Appendix I: Proof of Lemma 1

The proof will rely on the following claim.

Claim 1 *Let $n \geq 3$ and suppose that G is imperfect. Then, there exists $k \in \{3, ..., n\}$ such that $Var_G(x_k) \neq 1$ for almost all correlation submatrices $(\rho_{ij})_{i,j=1,...,k-1}$ (and therefore, for almost all correlation matrices $(\rho_{ij})_{i,j=1,...,n}$).*

Proof. For notational convenience, list the variables $x_1, ..., x_n$ such that $R(i) \subseteq \{1, ..., i-1\}$ for every i. This is without loss of generality, because as far as this claim is concerned, the variables x_1 and x_n do not have a fixed meaning. Consider the lowest k for which $R(k)$ is not a clique. This means that there exist two nodes $h, l \in R(k)$ that are unlinked in G, whereas for every $k' < k$ and every $h', l' \in R(k')$, h' and l' are linked in G.

Our goal is to show that $Var_G(x_k) \neq 1$ for almost all correlation submatrices $(\rho_{ij})_{i,j=1,...,k-1}$. Since none of the variables $x_{k+1}, ..., x_n$ appear in the equations for $x_1, ..., x_k$, we can ignore them and treat x_k as the terminal node in G without loss of generality, such that G is defined over the nodes $1, ..., k$, and p is defined over the variables $x_1, ..., x_k$.

Let $(\hat{\rho}_{ij})_{i,j=1,...,k-1}$ denote the correlation matrix over $x_1, ..., x_{k-1}$ induced by p_G - i.e., $\hat{\rho}_{ij}$ is the estimated correlation between x_i and x_j, whereas ρ_{ij} denotes their objective correlation. By assumption, the estimated marginals of $x_1, ..., x_{k-1}$ are correct, hence $\hat{\rho}_{ii} = 1$ for all $i = 1, ..., k-1$.

Furthermore, observe that in order to compute $\hat{\rho}_{ij}$ over $i, j = 1, ..., k-1$, we do not need to know the value of ρ_{hl} (i.e. the objective correlation between x_h and x_l). To see why, note that $(\hat{\rho}_{ij})_{i,j=1,...,k-1}$ is induced by $(p_G(x_1, ..., x_{k-1}))$. Each of the terms in the factorization formula for $p_G(x_1, ..., x_{k-1})$ is of the form $p(x_i \mid x_{R(i)})$, $i = 1, ..., k-1$. To compute this conditional probability, we only need to know $(\rho_{jj'})_{j,j' \in \{i\} \cup R(i)}$. By the definition of k, h and l, it is impossible for both h and l to be included in $\{i\} \cup R(i)$. Therefore, we can compute $(\hat{\rho}_{ij})_{i,j=1,...,k-1}$ without knowing the objective value of ρ_{hl}. We will make use of this observation toward the end of this proof.

The equation for x_k is

$$x_k = \sum_{i \in R(k)} \beta_{ik} x_i + \varepsilon_k \tag{10}$$

Let β denote the vector $(\beta_{ik})_{i \in R(k)}$. Let A denote the correlation sub-matrix $(\rho_{ij})_{i,j \in R(k)}$ that fully characterizes the objective joint distribution $(p(x_{R(k)}))$. Then, the objective variance of x_k can be written as

$$Var(x_k) = 1 = \beta^T A \beta + \sigma^2 \tag{11}$$

where $\sigma^2 = Var(\varepsilon_k)$.

In contrast, the estimated variance of x_k, denoted $Var_G(x_k)$, obeys the equation

$$Var_G(x_k) = \beta^T C \beta + \sigma^2 \tag{12}$$

where C denotes the correlation sub-matrix $(\hat{\rho}_{ij})_{i,j \in R(k)}$ that characterizes $(p_G(x_{R(k)}))$. In other words, the estimated variance of x_k is produced by replacing the objective joint distribution of $x_{R(k)}$ in the regression equation for x_k with its estimated distribution (induced by p_G), without changing the values of β and σ^2.

The undistorted-marginals constraint requires $Var_G(x_k) = 1$. This implies the equation

$$\beta^T A \beta = \beta^T C \beta \tag{13}$$

We now wish to show that this equation fails for generic $(\rho_{ij})_{i,j=1,...,k-1}$.

