Chapter 1

Results

en intro til hvad der kommer til at være resultater ift.

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 graphisce 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

1.1 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 G_{obs} based on the above assumption of how similarities are *convolved* which we equate with the noise N from Subsection 1.3.3, 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 1.1. *Given a bivariate normal distribution* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *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 $I(X_1, X_2) = -\frac{1}{2}\ln\left(1-\rho^2\right)$.

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

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 $\boldsymbol{X}$ be a d-dimensional Gaussian random vector, the $\boldsymbol{X}$ 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 $\boldsymbol{X}$ that permits the following structure.

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= \rho_{1,2}X_1 + \sqrt{1-\rho_{1,2}^2}Z_2 \\ X_3 &= \rho_{2,3}X_2 + \sqrt{1-\rho_{2,3}^2}Z_3 \\ &\vdots \\ X_d &= \rho_{d-1,d}X_{d-1} + \sqrt{1-\rho_{d-1,d}^2}Z_d \end{aligned} \tag{1.1}$$

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

factorization of the correlation matrix for $\boldsymbol{X}$, namely

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

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

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

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 1.1 the random variable $\boldsymbol{X}$ 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. $f\left(X_i \mid X_{i-1}, X_{i-1}, \ldots, X_1\right) = f\left(X_i \mid X_{i-1}\right)$. Thus, if the algorithm works as intended, we should observe that the deconvolved network is a *chain* of variables as shown in the Figure 1.1 Thus, we now have the expected result, and we

Figure 1.1: 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

1.1.1 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 1.5 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

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

Now, let G_{dir} be given as follows

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

then G_{dir}^2 is given by

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

It is not hard to show that in fact $\sum_{k\geq} G_{dir}^k = \sum_{k=1}^{d} G_{dir}^k = G_{obs}$. 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 1.1. 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 1.2 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 $X_2 \mid X_1 \sim \mathcal{N}\left(1-4\left(x_1-1/2\right)^2, \sigma^2\right)$ i.e. the joint distribution function is a parabola with a Gaussian noise added along the second dimension. In Figure 1.2, 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

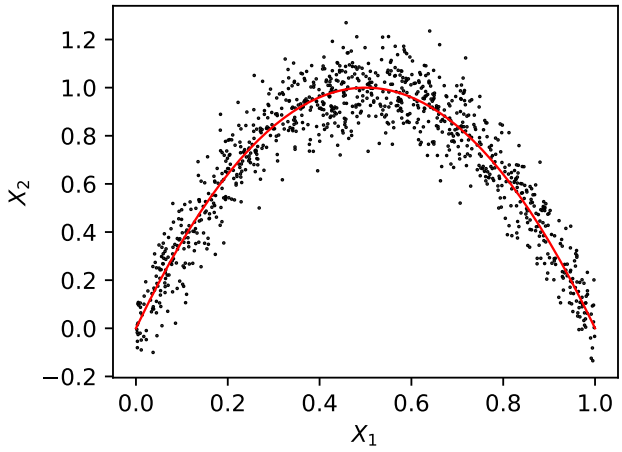


Figure 1.2: 1000 samples generated from $X_1 \sim \mathcal{U}(0,1)$ and $X_2 \mid X_1 \sim \mathcal{N}\left(1-4\left(x_1-1/2\right)^2, \sigma^2\right)$ with $\sigma = 1/10$. The mutual information is calculated theoretically to be $I(X_1, X_2) \approx 1.030$ amd repeated simulations show that the empirical correlation is symmetric around 0 supporting the claim that the underlying correlation is in fact 0

mutual information is a measures 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.

... changing alpha did not do a lot as we expect from how G obs looks. Hpwever, we can get closer by applying a larger threshhold. have chosen such that looks fairly sparse but is likely biased from my own prior knowledge

