

Cheating with Models*

Kfir Eliaz[†], Ran Spiegler[‡] and Yair Weiss[§]

September 24, 2020

Abstract

Beliefs and decisions are often based on confronting models with data. What is the largest “fake” correlation that a misspecified model can generate, even when it passes an elementary misspecification test? We study an “analyst” who fits a model, represented by a directed acyclic graph, to an objective (multivariate) Gaussian distribution. We characterize the maximal estimated pairwise correlation for generic Gaussian objective distributions, subject to the constraint that the estimated model preserves the marginal distribution of any individual variable. As the number of model variables grows, the estimated correlation can become arbitrarily close to one, regardless of the objective correlation.

*A longer working-paper version of this paper has appeared under the title “Cheating with (Recursive) Models” (Eliaz et al. (2019)). We gratefully acknowledge financial support from ISF grant no. 470/19 (Eliaz), ERC Advanced Investigator grant no. 692995 (Spiegler) and the Gatsby Charitable Trust (Weiss). Eliaz and Spiegler thank Briq and the Economics Department at Columbia University for their generous hospitality while this paper was written. We also thank Armin Falk, Irit Gat-Wiks, Xiaosheng Mu, Tal Pupko, Martin Weidner, seminar audiences and especially Heidi Thysen, for helpful comments.

[†]School of Economics, Tel-Aviv University and David Eccles School of Business, University of Utah. E-mail: kfire@tauex.tau.ac.il.

[‡]School of Economics, Tel Aviv University; Department of Economics, UCL; and CFM. E-mail: rani@tauex.tau.ac.il.

[§]School of Computer Science and Engineering, Hebrew University. E-mail: yweiss@cs.huji.ac.il.

1 Introduction

Quantifying the correlation between random variables is a problem that preoccupies decision makers and scientific researchers alike, for purposes of diagnosis, prediction or causal inference. In some cases, agents are capable of (and comfortable with) learning correlations directly from observational data. In other cases, they estimate them indirectly with the help of *models*. The use of models - whether informally by everyday decision makers or more formally by researchers - has several motivations. Belief in a model allows us to extrapolate from incomplete or noisy data. Models are instrumental in drawing causal inferences from observational data. Finally, as simplified representations of complex empirical regularities, models help perceiving and communicating them.

And yet, just as a correct model is valuable for all these reasons, a *wrong* model can derail decision makers and scientific researchers (or their audiences). This paper poses the following theoretical question: *To what extent can a misspecified model lead to a distorted estimate of pairwise correlations?*

There are several reasons to be interested in this question. First, individuals and policy makers are often guided by subjective models - whether these models are their own or fed to them by experts. When such models are wrong, we would like to get a bound on the magnitude of the resulting decision errors. Models that generate larger distortion of correlations between certain key variables will tend to induce larger decision errors. Our paper thus introduces worst-case analysis into the literature on decision making under misspecified models (for a few milestones in this growing literature, see Piccione and Rubinstein (2003), Jehiel (2005), Eyster and Piccione (2013) and Esponda and Pouzo (2016)).

Second, politicians and pundits often use false narratives (which Eliaz and Spiegler (2018) formalize as misspecified causal models) to exaggerate the perceived impact of policies and attribute spurious credit/blame for social outcomes. Our exercise helps quantifying the extent to which they can do so. Relatedly, when multiple contending models address a social issue, those that predict extreme effects are more likely to grab public attention. Models

that maximize distorted correlations survive this kind of “natural selection” (this idea is close in spirit to the notion of “competing models” in Montiel Olea et al. (2018)).

Finally, scientific researchers often aim to persuade their audience of diagnostic, predictive or causal relations between variables. Their motive could be that they serve a policy maker with a particular agenda (e.g. showing that cutting taxes fosters economic growth); or that they had staked their reputation on a claim that strongly links two variables; or that they may want to make a splash with a strong finding. Such incentives may lead to (possibly subconscious) self-serving model selection. For purely expositional purposes, we will focus on this “bad researcher” metaphor.

In our model, an *analyst* wishes to demonstrate to a lay audience that two given variables are strongly correlated (we will bounce between the diagnostic, predictive and causal interpretations of this correlation). The analyst has statistical data about the joint behavior of many variables in addition to the two target variables. His method is to propose a model, fit it to the data and use the estimated model to compute pairwise correlations. The analyst is unable (or unwilling) to tamper with the data, and his method of fitting the model to the data is “legitimate”. However, he is free to choose the variables that enter the model and how they operate in it. Thus, the researcher “does everything right” given the model; his only vehicle for “cheating” is model misspecification.

Of course, in order for our exercise to have content, it must define the domain of models our analyst can use. We assume that the analyst is restricted to use models that take the form of *directed acyclic graphs* (DAGs), a.k.a Bayesian networks. This class of models is widely used in various scientific areas (see Morgan and Winship (2015)) and Artificial-Intelligence systems (see Koller and Friedman (2009)). In Gaussian environments, DAGs are equivalent to recursive systems of linear-regression equations - a special case of simultaneous-equations models, which are familiar to economists from their introductory econometrics courses. DAGs have a natural interpretation as *causal models* (Pearl (2009), Sloman (2005)). Thus, when our analyst fits a DAG to objective data, he essentially interprets the data

through a causal model. Finally, Spiegel (2016,2020) showed how important families of misspecified models in the economics literature (e.g., Jehiel and Koessler’s (2008) analogy-based expectations in static games, Eyster and Rabin’s (2005) “cursed” beliefs, or Mailath and Samuelson’s (2020) “model-based inference”) can be recast in the language of DAGs. From this point of view, DAGs are a convenient “model of misspecified models”.

We can now sharpen our original question. Given that the objective distribution is Gaussian; that the objective (Pearson) correlation between the target variables is r ; and that the analyst can use a DAG-represented model with up to n variables; how large can the *estimated* correlation between the target variables be? We impose one constraint on the analyst: *His estimated model cannot distort the marginal distribution of any individual variable.* We interpret this constraint as an elementary “misspecification test” that an unsophisticated audience could implement, since monitoring individual variables (unlike their comovement) is relatively straightforward.