For any subsets $B, B' \subset \{1, ..., k-1\}$, use $\Sigma_{B \times B'}$ to denote the submatrix of $(\hat{\rho}_{ij})_{i,j=1,...,k-1}$ in which the selected set of rows is B and the selected set of columns is B'. By assumption, $h, l \in R(k)$ are unlinked. This means that according to G, x_h and x_l are independent conditional on x_M, where $M \subset \{1, ..., k-1\} - \{h, l\}$. Therefore, by Drton et al. (2008, p. 67),

$$\Sigma_{\{h\} \times \{l\}} = \Sigma_{\{h\} \times M} \Sigma^{-1}_{M \times M} \Sigma_{M \times \{l\}} \tag{14}$$

Note that equation (14) is precisely where we use the assumption that G

is imperfect. If G were perfect, then all nodes in $R(k)$ would be linked and therefore we would be unable to find a pair of nodes $h, l \in R(k)$ that necessarily satisfies (14).

The L.H.S of (14) is simply $\hat{\rho}_{hl}$. The R.H.S of (14) is induced by $p_G(x_1, ..., x_{k-1})$. As noted earlier, this distribution is pinned down by G and the entries in $(\rho_{ij})_{i,j=1,...,k-1}$ except for ρ_{hl}. That is, if we are not informed of ρ_{hl} but we are informed of all the other entries in $(\rho_{ij})_{i,j=1,...,k-1}$, we are able to pin down the R.H.S of (14).

Now, when we draw the objective correlation submatrix $(\rho_{ij})_{i,j=1,...,k-1}$ at random, we can think of it as a two-stage lottery. In the first stage, all the entries in this submatrix except ρ_{hl} are drawn. In the second stage, ρ_{hl} is drawn. The only constraint in each stage of the lottery is that $(\rho_{ij})_{i,j=1,...,k-1}$ has to be positive-semi-definite and have 1's on the diagonal. Fix the outcome of the first stage of this lottery. Then, it pins down the R.H.S of (14). In the lottery's second stage, there is (for a generic outcome of the lottery's first stage) a continuum of values that ρ_{hl} could take for which $(\rho_{ij})_{i,j=1,...,k-1}$ will be positive-semi-definite. However, there is only value of ρ_{hl} that will coincide with the value of $\hat{\rho}_{hl}$ that is given by the equation (14). We have thus established that $A \neq C$ for generic $(\rho_{ij})_{i,j=1,...,k-1}$.

Recall once again that we can regards β as a parameter of p that is independent of A (and therefore of C as well), because A describes $(p(x_{R(k)}))$ whereas β, σ^2 characterize $(p(x_k \mid x_{R(k)}))$. Then, since we can assume $A \neq C$, (13) is a non-tautological quadratic equation of β (because we can construct examples of p that violate it). By Caron and Traynor (2005), it has a measure-zero set of solutions β. We conclude that the constraint $Var_G(x_k) = 1$ is violated by almost every (ρ_{ij}). ■

By the claim, for every imperfect DAG G, the set of covariance matrices (ρ_{ij}) for which p_G preserves the mean and variance of all individual variables has measure zero. The set of imperfect DAGs over $\{1, ..., n\}$ is finite, and the finite union of measure-zero sets has measure zero as well. It follows that for almost all (ρ_{ij}), the property that p_G preserves the mean and variance of individual variables is violated unless G is perfect.

References

[1] Caron, R. and T. Traynor (2005), The Zero Set of a Polynomial, WSMR Report: 05-02.

[2] Drton, M., B. Sturmfels and S. Sullivant (2008), Lectures on Algebraic Statistics, Vol. 39, Springer Science & Business Media.

Appendix II: Uniform Binary Variables

In this appendix, we consider the case in which each variable x_i, $i = 1, ..., n$, takes values in $\{-1, 1\}$, and the marginal distribution over each x_i induced by p is uniform. This can be viewed as a coarsening of an underlying Gaussian distribution, such that x_i records the sign of a Gaussian variable.

We do not have a complete analysis of our problem for this specification of p, and focus on the chain model $1 \to 2 \to \cdots \to n$. In Eliaz et al. (2019), we provided a characterization of the maximal estimated correlation that such a model can generate in a uniform-binary environment. The proof was by induction on n. Here we give a constructive proof that emphasizes the analogy with the Gaussian case. Our analysis is based on a few preliminary observations.

Definition 2 *A $n \times n$ matrix C is called "Binary Factorizable" (BF) if it can be written as*

$$C = \lim_{M \to \infty} \frac{1}{M} A_M A_M^T$$

Where each A_M is a $n \times M$ matrix whose elements are all ± 1 and each row of A_M is zero mean.

