is uniformly distributed. We thus define a copula in probabilistic terms as

Definition 3.1 (Copula). A function $C : [0, 1]^d \rightarrow [0, 1]$ is called a copula if it has uniform marginals and is a distribution function for a d -dimensional random vector \mathbf{X} .

An important and fundamental theorem of copulas for especially continuous random variables where the marginals are also continuous functions is stated by Sklar:

Theorem 3.2 (Sklar's theorem). For a random vector \mathbf{X} with CDF F and univariate marginal CDFs F_1, \dots, F_d . There exists a copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \quad (3.6)$$

If X is continuous, C is unique; otherwise C is uniquely determined on the Cartesian product of the ranges of distribution functions F_i , $\prod \text{Ran}(F_i)$.

Note that the last statement for non-continuous random variables can be made unique by instead using subcopulas, a generalization of copulas with domain I only a subdomain of the unit hypercube $\mathbb{I}^d = [0, 1]^d$ containing all faces of the unit hyper cube. However, there are infinitely many ways of extending such a subcopula to a copula C [4]. In our case, this means that for discrete and/or mixed variables, we will later have to work around this non-uniqueness when calculating mutual information. The example made by Geenens[4] is a bivariate random vector of independent variables $X \sim \text{Bern}(\pi_X)$ and $Y \sim \text{Bern}(\pi_Y)$. The support of F_X and F_Y is then $\{0, 1 - \pi_X\}$ and $\{0, 1 - \pi_Y\}$ respectively. Due to the restriction on the boundary of the unit square, the only unique point of a copula C is then $(1 - \pi_X, 1 - \pi_Y)$, and by independence we must have

$$C(1 - \pi_X, 1 - \pi_Y) = (1 - \pi_X)(1 - \pi_Y)$$

Geenens then proceed to define an uncountable set of copulas that fulfill the above criterion which further illustrates that the basic concepts of copulas are not well suited for discrete random vectors. Note that in the article it is however argued how one can extend the concept to a more general concept that works for mixed variables.

From Equation 3.6 we see that a copula is thus simply just a function that couples the marginals of a random vector to the joint distribution. The following corollary follows immediately

Corollary 3.2.1 (Coordinate transformation). Under the assumptions of Theorem 3.2, given any set (T_1, \dots, T_d) of strictly increasing functions, if C is a copula of (X_1, \dots, X_d) then it is also a copula of $(T_1(X_1), \dots, T_d(X_d))$.

Proof. Suppose (X_1, \dots, X_d) permits a copula C and let T_i be given as stated. Consider coordinate wise the result of the transformation $Y_i = T_i(X_i)$ and consider the CDF $F_{Y_i}(y_i)$

$$\begin{aligned} F_{Y_i}(y_i) &= \mathbb{P}(Y_i \leq y_i) \\ &= \mathbb{P}(T_i^{-1}(Y_i) \leq T_i^{-1}(y_i)) \\ &= \mathbb{P}(X_i \leq T_i^{-1}(y_i)) \\ &= F_{X_i}(T_i^{-1}(y_i)) \end{aligned}$$

The above is easily generalized for a joint distribution as well. Thus, by the existence of a Copula C for \mathbf{X}

$$\begin{aligned} F_{\mathbf{Y}}(y_1, \dots, y_d) &= F_{\mathbf{X}}(T_1^{-1}(y_1), \dots, T_d^{-1}(y_d)) \\ &= C(F_{X_1}(T_1^{-1}(y_1)), \dots, F_{X_d}(T_d^{-1}(y_d))) \\ &= C(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d)) \end{aligned}$$

where Sklar's theorem have been used for the second equality. The above shows that C is indeed also a Copula for $\mathbf{Y} = (T_1(X_1), \dots, T_d(X_d))$. \square

The above corollary is actually equivalent with a seemingly stronger statement and follows easily

Proposition 3.3. *Since T_i is strictly increasing, the inverse T_i^{-1} exists and is also strictly increasing. Thus, the above implication is bidirectional and hence for strictly increasing functions T_i , C is a copula of (X_1, \dots, X_d) if and only if it is a copula of $(T_1(X_1), \dots, T_d(X_d))$.*

3.2.2 Mutual information and Copula entropy

In this section we introduce Copula entropy as done in [5] and see how it actually is equal to the well known mutual information (multiplied by -1) and hence as a corollary that mutual information is independent of marginals. The name comes from the general definition of (differential) entropy as we shall see shortly. However, first we define mutual information between a set of random variables

Definition 3.4 (Mutual information). *For a random vector $\mathbf{X} = \{X_i\}$, we define the mutual information as*

$$I(\mathbf{X}) = \mathbb{E} \left[\log_b \left(\frac{f(\mathbf{X})}{\prod_i f_i(X_i)} \right) \right]$$

where f is the joint density function with marginals f_i of the random vector \mathbf{X} . The base of the logarithm b is often chosen to be 2, e or 10 although the choice is unimportant as all logarithms are equivalent up to a scaling factor.

We note that later on, as the choice of b will result in a scaling of G_{obs} , but we will also introduce a scaling parameter α for G_{obs} to both ensure the convergence of the algorithm and to control higher order effects, we shall in general choose $b = e$.

An important property of mutual information is that the continuous version is the limit of the discrete mutual information for random (continuous) vector discretized as the mesh size goes to zero i.e. recovering the continuity of the random vector. This is discussed in Subsection 3.2.3. For now, we proceed with the definition of (joint) entropy for both discrete and continuous random vectors.

Definition 3.5 (Entropy). *The (joint) entropy of a random vector \mathbf{X} is defined as*

$$H(\mathbf{X}) = -\mathbb{E}[\log_b f(\mathbf{X})]$$

In case of a discrete random vector, this is called the Shannon entropy while for continuous random vectors, this is called differential entropy and is often denoted as $h(\mathbf{X})$ instead of $H(\mathbf{X})$.

We note the need for two separate notations of entropy as differential entropy is not the limit of Shannon entropy in the way mutual information is. Again, this is further discussed in Subsection 3.2.3.

Before discussing Copula entropy (CE), we note a very useful relation between entropy and mutual information. Indeed, we shall later use this to show that mutual information in the continuous version is the limit of the discretization.

Lemma 3.6 (Mutual information and entropy relation). *For a continuous random vector \mathbf{X} , the (joint) mutual information $I(\mathbf{X})$ can be decomposed into a sum of differential entropies as*

$$I(\mathbf{X}) = \sum_{i=1}^d h(X_i) - h(\mathbf{X})$$

where d is the dimension of \mathbf{X} . The same is true for discrete variables but with entropy H instead of differential entropy h .

Proof. This follows immediately from the definition of mutual information and entropy:

$$\mathbb{E} \left[\log_b \frac{f(\mathbf{X})}{\prod_i f_i(X_i)} \right] = \mathbb{E} [\log_b f(\mathbf{X})] - \sum_i \mathbb{E} [\log_b f_i(X_i)]$$

□

With the definitions of mutual information and entropy we are finally ready to introduce Copula entropy.

Definition 3.7 (Copula entropy). *For a continuous random vector \mathbf{X} with a uniquely defined Copula C , and Copula density c , we define the Copula entropy CE of \mathbf{X} as*

$$CE(\mathbf{X}) = h(\mathbf{U})$$

where \mathbf{U} has density c . In particular,

$$CE(\mathbf{X}) = -\mathbb{E} [\log_b c]$$

As stated above, Copula entropy is actually equal to the negative mutual information which we state as a theorem

Theorem 3.8 (Equality of Copula entropy). *For a continuous random vector \mathbf{X} , the Copula entropy CE is equal to the negative joint mutual information of \mathbf{X}*

$$CE(\mathbf{X}) = -I(\mathbf{X})$$

Proof. By Theorem 3.2, letting $x_i = F_i^{-1}(u_i)$, we can relate the copula density to the joint density of \mathbf{X} and its marginals

$$\begin{aligned} c(u_1, \dots, u_n) &= \frac{\partial}{\partial \mathbf{u}} C(u_1, \dots, u_n) \\ &= \frac{\partial}{\partial \mathbf{u}} F(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)) \\ &= f(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)) \frac{1}{f_1(F_1^{-1}(u_1)) \dots f_n(F_n^{-1}(u_n))} \end{aligned}$$

It follows directly that

$$\begin{aligned}
-CE(\mathbf{X}) &= \int_{[0,1]^d} c(\mathbf{u}) \log c(\mathbf{u}) \, d\mathbf{u} \\
&= \int_{\mathcal{X}} \frac{f(\mathbf{x})}{\prod_{i=1}^d f_i(x_i)} \log \left(\frac{f(\mathbf{x})}{\prod_{i=1}^d f_i(x_i)} \right) \prod_{i=1}^d f_i(x_i) \, d\mathbf{x} \\
&= \int_{\mathcal{X}} f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{\prod_{i=1}^d f_i(x_i)} \right) \, d\mathbf{x} \\
&= I(\mathbf{X})
\end{aligned}$$

where the third equality follows from a change of variables with the trivial substitution $u_i = F_i(x_i)$ such that $du_i = f_i(x_i) dx_i$. This concludes the proof.

□

Finally, before moving on to correlation as a measure of similarity, we discuss what happens in the limit of mutual information and entropy as we shall later need this as arguments for numerical stability.

3.2.3 Entropy and mutual information in the limit

In this section, we shall discuss the differences between entropy and differential entropy and observe how this difference cancels when computing mutual information. In fact, we shall see that mutual information defined for continuous random vectors is the limit of the discrete version which will be useful later when implementing the algorithm.

First, although one may think differential entropy this as the limit of (discrete) entropy, this is not the case. Namely, consider the support of $f(x)$ (here assumed to be the entire real line) binned into intervals i.e. a discretization of the continuous random variable X , which we shall denote X^Δ . To make notation simpler, we shall bin into equal-sized intervals of width Δ . Then, for each interval $[i\Delta, (i+1)\Delta]$ for $i \in \mathbb{Z}$, there exists an x_i such that the probability mass on this interval is represented by this x_i :

$$\mathbb{P}(X^\Delta = x_i) = f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx \quad (3.7)$$

Clearly, this discretization is a valid distribution as

$$\sum_{i \in \mathbb{Z}} f(x_i)\Delta = \int_{\mathbb{R}} f(x) \, dx = 1$$

and in the limit, as $\Delta \rightarrow 0$ we recover the original distribution $f(x)$. However, if we try to calculate the entropy of this discretization, denoted by H^Δ , we get a diverging limit

$$\begin{aligned} H^\Delta &= \sum_{i \in \mathbb{Z}} f(x_i) \Delta \log f(x_i) \Delta \\ &= \sum_{i \in \mathbb{Z}} f(x_i) \Delta \log f(x_i) + \sum_{i \in \mathbb{Z}} f(x_i) \Delta \log \Delta \\ &= \sum_{i \in \mathbb{Z}} f(x_i) \Delta \log f(x_i) + \log \Delta \end{aligned}$$

Clearly, the first term in the above expression converges to the differential entropy $h(X)$ as $\Delta \rightarrow 0$ whereas $\log \Delta \rightarrow -\infty$ i.e. the expression diverges altogether when differential entropy is well-defined.

A similar argument for the joint entropy between the discretization of X_1 and X_2 (and in principle to any number of dimensions), denoted by H_{12}^Δ , results in

$$H_{12}^\Delta = \sum_{i,j \in \mathbb{Z}} f(x_1^{(i)}, x_2^{(j)}) \Delta_1 \Delta_2 \log f(x_1^{(i)}, x_2^{(j)}) + \log \Delta_1 + \log \Delta_2$$

where $x_1^{(i)} \in [i\Delta_1, (i+1)\Delta_1]$ and $x_2^{(j)} \in [j\Delta_2, (j+1)\Delta_2]$ are defined such that

$$f(x_1^{(i)}, x_2^{(j)}) \Delta_1 \Delta_2 = \int_{j\Delta_2}^{(j+1)\Delta_2} \int_{i\Delta_1}^{(i+1)\Delta_1} f(x_1, x_2) dx_1 dx_2, \quad \forall i, j \in \mathbb{Z}$$

Note that clearly $(x_1^{(i)}, x_2^{(j)})$ exists for all $i, j \in \mathbb{Z}$. Again, the joint entropy diverges however, when computing the mutual information, we see that the diverging terms cancel. Namely, from Lemma 3.6

$$\begin{aligned} I_{12}^\Delta &= H_1^\Delta + H_2^\Delta - H_{12}^\Delta \\ &= \sum_{i \in \mathbb{Z}} f_1(\tilde{x}_1^{(i)}) \Delta_1 \log f_1(\tilde{x}_1^{(i)}) + \log \Delta_1 \\ &\quad + \sum_{j \in \mathbb{Z}} f_2(\tilde{x}_2^{(j)}) \Delta_2 \log f_2(\tilde{x}_2^{(j)}) + \log \Delta_2 \\ &\quad - \sum_{i,j \in \mathbb{Z}} f(x_1^{(i)}, x_2^{(j)}) \Delta_1 \Delta_2 \log f(x_1^{(i)}, x_2^{(j)}) - \log \Delta_1 \Delta_2 \\ &= \sum_{i \in \mathbb{Z}} f_1(\tilde{x}_1^{(i)}) \log f_1(\tilde{x}_1^{(i)}) \Delta_1 + \sum_{j \in \mathbb{Z}} f_2(\tilde{x}_2^{(j)}) \log f_2(\tilde{x}_2^{(j)}) \Delta_2 \\ &\quad - \sum_{i,j \in \mathbb{Z}} f(x_1^{(i)}, x_2^{(j)}) \log f(x_1^{(i)}, x_2^{(j)}) \Delta_1 \Delta_2 \\ &\rightarrow h(X_1) + h(X_2) - h(X_1, X_2) \text{ as } \Delta_1, \Delta_2 \rightarrow 0 \end{aligned}$$

Thus, the limit of the mutual information for discrete random variables is indeed the mutual information defined for continuous random variables and can be computed either as the limit of discretizing the probability density function and then computing entropies or just using the initial definition for (discrete) mutual information in Definition 3.4.

Before continuing, we discuss the case where X_1 is equal to X_2 . In this case, discretizing with a common Δ we have that

$$f(x_1^{(i)}, x_2^{(j)}) \Delta^2 = \int_{j\Delta}^{(j+1)\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x_1, x_2) dx_1 dx_2, \quad \forall i, j \in \mathbb{Z}$$

Clearly, the above integral is 0 for $i \neq j$. Although $f(x_1, x_2)$ is not well-defined in the usual functional sense, extending to distribution, we might write $f(x_1, x_2) = f(x_2|x_1)f(x_1)$. In terms of distributions, it works to put $f(x_2|x_1) = \delta(x_2 - x_1)$ where δ is the *Dirac delta* distribution, as then $\int_{\mathbb{R}} f(x_1, x_2) dx_2 = f(x_1)$ and $f(x_1, x_2)$ is "0" when $x_1 \neq x_2$. I.e. the right marginals and probability mass 1. Then, when calculating the above integral, we get that

$$\begin{aligned} f(x_1^{(i)}, x_1^{(i)}) \Delta^2 &= \int_{i\Delta}^{(i+1)\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x_1, x_2) dx_1 dx_2 \\ &= \int_{i\Delta}^{(i+1)\Delta} f(x_1) dx_1 \\ &= f(\tilde{x}_1^{(i)}) \Delta \end{aligned}$$

Thus, when calculating $I_{1,2}^\Delta$ we obtain

$$\begin{aligned} I_{1,2}^\Delta &= \sum_{i \in \mathbb{Z}} f_1(\tilde{x}_1^{(i)}) \log f_1(\tilde{x}_1^{(i)}) \Delta + \sum_{j \in \mathbb{Z}} f_2(\tilde{x}_2^{(j)}) \log f_2(\tilde{x}_2^{(j)}) \Delta \\ &\quad - \sum_{i \in \mathbb{Z}} f_1(\tilde{x}_1^{(i)}) \log f_1(\tilde{x}_1^{(i)}) \Delta - \log \Delta \\ &\rightarrow \infty \text{ as } \Delta \rightarrow 0 \end{aligned}$$

Thus in practice, it would not make much sense to compare equal variables or even a random vector only defined on a lower dimensional manifold as we would get an infinite Copula entropy.

3.2.4 Correlation

At this point, we have a good understanding of Copula entropy/mutual information for calculations later on. However, another typical measure of similarity is

correlation which is easily estimated from sample data. However, in this section we show that in general, we can not compute the correlation coefficient from a Copula which we saw above is the case for mutual information. Namely, given a copula C for some set of random variables $\{X_i\}_{i \in I}$ indexed by finite I , one can not calculate ρ between any pair (X_i, X_j) , $i \neq j$ from the copula. This is easily shown by the following argument.

First, note that from Corollary 3.2.1, C is also a copula for $Z_i := (X_i - \mu_i) / \sigma_i$ for $i \in I$ where $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i = \sqrt{\text{Var } X_i}$. Clearly, the correlation coefficient for Z_i and Z_j is the same as between X_i and X_j . We thus proceed trying to calculate the correlation between any pair Z_i and Z_j .

$$\begin{aligned}\rho_{ij} &= \int \int_{\mathbb{R}^2} z_i z_j f_{ij}(z_i, z_j) dz_i dz_j \\ &= \int \int_{[0,1]^2} F_i^{-1}(u_i) F_j^{-1}(u_j) c_{ij}(u_i, u_j) du_i du_j\end{aligned}$$

where c_{ij} density version of the copula defined for X_i and X_j and F_i and F_j are the marginals of Z_i and Z_j with mean 0 and variance 1. From the above, it is then clear for a fixed, non-constant copula C , the correlation depends on the marginals of X_i and X_j . Also, we see that a constant copula density (only admissible if $c \equiv 1$ on $[0,1]^2$ and 0 elsewhere) always results in $\rho_{ij} = 0$ as

$$\int_0^1 F^{-1}(u) du = \int_{\mathbb{R}} z f(z) dz = 0$$

again, under the assumption that Z_i has mean 0.

Thus, we conclude that indeed mutual information and correlation is very different measures of codependency (as correlation depends on the marginals whereas mutual information does not) and that it does not make much sense to introduce copulas in the setting of correlation albeit at this point we do not favor one measure above the other except if marginals should be insignificant to the network, Copula entropy is preferred.

3.3 Copula based network discovery

In this section, we will present the general algorithm and discuss some of its properties regarding uncertainty and convergence. We will focus on using mutual information i.e. Copula entropy as the measure of similarity but other measures such as correlation can be interchanged at will in the general algorithm.

By Theorem 3.8 we can compute the mutual information from observed data from the copula. Namely, let CE_{ij} denote the (pairwise) Copula entropy of variables X_i and X_j . We shall then set

$$G_{obs} = \begin{bmatrix} 0 & -CE_{12} & \dots & -CE_{1n} \\ -CE_{21} & 0 & \dots & -CE_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -CE_{n1} & -CE_{n2} & \dots & 0 \end{bmatrix} \quad (3.8)$$

where n is the number of nodes in the graph i.e. random variables that we have observed. Notice that we have chosen the diagonal elements as 0 since information between a random variable X and itself is not really well-defined and when trying to compute this numerically, we observe diverging results as also discussed in the previous section. Furthermore, only the information that propagates through the network is of interest and so setting 0 in the diagonal avoids a bias when (de)convolving the information or any similarity in general. Especially for mutual information where the information between a variable and itself diverges to ∞ thus in the limit, from Equation 3.5, we would get the identity matrix which does not tell us much about the direct dependencies.

Algorithm 1 then follows immediately from Equation 3.8

Algorithm 1 G_{obs} computation

```

Require:  $n > 0$  ▷ Number of variables
 $G_{obs} \leftarrow \mathbf{0}$ 
for  $1 \leq i, j \leq n \mid i \neq j$  do
    Estimate  $F_i$  and  $F_j$  from  $x_i^{\mathcal{D}}$  and  $x_j^{\mathcal{D}}$ 
     $u_i^{\mathcal{D}} \leftarrow F_i(x_i^{\mathcal{D}})$ 
     $u_j^{\mathcal{D}} \leftarrow F_j(x_j^{\mathcal{D}})$ 
    Estimate  $c_{ij}$  from  $u_i^{\mathcal{D}}$  and  $u_j^{\mathcal{D}}$ 
    Compute  $NCE_{ij}$ 
     $[G_{obs}]_{ij} \leftarrow -NCE_{ij}$ 
end for
return  $G_{obs}$ 

```

Namely, for each entry in G_{obs} , except for the diagonal elements, first estimate the cumulative distributions of X_i and X_j based on samples $x_i^{\mathcal{D}}$. Then, transform the samples by the estimated distribution function to obtain corresponding uniform samples. This may be done outside the loop to increase computational efficiency. From the paired samples $(x_i^{\mathcal{D}}, x_j^{\mathcal{D}})$, estimate the Copula density c_{ij} and finally use this to compute the mutual information/Copula entropy. Methods for estimating the densities and in continuation hereof the

distribution functions are presented in ???. The negative Copula entropy is then recorded in (i, j) entry of G_{obs} . We note that the algorithm can be optimized for symmetric measures such as Copula entropy itself, to only loop through $i < j$ and saving the computed entropy in the (j, i) entry as well. Also, as Copula entropy diverges as X_i and X_j are jointly distributed closer to a one-dimensional manifold, ideally there should be a check for such or the user should check the paired observations to exclude such variable combinations.