We show that subject to this constraint, the upper bound on the estimated correlation for generic Gaussian objective distributions is

$$\left(\cos \left(\frac{\arccos r}{n-1} \right) \right)^{n-1} \tag{1}$$

This upper bound is attained by a simple, n -length causal chain that connects the two target variables; and a selection of variables that are all different linear combinations of two, independent Gaussian variables. Formula (1) is strictly increasing in n . That is, cheating with larger models is easier. Indeed, when $n \rightarrow \infty$, the expression converges to 1 for *any* $r \in [-1, 1)$.

An Example: “Biomarker hacking”

To illustrate our exercise, imagine an analyst who has access to an arbitrarily large sample documenting the joint distributions of (x_1, x_2) and (x_2, x_3) , yet lacks direct data about the joint distribution of (x_1, x_3) . He assumes that x_1 has an effect on x_3 only through a *mediator* x_2 , such that x_1 and x_3 are statistically independent conditional on x_2 . This assumption is represented by the DAG $x_1 \rightarrow x_2 \rightarrow x_3$.

For a real-life situation behind this example, consider a pharmaceutical company that introduces a new drug, and therefore has a vested interest in demonstrating a large correlation between drug dosage (x_1) and the ten-year survival rate associated with some disease (x_3). This correlation cannot be directly measured in the short run. However, past experience reveals the correlations between the ten-year survival rate and the levels of various biomarkers - each of which can serve as the intermediate variable x_2 . The correlation between these markers and drug dosage can be measured experimentally in the short run. The model $x_1 \rightarrow x_2 \rightarrow x_3$ captures a research strategy that treats x_2 as a “surrogate marker” for x_3 . Yet the company’s R&D unit may select the marker x_2 opportunistically, in order to get a large estimated effect. Fleming and DeMets (1996) and Katz (2004) discuss pitfalls in interpreting correlation estimates from drug effects on biomarkers. Our exercise is (to our knowledge) the first to put an upper bound on the false correlation that can result from “biomarker hacking”.

Suppose the objective joint distribution over the three variables is Gaussian. Let ρ_{ij} denote the objective correlation between x_i and x_j . Suppose $\rho_{13} = 0$ - i.e., x_1 and x_3 are objectively uncorrelated. In contrast, the *estimated* correlation $\hat{\rho}_{13}$, given the analyst’s model, is

$$\hat{\rho}_{13} = \rho_{12} \cdot \rho_{23}$$

It is easy to see from this expression how the model can generate spurious estimated correlation between x_1 and x_3 . All the analyst has to do is select a variable x_2 that is positively correlated with both x_1 and x_3 , such that $\rho_{12}\rho_{23} > 0$. Yet how large can $\hat{\rho}_{13}$ be? Intuitively, since x_1 and x_3 are objectively uncorrelated, selecting x_2 to raise ρ_{12} will lower ρ_{23} . Formally, consider the objective correlation matrix:

$$\begin{matrix} 1 & \rho_{12} & 0 \\ \rho_{12} & 1 & \rho_{23} \\ 0 & \rho_{23} & 1 \end{matrix}$$

By definition, this matrix is positive semi-definite. This property is charac-

terized by the inequality $(\rho_{12})^2 + (\rho_{23})^2 \leq 1$. The maximal value of $\rho_{12}\rho_{23}$ subject to this constraint is $\frac{1}{2}$, hence this is the maximal false correlation that the model can generate. This bound is attained if we define x_2 to be a deterministic function of x_1 and x_3 , given by $x_2 = (x_1 + x_3)/\sqrt{2}$. Thus, while a given misspecified DAG-represented model may generate spurious correlation between objectively uncorrelated variables, there is a limit to how far it can go.

What is the significance of this upper bound on $\hat{\rho}_{13}$? As the “biomarker hacking” scenario suggests, we have in mind situations in which the analyst can select x_2 from a *large pool* of potential auxiliary variables. In the current age of “big data”, analysts have access to datasets involving a huge number of covariates. When the analyst has discretion over which variables will enter the model, he can generate a false correlation that approaches the theoretical upper bound.

To illustrate this claim, we examined a database compiled by the World Health Organization and collected by Reshef et al. (2011).¹ This database contains hundreds of health and socioeconomic variables about all countries. Let the target variables x_1 and x_3 be urban population and liver cancer deaths. Their objective correlation in the database is 0.05. If we select x_2 to be the variable coal consumption, we obtain $\hat{\rho}_{13} = 0.43$, far above the objective value and close to the theoretical upper bound. This selection of x_2 has the added advantage that the model suggests a plausible-sounding causal mechanism: Urbanization causes cancer deaths via its effect on coal consumption. The numerical illustration shows that the upper bound on $\hat{\rho}_{13}$ may have real-life relevance when researchers have substantial freedom to “fish for mediating variables”.

2 The Model

Let p be an objective probability measure over n variables, x_1, \dots, x_n . Assume p is multivariate normal. For every $A \subset \{1, \dots, n\}$, denote $x_A = (x_i)_{i \in A}$.

¹The variables are collected on all countries in the WHO database (see www.who.int/whosis/en/) for the year 2009.

Without loss of generality, the marginal of p on every variable has *zero mean and unit variance*. We use ρ_{ij} to denote the Pearson coefficient of correlation between the variables x_i, x_j , according to p . In particular, denote $\rho_{1n} = r$. The distribution p is fully identified by the covariance matrix (ρ_{ij}) . Note that we have presented the n variables and the distribution p as given, for the sake of expositional convenience. However, we have in mind situations in which the variables x_2, \dots, x_{n-1} are chosen opportunistically by an analyst from some large pool. This means that effectively, the analyst chooses p from a large set of Gaussian distributions for which $\rho_{1n} = r$.