Note that any BF matrix is symmetric, positive semi-definite, and has ones on the diagonal. Note also that any covariance matrix of zero-mean binary random variables must be BF, since we can define the matrix A_M as a sample covariance matrix, where the sample consists of M $i.i.d$ draws from the underlying distribution. The converse is also true: any BF matrix corresponds to the covariance matrix of zero-mean binary random variables. This can be seen by defining a distribution over n binary variables by randomly picking (with probability $1/M$) one of the columns of A_M.

Somewhat surprisingly, however, there exist symmetric, positive semi-definite matrices which are *not* BF. For example, the reader may recall the following correlation matrix from the example in the Introduction, where it gave the maximal false correlation for $n = 3$ in the Gaussian environment:

$$C = \begin{pmatrix} 1 & b & 0 \\ b & 1 & b \\ 0 & b & 1 \end{pmatrix}$$

with $b = \sqrt{1/2}$. This matrix is *not* BF. As we will see below, the largest value of b for which C is BF is $\frac{1}{2}$.

Proposition 2 *Suppose all variables take values in* $\{-1, 1\}$ *and the objective distribution* p *induces a uniform marginal over each variable. Let the objective (Pearson) coefficient of correlation between* x_1 *and* x_n*, according to* p*, is* r*. Then, the maximal estimated correlation that can be achieved by a linear DAG* $G : 1 \to 2 \to \cdots \to n$ *is given by:*

$$\rho^*_{1n} = \max_{\substack{\rho_{ij}=\rho_{ji} \text{ for all } i,j \\ (\rho_{ij}) \text{ is BF} \\ \rho_{ii}=1 \text{ for all } i \\ \rho_{1n}=r}} \prod_{i=1}^{n-1} \rho_{i,i+1}$$

Proof. The constraints are self-evident. We only need to show that for a linear DAG defined over uniformly distributed binary variables, the estimated correlation between x_1 and x_n is given by the product of the objective pairwise correlations of adjacent variables (as in the Gaussian case). We can show this by viewing $p_G(x_1, ..., x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1})$ as a Markov chain. The conditional probability $p_G(x_n \mid x_1)$ is thus given by a matrix product - specifically, the product of all the transition matrices defined by $p(x_{i+1} \mid x_i)$. Since all variables are uniformly distributed, the transition matrices are doubly stochastic, which means that they have the same eigenvectors. The top eigenvalue is always 1 and the second eigenvalue gives the correlation. Since all matrices have the same eigenvectors, the eigenvalues just multiply. ■

Note that Proposition 2 is exactly the same as the intermediate result we established at the beginning of Section 4.3 for the Gaussian environment. The only difference is that we replace the requirement that ρ be positive semi-definite with the requirement that ρ be BF. As mentioned above, the set of BF matrices is smaller than the set of positive semi-definite matrices. Therefore, we should expect a more stringent upper bound on the maximal false correlation.

Proposition 3 *Suppose all variables take values in* $\{-1, 1\}$ *and the objective distribution* p *induces a uniform marginal over each variable. Let the objective (Pearson) coefficient of correlation between* x_1 *and* x_n*, according to* p *be equal to* r*. Then, the maximal estimated correlation that can be generated by the DAG* $1 \to 2 \to \cdots \to n$ *is given by:*

$$\rho^*_{1n} = \left(1 - \frac{1}{n-1}(1-r)\right)^{n-1} \tag{15}$$

Proof. From Proposition 2, we know that the maximal estimated correlation is obtained by multiplying elements in a BF correlation matrix (ρ_{ij}) such that $\rho_{1n} = r$. For any $n \times M$ matrix A_M, let $a_i^{(M)}$ denote its i^{th} row. Then, we can rewrite the estimated correlation induced by $C_M = \frac{1}{M} A_M A_M^T$ as:

$$\prod_{i=1}^{n-1} \frac{1}{M} a_i^{(M)T} a_{i+1}^{(M)}$$

As we discussed following the definition of BF matrices, the dot product between the i^{th} and j^{th} rows of A_M is proportional to the empirical correlation of x_i and x_j in a sample consisting of M *i.i.d* draws from the underlying distribution.

Given a matrix A_M that gives an objective correlation of $\rho_{1n} = r$, we can always attempt to improve the estimated correlation by optimizing all other rows of the matrix $a_2, ..., a_{n-1}$. This implies that for any M:

$$\rho^*_{1n} \leq \max_{a_2,\cdots,a_{n-1}\in\{-1,1\}^M, a_1=a_1^{(M)}, a_n=a_n^{(M)}} \prod_{i=1}^{n-1} \frac{1}{M} a_i^T a_{i+1} \tag{16}$$

This is an upper bound for two reasons. First, we are not enforcing the constraint that the binary vectors a_i are zero mean. Second, if $C = \frac{1}{M} A_M A_M^T$ for some finite M, then C is BF.