From Subsection 3.2.3, to calculate the (joint) copula entropy of a continuous random vector, we simply discretize the domain of each random variable and use the estimated copula density evaluated at these points to estimate the total Copula entropy. Furthermore, if one or more elements of the random vector are mixed random variables, we choose the discrete events to be their own bins and discretize the rest or in the context of Algorithm 1 only estimate the distribution functions for the continuous component of the random variable. This works due to the Copula entropy for continuous random variables being the limit of the discretization and as such, the Copula entropy is well-defined for mixed random variables as well.

We continue with an example of how this discretization of a mixed random variable would work. Notice that we only have a discrete event (an atom) at 0 as this resembles the observed behavior of the delays, although the example could be extended to more complex discrete distributions.

Example 3.1 (Discretization of mixed random variable). *Let X be a mixture of an atom in e.g. 0 and an exponential with parameter λ with proportions p and $1-p$. Then, a discretization of X is 0 with probability mass p and the remaining support $(0, \infty)$ discretized in some way with total probability mass $1-p$ and each bin having probability according to Equation 3.7 scaled with $1-p$. If the bin size is a constant Δ , then for the discretized variable X^Δ , we have $\mathbb{P}(X^\Delta = 0) = p$ and $\mathbb{P}(X^\Delta = x_i) = (1-p) \exp(-\lambda i \Delta) (1 - \exp(-\lambda \Delta))$, where x_i is given by*

$$x_i = i\Delta + \frac{1}{\lambda} (\log(\lambda\Delta) - \log(1 - e^{-\lambda\Delta})) , \quad i \in \mathbb{N}_0$$

3.3.1 Network deconvolution

At this point, we have obtained a convolved matrix of information G_{obs} and are ready to use Equation 3.5. We present the original algorithm from [2] in the case G_{obs} is symmetric and hence diagonalizable by an orthogonal matrix U . The original **Matlab** implementation was translated to **Python** and is summarized in the following pseudocode.

Algorithm 2 (ND) Network Deconvolution**Require:** G_{obs}, α, β

$[G_{obs}]_{ii} \leftarrow 0, \forall i \in \{1, \dots, N\}$	\triangleright ensure zero-diagonal
$[G_{obs}]_{ij} \leftarrow 0, \text{ when } [G_{obs}]_{ij} < Q_\alpha(G_{obs})$	
Compute eigendecomposition U, Λ of G_{obs}	
$\lambda^+ \leftarrow \max(\lambda^{\max}, 0)$	
$\lambda^- \leftarrow \min(\lambda^{\min}, 0)$	
$c_s^+ \leftarrow \frac{1-\beta}{\beta} \lambda^+$	
$c_s^- \leftarrow \frac{1+\beta}{\beta} \lambda^-$	
$c_s \leftarrow \max(c_s^+, -c_s^-)$	
$\hat{\Lambda} \leftarrow \Lambda (c_s^{-1} I + \Lambda)^{-1}$	
return $U \hat{\Lambda} U^T$	

where $Q_\alpha(G_{obs})$ denotes the α quantile of the strictly upper (or lower due to symmetry) triangular part of G_{obs} . We note the two extra parameters *alpha* and β which we will discuss shortly. In particular, the paper contains conflicting information on how to find β from how it is defined. Furthermore, they include some analysis on the robustness of the above deconvolution algorithm but only in a somewhat particular case and with some confusion on matrix norms and spectral radius. This analysis on robustness, we will extend and clarify in the following Subsection 3.3.3.

From the definition of $Q_\alpha(G_{obs})$ it is clear that the α parameter is a filter on the observed edges and is useful if one wants to filter out insignificant observations. However, in practice, as we will see, it is often not very influential except for large α (corresponding to many edges set to 0) as small perturbations from e.g. imperfect calculations should not influence the results for fairly conditioned matrices as we shall observe in Section 4.3. Thus, setting $\alpha = 0$ retains all values in G_{obs} after setting the diagonal equal to 0. As a technical detail, we note that the `quantile` function from NumPy (v. 1.26.4) has been used to find this quantile as quantiles can be defined in many ways from a data set.

Finally, we note that the $\beta \in (0, 1)$ parameter corresponds to a scaling of G_{obs} such that the resulting spectral norm of G_{dir} is β . From Algorithm 2 it is seen that it serves as a regularization on the eigenvalues of G_{obs} and although this is discussed in [2], their results do not conform with their implementation, and we thus comment on this and what else could be done to ensure convergence of the algorithm in the following section. Also, in practice we choose a threshold t on the elements of G_{dir} returned from Algorithm 2 to further filter out insignificant direct dependencies.

3.3.2 Ensuring convergence and the effect of β

In this section we will further discuss the effect of β and how the steps for rescaling the observed similarity matrix G_{obs} are derived. In particular, we will reformulate the original derivation from [2] as there is a discrepancy between their code¹ and their proof of choosing a scaling parameter c_s of G_{obs} . Namely, denote \tilde{G}_{obs} as the rescaled G_{obs} such that $\tilde{G}_{obs} = c_s G_{obs}$. Choosing c_s as in Algorithm 2 i.e. $c_s = \max\left(\frac{1-\beta}{\beta}\lambda^+, -\frac{1+\beta}{\beta}\lambda^-\right)$ where λ^+ is the largest positive eigenvalue of G_{obs} (and 0 if no eigenvalue is positive) and λ^- is the most negative eigenvalue of G_{obs} (and 0 if no eigenvalue is negative) then implies \tilde{G}_{dir} obtained from the new \tilde{G}_{obs} has spectral radius $\beta < 1$ i.e. a proper G_{dir} with the largest numerical eigenvalue equal to β . This holds in general and not only for symmetric G_{obs} as we will see in the following. However, when G_{obs} is symmetric the resulting \tilde{G}_{dir} can easily be expressed through the eigendecomposition of G_{obs} , U , Λ as

$$\begin{aligned}\tilde{G}_{dir} &= \tilde{G}_{obs} \left(I + \tilde{G}_{obs} \right)^{-1} \\ &= c_s G_{obs} \left(I + c_s G_{obs} \right)^{-1} \\ &= U c_s \Lambda U^T \left(U U^T + U c_s \Lambda U^T \right)^{-1} \\ &= U c_s \Lambda U^T U \left(I + c_s \Lambda \right)^{-1} U^T \\ &= U \Lambda \left(c_s^{-1} I + \Lambda \right)^{-1} U^T\end{aligned}$$

which can also be seen in Algorithm 2. Thus, with everything else explained about the algorithm, we show that the resulting \tilde{G}_{dir} in general have spectral radius β .

Let (λ, v) be an eigenpair of G_{obs} with $\lambda \neq 0$, it then follows that $\left(\frac{\lambda}{c_s^{-1} + \lambda}, v\right)$ is an eigenpair of \tilde{G}_{dir} . Then, following the arguments in [2] (which we have redone to know why the original implementation and derivation differs), we obtain that for a λ in $[0, \infty)$, $(-c_s^{-1}, 0)$ or $(-\infty, 0]$ to be mapped to $[-\beta, \beta]$ we have that

$$c_s^{-1} \geq \frac{1-\beta}{\beta} \lambda^+, \quad c_s^{-1} \geq -\frac{1+\beta}{\beta} \lambda^-$$

Thus, the smallest c_s^{-1} we can choose to ensure that $\rho(\tilde{G}_{dir}) \leq \beta$ is by $c_s = \min\left(\frac{1-\beta}{\beta}\lambda^+, -\frac{1+\beta}{\beta}\lambda^-\right)$ which also implies that $\rho(\tilde{G}_{dir}) = \beta$ as either the most negative or most positive eigenvalue is mapped to β or $-\beta$ respectively. This coincides with the original implementation, noting that some error has been

¹<https://compbio.mit.edu/nd/>

made in the original discussion of the parameter β in [2]. Furthermore, we note that if we just want the algorithm to converge, as we discussed before, this is equivalent to $\sigma(\tilde{G}_{obs}) \subseteq (-1/2, \infty)$, so really, we can just choose $c_s^{-1} = -(2 + \delta)\lambda^-$ for some small δ if $\lambda^- < -1/2$ and otherwise not scale G_{obs} to preserve the structure. Finally, we note that as β tends to 0, higher order interactions become less significant as can clearly be seen from Equation 3.4. Thus, β also allows us to tune how much influence higher order interactions should have and one should try different β to see how influenced results are to higher order effects.

3.3.3 Robustness to noise

Finally, before discussing how to compute and estimate the mutual information between two random variables based on observations, we turn our heads to error analysis of the deconvolution algorithm. It is important to understand how well the algorithm performs subject to noise and errors. Namely, in the case of mutual information, the assumption that higher order effects can be calculated as a sum of matrix powers of the direct effects does not hold. Thus, if we can quantify the error in G_{obs} , we can from the following analysis quantify the resulting error in G_{dir} . We shall first discuss the original result from [2], correcting some errors in terms of definitions and see how their result can also be expressed as an absolute upper bound on the error instead of only how this error behaves for small perturbations. Furthermore, we shall extend their result to not only hold when $\rho(G_{obs}) < 1$ and $\rho(G_{obs} + N) < 1$ where N is some noise e.g. from computation or assumptions that does not completely hold.

The original result states that $\|G_{dir} - \tilde{G}_{dir}\|_2 \leq \gamma + \mathcal{O}(\delta^2 + \gamma^2 + \delta\gamma)$ where $\|\cdot\|_2$ is the Euclidean norm also known as the spectral norm as this is equal to the largest singular value of the input matrix. However, they note that the Euclidean norm of a matrix M is equal to $\sqrt{\sum_{i,j} m_{ij}^2}$ which is incorrect. This is the Frobenius norm, and instead it should have been defined as

$$\|M\|_2 = \sup_{\|x\|_2=1} \|Mx\|_2 = \sigma_{\max}(M)$$

They then proceed to let γ be the largest absolute eigenvalue of N and δ the largest absolute eigenvalue of $\tilde{G}_{obs} = G_{obs} + N$ however as the noise may be both positive and negative, it is easier to define δ as the largest absolute eigenvalue of G_{obs} instead which we will do in the following. We note that γ and δ are not the spectral/Euclidian norm of N and G_{obs} respectively as in general, we only have $\rho(M) \leq \|M\|_2$. However, if G_{obs} and N are both (real) symmetric

matrices, then the spectral norms are equal to the largest absolute eigenvalues of G_{obs} and N respectively. Thus, if instead one wanted to measure the difference in the direct dependency matrices in terms of e.g. the Frobenius norm, it is important to differentiate between the spectral radius and the norm that is actually being used. Finally, before constructing the actual upper bound on the error instead of quantizing the asymptotic behavior for small γ , we note that $\|\cdot\|_2$ is a sub-multiplicative matrix norm defined as below ([6]), and that we shall assume that $\rho(G_{obs}), \rho(\tilde{G}_{obs}) < 1$.

Definition 3.9 (Sub-multiplicative Matrix norm). *A matrix norm $\|\cdot\|$ is said to be sub-multiplicative, if for every $A, B \in \mathbb{F}^{n \times n}$ where \mathbb{F} is either the real or complex field:*

$$\|AB\| \leq \|A\| \cdot \|B\|$$

As we do not use any property of the spectral norm except that it is sub-multiplicative, we shall consider any norm $\|\cdot\|$ in general that is also sub-multiplicative. Thus, consider the norm of the difference $G_{dir} - \tilde{G}_{dir}$:

$$\begin{aligned} \|G_{dir} - \tilde{G}_{dir}\| &= \left\| G_{obs} (I + G_{obs})^{-1} - \tilde{G}_{obs} (I + \tilde{G}_{obs})^{-1} \right\| \\ &= \left\| - \sum_{k \geq 1} (-G_{obs})^k + \sum_{k \geq 1} (-\tilde{G}_{obs})^k \right\| \\ &\leq \sum_{k \geq 1} \left\| G_{obs}^k - (G_{obs} + N)^k \right\| \\ &\leq \sum_{k \geq 1} \sum_{i=1}^k \binom{k}{i} \|N\|^i \|G_{obs}\|^{k-i} \tag{3.9} \\ &= \sum_{k \geq 1} \sum_{i=1}^k \binom{k}{i} \gamma^i \delta^{k-i} \\ &= \sum_{k \geq 1} ((\gamma + \delta)^k - \delta^k) \\ &= \frac{\gamma + \delta}{1 - \gamma - \delta} - \frac{\delta}{1 - \delta} \\ &= \frac{\gamma}{(1 - \gamma - \delta)(1 - \delta)} \end{aligned}$$

where in the second to last inequality, we assume that $\gamma + \delta < 1$ as then both $\sum (\gamma + \delta)^k$ and $\sum \delta^k$ converges as $\gamma + \delta \geq \delta \geq 0$ and hence also the difference of the sums converges. Also, the second equality uses that the spectral norm of G_{obs} and \tilde{G}_{obs} is less than 1 in order to express the inverses as infinite series.

Thus, the above bound on the difference $G_{dir} - \tilde{G}_{dir}$ does not hold in every case and we observe for fixed γ , the bound tends to ∞ as $\delta \rightarrow 1$. Furthermore, we note that the final infinite sum diverges whenever $\gamma + \delta > 1$ through the following argument using the ratio test for infinite sums which is needed because we can not conclude on the convergence of a difference of diverging sums solely from the fact that the individual sums diverge:

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{(\gamma + \delta)^{n+1} - \delta^{n+1}}{(\gamma + \delta)^n - \delta^n} \right| &= \lim_{n \rightarrow \infty} \left| \frac{(\gamma + \delta) \left(1 + \frac{\gamma}{\delta}\right)^n - \delta}{\left(1 + \frac{\gamma}{\delta}\right)^n - 1} \right| \\ &= \lim_{n \rightarrow \infty} \left| \delta + \gamma \frac{\left(1 + \frac{\gamma}{\delta}\right)^n}{\left(1 + \frac{\gamma}{\delta}\right)^n - 1} \right| \\ &= \lim_{n \rightarrow \infty} \left| \delta + \gamma \frac{1}{1 - \left(1 + \frac{\gamma}{\delta}\right)^{-n}} \right| \\ &= |\gamma + \delta| = \gamma + \delta \end{aligned}$$

assuming that $\gamma, \delta > 0$ corresponding to neither N nor G_{obs} is the zero matrix in which case the above analysis is nonsensical.

Before continuing with a more general bound on the error, we first note that examples of sub-multiplicative matrix norms are every induced norm such as the spectral norm and the Frobenius norm which is often useful when interpreting error. Also, the max norm is *not* sub-multiplicative, but a scaled version is (which is true for any matrix norm from the fact that all matrix norms are equivalent).

Now, consider the general case, where we do not restrict the spectral radius of either G_{obs} or N except such that G_{obs} and \tilde{G}_{obs} admits direct similarity matrices G_{dir} and \tilde{G}_{dir} (with spectral radius less than 1). To obtain a more general result, we shall use the following result from [7], which is very useful when doing matrix perturbation analysis.

Theorem 3.10 (Inverse of sum of matrices). *Let $A, B \in \mathbb{R}^{n \times n}$ such that A and $A + B$ are invertible. Then the inverse of $A + B$ can be expressed as*

$$(A + B)^{-1} = A^{-1} - A^{-1}B(A + B)^{-1}$$

The proof of the above is simple through direct computation. Hence, we continue

to once again consider the difference $G_{dir} - \tilde{G}_{dir}$

$$\begin{aligned} G_{dir} - \tilde{G}_{dir} &= G_{obs} (I + G_{obs})^{-1} - (G_{obs} + N) (I + G_{obs} + N)^{-1} \\ &= G_{obs} \left((I + G_{obs})^{-1} - (I + G_{obs} + N)^{-1} \right) - N (I + G_{obs} + N)^{-1} \\ &= G_{obs} (I + G_{obs})^{-1} N (I + G_{obs} + N)^{-1} - N (I + G_{obs} + N)^{-1} \\ &= - (I + G_{obs})^{-1} N (I + G_{obs} + N)^{-1} \end{aligned}$$

where the third equality follows from Theorem 3.10. This way, we have a simple exact expression for the difference without any further assumptions on G_{obs} and N . Now, under a sub-multiplicative norm $\|\cdot\|$ we can bound the norm of the difference in the following way.

$$\|G_{dir} - \tilde{G}_{dir}\| \leq \|N\| \left\| (I + G_{obs})^{-1} \right\| \left\| (I + G_{obs} + N)^{-1} \right\| \quad (3.10)$$

We note that if once again, we assume that the spectral radius of G_{obs} and $G_{obs} + N$ are smaller than 1, we rediscover Equation 3.9. Equation 3.10 also shows that in general, if N is small or G_{obs} is large we should observe small errors which is also what we would expect intuitively. The above result is also very useful when later on in Section 4.3 we discuss the error from using mutual information in the case of a multi-variate Gaussian.

From Equation 3.10, by another application of Theorem 3.10, we find the relative error in general to be bounded as follows

$$\frac{\|G_{dir} - \tilde{G}_{dir}\|}{\|G_{dir}\|} \leq \|N\| \left| 1 - \frac{\|I\|}{\|G_{obs} (I + G_{obs})^{-1}\|} \right| \left\| (I + G_{obs} + N)^{-1} \right\|$$

Finally, before discussing the methods for estimating the copula density, we comment on some frequently used matrix norms and show some explicit bounds on the error only using the difference of G_{obs} and \tilde{G}_{obs} , N . Namely, we shall consider the max norm and Frobenius norm of the difference $G_{dir} - \tilde{G}_{dir}$ and note that from [8], we can relate the Euclidean norm to the Frobenius and max norm in the following way. Namely, for any matrix $A \in \mathbb{R}^{n \times n}$ it holds that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2$$

$$\|A\|_{\max} \leq \|A\|_2 \leq n \|A\|_{\max}$$

Finally, if G_{obs} and N are symmetric, the singular values are equal to the absolute eigenvalues for G_{obs} and \tilde{G}_{obs} and because $\sigma(I + G_{obs}), \sigma(I + \tilde{G}_{obs}) \subseteq (1/2, \infty)$ implies $\sigma((I + G_{obs})^{-1}), \sigma((I + \tilde{G}_{obs})^{-1}) \subseteq (0, 2)$ we infer that

$\left\| (I + G_{obs})^{-1} \right\|_2, \left\| (I + \tilde{G}_{obs})^{-1} \right\|_2 \leq 2$. Using this with the above equivalence on the Euclidean norm with Equation 3.10, we conclude that

$$\begin{aligned} \left\| G_{dir} - \tilde{G}_{dir} \right\|_F &\leq 4\sqrt{d} \|N\|_F \\ \left\| G_{dir} - \tilde{G}_{dir} \right\|_{\max} &\leq 4d \|N\|_{\max} \end{aligned} \quad (3.11)$$

This clearly shows us that for small networks (thus small d) we risk smaller errors in terms of the Frobenius and max norm (which is not surprising) which are clearly interpreted through the difference of individual element of G_{dir} and \tilde{G}_{dir} and that the max norm scales linearly with the number of nodes while the Frobenius difference only scales with the square root of the number of nodes.

3.4 Estimation of mutual information

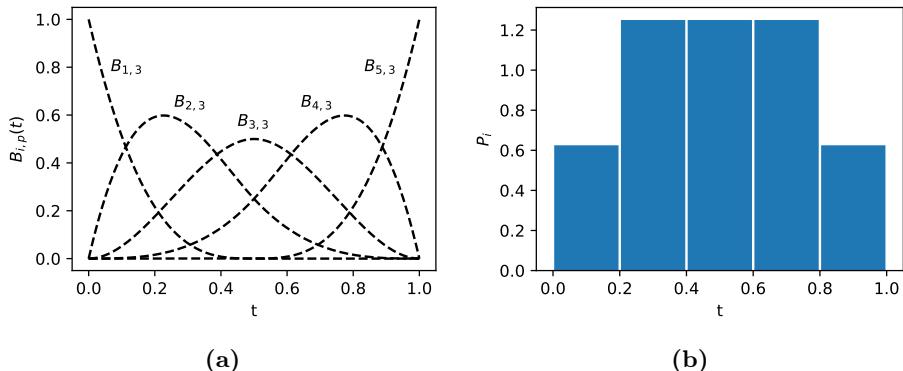
3.4.1 B-splines

B-spline - non-uniform marginals

[9] Cox-de Boor recursion formula

$$B_{i,0}(t) = 1, \quad t \in [t_i, T_{i+1}) \quad (3.12)$$

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i} B_{i,k-1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(t) \quad (3.13)$$



Have the useful property that $\sum_i B_{i,k}(t) = 1$ for all k and t . However, as we will be using this when trying to estimate mutual information from a copula, we

have a very specific problem, namely, the marginals are not discrete uniform. $i \in \{1, \dots, n\}$ basis functions

Let \mathbf{X} be a random vector of d , i.i.d uniformly distributed random variables X_j . Then the random variable corresponding to the probability mass assigned to bin i , P_i , is given by

$$P_i = \frac{1}{d} \sum_{j=1}^d B_{i,p}(X_j)$$

thus, the expected value of P_i is (independently of d)

$$\mathbb{E}[P_i] = \mathbb{E}[B_{i,p}(X_j)] = \int_0^1 B_{i,p}(x) dx$$

Ideally, this should be equal for all i as we want the marginal to be uniform as per definition of a Copula density. However, this is not the case. In fact,

In particular, we can calculate the expected probability mass for each bin i as

3.4.2 M-splines

M-spline - need to normalize distributions

[10] a set of n basis functions indexed by i . Letting the order of the basis function equal $p + 1$ such that the basis functions are polynomials of degree p . Define knots to be $t_1 = \dots = t_k = 0$ and $t_{n+1} = \dots = t_{n+k} = 1$. furthermore, we choose the remaining $n - k$ (interior) knots to be equally distributed on $[0, 1]$ i.e. if 3 interior points, they are $1/4$, $1/2$ and $3/4$ respectively. (slightly reformulated to agree with the notation for B splines)

$$M_{i,0}(t) = \frac{1}{t_{i+1} - t_i}, \quad t \in [t_i, T_{i+1}) \quad (3.14)$$

$$M_{i,k}(t) = \frac{k((t - t_i) M_{i,k-1}(t) + (t_{i+k} - t) M_{i+1,k-1}(t))}{(k-1)(t_{i+k} - t_i)} \quad (3.15)$$

such that $M_{i,p}$ is a degree p polynomial on $[t_i, t_{i+p+1}]$

it follows that each of these $M_{i,p+1}$ basis splines integrate to 1 which, as we saw before is important for uniform marginals. In particular, if we instead let the basis functions be $\tilde{M}_{i,p} = \frac{1}{n} M_{i,p}$, The marginal is uniform. However now there is the downside that basis functions do not sum to 1 as with the B-spline. Namely, points near the boundary are weighted more than points on the inside as is illustrated below

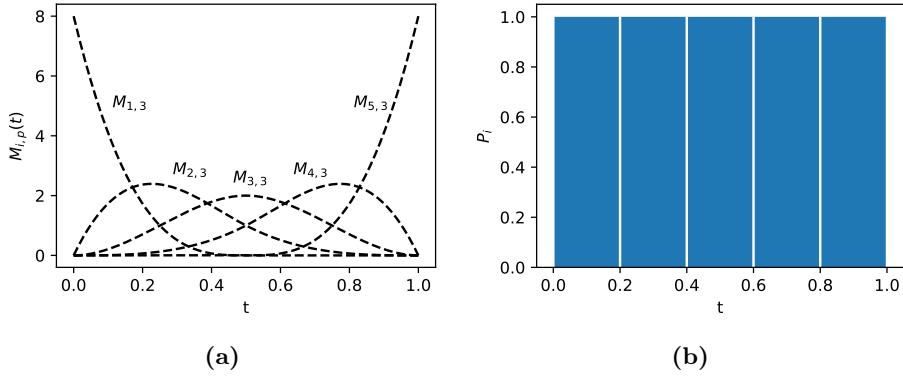


Figure 3.3: P_i is area of each rectangle i.e. 0.2.