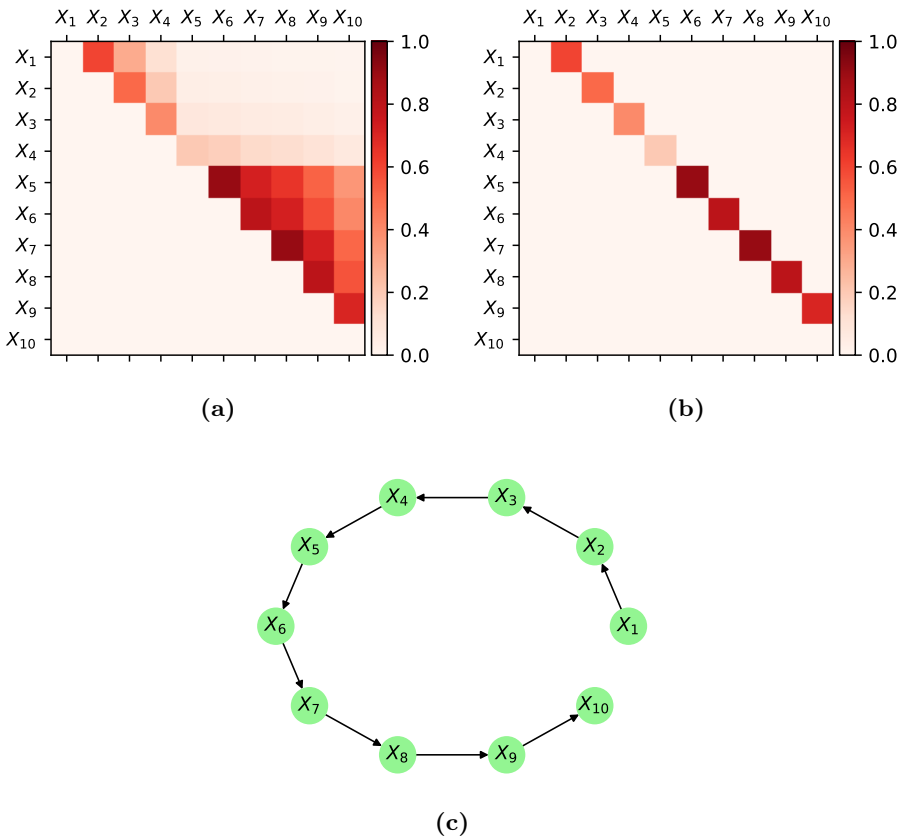


Figure 1.3: Triangular, correlation

1.1.2 Gaussian chain deconvolution using mutual information

1.2 General Gaussian graph

...correlation is positive semi definite, so always eig greater than 0

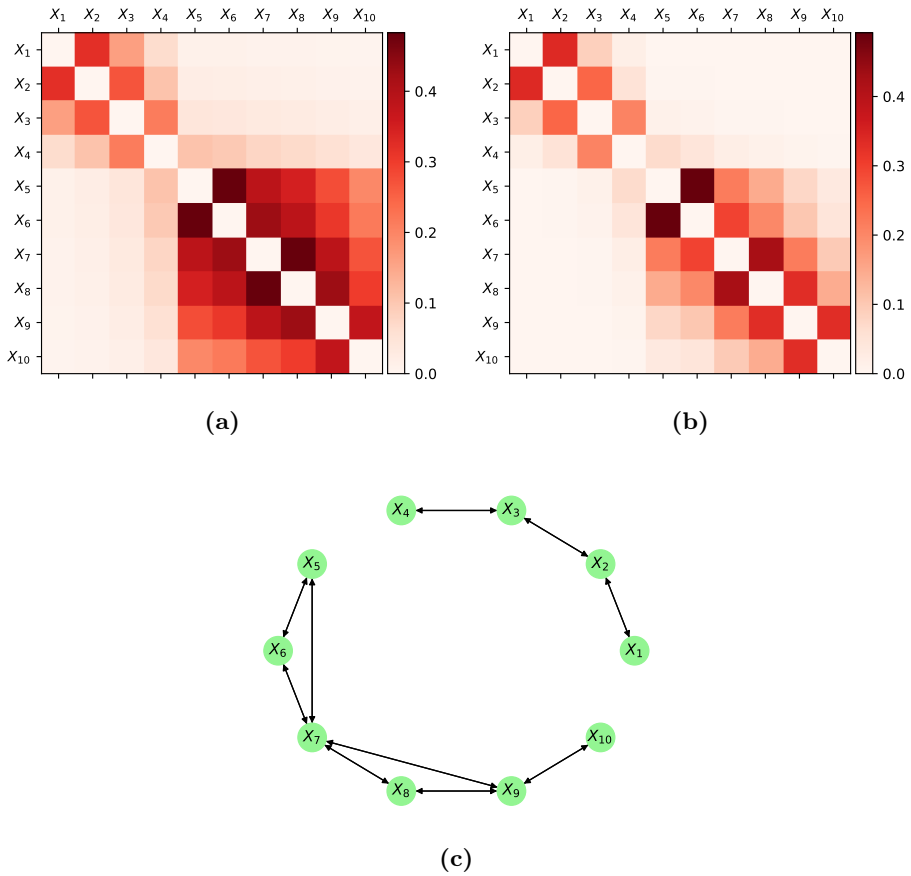


Figure 1.4: Symmetric, correlatoin

1.3 CE computation

Gaussian example Eksempel medd normal fordeling og fejlen der bliver lavet vha. algoritmen.

