Chapter 1

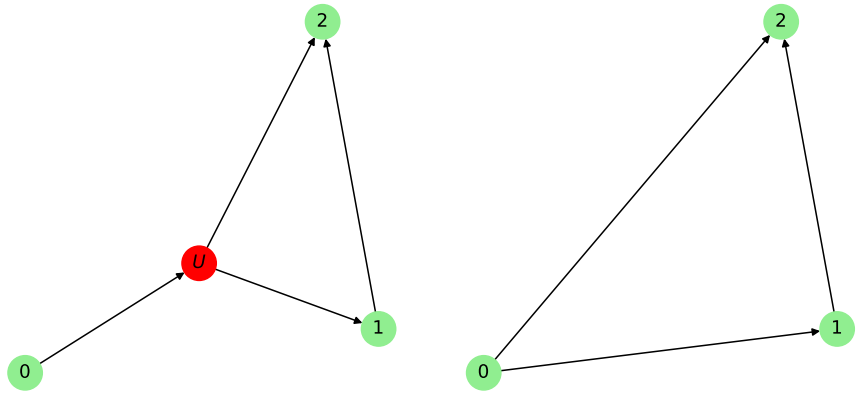
Method

General introduction to metodeafsnittet og causalitet

1.1 Causal discovery

In this section, we shall discuss the method for network deconvolution, originally proposed by [**?**]. 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 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 of 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 1.1a 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 1.1b. With these initial comments, we proceed



(a) **(b)**

Figure 1.1: (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.

1.1.1 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 an undirected graph as in the previous section. 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 [**?**] 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. 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 1.2 for a further discussion on these two measures and Section 1.3 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*.

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 1.3.2, 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). This hints to that a threshold on the inferred weights on the edges of the network might be a good idea which 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

$$G_{obs} = G_{dir} + G_{indir} \tag{1.1}$$

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

$$G_{indir} = G_{dir}^2 + G_{dir}^3 + \dots \tag{1.2}$$

i.e. that the observed *information* exchanged on an edge e_{ij} between nodes X_i and X_j is the sum 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

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

where e_{ij} is the (i, j) element of G_{dir}. 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

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir}^3 + \cdots = G_{dir} + G_{dir}\, G_{obs} \tag{1.3}$$

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. $\rho\left(G_{dir}\right) < 1$. Thus, assuming convergence we can rewrite the infinite series as

$$G_{obs} = G_{dir}\left(I - G_{dir}\right)^{-1} \tag{1.4}$$

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

$$G_{dir} = G_{obs}\left(I + G_{obs}\right)^{-1} \tag{1.5}$$

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

We note that from the above one needs $\left(I + G_{obs}\right)^{-1}$ 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 $\rho\left(G_{dir}\right) < 1$, and that $I + G_{obs}$ is thus invertible we use Equation 1.4 and simplify

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

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 1.5, 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 1.7. 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, $T = D\left(I + D^{-1}T_u\right)$ and hence, we have that

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

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*.

1.2 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.

1.2.1 Copula

Given a set of N random variables $X_1, \ldots, X_d$, 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 $F_1, \ldots, F_N$ we define a copula C as

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

Letting $u_i = F_i(x_i) \in [0,1]$ it is clear that C is a distribution function as described above [**?**]. 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 1.1 (Copula)**.** *A function $C : [0,1]^d \to [0,1]$ 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 1.2 (Sklar's theorem)**.** *For a random vector $\boldsymbol{X}$ with CDF F and univariate marginal CDFs $F_1, \ldots, F_d$. There exists a copula C such that*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{1.6}$$

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

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[**?**]. 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[**?**] 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 1.6 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 1.2.1 (Coordinate transformation)**.** *Under the assumptions of Theorem 1.2, 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 $F_{Y_i}(y_i)$

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

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

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

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

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

Proposition 1.3. *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 $(X_1, \ldots, X_d)$ if and only if it is a copula of $(T_1(X_1), \ldots, T_d(X_d))$.*

1.2.2 Mutual information and Copula entropy

In this section we introduce Copula entropy as done in [**?**] 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 1.4 (Mutual information)**.** *For a random vector $\boldsymbol{X} = \{X_i\}$, we define the mutual information as*

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

where $\mathcal{X}$ is the domain of the random vector $\boldsymbol{X}$. 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 1.2.3. For now, we proceed with the definition of (joint) entropy for both discrete and continuous random vectors.