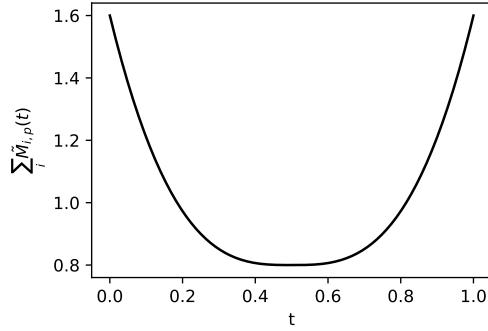


Figure 3.4

3.4.3 A better spline

In theory, we could make a set of splines ourselves, that has both of the desired properties of B-splines and M-splines. I.e. a set of piecewise polynomials, which we shall denote $Q_{i,p}$, that has both the property that

$$\sum_{i=1}^n Q_{i,p}(t) = 1$$

and

$$\int_0^1 Q_{i,p}(t) dt = \frac{1}{n}$$

Actually, both M- and B-splines satisfies this for $p = 0$ as they are then both piecewise constant and non-zero on disjoint intervals. From this, one can the

construct such sets of $Q_{i,p}$ in the following way. Namely, start with the piecewise constant functions, $Q_{i,0}$ depicted in Figure 3.5 as boxes in the case $n = 5$.

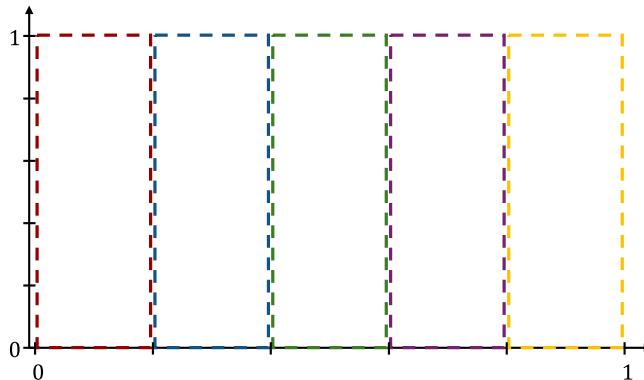


Figure 3.5

A better spline

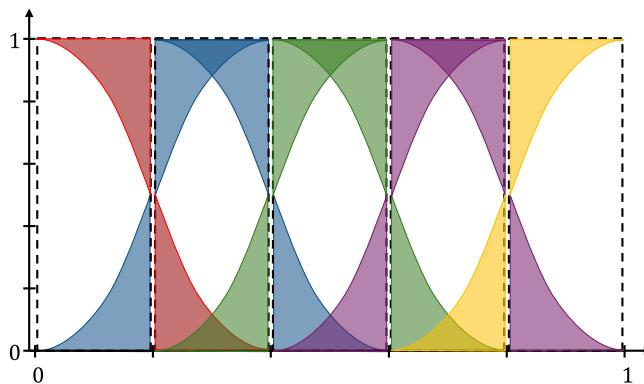


Figure 3.6

Cut off function has derivative 0 at top and is halfway down at the sides

This example does however not work well for many bins, in particular, due to the splines not overlapping that much i.e. a poor regularization on the smoothness. However, the method used for constructing this set of splines can be extended to bleed into multiple neighboring bins just as for B- and M-splines.

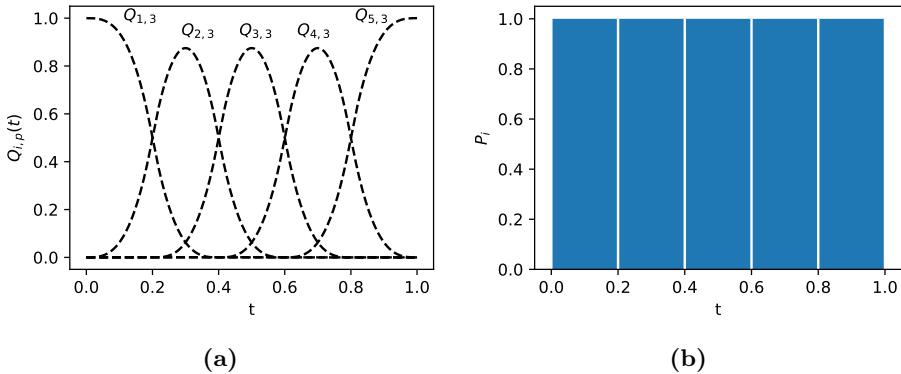


Figure 3.7: P_i is area of each rectangle i.e. 0.2.

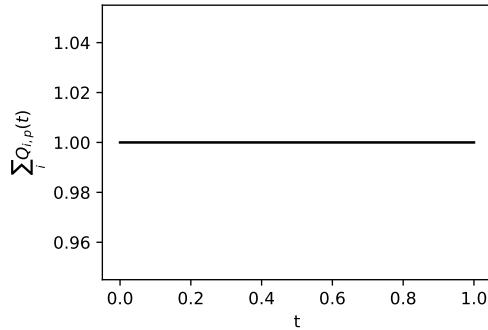


Figure 3.8

3.4.4 Naïve KDE

As we shall see in [REFERENCE TIL RESULTAT AFSNIT OM EST AF MIUD FRA PERFEKT DENSITET](#), if one can accurately determine the Copula density of X_1 and X_2 , then using an approximation of the integral, one can calculate the mutual information to any precision wanted. This clearly follows from the above analysis regarding the behavior of the discretization of X_1 and X_2 in the limit as the mesh gets more fine. Thus, if we can estimate the joint Copula density well, we obtain a good estimate of the mutual information of X_1 and X_2 . A widely used non-parametric method for density estimation is kernel density estimation. Namely, if $\{\mathbf{x}_i\}$ is a set of n , d -dimensional observations from a population, i.e. \mathbf{x}_i can be both scalars and vectors in the case of a multi-dimensional distribution, the kernel density estimator (KDE) of the probability

density function is in general given as

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\prod_{j=1}^d \mathbf{h}_{i,j}} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{\mathbf{h}_i}\right) \quad (3.16)$$

where \mathbf{h}_i is the bandwidth (vector) associated with observation \mathbf{x}_i and K is the kernel (function), defined on the domain of \mathbf{X} which is often \mathbb{R}^d . Often, the bandwidths \mathbf{h}_i are taken to be equal and initially, we shall do so as well. Furthermore, the kernel K is a non-negative function, and is in itself a density function i.e. integrates to 1 as shown below. This ensures that \hat{f} in Equation 3.16 integrates to 1 and is non-negative i.e. a proper distribution.

$$\int_{\mathbb{R}^d} K(\mathbf{x}) d\mathbf{x} = 1$$

In one dimension, a particular useful kernel is the Gaussian kernel given by $K(x) = \phi(x)$ where ϕ is the density function for the standard Normal distribution. ϕ is chosen due to its simple behavior and mathematical properties. In particular, we shall see in the following section, that the properties of the Gaussian kernel allows for simple expressions when correcting for a boundary such that computation is quick and efficient. For multiple dimensions, we often consider product kernels, which are kernels K of the form

$$K(\mathbf{x}) = \prod_{i=1}^d K_i(x_i) \quad (3.17)$$

I.e. just a product of kernels. In particular, we choose $K_i = \phi$ again due to the numerical properties. Thus, initially, we have a KDE \hat{f} of the following form where we once again note that $\mathbf{h}_i = \mathbf{h}$ for all $i \in \{1, \dots, n\}$ such that h_j denotes the bandwidth associated with the j 'th dimension.

$$\hat{f}(\mathbf{x}) = \frac{1}{n \prod_{j=1}^d h_j} \sum_{i=1}^n \prod_{j=1}^d \phi\left(\frac{x_j - x_{i,j}}{h_j}\right)$$

The choice of bandwidth \mathbf{h} is important regarding a trade-off between the variance and bias of the KDE. In general, we want to choose h as small as possible resulting in the least bias but a too small \mathbf{h} will result in large variance of the estimator. In particular, \mathbf{h} acts as a smoothing parameter like the number of bins from the previous section, but here, we can choose any $h > 0$ making the KDE a much more versatile tool. Often the *Mean Integrated Square Error* (MISE) is used which is the expected L^2 -norm of $\hat{f} - f$ i.e.

$$\text{MISE}(\hat{f}) = \mathbb{E}_f \left[\int_{\mathbb{R}^d} \left| \hat{f}(\mathbf{x}) - f(\mathbf{x}) \right|^2 d\mathbf{x} \right]$$

which of course depends on \mathbf{h} . The expectation \mathbb{E}_f denotes the expectation with respect to the samples $\{\mathbf{x}_i\}$ of \mathbf{X} with (true) density distribution function f . Expanding the above expression, we obtain a simple expression relating MISE to the integrated squared bias and integrated variance as shown below

$$\text{MISE}(\hat{f}) = \int_{\mathbb{R}^d} \left| \mathbb{E}_f [\hat{f}(\mathbf{x})] - f(\mathbf{x}) \right|^2 d\mathbf{x} + \int_{\mathbb{R}^d} \text{Var}[\hat{f}(\mathbf{x})] d\mathbf{x}$$

It is however quite complicated to optimize the above, and we shall thus often use a simple rule of thumb known as Scott's rule [11] for choosing \mathbf{h} . Namely, for product kernels, we let the bandwidths of each dimension j equal the following where $\hat{\sigma}_j$ is the standard deviation estimated from the observations of X_j

$$h_j^{Scott} = \hat{\sigma}_j n^{-1/(d+4)}, \quad j \in \{1, \dots, d\}$$

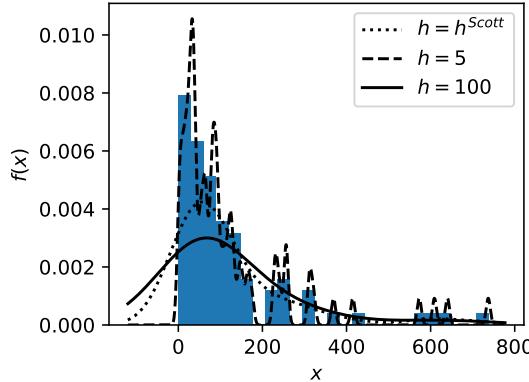


Figure 3.9: The suicide data from MHE. The 86 observations are shown as a histogram of densities along with Gaussian KDEs with bandwidth $h = 5$ and $h = 100$ and h chosen from Scott's rule $h = \hat{\sigma}n^{-1/5} \approx 59.86$.

In Figure 3.9, we have shown a basic example in one dimension with two different manual choices of the bandwidth h and h chosen by Scott's rule. We have used data from [12], tabulated in [13] which has been used in [14] which propose a method for correcting the KDE near a boundary which we shall discuss in the following section. The data consists of 86 observations regarding suicide and is known to be non-negative. In consideration of the reader, we have included the observations in Table 6.1. It is clear that using h^{Scott} results in what we qualitatively would deem a good estimate for the probability density function as $h = 100$ seen to be overly smoothed whereas $h = 100$ too under-smoothed. In particular, from repeated samples we would expect the estimator using $h = 100$ would have large bias but small variance whereas for $h = 5$ would have much

larger variance but smaller bias. However, a problem the estimators, $h = 100$ and $h = h^{Scott}$ especially, have is that they have probability mass below 0 which in this case is unwanted. I.e. when restricting \hat{f} to $[0, \infty)$ they are no longer proper probability distribution functions as they do not integrate to 1. A simple fix could be to simply rescale \hat{f} such that this is the case, but as seen from the example in Figure 3.10 where this method is applied to the same example as above, this tends to underestimate the peaks especially near the boundary.

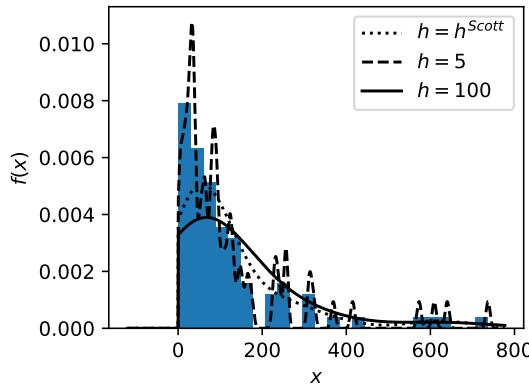


Figure 3.10: Using a rescaled version of \hat{f} on the interval $[0, \infty)$ and disregarding any probability mass below $x = 0$ we obtain proper probability distributions once again. However, neither of the methods capture the peak near the boundary $x = 0$. In particular, although h^{Scott} still seem to be a good choice for h , the KDE does not capture the tendency observed in the data.

We note that using a non-constant h would improve on this behavior, but simpler methods exists, and we thus proceed in the next section with a method that shows great promise regarding this seemingly fundamental issue with KDE. In particular, we refer to a systematic way of letting the shape of each of the kernels depend on the associated observation x_i .

3.4.5 Boundary corrected KDE

Before introducing the boundary corrected kernels presented by [14], we mention another simple method of boundary correction called reflection. Namely, suppose without loss of generality $x = 0$ is the lower boundary of the domain of \mathbf{X} and let \hat{f} be KDE as from the previous section. Then, the reflection boundary corrected KDE denoted \hat{f}_R is defined as

$$\hat{f}_R(x) = \hat{f}(x) + \hat{f}(-x)$$

Clearly, \hat{f}_R is non-negative, and it follows from the below that it is also a proper density function as the probability mass is 1

$$\int_0^\infty \hat{f}_R(x) dx = \int_0^\infty \hat{f}(x) + \hat{f}(-x) dx = \int_0^\infty \hat{f}(x) dx + \int_{-\infty}^0 \hat{f}(x) dx = 1$$

Also, the above is easily extended to two boundaries. Namely, if the domain is $[a, b]$, the reflection boundary corrected KDE is given by

$$\hat{f}_R(x) = \hat{f}(x) + \hat{f}(2a - x) + \hat{f}(2b - x), \quad x \in [a, b]$$

If we once again apply this to the suicide data, comparing to Figure 3.10 we see a big improvement near the boundary as shown in Figure 3.11a. However, we still proceed with the method originally presented in [14]. The reason for this is that when testing for distribution type using the Kolmogorov Smirnov test (on a 5% significance level) we reject that the observations originate from \hat{f}_R with $h = 100$. For $h = 5$ we do not but due to the above considerations

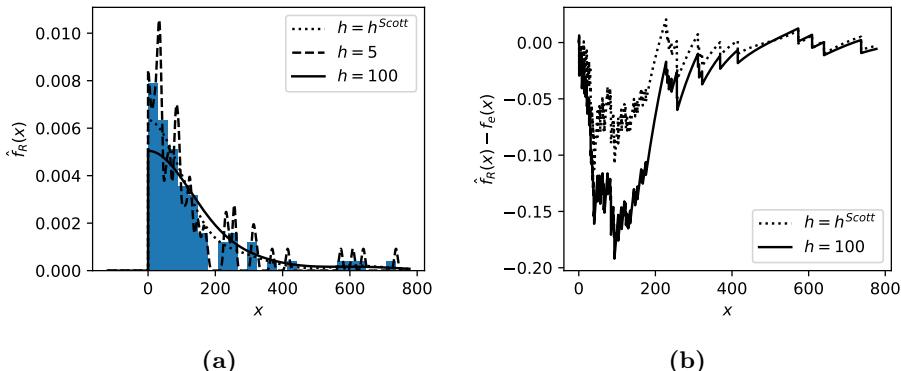


Figure 3.11

regarding the integrated point wise variance of the estimator, this choice of h is undesired in either case. For $h = h^{Scott}$ we do not reject the distribution but as

the shape of the error resembles the error for $h = 100$ we suspect that there is some systematic error which we also see from Figure 3.11b. Furthermore, the largest deviation from the empirical distribution to \hat{f}_R is close to the boundary (as expected). The test statistics (the largest absolute difference D between the distribution functions and the adjusted statistic) is shown in Table 3.1 where the adjusted test statistic should be compared to the critical value 1.358 on a 5% significance level. Furthermore, to really test the kernel density estimators, one should compute the MISE based on bootstrap and/or cross-validation. In particular, this way larger h i.e. more smoothing would be more favorable compared to the Kolmogorov Smirnov test results which shows small h best represent the empirical distribution.

h	D	Adjusted D
5	0.029042	0.27315
h^{Scott}	0.11262	1.0593
100	0.19194	1.8053

Table 3.1: $(\sqrt{n} + 0.12 + 0.11/\sqrt{n}) D$

We now turn our attention to the boundary corrected KDE from [14]. They shortly described this in [5] where it was used to estimate the mutual information in terms of Copula entropy. However, issues arise when using this KDE in terms of non-negativity of the KDE and the facts that it does not integrate to unity. All of these issues can however be handled in a general way without effecting the results regarding bias of the estimator as we shall also see from the above example on suicide data. In particular, [14] show that the bias of their estimator \hat{f}_L is of order $\mathcal{O}(h^2)$ whereas the reflection method discussed above has $\mathcal{O}(h)$ bias. The basic idea of \hat{f}_L is that it is a linear combination of a symmetric kernel K and xK i.e. a first order kernel. They do however only give explicit expressions when a lower bound at $x = 0$ is enforced but the math generalize nicely to 2 boundaries. In particular, we shall let x_u denote the upper bound of the domain and keep $x = 0$ as the lower bound. Before continuing with the derivation and results regarding implementation for the Gaussian kernel we note that in \mathbb{R}^2 , there is a library `evmix` which implements the boundary corrected KDE from [14] but only for 1 dimension and only with a lower bound at $x = 0$. We thus expand on this library (although in `Python`) to include both lower and upper bounds and furthermore, generalize to multiple dimensions using a product kernel as noted in Equation 3.17.