Example 1.1. *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*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0.9\sigma_1\sigma_2 & 0 \\ 0.9\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

It is clear that to Algorithm 1, the mean is of no importance as it simply corresponds to a scaling of the Y_i variables. Furthermore, because of Corollary 1.2.1, theoretically, due to the uniqueness of the Copula C (as $\boldsymbol{Y}$ is continuous) we should expect near equal or very similar results for $\boldsymbol{Y}$ and $\boldsymbol{X}$ from Algorithm 1. Additionally, different σ corresponds to different scaling of $\boldsymbol{X}$, and thus we should observe equal or near equal G_{dir} for all $\boldsymbol{Y}$. Initially, we shall see how this hypothesis holds up to the following three examples

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

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 $\boldsymbol{\sigma} = (1, 1, 1)$, Algorithm 1 and Algorithm 2 returns the following (using $\alpha = 1$ and $\beta = 0.99$)

$$G_{dir} = \begin{bmatrix} -0.33396 & 0.6660 & 0.02512 \\ 0.6660 & -0.3341 & 0.02730 \\ 0.02512 & 0.02730 & -0.0020583 \end{bmatrix} \tag{1.2}$$

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

$$G_{dir} = \begin{bmatrix} -0.3335 & 0.6665 & 0.01414 \\ 0.6665 & -0.3335 & 0.01418 \\ 0.01414 & 0.01418 & -0.00060124 \end{bmatrix} \tag{1.3}$$

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

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

For $\boldsymbol{\sigma} = (1, 1, 1)$ and $\boldsymbol{\sigma} = (0.07, 0.3, 0.9)$ 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, $\boldsymbol{\sigma} = (1, 2, 3)$, 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 1.5

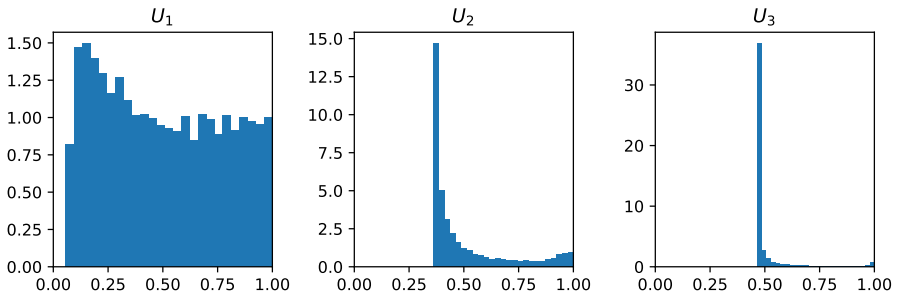


Figure 1.5: The samples transformed using $U_i = F_i(X_i)$ for $\boldsymbol{\sigma} = (1, 2, 3)$. These should be uniformly distributed, but clearly this is not the case for U_2 and U_3. Even U_1 does not quite resemble 10,000 samples from a uniform distribution.

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

Table 1.1: based on 10,000 samples for $\boldsymbol{\sigma} = (1, 2, 3)$.

Before handling this, the non-uniformity of U_1 in Figure 1.5 is likely also present in the case when $\boldsymbol{\sigma} = (1, 1, 1)$. Indeed, Figure 1.6 shows that this is indeed the case.

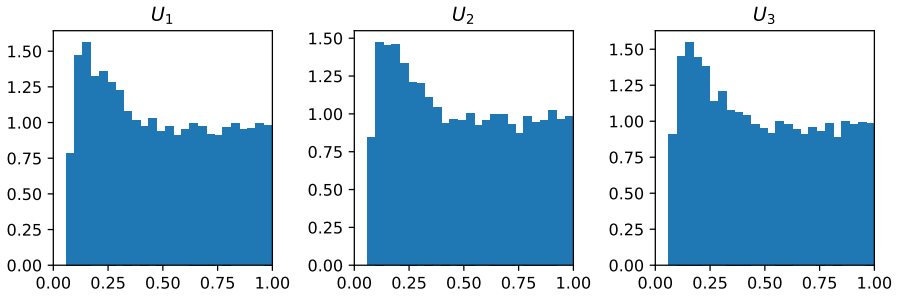


Figure 1.6: The samples transformed using $U_i = F_i(X_i)$ for $\boldsymbol{\sigma} = (1, 1, 1)$.