Definition 1.5 (Entropy)**.** *The (joint) entropy of a random vector $\boldsymbol{X}$ is defined as*

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

Where, often for continuous random vectors, we denote the (differential) entropy as $h(\boldsymbol{X})$ instead.

We note the need for two separate notations of entropy as differential entropy is not the limit of (discrete) entropy in the way mutual information is. Again, this is further discussed in Subsection 1.2.3.

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

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

where d is the dimension of $\boldsymbol{X}$. 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:

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

□

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

Definition 1.7 (Copula entropy)**.** *For a continuous random vector $\boldsymbol{X}$ with a uniquely defined Copula C, and Copula density c, we define the Copula entropy CE of $\boldsymbol{X}$ as*

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

where $\boldsymbol{U}$ *has density* c*. In particular,*

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

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

Theorem 1.8 (Equality of Copula entropy)**.** *For a continuous random vector* $\boldsymbol{X}$*, the Copula entropy* CE *is equal to the negative joint mutual information of* $\boldsymbol{X}$

$$CE\left(\boldsymbol{X}\right) = -I\left(\boldsymbol{X}\right)$$

Proof. By Theorem 1.2, letting $x_i = F_i^{-1}\left(u_i\right)$, we can relate the copula density to the joint density of $\boldsymbol{X}$ and its marginals

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

It follows directly that

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

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.

1.2.3 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:

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

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 $\Delta \to 0$ 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 $\Delta \to 0$ whereas $\log \Delta \to -\infty$ 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

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

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

$$f\left(x_1^{(i)}, x_2^{(j)}\right)\Delta_1\Delta_2 = \int_{j\Delta_2}^{(j+1)\Delta_2}\int_{i\Delta_1}^{(i+1)\Delta_1} f(x_1,x_2)\,dx_1dx_2, \quad \forall i,j \in \mathbb{Z}$$

Note that clearly $\left(x_1^{(i)}, x_2^{(j)}\right)$ 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 1.6

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

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 1.4.

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

$$f\left(x_1^{(i)}, x_2^{(j)}\right)\Delta^2 = \int_{j\Delta}^{(j+1)\Delta}\int_{i\Delta}^{(i+1)\Delta} f(x_1,x_2)\,dx_1dx_2, \quad \forall i,j \in \mathbb{Z}$$

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

$$\begin{aligned} f\left(x_1^{(i)}, x_1^{(i)}\right)\Delta^2 &= \int_{i\Delta}^{(i+1)\Delta}\int_{i\Delta}^{(i+1)\Delta} f(x_1,x_2)\,dx_1dx_2 \\ &= \int_{i\Delta}^{(i+1)\Delta} f\left(x_1\right)\,dx_1 \\ &= f\left(\tilde{x}_1^{(i)}\right)\Delta \end{aligned}$$

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

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

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.

1.2.4 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 infer the correlation coefficient from a copula. 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 1.2.1, C is also a copula for $Z_i \coloneqq (X_i - \mu_i)/\sigma_i$ for $\in\in I$ where $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i = \sqrt{\operatorname{Var} X_i}$. 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.

1.3 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 1.8 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

$$G_{obs} = \begin{bmatrix} 0 & -CE_{12} & \dots & -CE_{1n} \\ -CE_{21} & 0 & \dots & -CE_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -CE_{n1} & -CE_{n2} & \dots & 0 \end{bmatrix} \tag{1.7}$$

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.