To expand on the boundary corrected KDE to include an upper bound on the domain, we define the functions $a_m(x)$. Note that in [14], they define a_m as a

²<https://search.r-project.org/CRAN/refmans/evmix/html/bckden.html>

function of $p = x/h$ where h is the bandwidth, but to keep the expression later on easier to understand in terms of kernel centers etc. we instead define them as a function of $x \in [0, x_u]$. The reason for initially defining a_m in terms of h is to keep the bandwidth out of the expression, which we then define as follows, when there is no upper bound

$$a_m(x) = \int_{-\infty}^{\frac{x}{h}} u^m K(u) du, \quad x \in [0, \infty)$$

The above is actually equal to the part of the m th moment of the kernel centered at x (and width bandwidth h), that is inside the interval $[0, \infty)$ up to a difference in sign. In particular, using the change of variables $z = x - hu$ we have that

$$a_m(x) = (-1)^m \int_0^\infty \left(\frac{z-x}{h} \right)^m \frac{1}{h} K\left(\frac{z-x}{h}\right) dz$$

where we have used that K assumed to be symmetric. Indeed, the above is as described the part of the m th moment that comes from $[0, \infty)$ (up to a difference in sign) of the kernel that is centered at x with bandwidth h . From this, it is a natural extension to replace the upper bound of the integral with x_u which, when expanding on the initial definition in [14] as done in e.g. [15] turns out to be the correct adjustment. Thus, a_m can be understood as part of the moments which we shall then use as normalizing functions. Thus, for an upper bound x_u we find that instead a_m is defined as

$$a_m(x) = \int_{\frac{x-x_u}{h}}^{\frac{x}{h}} u^m K(u) du, \quad x \in [0, x_u]$$

The boundary adjusted KDE which we shall denote K^L and index depending on the i 'th kernel center is then given by

$$K_i^L(x) = \frac{1}{h} \frac{a_2(x) - a_1(x) \frac{x-x_i}{h}}{a_0(x) a_2(x) - a_1^2(x)} K\left(\frac{x-x_i}{h}\right) \quad (3.18)$$

For the Gaussian kernel, $a_m(x)$ for $m \in \{0, 1, 2\}$ are easily computed and implemented in code through standard routines as they have simple closed forms in terms of ϕ and Φ , i.e. the standard Gaussian density and distribution functions. Namely, for a_0 we simply have that by definition of Φ

$$\begin{aligned} a_0(x) &= \int_{\frac{x-x_u}{h}}^{\frac{x}{h}} \phi(u) du \\ &= \Phi\left(\frac{x}{h}\right) - \Phi\left(\frac{x-x_u}{h}\right) \end{aligned}$$

And similarly, for a_1 , using that $\int u \phi(u) du = -\phi(u) + C$

$$\begin{aligned} a_1(x) &= \int_{\frac{x-x_u}{h}}^{\frac{x}{h}} u\phi(u) du \\ &= \phi\left(\frac{x-x_u}{h}\right) - \phi\left(\frac{x}{h}\right) \end{aligned}$$

Finally, for a_2 we have, using integration by parts for the first step

$$\begin{aligned} a_2(x) &= \int_{\frac{x-x_u}{h}}^{\frac{x}{h}} u^2\phi(u) du \\ &= [-u\phi(u)]_{\frac{x-x_u}{h}}^{\frac{x}{h}} + \int_{\frac{x-x_u}{h}}^{\frac{x}{h}} \phi(u) du \\ &= \left(\frac{x-x_u}{h}\phi\left(\frac{x-x_u}{h}\right) - \frac{x}{h}\phi\left(\frac{x}{h}\right)\right) + \left(\Phi\left(\frac{x}{h}\right) - \Phi\left(\frac{x-x_u}{h}\right)\right) \\ &= \frac{x}{h}a_1(x) - \frac{x_u}{h}\phi\left(\frac{x-x_u}{h}\right) + a_0(x) \end{aligned}$$

From the improved bias of $\mathcal{O}(h^2)$ for K_i^L , we have then obtained an estimator that should perform better in terms of a smaller MISE as we have less bias and by choosing h optimally, we should expect small variance as well. Figure 3.12a shows the KDE \hat{f}_L using K^L as the kernel (with $x_u = \infty$) and is compared to the reflected kernel from above using the same bandwidth. We note that \hat{f}_L has been rescaled such that it integrates to unity on $[0, \infty)$. Also, from Figure 3.12b we see that the density is not statistically significant, and although we can not conclude from this alone, we observe that the deviation from the empirical distribution functions is less than that of the reflected KDE.

As noted above, we need to rescale \hat{f}_L to unity as the kernels in Equation 3.18 does not have unit integrals. Furthermore, \hat{f}_L is not even ensured to be non-negative. This can be handled in multiple ways. One way is to simply take the maximum of \hat{f}_L and 0 effectively cutting off any negative density however in [16], a method is given for KDE based on K_i^L specifically, ensuring the $\mathcal{O}(h^2)$ bias of the estimator. Namely, let K_i^N be given as

$$K_i^N(x) = \frac{1}{h a_0(x)} K\left(\frac{x-x_i}{h}\right)$$

which is then the kernel, locally renormalized using a_0 . We then define the related KDE as

$$\hat{f}_N(x) = \frac{1}{n} \sum_{i=1}^n K_i^N(x)$$

which is then non-negative everywhere as $a_0(x), K(x) > 0$. The non-negative boundary corrected KDE, denoted $\hat{f}_P(x)$ is then defined as

$$\hat{f}_P(x) = \hat{f}_N(x) e^{\frac{\hat{f}_L(x)}{\hat{f}_N(x)} - 1}$$

This \hat{f}_P is also shown in Figure 3.12 resulting in identical density functions. We note that \hat{f}_P works by multiplying the non-negative \hat{f}_N by a (large positive) constant when \hat{f}_L is larger than \hat{f}_N and thus drives \hat{f}_N towards \hat{f}_L . However, from implementation and trying on different distributions, sometimes we observe some odd behavior that can easily be derived from the definition of \hat{f}_P . Thus, we propose a modification to overcome these odd properties of \hat{f}_P . Namely, suppose both \hat{f}_L and \hat{f}_N are close to 0 at some point x . If \hat{f}_L is a magnitude of 10 larger than \hat{f}_N , then \hat{f}_P is approximately $8000 \cdot \hat{f}_N$ which even if \hat{f}_N is small may be large. Thus, it is possible even if both are close to 0, that $\hat{f}_P \gg 0$ which is in contrast to what we would want from the estimator. Thus, we propose a regularized KDE version of \hat{f}_P denoted \hat{f}_{regP} which is obtained from first rewriting \hat{f}_P as follows

$$\hat{f}_P(x) = \bar{f}(x) e^{\frac{\tilde{f}(x) - \hat{f}(x)}{\tilde{f}(x)}}$$

then, we introduce the regularizing parameters $\lambda \geq 0$ such that

$$\hat{f}_{regP}(x) = \bar{f}(x) e^{\frac{\tilde{f}(x) - \hat{f}(x)}{\tilde{f}(x) + \lambda}}$$

In practice, we have found that $\lambda = 0.001$ is a sufficient regularization whilst preserving the shape. A small λ is preferred as then we are ensured $\mathcal{O}(h^2)$ behavior of the bias. In Figure 3.12 we have also shown this regularized version with $\lambda = 0.001$ and observe that it is basically identical to \hat{f}_P on the domain while also behaving well numerically for large x , where the issue discussed above arise for \hat{f}_P . Before continuing with a few more interesting methods of density estimation and mutual information estimation in particular, we note that in the same R package there is a KDE based on the Gaussian Copula density which at first glance might seem a good choice especially if we are trying to estimate the density of a Copula that we know to be Gaussian. The KDE is based on [17], and trivially we get that the domain of the basis functions is already $[0, 1]$. In particular, the kernel for a given bandwidth h and kernel center x_i is given by

$$K_i^{GC}(x) = \frac{1}{h\sqrt{2-h^2}} e^{-\frac{(1-h^2)}{2h^2(2-h^2)} \left((1-h^2) \left((\Phi^{-1}(x))^2 + (\Phi^{-1}(x_i))^2 \right) - 2\Phi^{-1}(x)\phi^{-1}(x_i) \right)}$$

which is obtained by letting $h^2 = 1 - \rho$ in the original expression for the Gaussian Copula density. However, in practice, we shall see that this choice of kernel has some numerical instabilities especially for large correlations i.e. small h and does not perform as well as \hat{f}_{regP} even when using their optimal bandwidth.

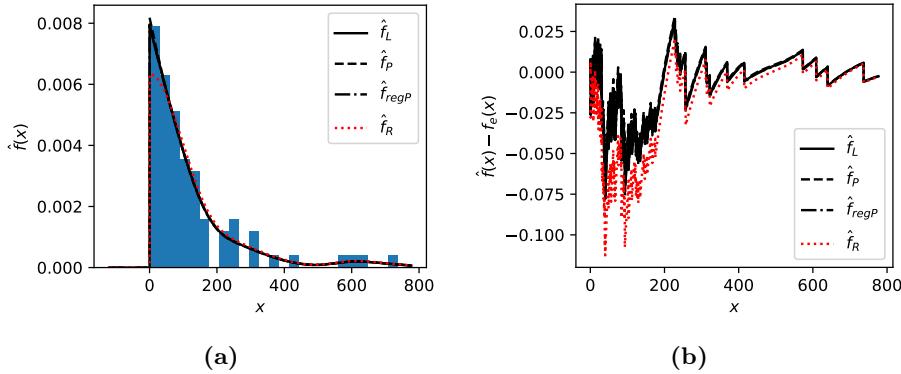


Figure 3.12

FUrthermore, we note that as observations are initially transformed such that they are approximately uniform on $[0, 1]$, the variance of these uniform observations $u_i = \hat{F}(x_i)$ will also be approximately $1/12$ and hence the Scott rule of thumb will give a near constant bandwidth. However, as we shall see later, the performance is drastically improved by choosing smaller bandwidth for sub-domains with high density. We thus note that a local bandwidth should be considered in the future. This could e.g. by using k -nearest neighbor and the distances of these which is the basic idea of [18]. Although we will not be using this method as it is out of scope for this paper, we note that their estimator seem to result in good estimates when $n \geq 300$. Using the core ideas of this paper, one could perhaps deduce a better algorithm for choosing h globally or locally.

Finally, we note that a recent method based on diffusion [19] [20] which shows great potential. Without going into too much detail, it works by considering the PDE

$$\frac{\partial}{\partial t} \hat{f}(x; t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} \hat{f}(x; t)$$

where $t = h^2$. The Gaussian kernel is then the unique solution when the domain is \mathbb{R} with condition $\hat{f}[x; 0] = \frac{1}{n} \sum_{i=1}^n \delta(x - x_i)$ i.e. the observations is the boundary condition at $h = 0$. If instead we add that the support should be e.g. $[0, 1]$ the Neumann boundary conditions on the boundary results in an analytical solution. Namely, from the boundary conditions

$$\left. \frac{\partial}{\partial x} \hat{f}(x; t) \right|_{x=0} = \left. \frac{\partial}{\partial x} \hat{f}(x; t) \right|_{x=1} = 0$$

the analytical solution to \hat{f} is

$$\hat{f}_D(x; t) = \frac{1}{n} \sum_{i=1}^n K_i^D(x)$$

where

$$K_i^D(x) = \frac{1}{h} \sum_{k=-\infty}^{\infty} \phi\left(\frac{x - (2k + x_i)}{h}\right) + \phi\left(\frac{x - (2k - x_i)}{h}\right)$$

which is similar to the reflection method from above. They even give an algorithm for calculating t which from the example below appear to work very well. Before presenting this example we note that although it is out of scope to use this method in the following chapter, it would be interesting to see how well it would perform as it is the only KDE here that does not use any transformation near the boundary while still being consistent at the boundary as they show in [20].

Example 3.2 (Diffusion based KDE). *Let $X \sim Beta(1, 4)$ such that $f(x) = 4(1-x)^3$ for $x \in [0, 1]$ and thus $F(x) = 1 - (1-x)^4$. Generating 1000 samples from the distribution and using the algorithm in [20] to estimate the bandwidth, we get a bandwidth $h^D = 0.05461$ whereas with Scott's rule we get $h^{Scott} = 0.04145$. From Figure 3.13 we see that diffusion and the regularized boundary corrected KDE \hat{f}_{regP} from above agree almost everywhere. Only near the lower boundary at $x = 0$ there is a significant difference and here \hat{f}_{regP} seem to fit the true distribution better. This is due to the Neumann boundary condition which enforce \hat{f}_D to have horizontal derivative at both boundaries. Note that we have used both the h^D and h^{Scott} bandwidth for the regularized boundary corrected KDE resulting in basically no difference.*

At this point, we thus have a complete set of methods for both estimating the mutual information from observations and from these, algorithms to estimate the (causal) structure depending on assumptions regarding direction i.e. a topological ordering of the variables. We thus proceed with numerical results in the following chapter and apply the framework on both fully controlled systems and the observations from chapter 2.

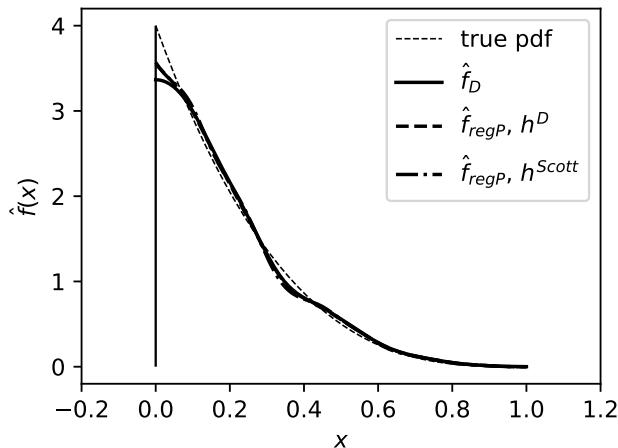


Figure 3.13: Using the diffusion based KDE on samples from a Beta(1, 4) distribution, we see that in general it fits very well. Only at the boundary $x = 0$ there seems to be a problem which on the other hand, \hat{f}_{regP} does not seem to suffer from to the same extend. In particular, repeating the experiment, \hat{f}_{regP} is on average 4 at $x = 0$ while \hat{f}_D constantly seem to undershoot.

CHAPTER 4

Results

An introduction to what is going to be included in this section. What results etc.

In this section, we will investigate how the algorithms Algorithm 1 and Algorithm 2 works in junction and individually. We shall observe how the algorithms can fail and what may be done to correct such cases.

Overordnet pointe er at genere forskellige mulige graphiske modeller, som senere ville kunne bruges til at lave PGM el.l. Er nok bedst som et ekspolartivt værktøj, og vi undersøger her forskellige situationer, og hvornår der kan ske fejl ud fra om det er lange kæder af kausalitet eller mere komplekse strukturer

4.1 Gaussian chains

In this section we discuss the errors made from the assumption that indirect effects can be computed as a sum of powers of the direct effects, i.e. $G_{indir} = \sum_{k \geq 1} G_{dir}^k$. In particular, on a theoretical level, we shall observe the error in \tilde{G}_{obs} based on the above assumption of how similarities are *convolved* which we equate with the noise N from Subsection 3.3.3, although it is a systematic error. To do this, we shall in this section use a multivariate Gaussian to be able to control the correlation and as an extension of this, the mutual information between pairs of random variables. As we already know, correlation and mutual information is independent of the mean and variance of each of the variables however for a bivariate Gaussian the mutual information is given by the correlation as stated in the following proposition.

Proposition 4.1. *Given a bivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Then the mutual information $I(X_1, X_2) = -\frac{1}{2} \ln(1 - \rho^2)$.

Proof. This follows by direct computation Using e.g. that $I(X_1, X_2) = h(X_1) + h(X_2) - h(X_1, X_2)$ \square

Thus, if we know a correlation structure of a Gaussian random vector, we also know the mutual information between every pair of variables which we shall now use in the following made up example. Namely, what we shall denote as a Gaussian chain defined as a Gaussian random vector in the following way. Let \mathbf{X} be a d -dimensional Gaussian random vector, the \mathbf{X} is a standard Gaussian chain if it can be written in the following way in terms of d independent standard normal variables Z_i up to a permutation i.e. there exists a permutation of the variables of the random vector \mathbf{X} that permits the following structure.

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= \rho_{1,2}X_1 + \sqrt{1 - \rho_{1,2}^2}Z_2 \\ X_3 &= \rho_{2,3}X_2 + \sqrt{1 - \rho_{2,3}^2}Z_3 \\ &\vdots \\ X_d &= \rho_{d-1,d}X_{d-1} + \sqrt{1 - \rho_{d-1,d}^2}Z_d \end{aligned} \tag{4.1}$$

It follows that the marginals have variance 1 as clearly $\text{Var}[X_1] = \text{Var}[Z_1] = 1$ and for $i > 1$, $\text{Var}[X_i] = \rho_{i-1,i}^2 \text{Var}[X_{i-1}] + (1 - \rho_{i-1,i}^2) \text{Var}[Z_i] = 1$ by independence of X_{i-1} and Z_i . Thus, the above structure also implies the Cholesky factorization of the correlation matrix for \mathbf{X} , namely

$$L = \begin{bmatrix} 1 & & & & & \\ \rho_{1,2} & \sqrt{1 - \rho_{1,2}^2} & & & & \\ \rho_{2,3}\rho_{1,2} & \rho_{2,3}\sqrt{1 - \rho_{1,2}^2} & \sqrt{1 - \rho_{2,3}^2} & & & \\ \vdots & & & \ddots & & \\ \prod_{i=2}^d \rho_{i-1,i} & \dots & \sqrt{1 - \rho_{j-1,j}^2} \prod_{i=j+1}^d \rho_{i-1,i} & \dots & \sqrt{1 - \rho_{d-1,d}^2} & \end{bmatrix}$$

Which will allow us to both sample from such a chain and calculate G_{dir} and G_{obs} theoretically. However, in this example, it is easier to calculate the correlation between the variable X_i and X_j directly. As the variance of each variable is 1 we simply calculate the covariance. We assume without loss of generality that $i < j$ whence

$$\text{Cov}[X_i, X_j] = \text{Cov}\left[X_i, \rho_{j-1,j} X_{j-1} + \sqrt{1 - \rho_{j-1,j}^2} Z_j\right] = \rho_{j-1,j} \text{Cov}[X_i, X_{j-1}]$$

which by induction implies $\rho_{i,j} = \prod_{k=i+1}^j \rho_{k-1,k} = \rho_{j,i}$. At this point, we are almost ready to use the algorithms from the previous chapter. First, we will only use Algorithm 2 to deconvolve the network based on theoretical correlations and later mutual information. However, before doing so, we note that from the definition in Equation 4.1 the random variable \mathbf{X} exhibits a Markovian property. Namely, the X_i above can be understood discrete stochastic process as they are successively drawn based only on the previous variable X_{i-1} i.e. $f(X_i | X_{i-1}, X_{i-1}, \dots, X_1) = f(X_i | X_{i-1})$. Thus, if the algorithm works as intended, we should observe that the deconvolved network is a *chain* of variables as shown in the Figure 4.1. Thus, we now have the expected result, and we



Figure 4.1: The graphical representation of a Gaussian chain. Arrows signify a possible causal structure. If furthermore, one assumes that X_1 is generated first, then X_2 and so on, this is the only causal structure that would make sense.

proceed with using correlation and mutual information to try and rediscover this structure in the following two sections

4.1.1 Gaussian chain deconvolution using correlation

In this section, we will use the observed correlation i.e. $\rho_{i,j} = \prod_{k=i+1}^j \rho_{k-1,k}$ for the elements of G_{obs} with $i < j$. Note that although it makes sense to consider the correlation between a variable and itself, we shall as discussed before set the diagonal to 0. Furthermore, we have the choice of using either a symmetrical G_{obs} or a (upper or lower) triangular G_{obs} . We shall first use an upper triangular G_{obs} but before using deconvolving using Equation 3.5 we note that we can actually get the result theoretically. Also, as G_{obs} is in this case strictly upper triangular, the spectral radius is 0 and hence we have no problems with converge of the infinite sum of powers of (the uniquely defined) G_{dir} . From the above, it is clear that G_{obs} is given as follows

$$G_{obs} = \begin{bmatrix} 0 & \rho_{1,2} & \rho_{1,2}\rho_{2,3} & \dots & \prod_{k=2}^d \rho_{k-1,k} \\ 0 & 0 & \rho_{2,3} & \dots & \prod_{k=3}^d \rho_{k-1,k} \\ & \ddots & & & \vdots \\ & & 0 & \rho_{d-1,d} & \\ & & & & 0 \end{bmatrix} \quad (4.2)$$

Now, let G_{dir} be given as follows

$$G_{dir} = \begin{bmatrix} 0 & \rho_{1,2} & & & \\ 0 & 0 & \rho_{2,3} & & \\ & \ddots & \ddots & & \\ & & 0 & \rho_{d-1,d} & \\ & & & & 0 \end{bmatrix}$$

then G_{dir}^2 is given by

$$G_{dir}^2 = \begin{bmatrix} 0 & 0 & \rho_{1,2}\rho_{2,3} & & \\ 0 & 0 & 0 & \rho_{2,3}\rho_{3,4} & \\ & \ddots & \ddots & \ddots & \\ & & 0 & 0 & \rho_{d-2,d-1}\rho_{d-1,d} \\ & & & 0 & 0 \end{bmatrix}$$

It is not hard to show that in fact $\sum_{k \geq 1} G_{dir}^k = \sum_{k=1}^d G_{dir}^k = G_{obs}$. Thus, if we know a graph topological ordering of the variables corresponding to the structural causal model, we completely recover (without any error) the direct dependencies/correlation from to the initial definition in Equation 4.1. This actually holds for a general *chain* where Z_i can follow any distribution as long as they are independent as the above computations did not use the fact that Z_i follows a standard Gaussian. From this, we might think that if we have a topological ordering of the variables this is the preferred method, and it is as

long as correlation is a good enough measure of similarity/codependency. Albeit this is only shown for the special case of a chain, in Section 4.2 we consider the more general case and conclude that this indeed holds. Regarding the comment on correlation being a good enough measure of similarity, a prototypical case is when joint probability density function of two variables resemble a parabola. Namely, let $X_1 \sim \mathcal{U}(0, 1)$ and $X_2 | X_1 \sim \mathcal{N}\left(1 - 4(x_1 - 1/2)^2, \sigma^2\right)$ i.e. the joint distribution function is a parabola with a Gaussian noise added along the second dimension. In Figure 4.2, 1000 samples from this distribution is shown for $\sigma = 1/10$ along with the expectation $\mathbb{E}[X_2|X_1]$. It is not hard to show that the covariance between X_1 and X_2 is 0 however we clearly see a relationship between the two variables. In fact, computing the mutual information results in $I(X_1, X_2) \approx 1.030$ implying $X_1 \not\perp X_2$ i.e. there exists a higher order (non-linear) dependency. Thus, if the algorithm permits, we would prefer mutual information to correlation as we can then use observed higher order relationships to infer a causal structure. On a more technical point of view, we note that

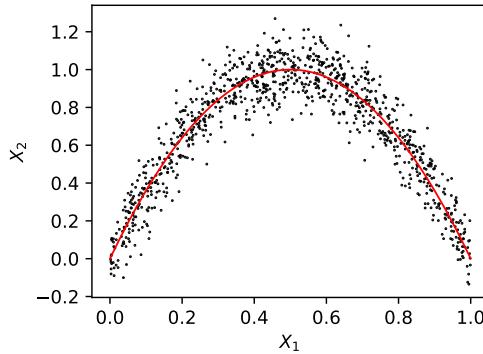


Figure 4.2: 1000 samples generated from $X_1 \sim \mathcal{U}(0, 1)$ and $X_2 | X_1 \sim \mathcal{N}\left(1 - 4(x_1 - 1/2)^2, \sigma^2\right)$ with $\sigma = 1/10$. The mutual information is calculated theoretically to be $I(X_1, X_2) \approx 1.030$ and repeated simulations show that the empirical correlation is symmetric around 0 supporting the claim that the underlying correlation is in fact 0

mutual information is a measure of how dense the joint distribution is, invariant to scale. In a way, it is a measure of how close the joint distribution is to a lower dimensional manifold.