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

Table 1.2: based on 10,000 samples for $\boldsymbol{\sigma} = (1, 1, 1)$.

Finally, just to be sure, $\boldsymbol{\sigma} = (0.07, 0.3, 0.9)$ *is also shown in Figure 1.7 and seems very reasonable, except for* U_3*.*

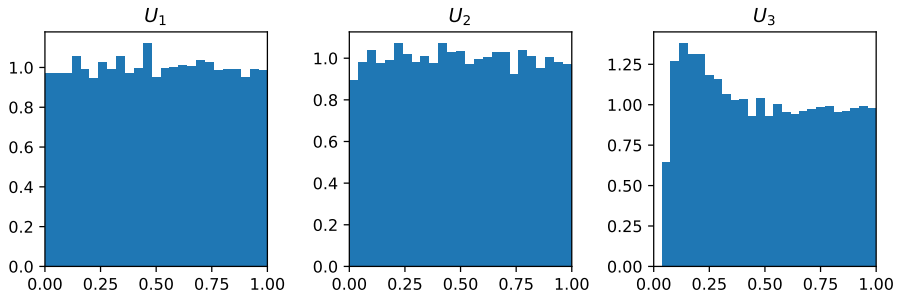


Figure 1.7: The samples transformed using $U_i = F_i(X_i)$ for $\boldsymbol{\sigma} = (0.07, 0.3, 0.9)$.

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

Table 1.3: based on 10,000 samples for $\boldsymbol{\sigma} = (0.07, 0.3, 0.9)$.

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 $u_i^{(j)}$ *form* $y_i^{(j)}$ *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 1.8 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.*

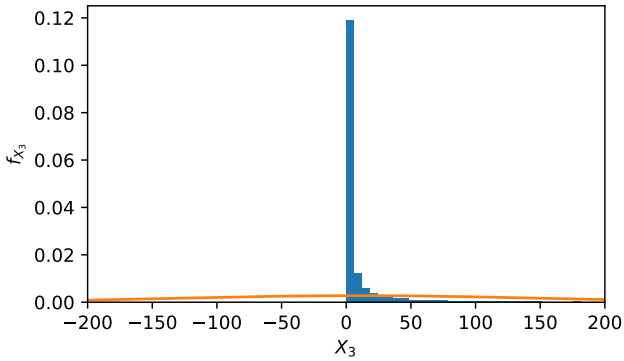


Figure 1.8

The poor fit also explains the high concentration of U_3 *around* 0.5 *in Figure 1.5 as only* 54.5% *of the probability mass lies above* 0.

However, also here Corollary 1.2.1 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* $\boldsymbol{Y}$ *prior to running Algorithm 1 and the resulting* $u_i^{(j)}$ *is shown in Figure 1.9.*

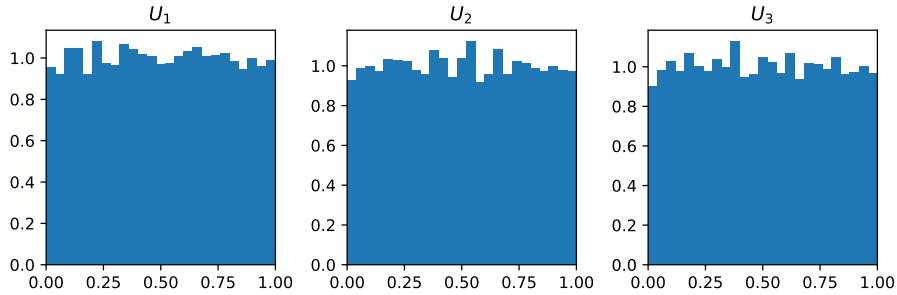


Figure 1.9

The resulting $u_i^{(j)}$ *now seem to follow a uniform distribution and indeed the KDE fits much better as seen in Figure 1.10.*

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

Table 1.4: based on 10,000 samples for $\boldsymbol{\sigma} = (1, 2, 3)$ with power transform.