Algorithm 1 G_{obs} computation

Require: $n > 0$ ▷ Number of variables
  for $1 \leq i < j \leq n$ **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}
    $G_{ij}, G_{ji} \leftarrow -NCE_{ij}$
  end for

Finally, from the deconvoluted information matrix D_{dir} we may choose a threshold t for choosing which edges are significant. The choice of t

Algorithm 2 (ND) Network Deconvolution

Require: G_{obs} ▷ Input observational matrix
  $[G_{obs}]_{ii} \leftarrow 0,\ \forall 1 \leq i \leq N$ ▷ zero-diagonal
  $Q_p \leftarrow G_{[1-\alpha]}$
  Set G_{obs} 0, where $G_{obs} < Q_p$
  Compute eigendecomposition Q, Λ of G_{obs}
  $\lambda^+ \leftarrow \max(\lambda^{\max}, 0)$
  $\lambda^- \leftarrow -\min(\lambda^{\min}, 0)$
  $m^+ \leftarrow \frac{1-\beta}{\beta}\lambda^+$
  $m^- \leftarrow \frac{1+\beta}{\beta}\lambda^-$
  $m \leftarrow \max(m^+, m^-)$
  $\hat{\Lambda} \leftarrow \Lambda(mI + \Lambda)^{-1}$
  return $Q\hat{\Lambda}Q^T$

Remark. The $1-\alpha$ quantile, denoted by $G_{[1-alpha]}$ (of the upper triangular matrix), can be computed in many ways. As it is only used to filter the G_{obs} matrix, its precise value does not matter. Only the property α part of the observations are below $Q_{1-\alpha}$ and $1-\alpha$ are above (or equal to) $Q_{1-\alpha}$. This property is for example fulfilled by the quantile function `quantile` from `NumPy` (v. 1.26.4). Thus setting $\alpha = 1$ will result in G_{obs} retaining all entries (except for the diagonal entries).

Remark. The $\beta \in (0, 1)$ parameter serves as a sort of regularization. The algorithm above also maps the maximum absolute value of the eigenvalues (the spectral radius) to β as is also pointed out in the implementation of Source code from Nature paper at https://compbio.mit.edu/nd/index.html. The proof of this is shown below

Proof. To show that eigenvalues i.e. the diagonal elements of Λ resulting from Algorithm 2 all fall in the interval $[-\beta, \beta]$ (i.e. $\sigma\left(Q\Lambda Q^T\right) \subseteq [-\beta, \beta]$) where at least one λ is mapped to either $-\beta$ or β, first notice that clearly the resulting eigenvalues of $G_{dir} = Q\hat{\Lambda}Q^T$ are clearly given by $\frac{\lambda_i}{m+\lambda_i}$ where $(\lambda_i)_{\{1,\dots,N\}}$ are the (real) eigenvalues of G_{obs} from the definition of $\hat{\Lambda}$. We will show the above by first considering $\lambda \geq 0$ and $\lambda < 0$.

For $\lambda \geq 0$, clearly $m \geq \frac{1-\beta}{\beta}\lambda^+$, thus

$$\frac{\lambda}{m+\lambda} = \frac{1}{1+m/\lambda} \leq \frac{1}{1+\frac{\lambda^+}{\lambda}\frac{1-\beta}{\beta}} \leq \frac{1}{1+\frac{1-\beta}{\beta}} = \beta$$

where the final inequality follows from $\lambda \leq \lambda^+$. Hence $[0, \lambda^+] \to [0, \beta]$.

Furthermore, for $0 > \lambda \geq -\lambda^-$, note that also $m \geq \frac{1+\beta}{\beta}\lambda^-$. Since $\beta \in (0, 1]$, $m + \lambda \geq \frac{1+\beta}{\beta}\lambda^- + \lambda > 0$ and thus $\frac{\lambda}{m+\lambda} < 0$ which implies

$$-\frac{\lambda}{m+\lambda} \leq \frac{-\lambda}{\frac{1+\beta}{\beta}\lambda^- + \lambda} = \frac{1}{\frac{1+\beta}{\beta}\frac{\lambda^-}{-\lambda} - 1} \leq \frac{1}{\frac{1+\beta}{\beta} - 1} = \beta$$