We proceed with a 10-Gaussian chain defined by the following correlations:

$$\begin{aligned} \rho_{1,2} &= 0.6, & \rho_{2,3} &= 0.5, & \rho_{3,4} &= 0.4 \\ \rho_{4,5} &= 0.2, & \rho_{5,6} &= 0.9, & \rho_{6,7} &= 0.8 \\ \rho_{7,8} &= 0.9, & \rho_{8,9} &= 0.8, & \rho_{9,10} &= 0.7 \end{aligned} \tag{4.3}$$

We have chosen correlations of different sizes to check if the deconvolution is robust in presence of both strong and weak links. In particular, X_5 is only $\rho_{4,5}^2 = 4\%$ of X_4 and the remaining 96% is noise/indescribable variance i.e. a very weak link between the first part of the chain up to and including X_4 and the rest. However, as discussed above, if let G_{obs} be upper triangular, we should completely rediscover these direct relations which is indeed also the case. In particular, from Figure 4.3 we observe that the inferred network, represented by G_{dir} , is indeed a chain of variables and is exactly equal to the theoretical G_{dir} as we would expect (up to very small rounding errors of the size 10^{-16}). The estimated G_{dir} is also shown as a directed graph which the initial topological assumption implies, with edges wherever G_{dir} is non-zero. We now proceed to in-

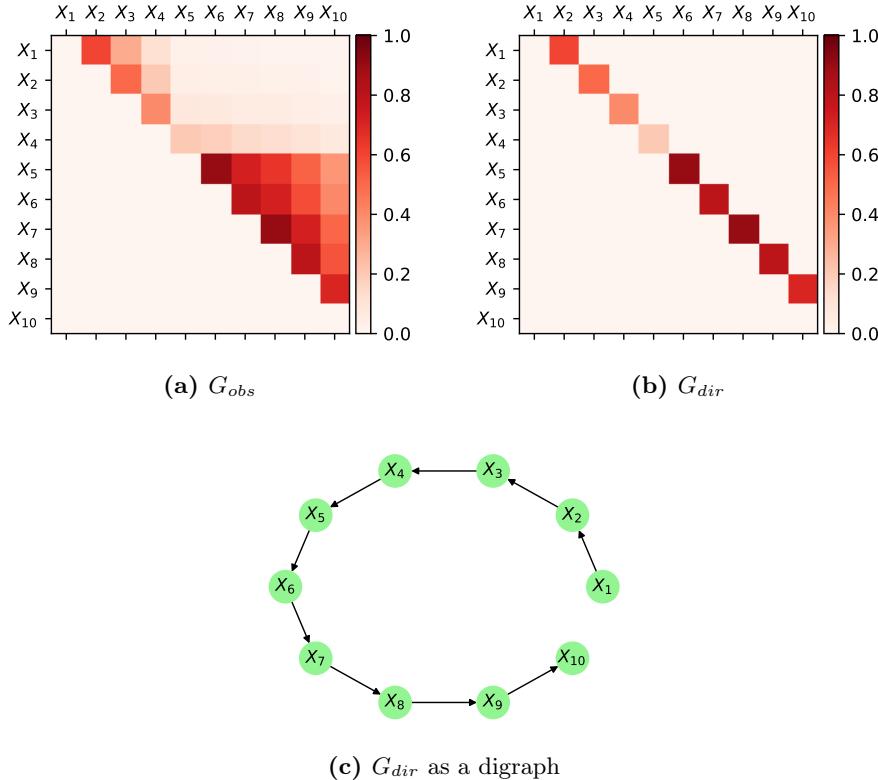


Figure 4.3: Results from using an upper triangular G_{obs} and correlation to infer the causal network structure. (a) shows the upper triangular G_{obs} with the correlation between every pair of variables. (b) shows the deconvolved G_{obs} and as we expect, the superdiagonal contains the original correlations given in Equation 4.3. (c) shows G_{dir} represented as a digraph and matches the expected result.

vestigate what happens when we remove the prior information of the topological ordering. Namely, if G_{obs} is no longer triangular but symmetric. In particular, let T_{dir} be given as G_{dir} above. We then have that G_{dir} in the symmetric case is $T_{dir} + T_{dir}^T$ and similarly for G_{obs} , $G_{obs} = T_{obs} + T_{obs}^T$. Clearly, $I + G_{obs}$ is positive definite as it is a proper correlation matrix. However, that also implies that we might have eigenvalues of G_{obs} less than or equal to $-1/2$ which we know from Subsection 3.1.1 is not the result of a G_{dir} such that Equation 3.4 holds as then the infinite sum diverges. However, as -1 is not an eigenvalue of G_{obs} , we will investigate what happens if one tries to use Equation 3.5 anyway.

First, we shall however discuss the errors being made using the symmetric G_{obs} and G_{dir} . Namely, we investigate the powers of G_{dir} :

$$G_{dir}^2 = (T_{dir} + T_{dir}^T)^2 = T_{dir}^2 + (T_{dir}^T)^2 + T_{dir}T_{dir}^T + T_{dir}^TT_{dir}$$

Higher power can be calculated similarly, but for the second power we already observe an error. The first two terms corresponds to a reflection of the second order effects that we saw above and know to be true, whence the final two terms, that add to a diagonal matrix, is an error and will propagate with higher order powers of G_{dir} . Through simple calculation the resulting error is

$$T_{dir}T_{dir}^T + T_{dir}^TT_{dir} = \begin{bmatrix} \rho_{1,2}^2 + \rho_{2,3}^2 & & & \\ & \rho_{2,3}^2 + \rho_{3,4}^2 & & \\ & & \ddots & \\ & & & \rho_{d-2,d-1}^2 + \rho_{d-1,d}^2 \end{bmatrix}$$

Thus, for chains, we expect larger errors for sub-chains with strong links i.e. a subgraph of a chain that is also a chain where the correlation from one variable to the next is large. Using $G_{obs} = T_{obs} + T_{obs}^T$ we have that the smallest eigenvalue is approximately $\lambda_{\min} \approx -0.92263$ thus, multiplying G_{obs} with a constant $c_s < 0.54192$ will make G_{dir} have spectral radius at most 1. The results vary with one or two edges for the choice of c_s and in the following we have chosen $c_s = 0.53651$ resulting in $\rho(G_{dir}) \approx 0.98020$ and \tilde{G}_{obs} and \tilde{G}_{dir} as seen in Figure 4.4. From Figure 4.4b we see that some correlation/association seem to bleed to variables 2 or 3 edges away which we of course know is not true given the Markov property discussed above. However, it is also clear that the error here is that the original assumption does not hold since using a symmetric G_{obs} imply that the measure of similarity flows both ways where in this case it is very much unidirectional.

From Figure 4.4 conclude that we are somewhat able to rediscover the causal structure. Not surprisingly, we observe that the weak link between X_4 and X_5 is one of the first to break and that we observe some extra edges between the later more strongly linked sub-chain as by the above discussion. Finally, before presenting the results for the unscaled G_{obs} (where the smallest eigenvalue is

smaller than $-1/2$) we note that changing the parameter α in Algorithm 2 did not have much of an effect indicating that the network is quite sparse (as we also know it to be) as even removing 65% of the smallest correlations from G_{obs} did not have any effect. The chosen threshold of $t = 0.2$ on G_{dir} seemed to be the best compromise of a connected graph and the density of the edges (although this is somewhat biased from prior knowledge of the true graphical structure). Finally, we try using the unscaled G_{obs} in Equation 3.5. Interestingly, we find

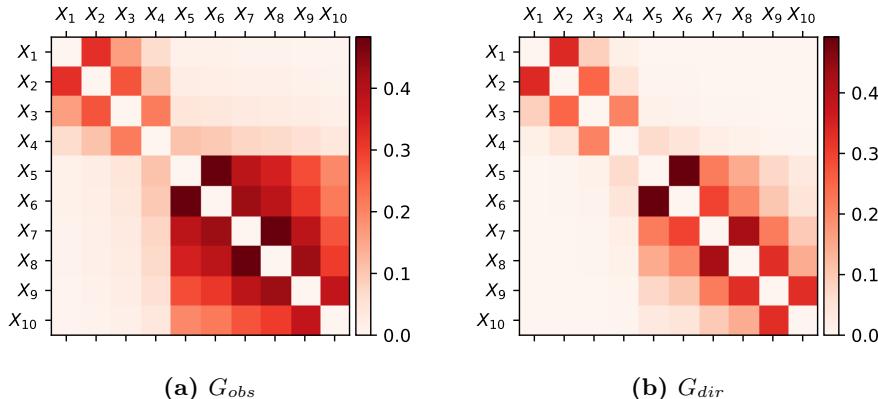


Figure 4.4: Using a symmetric G_{obs} as shown in (a), we observe that higher order effects start to emerge as can be seen in (b). The main response is still in the superdiagonal and subdiagonal as we expect, where some similarity seems to bleed to nearby nodes/variables thus making the threshold used important for the resulting graph. For (c), a threshold $t = 0.2$ was used to obtain a decent compromise between connectedness and denseness of the direct association.

that the true structure emerges as can be seen from Figure 4.5. Although the *correlations* in Figure 4.5b are not really correlations they do resemble those discovered in Figure 4.3b. On closer inspection, it is not apparent how they

are related except that it is a non-linear relationship. Although in this case it seemed to work not rescaling G_{obs} in order to discover the causal structure we will in general not apply this to real world scenarios as the method is not well-defined in terms of assumptions and what the resulting G_{dir} should be interpreted as. Thus, at this point we have a rather good understanding of how

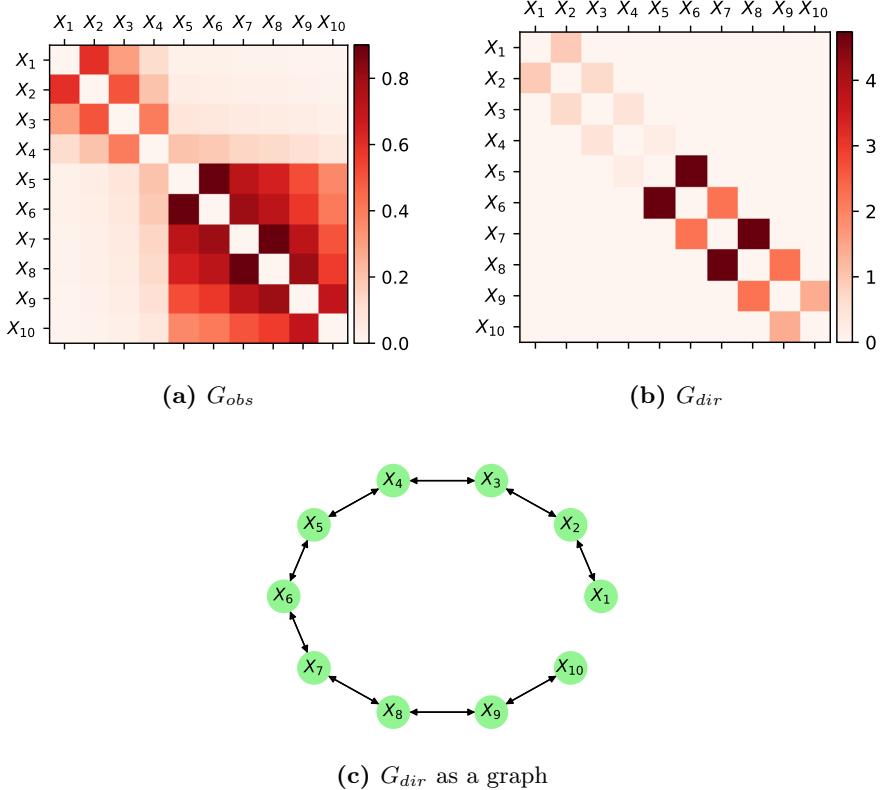


Figure 4.5: Using an unscaled (symmetric) G_{obs} results in a good recovery of the causal structure as seen in (b) and (c). However, at this point it is not clear whether it holds only for chains and using correlation or if it is a more general phenomenon.

the method works on Gaussian chains if one uses correlation as a measure of association. Furthermore, if one knows (a plausible) topological ordering of the variables, we are able to perfectly rediscover the network of direct dependencies. However, as noted above, correlation is not always a good measure of similarity. Thus, we proceed experimenting with mutual information on the same Gaussian chain.

4.1.2 Gaussian chain deconvolution using mutual information

In this section, we continue the example from the previous section but instead of using correlation as a measure of similarity, we will use mutual information. Immediately, we note that mutual information or Copula entropy does not propagate as assumed in Equation 3.4. As an example, from Proposition 4.1, we know that the mutual information in the case of a Gaussian chain between a variable X_i and the next variable X_{i+1} is $-1/2 \log(1 - \rho_{i,i+1}^2)$ and similarly, using Equation 4.2, we have that

$$I(X_i, X_{i+2}) = -\frac{1}{2} \log(1 - \rho_{i,i+1}^2 \rho_{i+1,i+2}^2)$$

Thus, if G_{dir} is triangular, using Equation 3.4 we should observe the following at the $(i, i+2)$ entry of G_{obs} instead

$$\frac{1}{4} \log(1 - \rho_{i,i+1}^2) \log(1 - \rho_{i+1,i+2}^2)$$

I.e. we make an error (which we could take to be the noise N from Subsection 3.3.3) for second order effects equal to

$$-\frac{1}{2} \log(1 - \rho_{i,i+1}^2 \rho_{i+1,i+2}^2) - \frac{1}{4} \log(1 - \rho_{i,i+1}^2) \log(1 - \rho_{i+1,i+2}^2)$$

In general, for a Gaussian chain, we have that

$$N_{i,j} = -\frac{1}{2} \log \left(1 - \prod_{k=i+1}^j \rho_{k-1,k}^2 \right) - \left(-\frac{1}{2} \right)^{j-i} \prod_{k=i+1}^j \log(1 - \rho_{k-1,k}^2)$$

As we will see in Figure 4.6 and Figure 4.7, for Gaussian chains we can expect some of the same bleeding behavior as observed in Figure 4.4 where we did not use the topological ordering but based the deconvolution on correlation. In particular, from the figures below, we see that for 3-chains, the error is in many cases close to 0 and for most combinations of $\rho_{1,2}$ and $\rho_{2,3}$ less than 0.1. Furthermore, we note that the errors are the largest when it is a strongly connected 3-chain i.e. if both $\rho_{1,2}$ and $\rho_{2,3}$ are close to 1 which again resemble the behavior seen in the case of a symmetrical G_{obs} using correlation as the measure of association although in this case, the error does not propagate to the same extend which we shall also see shortly, when applying the deconvolution algorithm. Notice that as as only the absolute value of the correlation matters, we only show the error for $\rho_{1,2}, \rho_{2,3} \geq 0$.

We extend the above discussion to 4- and 5-chains (i.e. $j = i + 3$ and $j = i + 4$ in the above expression for N_{ij}) to see how the error propagates in more detail.

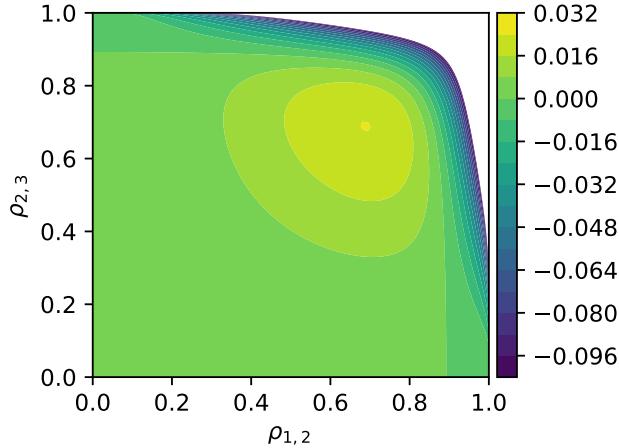


Figure 4.6: The error made by the assumption of G_{obs} and G_{dir} for second order observed effect. Although mutual information does not comply with the underlying assumptions, we observe that in the case of a Gaussian 2-chain, we can expect the error to be relatively small.

This is shown in Figure 4.7 for three different scenarios of a 4-chain and a single 5-chain. In particular, as the error $N_{i,j}$ is symmetric in $\rho_{1,2}$, $\rho_{2,3}$ and $\rho_{3,4}$ (and $\rho_{4,5}$ in the case of a 5-chain) and because it is hard to accurately show four or five dimensional surfaces, we keep to a fixed set of $\rho_{3,4}$ and $\rho_{4,5}$ when investigating. For the 4-chain, choosing $\rho_{2,3} = 0.9$ (corresponding to mutual information about 0.8304) approximately results in the same error as in Figure 4.6 and if $\rho_{2,3}$ is above e.g. 0.95, we get a worse propagation of errors compared to the 3-chain. Finally, from Figure 4.7d, we see the same picture i.e. that keeping the correlations and hence information between subsequent variable low results in smaller errors in G_{obs} and hence the inferred G_{dir} . Note that under the assumption of a topological ordering such that G_{obs} is strictly upper triangular results in $\rho(G_{obs}) = 0$ such that no rescaling is necessary (although different choices of the base of the logarithm would have an effect on how much higher order associations influence G_{dir}).

Having obtained a good understanding of how shifting to mutual information instead of correlation in the case of Gaussian chains, we continue with the above example now using mutual information as the elements of G_{obs} based on the correlation matrix from the previous section. Using a triangular G_{obs} we observe similar behavior to that of original example using a triangular G_{obs} but with correlation as can be seen from Figure 4.8. In particular, we do not observe the same magnitude of bleeding effects as in Figure 4.4. However, we observe

the same tendency to miss weak connections as was also observed in Figure 4.4.

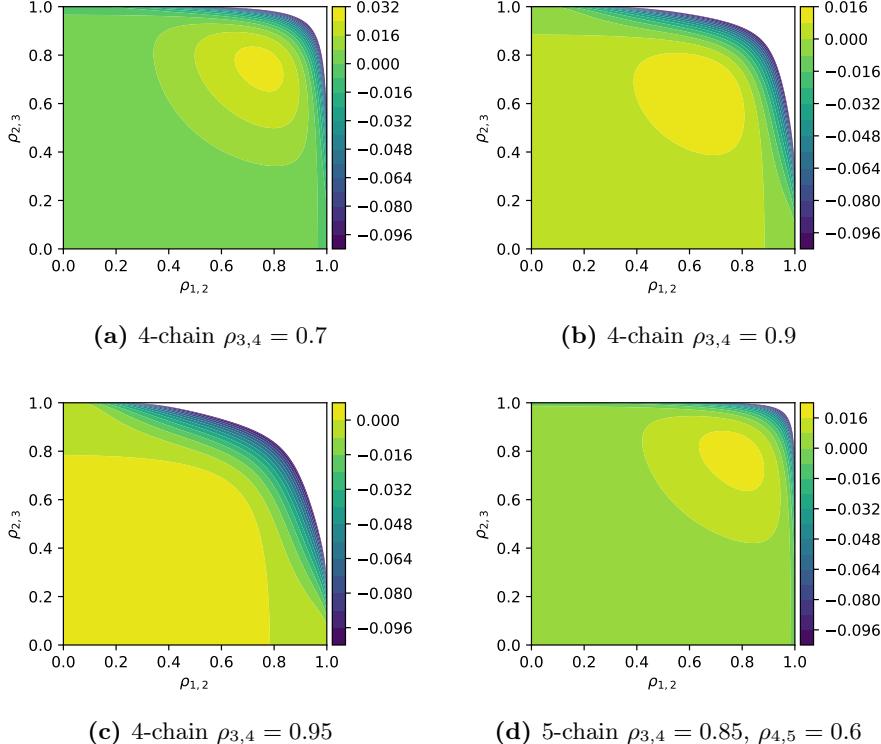


Figure 4.7: Errors of convolving mutual information along a 4-chain (a), (b), (c) and a 5-chain (d). Due to symmetry in the expression of the error, only the first 2 links i.e. $\rho_{1,2}$ and $\rho_{2,3}$ are varied on $[0, 1]$ respectively. Only positive correlations are shown as the sign of the correlation cancels in the expression for the error. We note that large correlations and hence large mutual information on each edge results in larger error. In particular, when not too many of the links are strong, we have almost 0 error.