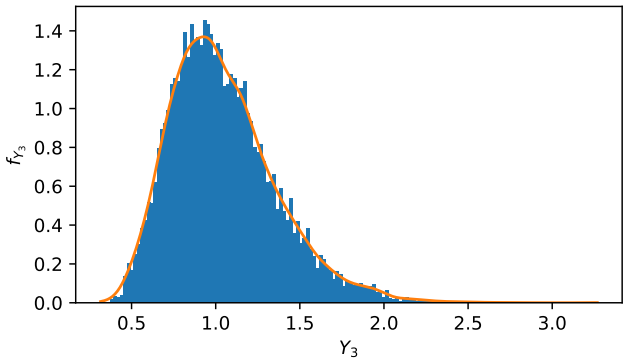
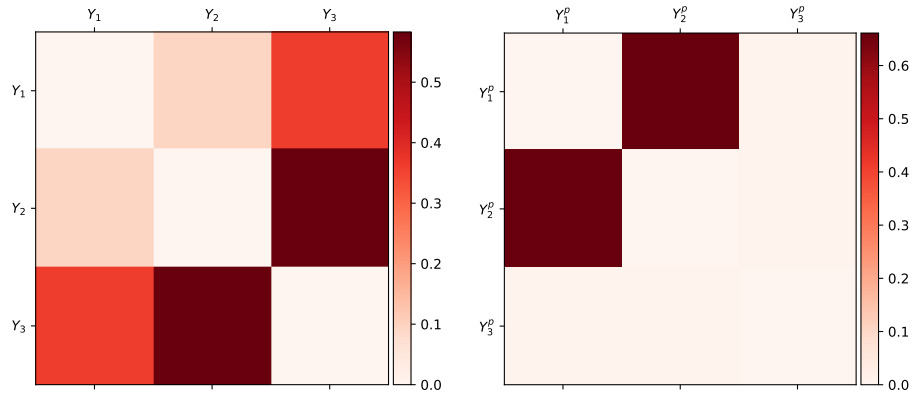


Figure 1.10

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 1.2 and Equation 1.3. The difference between G_{dir} from $\boldsymbol{Y}$ and $\boldsymbol{Y}^p$ is clearly visible in Figure 1.11 and also Figure 1.11b resembles the original correlation structure.



(a) **(b)**

Figure 1.11: G_{dir} resulting from 10,000 samples from multi variate Gaussian with $\boldsymbol{\sigma} = (1, 2, 3)$ in **(a)** with raw samples from $\boldsymbol{Y}$ and in **(b)** the transformed data corresponding to $\boldsymbol{Y}^p$.

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 1.1 which permits a theoretical result, namely

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

Similarly, prior to deconvolution, using just the sampled $\boldsymbol{X}$ (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

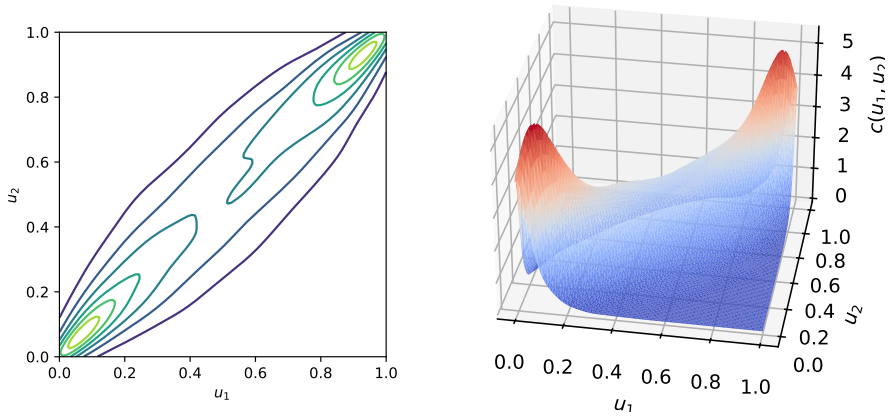
r is given by

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

from the mutual information, we can calculate the absolute correlation. Notice that the density does not change when reversing both r and ρ simultaniously, thus, wlog, 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 absoulte correlations and see if they are the same (by checking if the CI contains the theoretical correlation) by [?]. 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 1.1