A directed graph is a pair $G = (N, R)$, where N is a set of nodes and $R \subset N \times N$ is a pair of directed links. We assume throughout that $N = \{1, \dots, n\}$. With some abuse of notation, $R(i)$ is the set of nodes j for which the DAG includes a link $j \rightarrow i$. We restrict attention to *acyclic* directed graphs (DAGs) - i.e. graphs that do not include directed paths $i \rightarrow j \rightarrow \dots \rightarrow i$.

A node $i \in N$ is ancestral if $R(i)$ is empty. A node $i \in N$ is terminal if there is no $j \in N$ such that $i \in R(j)$. A DAG (N, R) is *perfect* if whenever $i, j \in R(k)$ for some $i, j, k \in N$, it is the case that $i \in R(j)$ or $j \in R(i)$. A subset of nodes $C \subseteq N$ is a *clique* if for every $i, j \in C$, $i \in R(j)$ or $j \in R(i)$. We say that a clique is *maximal* if it is not contained in another clique. We use \mathcal{C} to denote the collection of maximal cliques in a DAG. Observe that if (N, R) is perfect, then $R(i)$ is a clique for every $i \in N$.

We consider an analyst who fits a DAG-represented model to the objective distribution. Following the literature on graphical probabilistic models (Cowell et al. (1999), Koller and Friedman (2009), Pearl (2009)), there are two primary interpretations for such models. First, a DAG can be viewed as a representation of conditional-independence assumptions. E.g., the DAG $x_1 \rightarrow x_2 \rightarrow x_3$ represents the assumption that x_1 and x_3 are independent conditional on x_2 , whereas the DAG $x_1 \rightarrow x_3 \leftarrow x_2$ represents the assumption that x_1 and x_2 are unconditionally independent. Second, a DAG has a natural interpretation as a *causal model*, such that directed links represent postulated direct causal influences. For this reason, Pearl (2009) used DAGs as a platform for a systematic theory of causal inference, while psychologists (see Sloman (2005)) used them as representations of people's intuitive

causal perceptions. Thus, a DAG can be viewed as a model that imposes assumptions about the causal relations between the model’s variables.

Given an objective distribution p over x_1, \dots, x_n and a DAG G , define the Bayesian-network factorization formula:

$$p_G(x_1, \dots, x_n) = \prod_{k=1}^n p(x_k \mid x_{R(k)}) \quad (2)$$

For instance, when $G : x_1 \rightarrow x_2 \rightarrow x_3$,

$$p_G(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)$$

Note that p_G is a well-defined probability distribution. We say that p is consistent with G if $p_G = p$.

The distribution p_G formalizes the notion of imposing a DAG-represented model on objective data (see Cowell et al. (1999) or Koller and Friedman (2009)). Hence, we refer to p_G as the “estimated model”. If p were consistent with G , it would be legitimate to write it according to the R.H.S of (2). When p is inconsistent with G , p_G distorts p .² As long as p has full support, p_G induces well-defined marginal and conditional distributions, as well as the estimated (Pearson) coefficient of correlation between any pair of variables x_i and x_j , which we denote $\hat{\rho}_{ij}$. We use $Var_G(x_k)$ and $Cov_G(x_i, x_j)$ to denote the variance of x_k and covariance between x_i and x_j that are induced by p_G .

We say that p_G satisfies the *Undistorted Marginals Constraint* (UMC) if the induced marginal distribution $p_G(x_i)$ coincides with the objective marginal distribution $p(x_i)$ for every $i = 1, \dots, n$. Since p is assumed to be Gaussian, UMC is equivalent to requiring that $p_G(x_i)$ has zero mean and unit variance for every i .

To motivate this constraint, suppose the analyst’s interest in pairwise correlations arises from diagnostics or prediction tasks. An *unsophisticated* audience cannot be expected to discipline the analyst’s opportunistic model selection with elaborate tests for model misspecification that involve

²Thus, as in Spiegler (2016), G can be viewed as a function that systematically transforms any objective distribution p into a subjective distribution p_G .

conditional or unconditional correlations. However, monitoring *individual* variables is a much simpler task than monitoring correlations. E.g., it is relatively easy to disqualify an economic model that predicts highly volatile inflation if observed inflation is relatively stable. Likewise, a climatological model that underpredicts temperature volatility loses credibility, even for a lay audience. Beyond this justification, we simply find it intrinsically interesting to know the extent to which misspecified models can distort pairwise correlations without distorting marginals.

Comment: The Gaussianity assumption

Each term $p(x_k | x_{R(k)})$ in (2) records a conditional distribution. The Gaussianity assumption means that this term can be written as a linear-regression equation:

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

where the β_{ik} 's are obtained by applying Ordinary Least Squares to p .

This observation means that we could equivalently describe the analyst's procedure as estimating a recursive system of linear regression equations (by "recursive" we mean that a R.H.S variable on one equation cannot appear as the L.H.S variable of a subsequent equation). Under this description, (ρ_{ij}) is a sufficient statistic for p as a result of the analyst's linearity assumption, and so we need not assume that p is Gaussian. The proof of our main result will make use of this equivalence (for more on Gaussian Bayesian networks, see Koller and Friedman (2009, Ch. 7)).

To see how this equivalence enables us to calculate $\hat{\rho}_{ij}$, revisit the example from the Introduction, where $G : x_1 \rightarrow x_2 \rightarrow x_3$. The conditional-probability terms $p(x_2 | x_1)$ and $p(x_3 | x_2)$ are given by OLS estimators of the linear-regression equations

$$\begin{aligned} x_2 &= \alpha x_1 + \varepsilon_2 \\ x_3 &= \beta x_2 + \varepsilon_3 \end{aligned}$$

such that

$$x_3 = \alpha\beta x_1 + \beta\varepsilon_2 + \varepsilon_3$$

The model's conditional-independence assumption means that ε_3 and x_1 are treated as independent variables (ε_2 and x_1 are independent by the OLS method). It is now easy to confirm that $\hat{\rho}_{13} = \alpha\beta = \rho_{12}\rho_{23}$, as stated in the Introduction.

3 The Main Result

For every r, n , denote

$$\theta_{r,n} = \frac{\arccos r}{n-1}$$

We are now able to state our main result.