All in all, we get very good results using a triangular G_{obs} even though mutual information does not have the same properties as correlation. In particular, this is what we expected as we have only used $\rho_{i,i+1} \leq 0.9$, from the above investigation of the error.

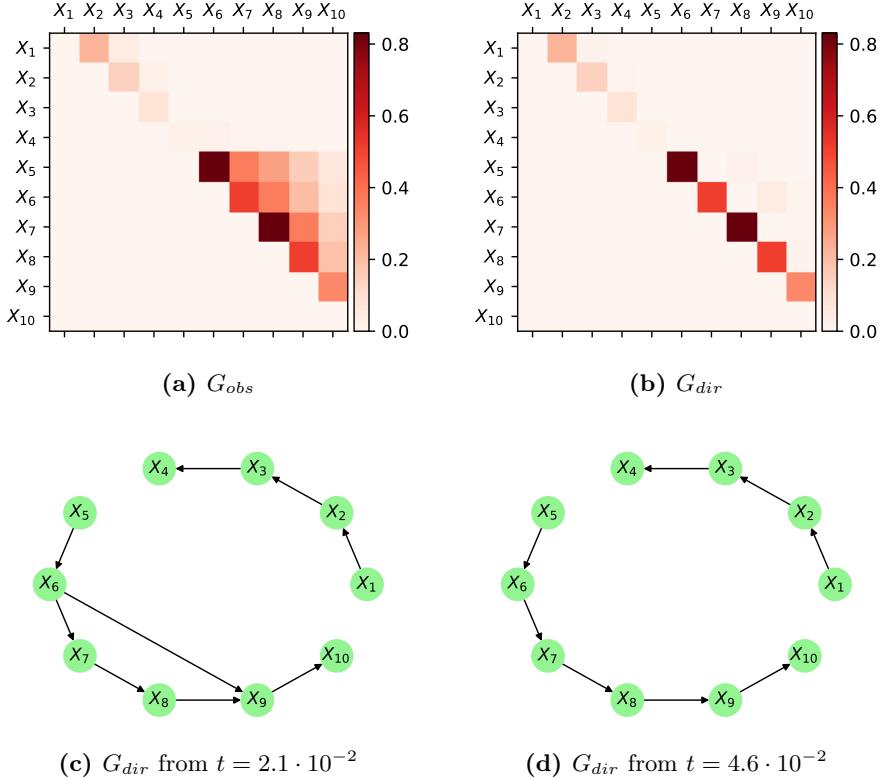


Figure 4.8: Using mutual information as the measure of similarity as well as assuming a topological order i.e. making G_{obs} strictly triangular as seen in (a) we almost perfectly infer G_{dir} as seen in (b) except for $[G_{dir}]_{6,9}$. Choosing cutoffs $t = 2.1 \cdot 10^{-2}$ (c) and $t = 4.6 \cdot 10^{-2}$ (d) it is clear that adjusting the threshold we can get a better result than using a symmetric G_{obs} with correlation.

Finally, we use the corresponding symmetric G_{obs} (rescaled such that the largest absolute eigenvalue of G_{dir} is 0.99) which results in G_{dir} and the graph using a threshold $t = 4.88 \cdot 10^{-2}$ shown in Figure 4.9. Again, we observe some bleeding on the more strongly connected sub-chain as with the symmetric G_{obs} using correlation in Figure 4.4. Again, we observe comparable results and note that increasing the threshold would disconnect X_3 and X_4 before removing the higher order effects.

In conclusion, we have seen what errors can arise in the discovered network using both correlation and mutual information as the measure of association. Namely, long strongly connected chains seem to be a problem if one does not know a

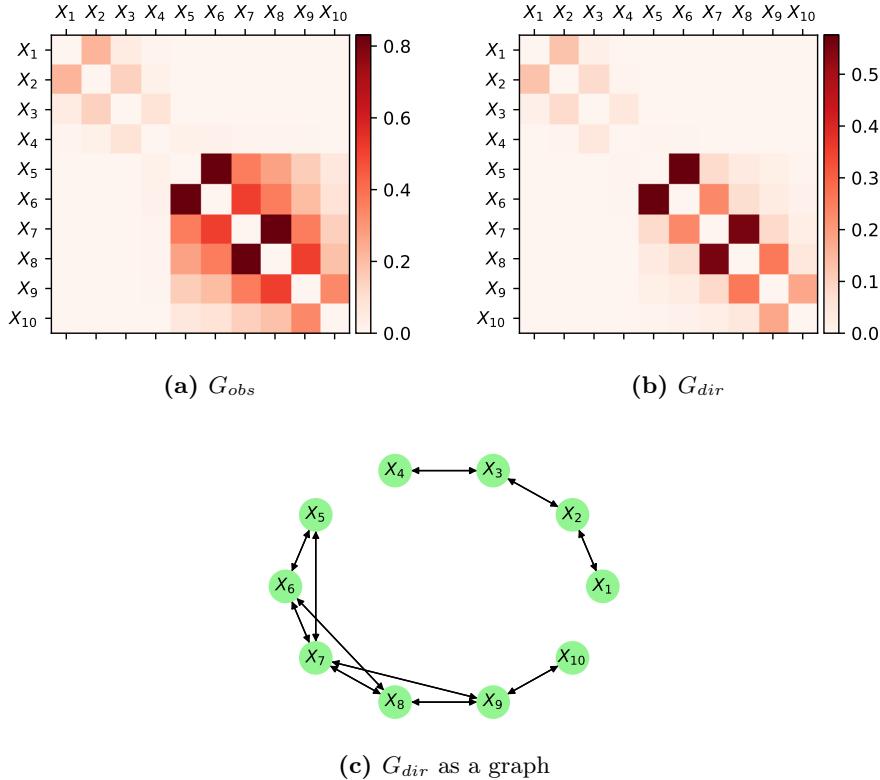


Figure 4.9: Using a symmetric G_{obs} containing the observed mutual information (a) we infer a G_{dir} (b) comparable to that if we had used correlation instead. Choosing the threshold $t = 4.88 \cdot 10^{-2}$ seem a good compromise between connectedness and density resulting in an almost identical discovered network structure to that of using a symmetric correlation G_{obs} .

topological ordering of the variables, in which case these are heavily reduced as seen in Figure 4.3 and Figure 4.8. Thus, we proceed in the next section by considering a more complicated underlying (Gaussian) network to observe if other unwanted effects can occur and if a topological ordering is necessary if the network is not simply a path.

4.2 Directed acyclic Gaussian graphs

In this section, we will expand on the results from the previous section by considering a more general structure. In particular, let \mathcal{G} be a directed acyclic graph with nodes corresponding to variables from a random vector \mathbf{X} with directed edges indicating direct dependencies. Clearly, such a DAG has a topological ordering and as such we shall index the variables 1 through d such that if the index of a variable is i , and j is the index of another element of the random vector \mathbf{X} , then $i < j$ implies there is no (directed) path from j to i . Note that since a topological ordering is not necessarily unique, we can not infer that there is a (directed) path from i to j or even if k is reachable from j (i.e. there exists a path from j to k) it does not follow that k is reachable from i . In Figure 4.10 a subset of such a DAG is shown with a possible labelling where $i < j$ and $k_m < k_n$ when $m < n$. It is then the weights along these directed edges which we will once again call G_{dir} that we wish to infer based on the transitive closure. As an example, from Figure 4.10, the transitive closure would result in an observed similarity between i and j although no 1 path i.e. single direct edge connects the two variables. From the definition of the labels, it is clear

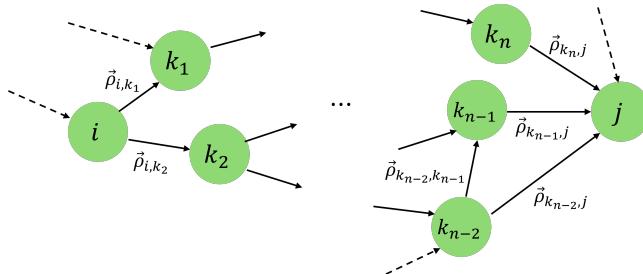


Figure 4.10: A general (linear) network represented as a DAG. The directed edge weights $\vec{\rho}_{k,l}$ specify how much the variable index k make up of variable l . Although i and j are not directly connected, multiple paths may exist between the two nodes, making the propagation of similarity more complex from that of a chain.

that G_{dir} is once again strictly upper triangular as entries below the diagonal corresponds to edges going from a random variable with an index i to another random variable with index j such that $i > j$ which is clearly a contradiction. Also, the diagonal elements are 0 as there can not be any loops in DAGs.

Similarly to the definition of (Gaussian) chains, based on d independent (or even just pairwise uncorrelated) random variables Z_i we can define a general network

of random variables X_i based on \mathbf{Z} in the following way

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= \vec{\rho}_{1,2}X_1 + \sqrt{1 - \vec{\rho}_{1,2}^2}Z_2 \\ X_3 &= \vec{\rho}_{1,3}X_1 + \vec{\rho}_{2,3}X_2 + c_3Z_3 \\ &\vdots \\ X_d &= \sum_{k < d} \vec{\rho}_{k,d}X_k + c_dZ_d \end{aligned} \tag{4.4}$$

where c_i is chosen such that $\text{Var}(X_i) = 1$ to make the analysis later on carries out simpler as then $\vec{\rho}_{i,j}$ is actually the *direct* correlation between the variables indexed i and j as shown in Figure 4.10. Of course, for the variance of each random variable to be 1 there must be some constraints on the chosen $\vec{\rho}_{i,j}$ such as neither one of them can exceed 1 in absolute value. Furthermore, consider the following bound on the variance of X_i assuming c_k for $k < i$ have been chosen such that $\text{Var}(X_k) = 1$.

$$\begin{aligned} \text{Var}[X_i] &= \sum_{k < i} \vec{\rho}_{k,i}^2 + 2 \sum_{k < l < i} \vec{\rho}_{k,i}\vec{\rho}_{l,i} \text{Cov}[X_k, X_l] + c_i^2 \\ &\leq \sum_{k < i} \vec{\rho}_{k,i}^2 + 2 \sum_{k < l < i} \vec{\rho}_{k,i}\vec{\rho}_{l,i} + c_i^2 \\ &= \left(\sum_{k < i} \rho_{k,i} \right)^2 + c_i^2 \end{aligned} \tag{4.5}$$

where we have used that Z_i is uncorrelated with X_k for $k < i$ and that the covariance between variables with variance 1 is at most 1 to obtain the inequality. Hence, choosing the sum of the ingoing edges to be at most 1 for every node ensures that the constants c_i for $i \in \{2, \dots, d\}$ exist in order to make the variance of each X_i 1. This, we will use in the following example to easily build a network such that $\vec{\rho}_{i,j}$ is the pure correlation.

However, before constructing an example and using bot correlation and mutual information we must determine the theoretical G_{obs} for both cases. To do this, we shall consider the (i, j) element of G_{obs} when using correlation as a measure of similarity and later use mutual information based on these correlations and Proposition 4.1 in the case of \mathbf{Z} being a Gaussian random vector. To calculate $[G_{obs}]_{i,j}$ we shall consider the immediate predecessors to node j in the graph \mathcal{G} corresponding to Equation 4.4. The immediate predecessors or *in-neighbors* of a node j is denoted $N^-(X_j)$ or in shorthand notation N_j^- . An example of this is shown in Figure 4.11 where the in-neighbors of j has been marked in red. With this notation, we proceed with the computation of the (i, j) entry of G_{obs} which is the covariance between X_i and X_j when $i < j$ and 0 elsewhere.

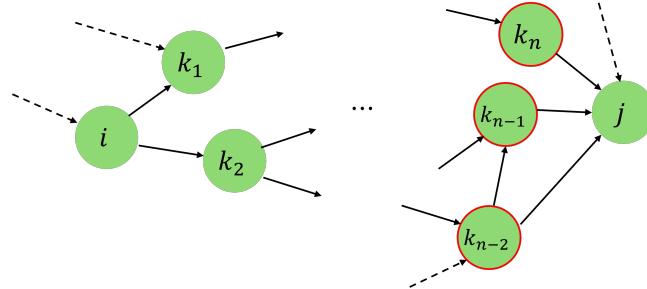


Figure 4.11: For node j , the set N_j^- is illustrated with red borders. It is exactly the set of nodes going directly into j . We note that an in-neighbor l of in-neighbor k of node j can also be an in-neighbor of j i.e. l can influence both k and j whilst k also directly influenced j . It is in particular these direct dependencies we want to be sure of as their existence makes the network more complex but failing to discover these can lead to a significant reduction in prediction accuracy.

$$\begin{aligned}
 [G_{obs}]_{i,j} &= \text{Cov} \left[X_i, \sum_{k \in N_j^-} \vec{\rho}_{k,j} X_k + c_j Z_j \right] \\
 &= \text{Cov} \left[X_i, \sum_{k \in N_j^-} \vec{\rho}_{k,j} X_k \right] \\
 &= \sum_{k \in N_j^-} \vec{\rho}_{k,j} \text{Cov}[X_i, X_k] \\
 &= \sum_{k=1}^{j-1} \vec{\rho}_{k,j} \text{Cov}[X_i, X_k] \\
 &= \vec{\rho}_{i,j} + \sum_{k=1}^d \vec{\rho}_{k,j} [G_{obs}]_{i,k}
 \end{aligned} \tag{4.6}$$

For the fourth equality, we have used that $\vec{\rho}_{k,j} = 0$ whenever $k \notin N_j^-$ which again for the fifth equality holds for any $k \geq j$. Furthermore, since $[G_{obs}]_{i,i} = 0$ we need to add $\vec{\rho}_{i,j}$ to make the final equality hold. It is clear that the above can also be expressed as a matrix equation, namely

$$G_{obs} = G_{obs} G_{dir} + G_{dir}$$

Hence, as G_{dir} is strictly upper triangular thus making $I - G_{dir}$ invertible, we can directly express G_{obs} in terms of G_{dir} . We find that

$$G_{obs} = G_{dir} (I - G_{dir})^{-1}$$

which we recognize as Equation 3.4 hence also for a general network (and not just a chain), using correlation and knowing/assuming a topological order of the random variables we are able to perfectly rediscover G_{dir} from G_{obs} .

With the above, we then define an example Gaussian network with the following weights and shown in Figure 4.12 to get a better understanding of this example hopefully should reappear after deconvolution using both correlation and mutual information respectively.

$$\begin{aligned} \vec{\rho}_{1,2} &= 0.7, & \vec{\rho}_{5,6} &= 0.5, & \vec{\rho}_{2,7} &= 0.3 \\ \vec{\rho}_{6,7} &= 0.3, & \vec{\rho}_{6,8} &= 0.7, & \vec{\rho}_{4,9} &= 0.3 \\ \vec{\rho}_{8,9} &= 0.3, & \vec{\rho}_{7,10} &= 0.4, & \vec{\rho}_{9,10} &= 0.2 \end{aligned} \quad (4.7)$$

In particular, from Equation 4.5 and Figure 4.6 and Figure 4.7, we suspect that the bleeding effects observed for the Gaussian chain won't appear to the same extent in this case.

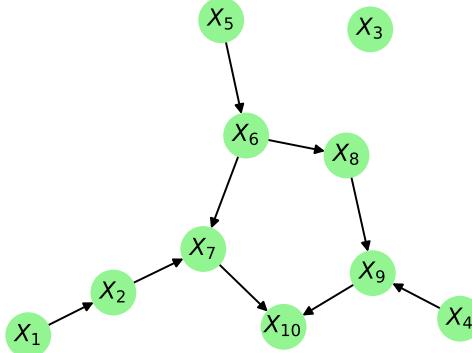


Figure 4.12: The graph defined in Equation 4.7. Note that X_3 is neither influenced nor influences any other variable. It is of course in our interest to accurately tell if X_3 should be considered if we try to infer a probability distribution on e.g. X_{10} given observations of the other variables.

Applying the deconvolution algorithm, we obtain the results in Figure 6.3 which trivially, from the above analysis on G_{obs} , results in a perfect reconstruction of the network. If instead, we do not assume a topological structure, we can also recover the structure, although we need to tune the threshold as can be seen from Figure 4.13. Tuning the α and β did not have much of an effect. Actually, decreasing β seemed to worsen the results which is also in line with our expectations as choosing smaller β skews the effects of higher order interactions.

Thus, it is primarily the threshold that we want to tune in this case and choosing $t = 1.18 \cdot 10^{-1}$ we accurately infer the network structure contrary to the results from the Gaussian chain. However, we still observe second order effects i.e. the edge between X_5 and X_8 which was also the case in Figure 4.4 Finally, before

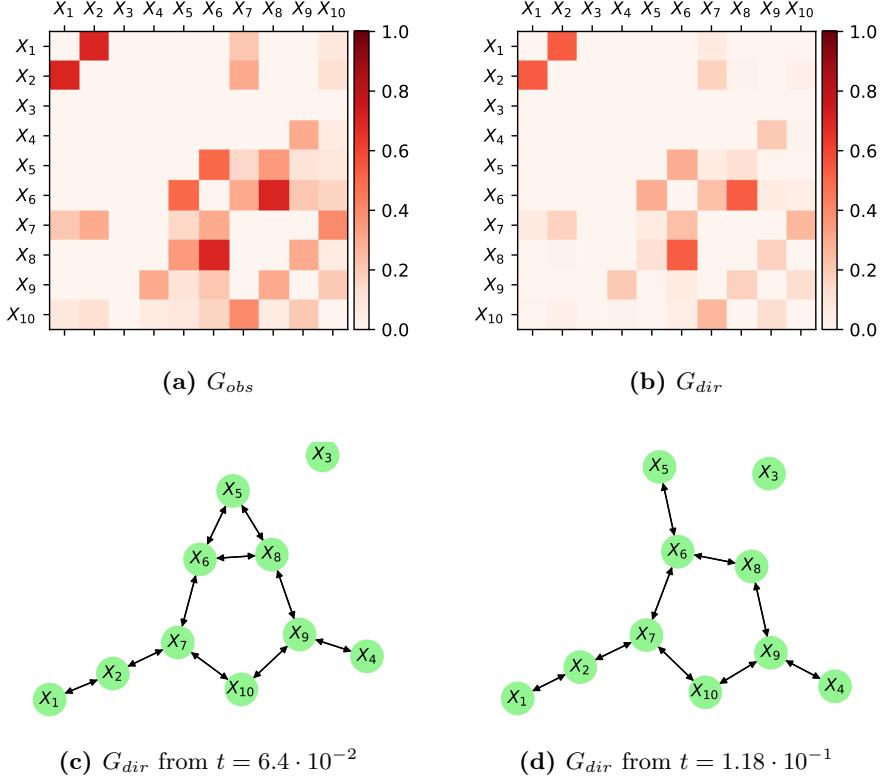


Figure 4.13: Not knowing the topological structure and thus using a symmetric G_{obs} (a) we obtain the G_{dir} in (b). Clearly, there is some bleeding, but choosing the threshold $t = 1.18 \cdot 10^{-1}$ we can accurately rediscover the network structure up to a direction on the edges. As with the previous example of Gaussian chains, we observe some tendency to inaccurately filter out second order effects as can be seen in (c) where X_5 and X_8 is connected.

continuing with results regarding the different methods for estimating mutual information, we present the results from above using mutual information instead of correlation as the measure of similarity. Namely, once again assuming the topological order such that G_{obs} is strictly upper triangular and hence no need for rescaling we get the results shown in Figure 4.14. As expected, we observe on par performance to using correlation. Only the edge from 5 to 8 being almost

as strong as 9 to 10 could be a problem i.e. choosing a threshold a little larger than $t = 1.7 \cdot 10^{-2}$ (which is quite small and has been used for Figure 4.14c) would have resulted in an edge from X_5 to X_8 . Hence, in a real world example we might have chosen to either leave out both edges which depending on the scenario may or may not be an acceptable error or include the both of them.

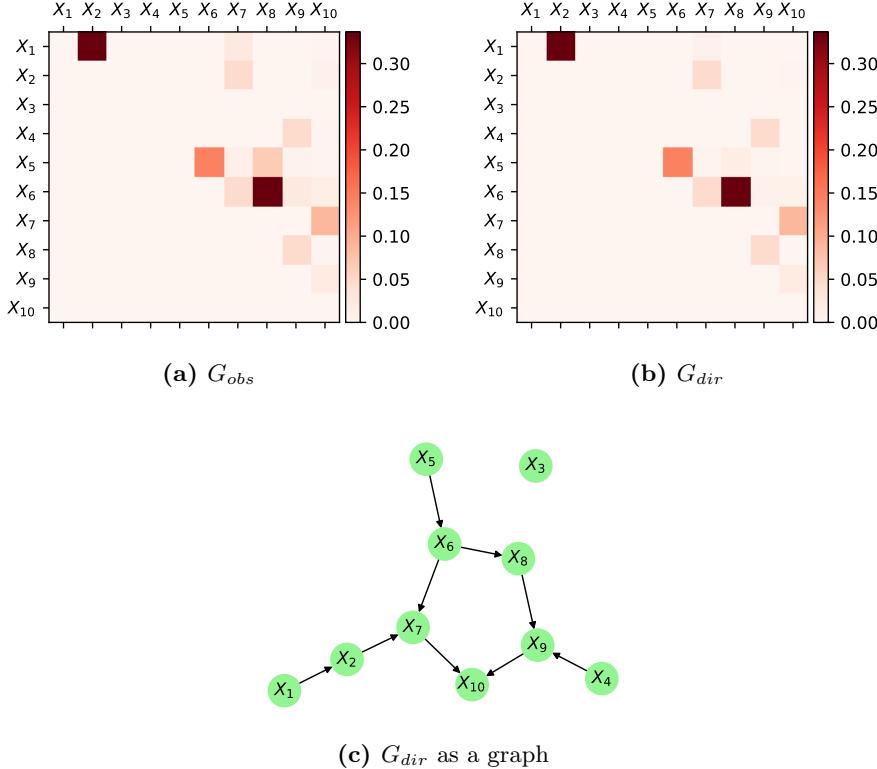


Figure 4.14: Using mutual information instead of correlation results in G_{obs} shown in (a). The non-linear map from correlation to mutual information only effects the resulting G_{dir} a little as shown in (b) when comparing to the $\vec{\rho}_{i,j}$ from Equation 4.7. Choosing the relatively small threshold $t = 1.7 \cdot 10^{-2}$ results in a perfect reconstruction of the graph structure.