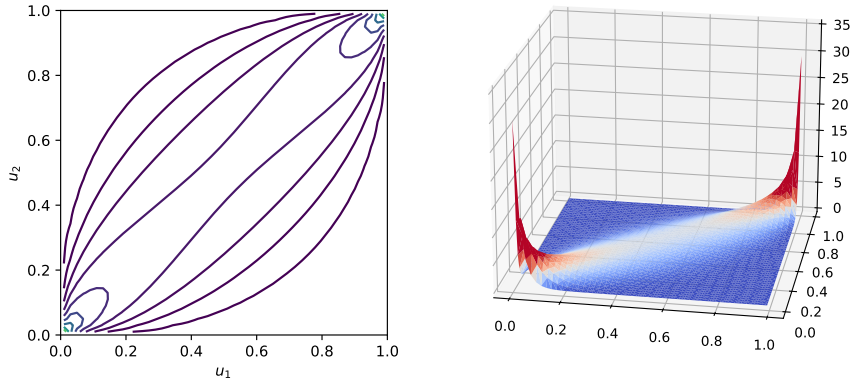
Clearly these are not equal, but in this case, the error is suspected to originate from the estimated joint density. For example, considering X_1 and X_2, we compare the estimated joint copula density and compare to the theoretical reference til et sted hvor gausisk copula står shown in Figure 1.12 and Figure 1.13 respectively.



(a) **(b)**

Figure 1.12: 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.*



(a) **(b)**

Figure 1.13: 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 1.2. *From Example 1.1 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* $\boldsymbol{X}$ *from a 10 dimensional*

1.4 sammenligning af metoder for at finde MI

Sammenligning af gammel metode og "min"

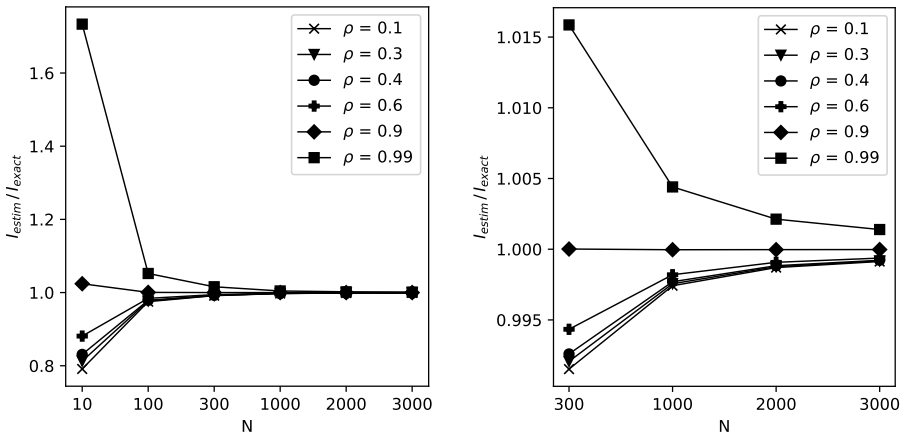


Figure 1.14: Evaluation of MI for new method for different N. Bør sammenlignes med artiekl fundet (har sat i bibtex) og original papers (ikke Kina)

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

Inkluder flere exempler end blot gaussian as done by [?]

1.4.1 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

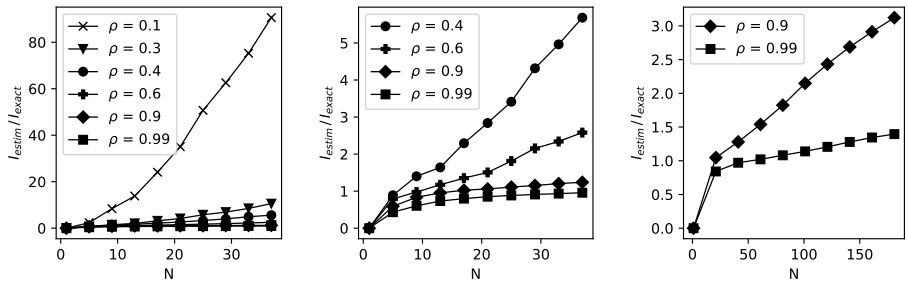


Figure 1.15: Evaluation of MI for old method for different N. Ligner der er knæk ved fporhold lig 1. Men ved næremere undersøgelse blev det fundet ud af at det ikke helt er tilfældet, og derudover vil der skulle laves en algoritmisk måde at finde dette knæk på. Savitzky–Golay filter kunne være en mulighed, 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 (tilnærmelses vist)