Theorem 1 *For almost every objective correlation matrix (ρ_{ij}) satisfying $\rho_{1n} = r$, if an n -variable DAG-represented model satisfies UMC, then*

$$\hat{\rho}_{1n} \leq (\cos \theta_{r,n})^{n-1}$$

Moreover, this upper bound can be implemented by the following pair:

- (i) A DAG $1 \rightarrow 2 \rightarrow \dots \rightarrow n$.
- (ii) A Gaussian distribution satisfying, for every $k = 1, \dots, n$:

$$x_k = s_1 \cos((k-1)\theta_{r,n}) + s_2 \sin((k-1)\theta_{r,n}) \quad (3)$$

where s_1, s_2 are independent standard normal variables.

Let us illustrate the upper bound given by Theorem 1 numerically for the case of $r = 0$, as a function of n :

n	2	3	4	5
upper bound on $\hat{\rho}_{1n}$	0	0.5	0.65	0.73

As we can see, the marginal contribution of an additional variable decays quickly. However, when $n \rightarrow \infty$, the upper bound converges to one, for any $r \in [-1, 1]$.

The DAG that implements the upper bound is a chain. Indeed, it is the *simplest* connected n -node DAG - whether one measures complexity by the number of links, the size of cliques or the number of ancestral nodes. The distribution over the auxiliary variables x_2, \dots, x_n in the upper bound's implementation has a simple structure, too: Every x_k is a different linear combination of two independent “factors”, s_1 and s_2 . We can identify s_1 with x_1 , without loss of generality. The closer the variable lies to x_1 along the chain, the larger the weight it puts on s_1 .

We interpret the genericity aspect of the theorem as follows. Recall that we think of our analyst as effectively choosing the variables x_2, \dots, x_{n-1} from some large, yet finite pool. Genericity means that for almost all such pools, our upper bound on $\hat{\rho}_{1n}$ will hold.

Causal interpretation of $\hat{\rho}_{1n}$

The fact that x_1 functions as an *ancestral* node in the DAG that implements the upper bound on $\hat{\rho}_{1n}$ means that the analyst can interpret $\hat{\rho}_{1n}$ as an estimated *causal* effect. If x_1 were not ancestral in the analyst's DAG, the model itself might interpret part of the correlation between x_1 and x_n as a consequence of confounding by some other variable. This limitation does not exist when x_1 is an ancestral node. Thus, although the analyst's problem was to find a model that maximizes an estimated pairwise correlation - regardless of whether this correlation is interpreted in predictive, diagnostic or causal terms - the solution turns out to enable a causal interpretation.

Variable selection vs. model selection

Suppose the analyst selects the variables x_2, \dots, x_{n-1} that will enter the model (in addition to x_1 and x_n), but then has no discretion over the model *given these variables*. Instead, he employs a standard procedure for “model discovery” that penalizes complexity (measured by the maximal size of $R(k)$). Specifically, suppose that he employs the Chow-Liu algorithm (Chow and Liu (1968)), which only admits models with $|R(k)| \leq 1$. Then, it is easy to show that if x_2, \dots, x_{n-1} are selected as in (3), the Chow-Liu algorithm will select the model given by the chain $1 \rightarrow 2 \rightarrow \dots \rightarrow n$. In this sense, the crucial assumption in our exercise is that the analyst chooses model variables,

whereas the selection of the model given these variables can be automatized.

Perfect DAGs

The Chow-Liu algorithm is a special case of procedures that restrict attention to perfect DAGs. The chain $1 \rightarrow 2 \rightarrow \dots \rightarrow n$ is trivially perfect. Perfect DAGs preserve marginals of individual variables for *every* objective distribution (see Spiegler (2017)). This enables us to state Theorem 1 more strongly for this subclass of models.

Proposition 1 *If the analyst’s model is represented by a perfect DAG, then $\hat{\rho}_{1n} \leq (\cos \theta_{r,n})^{n-1}$ for every objective correlation matrix (ρ_{ij}) satisfying $\rho_{1n} = r$.*

That is, when we require the analyst’s model to be represented by a *perfect* DAG, the upper bound on $\hat{\rho}_{1n}$ holds for *any* objective covariance matrix, and UMC is redundant.

4 Proof of Theorem 1

We first give a broad outline of our proof, which proceeds in three steps. First, we show that for generic Gaussian distributions, UMC forces the DAG to be perfect. This is the only place in the proof that relies on genericity. To give a rough intuition for this step, consider the imperfect DAG G :

$$\begin{array}{ccc} 1 & \rightarrow & 3 \\ \downarrow & & \downarrow \\ 2 & \rightarrow & 4 \end{array} \tag{4}$$

UMC requires $Var_G(x_4) = 1$. To satisfy this condition, it can be shown that $\hat{\rho}_{23}$ should coincide with ρ_{23} . But since G assumes that x_2 and x_3 are independent conditional on x_1 , $\hat{\rho}_{23} = \rho_{12}\rho_{13}$. The resulting equation $\rho_{23} = \rho_{12}\rho_{13}$ is violated by generic (ρ_{ij}) .

Second, we utilize the Gaussianity assumption and apply the tool of *junction trees* from the Bayesian-network literature to further shrink the relevant

domain from perfect DAGs to simple chains, by replacing clusters of nodes with single “mega-nodes”. For example, consider the perfect DAG G' that adds the link $2 \rightarrow 3$ to the DAG G given by (4), such that

$$p_{G'}(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2)p(x_4 | x_2, x_3)$$

Because p is Gaussian, the term $p(x_4 | x_2, x_3)$ is given by a regression equation $x_4 = \beta_{24}x_2 + \beta_{34}x_3 + \varepsilon_4$. Since x_2 and x_3 only affect x_4 through the linear combination $x_5 = \beta_{24}x_2 + \beta_{34}x_3$, we can replace (x_2, x_3) with the scalar variable x_5 , such that $p_{G'}(x_1, x_2, x_3, x_4)$ can be equivalently rewritten as $p(x_1)p(x_5 | x_1)p(x_4 | x_5)$, which is the factorization formula for the chain $1 \rightarrow 5 \rightarrow 3$.