i.e. $[-\lambda^-, 0) \to [-\beta, 0)$. This shows that indeed all the eigenvalues of G_{dir} is numerically less that or equal to β. Finally, assuming $m \neq 0$ or equivalently that $G_{obs} \neq \mathbf{0}$, either $m = \frac{1-\beta}{\beta}\lambda^+$ (and thus $\lambda^+ \neq 0$ is an eigenvalue of G_{obs}) for which the above shows that indeed λ^+ is mapped to β or $m = \frac{1+\beta}{\beta}\lambda^-$ (and hence $\lambda^- \neq 0$ and thus $-\lambda^-$ is an eigenvalue of G_{obs}) for which $-\lambda^-$ is mapped to $-\beta$. This shows that G_dir indeed has an eigenvalue which numerical value is β. □

1.3.1 Ensuring convergence and the effect of β

From Network Deconvolution - A General Method to Distinguish Direct Dependencies over Networks - Supplementary Notes and their implementation found at their webpage at https://compbio.mit.edu/nd/ there seem to be an inconsistency between code and theory. In this section, we shall thus investigate from where the discrepancy arises. Initially, from the formulation in Equation 1.3, for the right-hand side to converge, it must have spectral radius at most 1 and to ensure convergence, less than 1. In the latter case,

Noget introduktion til at vi først lige kører nogle basale definition afsted, som kommer til at blive brugt senere

Definition 1.9. *Induced Norm.*

A matrix norm $|||\cdot|||$ *is said to be induced by the vector norm* $||\cdot||$ *when*

$$|||A||| = \sup_{||x||=1} ||Ax||$$

See [**?**]

Definition 1.10. *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:*

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

1.3.2 Robustness to noise

They show that the procedure is robust noise by considering that the observed information matrix is influenced by some noise $N \in \mathbb{R}^{n\times n}$ and characterize the noise by its Euclidean norm $||N||_2 \coloneqq \sup_{||x||_2=1} ||Nx||_2$. They show that

$$||G_{dir} - \hat{G}_{dir}||_2 \leq \gamma + \mathcal{O}\left(\delta^2 + \gamma^2 + \delta\gamma\right)$$

where γ is the spectral norm of N and δ is the spectral norm of $\hat{G}_{obs}$. However, this upper bound can actually be computed when $\delta + \gamma < 1$ and diverges when $\delta + \gamma > 1$. Furthermore, the result can be generalized to other norms than the spectral norm. In particular, the Frobenius norm admits a similar upper bound on the difference. Consider any matrix norm $|||\cdot|||$ for which $|||AB||| \leq$

$|||A||| \cdot |||B|||$. It then follows that

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

Where the final equality assumes that $|||N||| + |||G_{obs}||| < 1$ and by splitting up the sum into 2 geometric series. However, we also show that the series diverge when $|||N||| + |||G_{obs}||| > 1$ which is not directly apparent as it is a difference geometric series. Namely, by the ratio test, letting $\gamma = |||N|||$ and $\delta = |||G_{obs}|||$ as by [?] supp. notes NetworkDeconvolution-AGeneralMethodtoD...

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

as $\gamma, \delta > 0$ unless they are the zero-matrix in which case the above is nonsensical from a perspective of interest. The above shows that indeed the above result diverges when $\gamma + \delta > 1$.

Thus, for the spectral norm, denoted by $|||\cdot|||_2$, one simply needs the fact that it is sub-multiplicative (Definition 1.10) which follows from the fact that the spectral norm is induced by the l_2 norm (see Definition 1.9) on $\mathbb{R}$, and hence

$$\sup_{||x||_2=1}||ABx|| \leq ||A||_2 \cdot \sup_{||x||_2=1}||Bx|| = ||A||_2\,||B||_2$$

which by definition means that $||AB||_2 \leq ||A||_2 ||B||_2$. The Frobenius norm $|||\cdot|||_F$ is also sub-multiplicative and depending on the use case may be very useful.

1.3.3 KDE methods

Det her mangler lige

1.3.4 Normal example error

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

Proposition 1.11. *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 $I(X_1, X_2) = h(X_1) + h(X_2) - h(X_1, X_2)$ □