For binary vectors $a_i, a_j \in \{-1, 1\}^M$, the dot product $\frac{1}{M} a_i^T a_j$ is a monotone function of the proportion q of components for which the two vectors agree: $\frac{1}{M} a_i^T a_j = 2q - 1$. Thus, maximizing the dot product between two binary vectors is equivalent to minimizing the number of components on which they disagree. This means that the R.H.S of (16) is a form of a *shortest path on a lattice*: we are given two points in $\{-1, 1\}^M$ (a_1 and a_n), and seek a set of intermediate points on this lattice that are as close as possible to each other. By analogy, in the third step of our proof for the Gaussian case, we were also given two vectors in a high-dimensional space (an n-dimensional unit sphere) and searched for a set of intermediate points on the sphere such that the intermediate points are as close as possible to one another (in terms of spherical distance).

To solve this "shortest path on a lattice" problem, we divide the M indices into two disjoint groups: M_1 indices k for which $a_1(k) = a_n(k)$ and M_{-1} indices k for which $a_1(k) \neq a_n(k)$. For any of the M_1 indices for which $a_1(k) = a_n(k)$, setting $a_i(k) = a_1(k)$ for all i can only increase the objective function (since this can only increase the dot product between consecutive vectors).

For the remaining M_{-1} indices k for which $a_1(k) \neq a_n(k)$, denote by m_i the number of indices k for which $a_i(k) = a_1(k)$ and $a_i(k) \neq a_n(k)$. Assuming $m_i > m_j$, the dot product between a_i and a_j can be written as follows:

$$a_i^T a_j = M - 2(m_i - m_j)$$

This enables us to rewrite (16) as:

$$\rho^*_{1n} \leq \max_{m_2,\cdots,m_{n-1}} \prod_{i=1}^{n-1} \frac{1}{M} \left(M - 2(m_{i-1} - m_i)\right) \tag{17}$$

The R.H.S. of (17) should be maximized subject to the constraint that $m_i \in \{0, 1, \cdots M_{-1}\}$, but we can get an upper bound by maximizing over real-valued m_i.

Taking the logarithm of the R.H.S of (17) and differentiating with respect to m_i yields that at an optimum, m_i should be linearly spaced between m_1 and m_n:

$$m_i - m_{i+1} = \frac{M_{-1}}{n-1}$$

Thus, the optimal shortest path is a set of binary vectors whose components agree with x_1 and x_n whenever they coincide, and the rest of the indices agree with x_1 with a fraction that decreases linearly with i.

Now, for large M, M_{-1}/M converges to the probability that $x_1 \neq x_n$, namely $\frac{1-r}{2}$, such that

$$\frac{1}{M} a_i^T a_{i+1} \to \left(1 - \frac{1}{n-1}(1-r)\right)$$

Since there are $n-1$ such dot products, we take their product, thus obtaining the R.H.S of (15).

To show that the upper bound is tight, given two uniform binary random variables x_1, x_n that satisfy $E(x_1 x_n) = r$, consider a set of variables x_i, whose distribution conditional on x_1, x_n is defined as follows:

- If $x_1 = x_n$, then $x_i = x_1 = x_n$.
- If $x_1 \neq x_n$, then $x_i = x_1$ with probability $1 - \frac{i}{n}$ and $x_i = x_n$ with probability $\frac{i}{n}$.

By construction, a vector of M random samples from x_i and x_{i-1} will generate a normalized dot product $\frac{1}{M} a_i^T a_{i+1}$ that converges to $\left(1 - \frac{1}{n-1}(1-r)\right)$ when $M \to \infty$, thus attaining the upper bound.

It is also worth noting that in Eliaz et al. (2019), we implement the upper bound by taking the n variables to be the sign of the Gaussian variables we used in the implementation of the upper bound of our the main theorem. ■

Let us illustrate the upper bound. For $n = 3$ and $r = 0$, the maximal estimated correlation between x_1 and x_3 using the chain model $1 \to 2 \to 3$ is $\frac{1}{4}$ (compared with the value $\frac{1}{2}$ in the Gaussian case). Finally, for any r, the maximal estimated correlation converges to e^{r-1} as $n \to \infty$ (compared with 1 in the Gaussian case).