Thus, in order to calculate the upper bound on $\hat{\rho}_{1n}$, we can restrict attention to chain models. The analyst’s objective function attains a simple explicit form:

$$\hat{\rho}_{1n} = \prod_{k=1}^{n-1} \rho_{k,k+1}$$

In the proof’s final step, we find the matrix (ρ_{ij}) that maximizes this expression, subject to the constraints that $\rho_{1n} = r$, $\rho_{ii} = 1$ for all i (capturing UMC) and that (ρ_{ij}) is positive semi-definite (the defining property of covariance matrices). This problem has a simple geometric interpretation: given the relative location of two points on a sphere, locate $n - 2$ intermediate points to minimize the average spherical distance between adjacent points. The solution is to place the points equidistantly along a great circle.

Let us now turn to the formal proof.

4.1 The First Step: From DAGs to Perfect DAGs

We first establish that for generic Gaussian p , perfection is necessary for the correct-marginal constraint. Because p is Gaussian, p_G is Gaussian as well (see Koller and Friedman (2009, Ch. 7)). Therefore, the marginal of p_G over x_k is given entirely by its mean and variance, denoted $E_G(x_k)$ and $Var_G(x_k)$.

Lemma 1 *For almost every (ρ_{ij}) , if a DAG G satisfies $E_G(x_k) = 0$ and $\text{Var}_G(x_k) = 1$ for all $k = 1, \dots, n$, then G is perfect.*

Proof. The proof is “technical” in nature and relegated to Appendix I. ■

4.2 The Second Step: From Perfect DAGs to Chains

Our next step shows that within the class of perfect DAGs, simple chains entail no loss of generality.

Definition 1 *A DAG (N, R) is linear if 1 is the unique ancestral node, n is the unique terminal node, and $R(i)$ is a singleton for every non-ancestral node.*

A linear DAG is thus a causal chain $1 \rightarrow \dots \rightarrow n$. Every linear DAG is perfect by definition.

Lemma 2 *For every Gaussian distribution with correlation matrix ρ and non-linear perfect DAG G with n nodes, there exist a Gaussian distribution with correlation matrix ρ' and a linear DAG G' with weakly fewer nodes than G , such that $\rho_{1n} = \rho'_{1n}$ and the estimated correlation induced by G' given ρ' is exactly the same as the estimated correlation induced by G given ρ .*

Proof. The proof proceeds in two main steps.

Step 1: Deriving an explicit form for the false correlation using an auxiliary “cluster recursion” formula

The following is standard material in the Bayesian-network literature. For any distribution $p_G(x)$ corresponding to a perfect DAG, we can rewrite the distribution as if it factorizes according to a tree graph, where the nodes in the tree are the maximal cliques of G . This tree satisfies the *running intersection property* (Koller and Friedman (2009, p. 348)): If $i \in C, C'$ for two tree nodes, then $i \in C''$ for every C'' along the unique tree path between C and C' . Such a tree graph is known as the “*junction tree*” corresponding

to G and we can write the following “cluster recursion” formula (Koller and Friedman (2009, p. 363)):

$$p_G(x) = p_G(x_{C_r}) \prod_i p_G(x_{C_i} | x_{C_{r(i)}}) = p(x_{C_r}) \prod_i p(x_{C_i} | x_{C_{r(i)}})$$

where C_r is an arbitrarily selected root clique node, and $C_{r(i)}$ is the upstream neighbor of clique i (the one in the unique path from C_i to the root C_r). The second equality is due to the fact that G is perfect, hence $p_G(x_C) \equiv p(x_C)$ for every clique C of G .

Let $C_1, C_K \in \mathcal{C}$ be two cliques that include the nodes 1 and n , respectively. Furthermore, for a given junction tree representation of the DAG, select these cliques to be minimally distant from each other - i.e., $1, n \notin C$ for every C along the junction-tree path between C_1 and C_K . We now derive an upper bound on K . Recall the running intersection property: If $i \in C_j, C_k$ for some $1 \leq j < k \leq K$, then $i \in C_h$ for every h between k and j . Since the cliques C_1, \dots, C_K are maximal, it follows that every C_k along the sequence must introduce at least one new element $i \notin \cup_{j < k} C_j$ (in particular, C_1 includes some $i > 1$). As a result, it must be the case that $K \leq n - 1$. Furthermore, since G is assumed to be *non-linear*, the inequality is *strict*, because at least one C_k along the sequence must contain at least three elements and therefore introduce at least *two* new elements. Thus, $K \leq n - 2$.

Since p_G factorizes according to the junction tree, it follows that the distribution over the variables covered by the cliques along the path from C_1 to C_K factorize according to a linear DAG $1 \rightarrow C_1 \rightarrow \dots \rightarrow C_K \rightarrow n$, as follows:

$$p_G(x_1, x_{C_1}, \dots, x_{C_K}, x_n) = p(x_1) \prod_{k=1}^K p(x_{C_k} | x_{C_{k-1}}) p(x_n | x_{C_K}) \quad (5)$$

where $C_0 = \{1\}$. Thus, we can regard 1 and n as *ancestral* and *terminal* nodes in this DAG, without loss of generality. Therefore, more generally, we can enumerate the variables such that lower-indexed variables belong to earlier nodes of the linear DAG. The length of this linear DAG is $K + 2 \leq n$.

While this factorization formula superficially appears to complete the proof, note that the variables x_{C_k} are typically *multivariate* normal variables, whereas our objective is to show that we can replace them with scalar (i.e. univariate) normal variables without changing $Cov_G(x_1, x_n)$.