Furthermore, using a symmetric G_{obs} instead i.e. no assumption on topology, does not seem to have much of an effect as seen from Figure 4.15. Although there still is a small weight on the edge from X_5 to X_8 , by choosing the threshold $t = 1.96 \cdot 10^{-2}$ we can accurately construct the true network structure.

In conclusion, we observe a useful property of more general networks that for

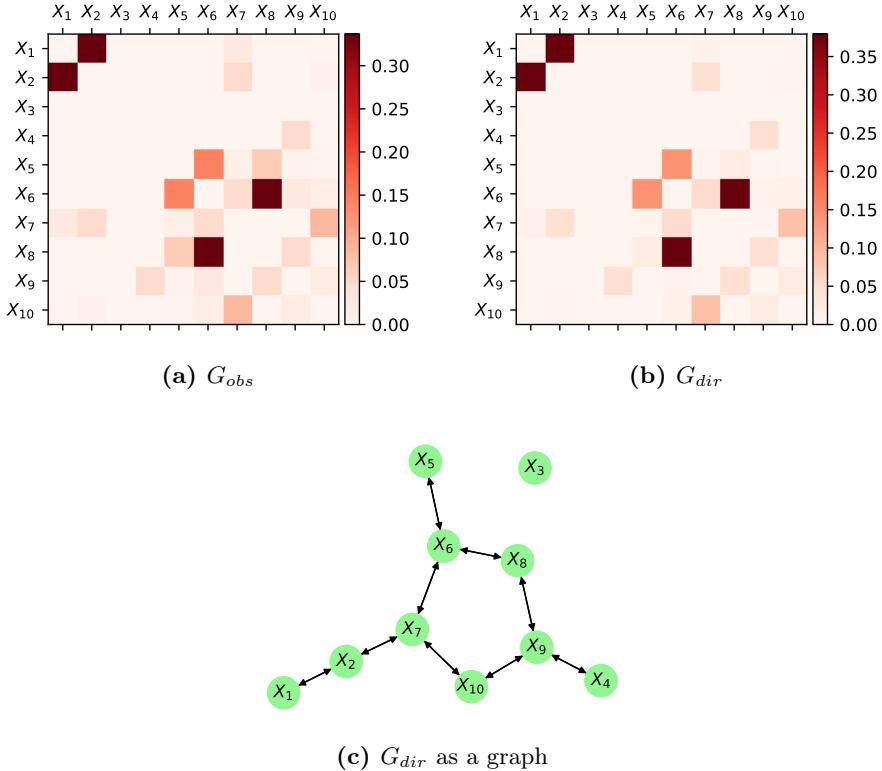


Figure 4.15: Using a symmetric G_{obs} instead of an upper triangular G_{obs} result in near identical G_{dir} in terms of relative weights on the edges. Namely, the G_{dir} shown in (b) seem to be almost a scaled version of the (reflected) G_{dir} derived from a triangular G_{obs} . Thus, as (c) also shows, we can accurately infer the structure of the network using a threshold $t = 1.96 \cdot 10^{-2}$.

both mutual information and correlation, the additional assumption of the topological order does not have much of an effect in these cases contrary to what we observed for Gaussian chains and linear chain models in general, when using correlation.

4.3 CE computation

Having discussed the strengths and weaknesses of Algorithm 2, we now turn our attention to Algorithm 1. Namely, in this section we shall discuss the different methods from Section 3.4 and how they perform on two examples. Once again, we shall base our results on two examples. The first is a simple case, where we shall see what to be aware of when initially the observations are transformed through estimated distribution functions as well how accurately the different methods estimate the Copula entropy i.e. mutual information. Continuing from the first example, we shall once again consider the network from Section 4.2 specified by Equation 4.7. In particular, we will see how well combined framework performs on an example we have already seen to be quite solvable if one uses accurate estimates of the mutual information which previously we calculated theoretically.

4.3.1 Exponentiated multivariate Gaussian

Let us consider a simple case with $\mathbf{Y} = e^{\mathbf{X}}$ (element wise exponentiation) where $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0.9\sigma_1\sigma_2 & 0 \\ 0.9\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} = \text{diag}(\boldsymbol{\sigma}) \begin{bmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{diag}(\boldsymbol{\sigma})$$

In particular, in terms of Equation 4.4, we have that for \mathbf{X} , $\vec{p}_{1,2} = 0.9$. It is clear that to Algorithm 1, the mean of \mathbf{X} is of no importance as it simply corresponds to a scaling of the Y_i variables. Furthermore, because of Corollary 3.2.1, theoretically, due to the uniqueness of the Copula C (as \mathbf{Y} is continuous) we should expect near equal or very similar results for \mathbf{Y} and \mathbf{X} from Algorithm 1. Additionally, different $\boldsymbol{\sigma}$ corresponds to different scaling of \mathbf{X} , and thus we should observe equal or near equal G_{dir} for all \mathbf{Y} independently of $\boldsymbol{\sigma}$. Initially, we shall see how this hypothesis holds up when considering the following three examples

$$\boldsymbol{\sigma} = (0.07, 0.3, 0.9), \quad \boldsymbol{\sigma} = (1, 1, 1), \quad \boldsymbol{\sigma} = (1, 2, 3)$$

To draw from this distribution, one can either use built-in functions or use the Cholesky factorization of the correlation matrix to generate proper correlated variables from 3 independent standard normal distributions and then scale with the chosen standard deviation to generate samples from all three cases based on the same seed. We shall do the latter and also generate a generous number of samples (10,000) such that the KDE based methods have the best possible prerequisites whilst also being numerically tractable later on.

In order for the sample size to not influence the results, we simulate a generous number of samples, namely, for the following results we have used $n = 10,000$ samples. For $\sigma = (1, 1, 1)$, Algorithm 1 and Algorithm 2 returns the following (using $\alpha = 1$ and $\beta = 0.99$)

$$G_{dir} = \begin{bmatrix} -0.33396 & 0.6660 & 0.02512 \\ 0.6660 & -0.3341 & 0.02730 \\ 0.02512 & 0.02730 & -0.0020583 \end{bmatrix} \quad (4.8)$$

Similarly, for $\sigma = (0.07, 0.3, 0.9)$:

$$G_{dir} = \begin{bmatrix} -0.3335 & 0.6665 & 0.01414 \\ 0.6665 & -0.3335 & 0.01418 \\ 0.01414 & 0.01418 & -0.00060124 \end{bmatrix} \quad (4.9)$$

Finally, for $\sigma = (1, 2, 3)$:

$$G_{dir} = \begin{bmatrix} -0.1490 & 0.09535 & 0.3599 \\ 0.09535 & -0.2989 & 0.5831 \\ 0.3599 & 0.5831 & -0.4037 \end{bmatrix}$$

For $\sigma = (1, 1, 1)$ and $\sigma = (0.07, 0.3, 0.9)$ we observe the most resemblance to the Σ , although the resulting G_{dir} deviate in the final column. The difference is likely produced by Algorithm 1 as if the resulting G_{obs} was the same, then so would G_{dir} and from the above argument, we know that theoretically this should be the case. For the final example, $\sigma = (1, 2, 3)$, we see a completely different result and immediately suspect that there must be some numerical errors. Investigating the partial results of Algorithm 1 we immediately see a flaw in the supposedly uniform variables U_i as shown in figure Figure 4.16

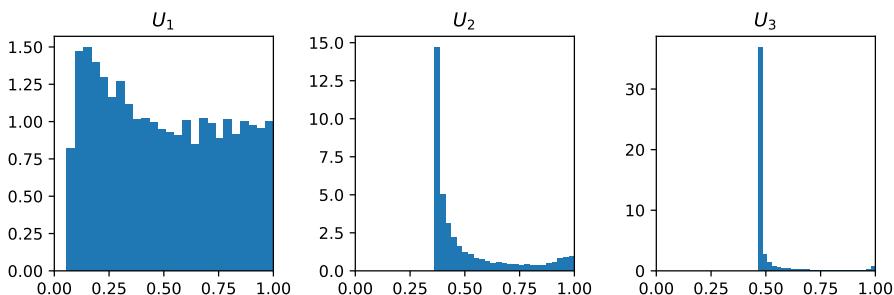


Figure 4.16: The samples transformed using $U_i = F_i(X_i)$ for $\sigma = (1, 2, 3)$. These should be uniformly distributed, but clearly this is not the case for U_2 and U_3 . Even U_1 does not quite resemble 10,000 samples from a uniform distribution.

	U_1	U_2	U_3
D_n	0.066209	0.36014	0.46285
p-value	0	0	0

Table 4.1: based on 10,000 samples for $\sigma = (1, 2, 3)$.

Before handling this, the non-uniformity of U_1 in Figure 4.16 is likely also present in the case when $\sigma = (1, 1, 1)$. Indeed, Figure 4.17 shows that this is indeed the case.

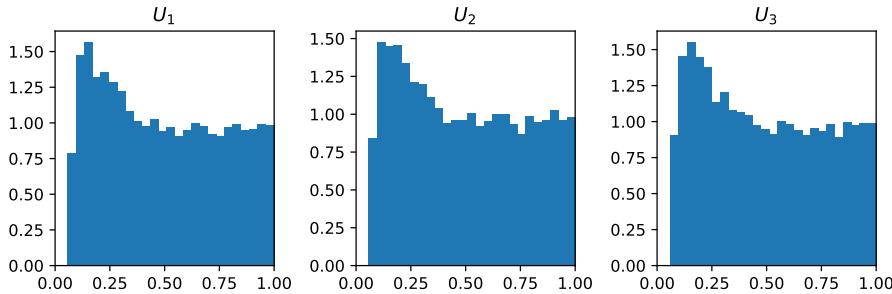


Figure 4.17: The samples transformed using $U_i = F_i(X_i)$ for $\sigma = (1, 1, 1)$.

	U_1	U_2	U_3
D_n	0.068408	0.066808	0.070809
p-value	0	0	0

Table 4.2: based on 10,000 samples for $\sigma = (1, 1, 1)$.

Finally, just to be sure, $\sigma = (0.07, 0.3, 0.9)$ is also shown in Figure 4.18 and seems very reasonable, except for U_3 .

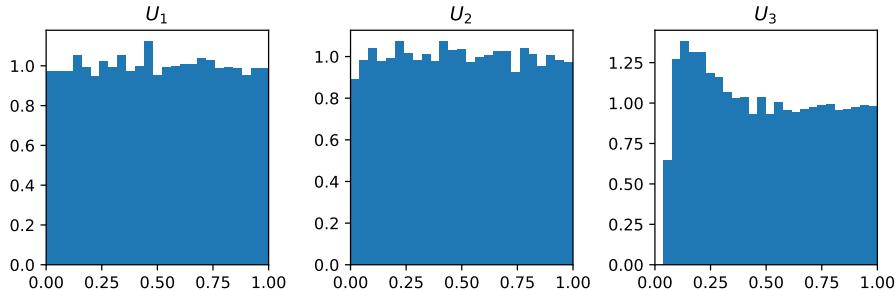


Figure 4.18: The samples transformed using $U_i = F_i(X_i)$ for $\sigma = (0.07, 0.3, 0.9)$.

	U_1	U_2	U_3
D_n	0.00581897	0.0068066	0.050908
p-value	0.88645	0.74179	0

Table 4.3: based on 10,000 samples for $\sigma = (0.07, 0.3, 0.9)$.

From the above examples, it seems that the larger the variance, the worse the uniforms turn out. Reasons for this could include numerical issues when trying to calculate $u_i^{(j)}$ from $y_i^{(j)}$ by $u_i^{(j)} = \int_{-\infty}^{y_i^{(j)}} f_i(y) dy$ and bad fitting of the kernel density estimate from observations. In particular, for values similar, which happens in the case for large σ such that we observe large negative realizations of X_i , $y_i^{(j)}$ are almost 0, and when computing the integral could result in identical values. Furthermore, from Figure 4.19 we see that indeed the fit is quite poor. Note that we have zoomed in on the interval $[-200, 200]$ which contains 96.2% of observations. The poor fit is primarily due to the use of Scott's Rule **as discussed above** which in this case overshoots the optimal bandwidth by a lot.

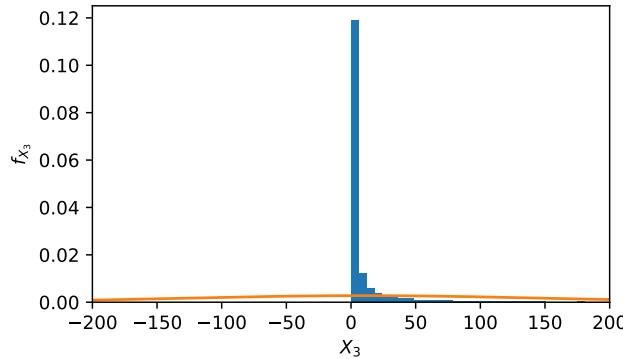


Figure 4.19

The poor fit also explains the high concentration of U_3 around 0.5 in Figure 4.16 as only 54.5% of the probability mass lies above 0.

However, also here Corollary 3.2.1 proves to be useful. Namely, we can get rid of the numerical issues by transforming Y_i using e.g. $\log(\cdot)$ or $(\cdot)^p$ for $p > 0$ to get even out the observations more. As the first simply inverts the initial transformation of X_i , we choose the latter as a more interesting case. In particular, choosing $p < 1$ will result in a more even distribution. In the following, $p = 1/10$ has been used to transform \mathbf{Y} prior to running Algorithm 1 and the resulting $u_i^{(j)}$ is shown in Figure 4.20.

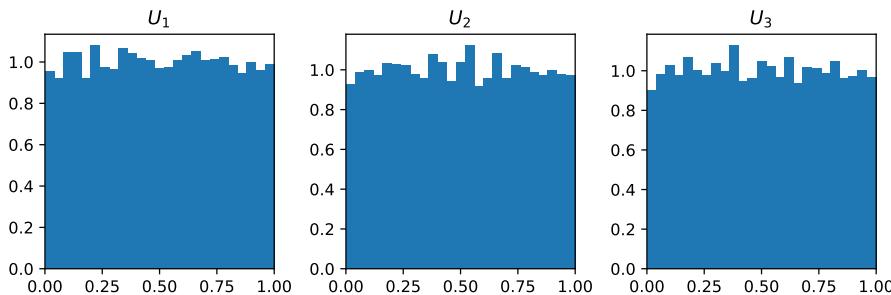


Figure 4.20

The resulting $u_i^{(j)}$ now seem to follow a uniform distribution and indeed the KDE fits much better as seen in Figure 4.21.

	U_1	U_2	U_3
D_n	0.0061099	0.0061435	0.0073148
p-value	0.84838	0.84368	0.65690

Table 4.4: based on 10,000 samples for $\sigma = (1, 2, 3)$ with power transform.

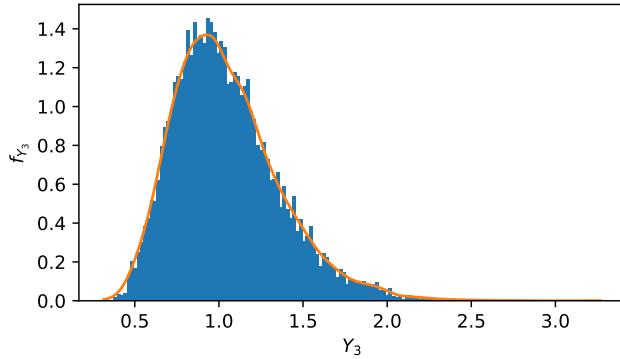


Figure 4.21

Turning to Algorithm 1 and Algorithm 2 we now find that G_{dir} is given by

$$G_{dir} = \begin{bmatrix} -0.3290 & 0.6610 & 0.008440 \\ 0.6610 & -0.3290 & 0.008150 \\ 0.008440 & 0.008150 & -0.0002061 \end{bmatrix}$$

Which is indeed much more comparable with the result from before in Equation 4.8 and Equation 4.9. The difference between G_{dir} from \mathbf{Y} and \mathbf{Y}^p is clearly visible in Figure 4.22 and also Figure 4.22b resembles the original correlation structure.

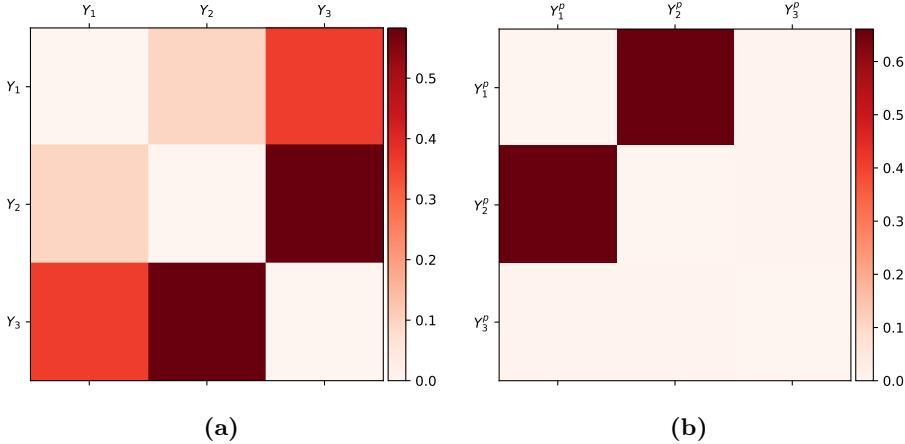


Figure 4.22: G_{dir} resulting from 10,000 samples from multi variate Gaussian with $\sigma = (1, 2, 3)$ in (a) with raw samples from \mathbf{Y} and in (b) the transformed data corresponding to \mathbf{Y}^p .

Finally, to end this example we shall compare with some theoretical results. Namely, the output G_{obs} of Algorithm 1 can also be calculated theoretically. For this, we shall use Proposition 4.1 which permits a theoretical result, namely

$$\begin{aligned} G_{obs} &= \begin{bmatrix} 0 & -\frac{1}{2} \ln(1 - \rho_{12}^2) & -\frac{1}{2} \ln(1 - \rho_{13}^2) \\ -\frac{1}{2} \ln(1 - \rho_{21}^2) & 0 & -\frac{1}{2} \ln(1 - \rho_{23}^2) \\ -\frac{1}{2} \ln(1 - \rho_{31}^2) & -\frac{1}{2} \ln(1 - \rho_{32}^2) & 0 \end{bmatrix} \\ &\cong \begin{bmatrix} 0 & 0.83037 & 0 \\ 0.83037 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Similarly, prior to deconvolution, using just the sampled \mathbf{X} (i.e. no exponential transform), Algorithm 1 returns

$$G_{obs} = \begin{bmatrix} 0. & 0.71841756 & 0.01781815 \\ 0.71841756 & 0. & 0.01769672 \\ 0.01781815 & 0.01769672 & 0. \end{bmatrix}$$

Test om denne G er lige den teoretiske. Eller nærmere, argumenter for hvorfor vi ikke laver en test, eller hvad man kunne gøre. Har samplet fra en simultan normalfordeling, så kan lave en til en mellem MI og korrelation.

From the confidence density for the correlation ρ given the emperical correlation

r is given by

$$f(\rho | r, \nu) = \frac{\nu(\nu - 1)\Gamma(\nu - 1)}{\sqrt{2\pi}\Gamma(\nu + \frac{1}{2})} \frac{(1 - r^2)^{\frac{\nu-1}{2}} (1 - \rho^2)^{\frac{\nu-2}{2}}}{(1 - r\rho)^{\frac{2\nu-1}{2}}} F\left(\frac{3}{2}, -\frac{1}{2}, \nu + \frac{1}{2}, \frac{1 + r\rho}{2}\right)$$

from the mutual information, we can calculate the absolute correlation. Notice that the density does not change when reversing both r and ρ simultaneously, thus, without loss of generality, assume $r \geq 0$, then we can calculate a CI for ρ (which will be negated if we had used $-r$ instead and thus would be identical when taking the absolute value). If the original CI $[a, b]$ contains 0 i.e. $a < 0$, we shall write the CI for the absolute correlation as $[0, b]$ instead. This way, we can compare the absolute correlations and see if they are the same (by checking if the CI contains the theoretical correlation) by [21]. Using numerical integration (fast enough with high numerical accuracy from many bins, 1 mil bins, yielding probability mass 1.0000000000008133), can compute CI for absolute correlation

Section 6.2

Clearly these are not equal, but in this case, the error is suspected to originate from the estimated joint density. For example, considering X_1 and X_2 , we compare the estimated joint copula density and compare to the theoretical reference til et sted hvor gausisk copula står shown in Figure 4.23 and Figure 4.24 respectively.