Since p is multivariate normal, for any two subsets of variables C, C' , the distribution of x_C conditional on $x_{C'}$ can be written $x_C = Ax_{C'} + \eta$, where A is a matrix that depends on the means and covariances of p , and η is a zero-mean vector that is uncorrelated with $x_{C'}$. Applying this property to the junction tree, we can describe $p_G(x_1, x_{C_1}, \dots, x_{C_K}, x_n)$ via the following recursion:

$$\begin{aligned}
 x_1 &\sim N(0, 1) & (6) \\
 x_{C_1} &= A_1 x_1 + \eta_1 \\
 &\vdots \\
 x_{C_k} &= A_k x_{C_{k-1}} + \eta_k \\
 &\vdots \\
 x_{C_K} &= A_K x_{C_{K-1}} + \eta_K \\
 x_n &= A_{K+1} x_{C_K} + \eta_n
 \end{aligned}$$

where each equation describes an objective conditional distribution - in particular, the equation for x_{C_k} describes $(p(x_{C_k}|x_{C_{k-1}}))$. The matrices A_k are functions of the vectors β_i in the original model. The η_k 's are all zero mean and uncorrelated with the explanatory variables $x_{C_{k-1}}$, such that $E(x_{C_k}|x_{C_{k-1}}) = A_k x_{C_{k-1}}$. Furthermore, according to p_G (i.e. the analyst's estimated model), each x_k is conditionally independent of x_1, \dots, x_{k-1} given $x_{R(k)}$. (Here we use the earlier result that we can enumerate the variables such that $R(k) \subseteq \{1, \dots, k-1\}$ for every k .) Since the junction-tree factorization (5) represents exactly the same distribution p_G , this means that every η_k is uncorrelated with all other η_j 's as well as with $x_1, \dots, x_{C_{k-2}}$. Therefore,

$$E_G(x_1 x_n) = A_{K+1} A_K \cdots A_1$$

Since p_G preserves the marginals of individual variables, $Var_G(x_k) = 1$ for all k . In particular $Var_G(x_1) = Var_G(x_n) = 1$. Then,

$$\rho_G(x_1, x_n) = A_{K+1}A_K \cdots A_1$$

Step 2: Defining a new distribution over scalar variables

For every k , define the variable

$$z_k = (A_{K+1}A_K \cdots A_{k+1})x_{C_k} = \alpha_k x_{C_k}$$

Plugging the recursion (6), we obtain a recursion for z :

$$z_k = \alpha_k x_{C_k} = \alpha_k (A_k x_{C_{k-1}} + \eta_k) = z_{k-1} + \alpha_k \eta_k$$

Given that p is taken to be multivariate normal, the equation for z_k measures the objective conditional distribution ($p_G(z_k | z_{k-1})$). Since p_G does not distort the objective distribution over cliques, ($p_G(z_k | z_{k-1})$) coincides with ($p(z_k | z_{k-1})$). This means that an analyst who fits a model given by the linear DAG $G' : x_1 \rightarrow z_1 \rightarrow \cdots \rightarrow z_K \rightarrow x_n$ will obtain the following estimated model, where every ε_k is a zero-mean scalar variable that is assumed by the analyst to be uncorrelated with the other ε_j 's as well as with z_1, \dots, z_k (and as before, the assumption holds automatically for z_k but is typically erroneous for $z_j, j < k$):

$$\begin{aligned} x_1 &\sim N(0, 1) \\ z_1 &= \alpha_1 A_1 x_1 + \varepsilon_2 \\ &\vdots \\ z_{k+1} &= z_k + \varepsilon_{k+1} \\ &\vdots \\ x_n &= z_K + \varepsilon_n \end{aligned}$$

Therefore, $E_{G'}(x_1, x_n)$ is given by

$$E_{G'}(x_1, x_n) = A_{K+1}A_K \cdots A_1$$

Since G' is perfect, $Var_{G'}(x_n) = 1$, hence

$$\rho_{G'}(x_1, x_n) = A_{K+1}A_K \cdots A_1 = \rho_G(x_1, x_n)$$

We have thus reduced our problem to finding the largest $\hat{\rho}_{1n}$ that can be attained by a linear DAG $G : 1 \rightarrow \cdots \rightarrow n$ of length n at most. ■

4.3 The Final Step: Solving the Reduced Problem

To solve the reduced problem we have arrived at, we first note that

$$\hat{\rho}_{1n} = \prod_{i=1}^{n-1} \rho_{i,i+1} \quad (7)$$

Thus, the problem of maximizing $\hat{\rho}_{1n}$ is equivalent to maximizing the product of terms in a symmetric $n \times n$ matrix, subject to the constraint that the matrix is positive semi-definite, all diagonal elements are equal to one, and the $(1, n)$ entry is equal to r :

$$\begin{array}{l} \max \\ \rho_{ij} = \rho_{ji} \text{ for all } i, j \\ (\rho_{ij}) \text{ is P.S.D} \\ \rho_{ii} = 1 \text{ for all } i \\ \rho_{1n} = r \end{array} \prod_{i=1}^{n-1} \rho_{i,i+1}$$

The positive semi-definiteness constraint is what makes the problem non-trivial. We can arbitrarily increase the value of the objective function by raising off-diagonal terms of the matrix, but at some point this will violate positive semi-definiteness. Since positive semi-definiteness can be rephrased as the requirement that $(\rho_{ij}) = AA^T$ for some matrix A , we can rewrite the constrained maximization problem as follows:

$$\begin{array}{l} \max \\ a_i^T a_i = 1 \text{ for all } i \\ a_1^T a_n = r \end{array} \prod_{i=1}^{n-1} a_i a_{i+1}^T \quad (8)$$

Denote $\alpha = \arccos r$. Since the solution to (8) is invariant to a rotation of all vectors a_i , we can set

$$\begin{aligned} a_1 &= e_1 \\ a_n &= e_1 \cos \alpha + e_2 \sin \alpha \end{aligned}$$

without loss of generality. Note that a_1, a_n are both unit norm and have dot product r . Thus, we have eliminated the constraint $a_1^T a_n = r$ and reduced the variables in the maximization problem to a_2, \dots, a_{n-1} .

Now consider some $k = 2, \dots, n-1$. Fix a_j for all $j \neq k$, and choose a_k to maximize the objective function. As a first step, we show that a_k must be a linear combination of a_{k-1}, a_{k+1} . To show this, we write $a_k = u + v$, where u, v are orthogonal vectors, u is in the subspace spanned by a_{k-1}, a_{k+1} and v is orthogonal to the subspace. Recall that a_k is a unit-norm vector, which implies that

$$\|u\|^2 + \|v\|^2 = 1 \tag{9}$$

The terms in the objective function (8) that depend on a_k are simply $(a_{k-1}^T u)(a_{k+1}^T u)$. All the other terms in the product do not depend on a_k , whereas the dot product between a_k and a_{k-1}, a_{k+1} is invariant to v : $a_{k-1}^T (u+v) = a_{k-1}^T u$.

Suppose that v is nonzero. Then, we can replace a_k with another unit-norm vector $u/\|u\|$, such that $(a_{k-1}^T u)(a_{k+1}^T u)$ will be replaced by

$$\frac{(a_{k-1}^T u)(a_{k+1}^T u)}{\|u\|^2}$$

By (9) and the assumption that v is nonzero, $\|u\| < 1$, hence the replacement is an improvement. It follows that a_k can be part of an optimal solution only if it lies in the subspace spanned by a_{k-1}, a_{k+1} . Geometrically, this means that a_k lies in the plane defined by the origin and a_{k-1}, a_{k+1} .

Having established that a_k, a_{k-1}, a_{k+1} are coplanar, let α be the angle between a_k and a_{k-1} , let β be the angle between a_k and a_{k+1} , and let γ be the (fixed) angle between a_{k-1} and a_{k+1} . Due to the coplanarity constraint,

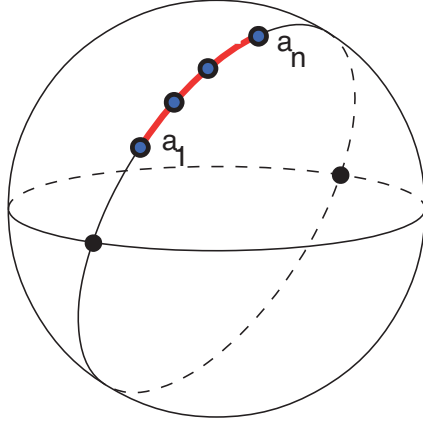


Figure 1: Geometric intuition for the proof.

$\alpha + \beta = \gamma$. Fixing a_j for all $j \neq k$ and applying a logarithmic transformation to the objective function, the optimal α must satisfy

$$\log \cos(\alpha) + \log \cos(\gamma - \alpha)$$

Differentiating this expression with respect to α and setting the derivative to zero, we obtain $\alpha = \beta = \gamma/2$. Since this must hold for *any* $k = 2, \dots, n-1$, we conclude that at the optimum, any a_k lies on the plane defined by the origin and a_{k-1}, a_{k+1} and is at the same angular distance from a_{k-1}, a_{k+1} . That is, an optimum must be a set of equiangular unit vectors on a great circle, equally spaced between a_1 and a_n . The explicit formulas for these vectors are given by (3).

The formula for the upper bound has a simple geometric interpretation (illustrated by the figure). We are given two points on the unit n -dimensional sphere (representing a_1 and a_n) whose dot product is r , and we seek $n - 2$ additional points on the sphere such that the geometric average of the successive points' dot product is maximal. Since the dot product for points on the unit sphere decreases with the spherical distance between them, the problem is akin to minimizing the geometric average of the spherical distances between adjacent points. The solution is to place all the additional points

equidistantly on the great circle that connects a_1 and a_n .

Since by construction, every neighboring points a_k and a_{k+1} have a dot product of $\cos \theta_{r,n}$, we have $\rho_{k,k+1} = \cos \theta_{r,n}$, such that $\hat{\rho}_{1n} = (\cos \theta_{r,n})^{n-1}$. This completes the proof.

5 Conclusion

Many real-life decisions - whether by policy makers, firms or individuals - are guided by models that estimate (informally or explicitly) the correlation between pairs of variables: immigration and unemployment, drug dosage and health outcomes, etc. Since “all models are wrong”, it is important to understand how badly a wrong model can distort correlations.

This paper showed that within the class of models that can be represented by Gaussian Bayesian networks, things can get quite bad indeed. When the model is allowed to include a moderately large number of variables, it can lead decision makers to the conclusion that two variables are almost perfectly correlated, even if in reality they move in opposite directions. Furthermore, such extreme distortions can be generated by models that pass an intuitive “misspecification test” that disqualifies a model if it distorts the marginal distribution of some variable.

The Gaussianity assumption (or, equivalently, the assumption that the analyst assumes linear relationships between variables) plays an important role in this extreme finding. By comparison, if we assume that x_1, \dots, x_n are all binary variables with uniform marginals, then the upper bound on the estimated correlation that chain models can generate is e^{r-1} . See Appendix II for this result.

More broadly, our analysis raises an important question, related to George Box’s famous dictum that while all models are wrong, some are useful: what is the right balance between the costs of cheating with models and the benefits of using simple (and therefore inherently wrong) models?

References

- [1] Chow, C. and C. Liu (1968), Approximating Discrete Probability Distributions with Dependence Trees, *IEEE transactions on Information Theory* 14.3, 462-467.
- [2] Cowell, R., P. Dawid, S. Lauritzen and D. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems*, Springer, London.
- [3] Eliaz, K. and R. Spiegler (2018), A Model of Competing Narratives, *American Economic Review*, forthcoming.
- [4] Eliaz, K., R. Spiegler and Y. Weiss (2019), Cheating with (Recursive) Models, mimeo.
- [5] Esponda, I. and D. Pouzo (2016), Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models, *Econometrica* 84, 1093-1130.
- [6] Eyster, E. and M. Piccione (2013), An Approach to Asset Pricing Under Incomplete and Diverse Perceptions, *Econometrica*, 81, 1483-1506.
- [7] Fleming, T. and D. DeMets (1996), Surrogate End Points in Clinical Trials: Are we Being Misled?, *Annals of internal medicine* 125, 605-613.
- [8] Jehiel, P. (2005), Analogy-Based Expectation Equilibrium, *Journal of Economic Theory*, 123, 81-104.
- [9] Jehiel, P. and F. Koessler (2008), Revisiting Games of Incomplete Information with Analogy-Based Expectations, *Games and Economic Behavior*, 62, 533-557.
- [10] Katz, R. (2004), Biomarkers and Surrogate Markers: An FDA Perspective, *NeuroRx* 1, 189-195.
- [11] Mailath, G. and L. Samuelson (2018), Learning under Diverse World Views: Model Based Inference, *American Economic Review*, forthcoming.