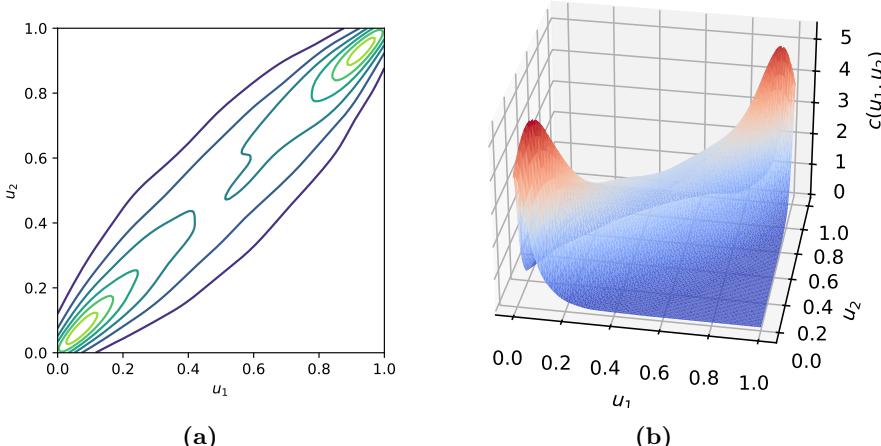


Figure 4.23: Estimated copula density c with $\rho = 0.9$ corresponding to X_1 and X_2 .

The noticeable difference is in the corners $(0, 0)$ and $(1, 1)$ where the theoretical copula density tends to infinity whereas the estimated density has modes at $(0.1, 0.1)$ and $(0.9, 0.9)$. In particular, simply rescaling the copula density in Algorithm 1 does not resemble the theoretical boundary which is a known issue reference til artikel om undershoot peaks og boundary conditions for KDE. A better approach may be to use jackknifing link til afsnit of jackknifing, som også indeholder reference til artikel hvor dette gøres.

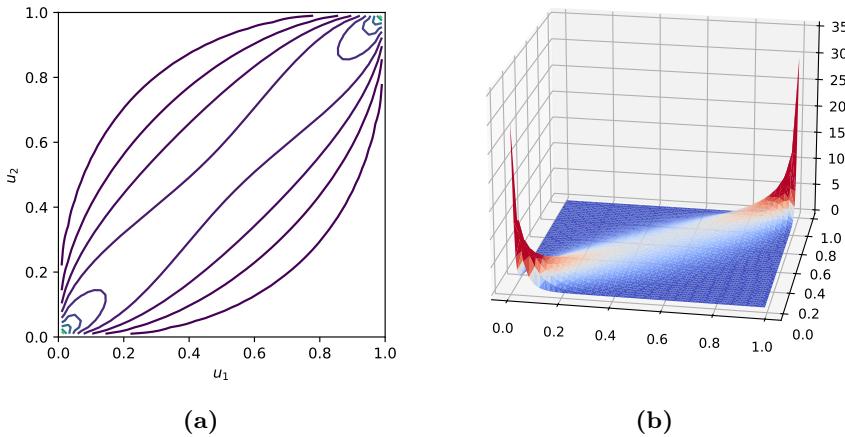


Figure 4.24: Theoretical copula density c with $\rho = 0.9$ corresponding to X_1 and X_2 .

We note however, that the underlying structure is still captured i.e. that Y_1 and Y_2 covary while Y_3 does not inform Y_1 or Y_2 and vice versa.

We continue with a similar example to the previous one. The key difference is the number of variables and a more complicated correlation structure to test the algorithms further.

Example 4.1. From Subsection 4.3.1 we saw how one could handle some numerical issues. Thus, in this example we shall not bother ourselves with such computations and merely focus on the correlation structure. In particular, we shall sample \mathbf{X} from a 10 dimensional

4.4 sammenligning af metoder for at finde MI

Sammenligning af gammel metode og "min"

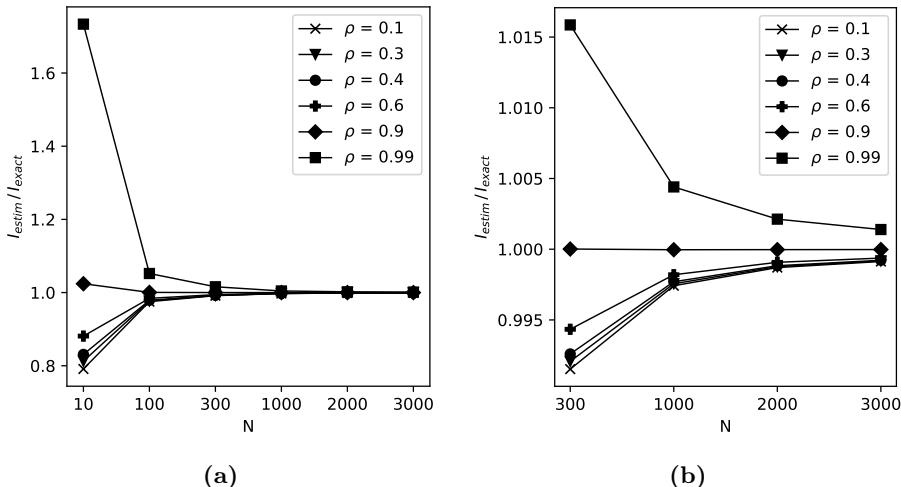


Figure 4.25: Evaluation of MI for new method for different N. **Bør sammenlignes med artielk fundet (har sat i bibtex) og original papers (ikke Kina)**

ved høj korrelation i.e. tæt på laver dimensionel manifold, skal der bruges mange, som i rigtig mange samples i mesh.

Inkluder flere exemplarer end blot gaussian as done by [18]

4.4.1 10D gaussian example

casuality svarer til at lave nedre/øvre trekant. Er der forskel i at gør edet før og efter for en symmetrisk matrix? - Ja, men begge metoder på 10 eksempel giver gode resultater. Kommenter at det er matematisk meget forskelligt at filtrere først og så ND efter og omvendt

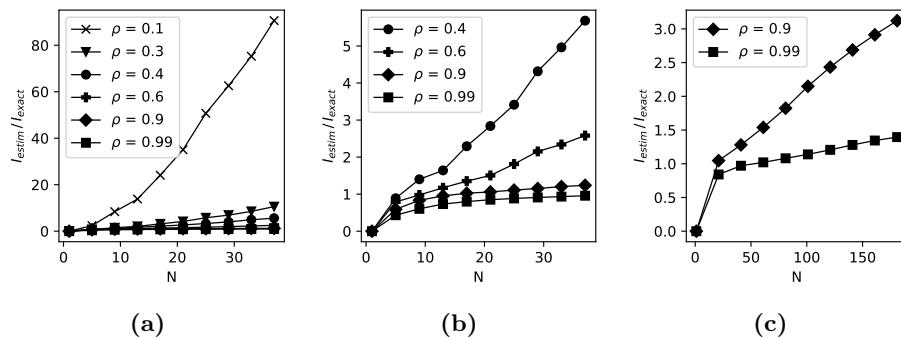


Figure 4.26: Evaluation of MI for old method for different N. Ligner der er knæk ved forholdet lig 1. Men ved næremere undersøgelse blev det fundet ud af at det ikke helt er tilfældet, og derudover vil der skulle laves en algoritmisk måde at finde dette knæk på. Savitzky–Golay filter kunne være en mulighed, eller gruppere e.g. 5 forskellige bins og tag gennemsnit. Efter smoothing kan anden afledte tæt på 0 bruges, til at finde hvornår stykket bliver fladt (tilnærmedes vist)

CHAPTER 5

Conclusion

CHAPTER 6

Appendix

6.1 Suicide data

1	25	40	83	123	256
1	27	49	84	126	257
1	27	49	84	129	311
5	30	54	84	134	314
7	30	56	90	144	322
8	31	56	91	147	369
8	31	62	92	153	415
13	32	63	93	163	573
14	34	65	93	167	609
14	35	65	103	175	640
17	36	67	103	228	737
18	37	75	111	231	
21	38	76	112	235	
21	39	79	119	242	
22	39	82	122	256	

Table 6.1: The length of treatment of control patients in suicide study. The data originates from the Mental Health Enquiry (MHE) of England of Wales and was published in 1967.

6.2 Confidence interval for absolute correlation in bivariate Gaussian

From [21], given a bivariate Gaussian, the confidence distribution of ρ given the empirical correlation r based on n observations is given by

$$f(\rho | r, \nu) = \frac{\nu(\nu - 1)\Gamma(\nu - 1)}{\sqrt{2\pi}\Gamma(\nu + \frac{1}{2})} \frac{(1 - r^2)^{\frac{\nu-1}{2}} (1 - \rho^2)^{\frac{\nu-2}{2}}}{(1 - r\rho)^{\frac{2\nu-1}{2}}} F\left(\frac{3}{2}, -\frac{1}{2}, \nu + \frac{1}{2}, \frac{1 + r\rho}{2}\right)$$

where $F(a, b, c, z)$ is the Gaussian hypergeometric function and $\nu = n - 1$. That is, given a sample correlation r , what is the confidence in ρ in terms of a distribution. In the following figure, a sample correlation $r = 0.8$ and $r = 0$ has been used with varying number of observations (degrees of freedom) in figures Figure 6.1a and Figure 6.1b respectively. A key property is that f is *even*

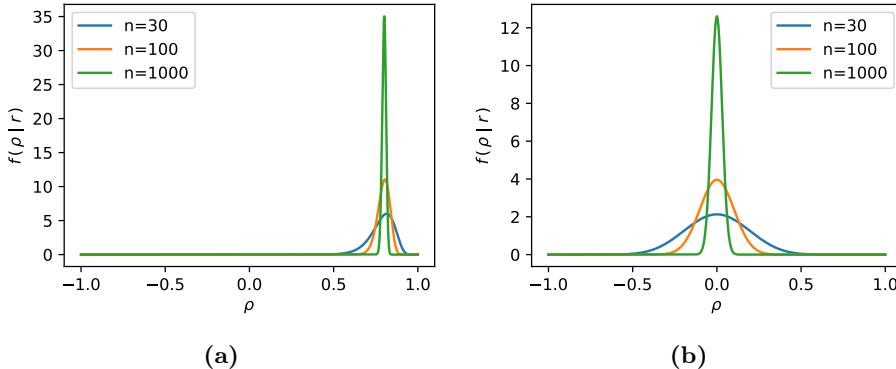


Figure 6.1: $f(\rho | r, \nu)$ shown for $r = 0.8$ and $r = 0$ in (a) and (b) with $n \in \{30, 100, 1000\}$. As one would expect, the power i.e. the width of the peak decreases with increasing n and for correlations closer to 0, the width is the largest.

symmetric in ρ, r . That is $f(\rho | r) = f(-\rho | -r)$. Thus, a confidence interval for ρ given r is the negative of the confidence interval given $-r$. In particular, if we only observe $|r|$, we can calculate a confidence interval for ρ up to the sign of the bounds of the interval. Furthermore, as we want a CI for $|\rho|$, it does not matter if r is negative or positive. Hence, without loss of generality, we assume that $r \geq 0$. At this point, to construct a confidence interval for $|\rho|$ we list the following desired properties. Firstly, it should be an exact confidence interval, meaning that for a given significance level α , the CI includes the true value exactly $1 - \alpha$ fraction of the times. Secondly, if for a given r , it can not be rejected that ρ is 0, 0 should also be contained in the interval. Finally, if we reject

that $\rho = 0$, we shall have $\alpha/2$ probability mass above and below the bounds of the interval. The above is enough to uniquely define a confidence interval in all cases. Before continuing with how this CI is calculated, we mention that as r is an unbiased estimator of ρ , we would preferably want $|r| \in CI_{1-\alpha}(|\rho|)$ (where $CI_{1-\alpha}(|\rho|)$ denotes the $1 - \alpha$ confidence interval for $|\rho|$). However, although this will in almost every scenario be the case, we can not be sure of this from the above properties and in fact examples with large α can be constructed such that $|r|$ lies just outside the constructed CI.

First, to conform with the second desired property, if it can not be rejected that $\rho = 0$ on a significance level α , we will initially compute a CI for ρ (not $|\rho|$) based on r (wlog chosen to be non-negative). This CI will just be a symmetric CI in the sense that $\alpha/2$ of the probability mass will lie below the lower bound of the CI and above the upper bound of the CI respectively. If 0 is contained in this CI, we can not reject that $\rho = 0$ and vice versa on an α significance level. Thus, if 0 is contained in this initial CI for ρ , we will start the CI for $|\rho|$ at 0 and determine and upper bound b such that α probability mass is above this b . Otherwise, we shall find a and b such that $\alpha/2$ probability mass is below a and above b respectively. Choosing a and b this way also conforms with the third property. Finally, to ensure that the CI contains exactly $1 - \alpha$, we define \tilde{f} as the reflected f in ρ such that

$$\tilde{f}(\rho_a | r_a, \nu) = f(\rho_a | r_a, \nu) + f(-\rho_a | r_a, \nu), \quad \rho_a, r \in [0, 1]$$

where ρ_a and r_a is the absolute correlation and empirical correlation respectively. With this \tilde{f} , the density at ρ_a is both the density for the negative and positive correlation ensuring that the \tilde{f} has probability mass 1. Thus, if $a = 0$ (i.e. the CI must contain 0), we find b as the $1 - \alpha$ percentile of \tilde{f} and if $a \neq 0$, we take a as the $\alpha/2$ percentile and b as the $1 - \alpha/2$ percentile of \tilde{f} .

As an example, suppose $r_a = 0.06$ with 1000 observations. Then a 95% CI for $|\rho|$ is $[0, 0.11164]$ whereas if one had observed $r_a = 0.07$ the CI would be $[0.01071, 0.1314]$. These CI could then be used to test the absolute correlation of a bivariate Gaussian i.e. for $r_a = 0.07$ based on 1000 observations would be rejected as stemming from a Gaussian with absolute correlation 0.01 on a 5% significance level.

6.3 Gaussian chain deconvolution

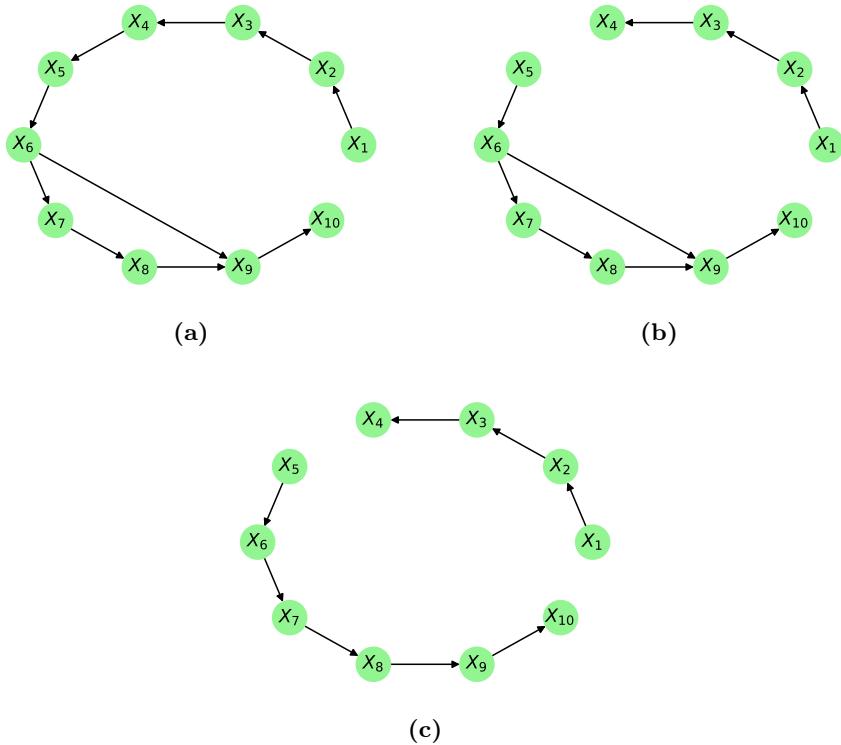


Figure 6.2: Triangular, mutual information, cutoff $2 \cdot 10^{-10}$, $2.1 \cdot 10^{-2}$ and $4.51 \cdot 10^{-2}$.

6.4 Gaussian network deconvolution

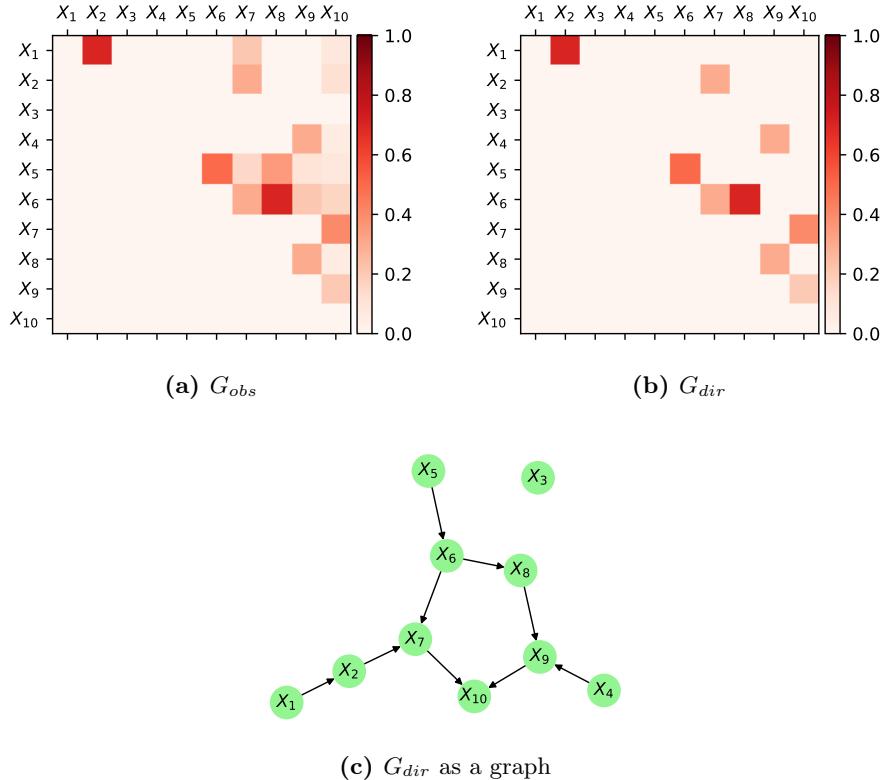


Figure 6.3: For the linear network defined in Equation 4.7, using a triangular G_{obs} (a) with the true topological structure we are able to perfectly rediscover the causal structure as seen in (b) and (c).

Bibliography

- [1] M. L. C. Vicente, “A benchmark model to generate batch process data for machine learning testing and comparison,” 2021.
- [2] S. Feizi, D. Marbach, M. Médard, and M. Kellis, “Network deconvolution as a general method to distinguish direct dependencies in networks,” *Nature biotechnology*, vol. 31, pp. 726 – 733, 2013.
- [3] D. Kurowicka and H. Joe, *Dependence Modeling: Vine Copula Handbook*. 12 2010.
- [4] G. Geenens, “Copula modeling for discrete random vectors,” *Dependence Modeling*, vol. 8, pp. 417–440, 12 2020.
- [5] Y.-N. Sun, W. Qin, and Z. Zhuang, “Nonparametric-copula-entropy and network deconvolution method for causal discovery in complex manufacturing systems,” *Journal of Intelligent Manufacturing*, vol. 33, 08 2022.
- [6] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2 ed., 2012.
- [7] H. V. Henderson and S. R. Searle, “On deriving the inverse of a sum of matrices,” *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.
- [8] G. H. Golub and C. F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, third ed., 1996.
- [9] C. d. Boor, *A Practical Guide to Splines*. New York: Springer Verlag, 1978.
- [10] J. O. Ramsay, “Monotone regression splines in action,” *Statistical Science*, vol. 3, no. 4, pp. 425–441, 1988.

- [11] D. W. Scott, “Multivariate density estimation: Theory, practice, and visualization,” in *Wiley Series in Probability and Statistics*, 1992.
- [12] J. B. Copas and M. J. Fryer, “Density estimation and suicide risks in psychiatric treatment,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 143, no. 2, pp. 167–176, 1980.
- [13] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.
- [14] M. C. Jones and T. Buch-Kromann, “Simple boundary correction for kernel density estimation,” *Statistics and Computing*, vol. 3, pp. 135–146, 1993.
- [15] J. Dai and S. Sperlich, “Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation,” *Computational Statistics & Data Analysis*, vol. 54, pp. 2487–2497, 08 2010.
- [16] M. Jones and P. Foster, “A simple nonnegative boundary correction method for kernel density estimation,” *Statistica Sinica*, vol. 6, 01 1996.
- [17] M. C. Jones and D. A. Henderson, “Kernel-type density estimation on the unit interval,” *Biometrika*, vol. 94, no. 4, pp. 977–984, 2007.
- [18] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.
- [19] Z. Botev, “A novel nonparametric density estimator,” *The University of Queensland, Tech Rep*, 01 2006.
- [20] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, “Kernel density estimation via diffusion,” *The Annals of Statistics*, vol. 38, no. 5, pp. 2916 – 2957, 2010.
- [21] G. Taraldsen, “Confidence in correlation,” 11 2020.