- [12] Koller, D. and N. Friedman. (2009). Probabilistic Graphical Models: Principles and Techniques, MIT Press, Cambridge MA.
- [13] Monteaal Olea, J., P. Ortoleva, M. Pai and A. Prat (2018), Competing Models, mimeo.
- [14] Pearl, J. (2009), Causality: Models, Reasoning and Inference, Cambridge University Press, Cambridge.
- [15] Piccione, M. and A. Rubinstein (2003), Modeling the Economic Interaction of Agents with Diverse Abilities to Recognize Equilibrium Patterns, Journal of the European Economic Association, 1, 212-223.
- [16] Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher and P. Sabeti (2011), Detecting Novel Associations in Large Data Sets, Science, 334(6062), 1518-1524.
- [17] Sloman, S. (2005), Causal Models: How People Think about the World and its Alternatives, Oxford University Press.
- [18] Spiegel, R. (2017), "Data Monkeys": A Procedural Model of Extrapolation From Partial Statistics, Review of Economic Studies 84, 1818-1841.
- [19] Spiegel, R. (2020), Behavioral Implications of Causal Misperceptions, Annual Review of Economics 12, 81-106.

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 $k \in \{3, \dots, n\}$ such that $\text{Var}_G(x_k) \neq 1$ for almost all correlation submatrices $(\rho_{ij})_{i,j=1,\dots,k-1}$ (and therefore, for almost all correlation matrices $(\rho_{ij})_{i,j=1,\dots,n}$).*

Proof. For notational convenience, list the variables x_1, \dots, x_n such that $R(i) \subseteq \{1, \dots, i-1\}$ 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 $\text{Var}_G(x_k) \neq 1$ for almost all correlation submatrices $(\rho_{ij})_{i,j=1,\dots,k-1}$. Since none of the variables x_{k+1}, \dots, x_n appear in the equations for x_1, \dots, 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 $1, \dots, k$, and p is defined over the variables x_1, \dots, x_k .

Let $(\hat{\rho}_{ij})_{i,j=1,\dots,k-1}$ denote the correlation matrix over x_1, \dots, 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, \dots, x_{k-1} are correct, hence $\hat{\rho}_{ii} = 1$ for all $i = 1, \dots, k-1$.

Furthermore, observe that in order to compute $\hat{\rho}_{ij}$ over $i, j = 1, \dots, k-1$, 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 $(\hat{\rho}_{ij})_{i,j=1,\dots,k-1}$ is induced by $(p_G(x_1, \dots, x_{k-1}))$. Each of the terms in the factorization formula for $p_G(x_1, \dots, x_{k-1})$ is of the form $p(x_i | x_{R(i)})$, $i = 1, \dots, k-1$. 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 $(\hat{\rho}_{ij})_{i,j=1,\dots,k-1}$ 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

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

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

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

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

In contrast, the estimated variance of x_k , denoted $\text{Var}_G(x_k)$, obeys the equation

$$\text{Var}_G(x_k) = \beta^T C \beta + \sigma^2 \quad (12)$$

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 $\text{Var}_G(x_k) = 1$. This implies the equation

$$\beta^T A \beta = \beta^T C \beta \quad (13)$$

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

For any subsets $B, B' \subset \{1, \dots, k-1\}$, use $\Sigma_{B \times B'}$ to denote the submatrix of $(\hat{\rho}_{ij})_{i,j=1,\dots,k-1}$ 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 $M \subset \{1, \dots, k-1\} - \{h, l\}$. Therefore, by Drton et al. (2008, p. 67),

$$\Sigma_{\{h\} \times \{l\}} = \Sigma_{\{h\} \times M} \Sigma_{M \times M}^{-1} \Sigma_{M \times \{l\}} \quad (14)$$

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

Now, when we draw the objective correlation submatrix $(\rho_{ij})_{i,j=1,\dots,k-1}$ 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 $(\rho_{ij})_{i,j=1,\dots,k-1}$ 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 $(\rho_{ij})_{i,j=1,\dots,k-1}$ 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 $(\rho_{ij})_{i,j=1,\dots,k-1}$.

Recall once again that we can regard 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 $(p(x_k | x_{R(k)}))$. 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 $\{1, \dots, n\}$ 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 , $i = 1, \dots, n$, 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 $1 \rightarrow 2 \rightarrow \dots \rightarrow n$. 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*

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

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 $G : 1 \rightarrow 2 \rightarrow \dots \rightarrow n$ 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 $p_G(x_1, \dots, x_n) = p(x_1)p(x_2 | x_1) \cdots p(x_n | x_{n-1})$ as a Markov chain. The conditional probability $p_G(x_n | x_1)$ is thus given by a matrix product - specifically, the product of all the transition matrices defined by $p(x_{i+1} | x_i)$. 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 $1 \rightarrow 2 \rightarrow \dots \rightarrow n$ is given by:*

$$\rho_{1n}^* = \left(1 - \frac{1}{n-1}(1-r)\right)^{n-1} \quad (15)$$

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, \dots, a_{n-1} . This implies that for any M :

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

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:

$$\rho_{1n}^* \leq \max_{m_2, \dots, m_{n-1}} \prod_{i=1}^{n-1} \frac{1}{M} (M - 2(m_{i-1} - m_i)) \quad (17)$$

The R.H.S. of (17) should be maximized subject to the constraint that $m_i \in \{0, 1, \dots, M-1\}$, 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

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

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 $M \rightarrow \infty$, 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 $1 \rightarrow 2 \rightarrow 3$ 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 $n \rightarrow \infty$ (compared with 1 in the Gaussian case).