AER: Insights 2021, 3(4): 417–434
https://doi.org/10.1257/aeri.20200635

Cheating with Models[†]

By Kfir Eliaz, Ran Spiegler, and Yair Weiss*

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. (*JEL* D83, C13, C46, C51)

Quantifying the correlation between random variables is a problem that preoccupies decision-makers and scientific researchers alike, for purposes of diagnostication, 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 phenomena, models help perceiving and communicating about 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?*

*Eliaz: School of Economics, Tel-Aviv University, and David Eccles School of Business, University of Utah (email: kfire@tauex.tau.ac.il); Spiegler: School of Economics, Tel Aviv University; Department of Economics, UCL; and CFM (email: rani@tauex.tau.ac.il); Weiss: School of Computer Science and Engineering, Hebrew University (email: yweiss@cs.huji.ac.il). Larry Samuelson was coeditor for this article. A longer working-paper version of this paper has appeared under the title "Cheating with (Recursive) Models" (Eliaz, Spiegler, and Weiss 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 the referees of this journal, and especially Heidi Thysen, for helpful comments.

[†]Go to https://doi.org/10.1257/aeri.20200635 to visit the article page for additional materials and author disclosure statement(s).

There are several reasons to be interested in this question. First, individuals and policymakers are often guided by models. 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 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 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 2020 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 policymaker with a particular agenda (for example, showing that cutting taxes fosters economic growth), had staked their reputation on a claim that strongly links two variables, or may want to make a splash with a strong finding. Such incentives may lead to (possibly subconscious) self-serving model selection. For expositional convenience, we 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 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 models that take the form of *directed acyclic graphs* (DAGs). 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, Spiegler (2017, 2020) showed how important families of misspecified models in the literature (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, the objective (Pearson) correlation between the target variables is r, and 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 $n \to \infty$, the expression converges to 1 for *any* $r \in [-1, 1)$.

An Example: "Marker 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 $x_1 \to x_2 \to x_3$.

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 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 $x_1 \to x_2 \to x_3$ 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.

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 $\rho_{13} = 0$—that is, 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 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 $\rho_{12}\rho_{23} > 0$.

We refer to this opportunistic selection of x_2 as "*marker hacking*." The literature on surrogate markers refers to the possibility that $\hat{\rho}_{13} > 0$ even though $\rho_{13} \leq 0$ as the "*surrogate paradox*" (for example, see VanderWeele 2015, pp. 217–28). Yet we are unaware of previous writings on the possibility that researchers will leverage

the paradox via marker hacking or on the magnitude of the errors to which marker hacking can lead.

This brings us to the following question: 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 semidefinite. This property is characterized by the inequality $(\rho_{12})^2 + (\rho_{23})^2 \leq 1$. The maximal value of $\rho_{12}\rho_{23}$ subject to this constraint is $1/2$, hence this is the maximal false correlation the model can generate. This bound is attained if we define x_2 to be a deterministic function of x_1 and x_3, $x_2 = (x_1 + x_3)/\sqrt{2}$. 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 marker 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 many 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."

I. The Model

Let p be an objective probability measure over n variables, $x_1, \ldots, x_n$. Assume p is multivariate normal. For every $A \subset \{1, \ldots, n\}$, denote $x_A = (x_i)_{i \in A}$. 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 $x_2, \ldots, x_{n-1}$ are chosen opportunistically by

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

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 $N = \{1,\ldots,n\}$. With some abuse of notation, $R(i)$ is the set of nodes j for which the DAG includes a link $j \to i$. We restrict attention to *acyclic* directed graphs (DAGs)—that is, graphs that do not include directed paths $i \to j \to \cdots \to i$.

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 Bayesian networks (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. For example, the DAG $x_1 \to x_2 \to x_3$ represents the assumption $x_1 \perp x_3 \,|\, x_2$. 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 $x_1,\ldots,x_n$ and a DAG G, define the Bayesian-network factorization formula:

$$p_G(x_1,\ldots,x_n) = \prod_{k=1}^{n} p\left(x_k \,|\, x_{R(k)}\right). \tag{2}$$

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

$$p_G(x_1,x_2,x_3) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_2).$$

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 right-hand side 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

[2] Thus, as in Spiegler (2017), G can be viewed as a function that systematically transforms any objective distribution p into a subjective distribution p_G.

$\text{var}_G(x_k)$ and $\text{cov}_G(x_i, x_j)$ 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 $i = 1, \ldots, n$. 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 diagnostication 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 conditional or unconditional correlations. However, monitoring *individual* variables is a much simpler task than monitoring correlations. For example, 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 $p\left(x_k \mid x_{R(k)}\right)$ 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} terms 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 right-hand-side variable on one equation cannot appear as the left-hand-side 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, chapter 7).

For instance, consider the DAG $G : x_1 \to x_2 \to x_3$. The terms $p(x_2 \mid x_1)$ and $p(x_3 \mid x_2)$ are given by OLS estimators of the linear regression equations

$$x_2 = \alpha x_1 + \varepsilon_2,$$

$$x_3 = \beta x_2 + \varepsilon_3,$$

such that

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

We will see in the proof of our main result (see Section IIIB) that this means $\hat{\rho}_{13} = \alpha\beta = \rho_{12}\rho_{23}$.

II. 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*

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

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

(*i*) *A DAG* $1 \to 2 \to \cdots \to n$.

(*ii*) *A Gaussian distribution satisfying*, *for every* $k = 1, \ldots, n$:

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

where s_1, s_2 *are independent standard normal variables.*

The following simple observation establishes that when n is allowed to be arbitrarily large, $\hat{\rho}_{1n}$ can get arbitrarily close to 1, irrespective of the objective correlation ρ_{1n}.

REMARK 1: *For any* $r \in [-1, 1]$, $\lim_{n\to\infty}\left(\cos\theta_{r,n}\right)^{n-1} = 1$.

The Gaussianity assumption plays a key role in this extreme finding. By comparison, if we assumed that $x_1, \ldots, x_n$ are all binary variables with uniform marginals, then the upper bound on $\hat{\rho}_{1n}$ that chain models can generate is e^{r-1} (see the online Appendix).

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.

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 $x_2, \ldots, x_n$ 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. (Using the same chain

model but a different selection of the variables $x_2, \ldots, x_n$, the analyst can generate any value of $\hat{\rho}_{1n}$ in the interval $\left[r, \left(\cos\theta_{r,n}\right)^{n-1}\right]$—see an explanation at the end of Section III.)

We interpret the genericity aspect of the theorem as follows. Recall that we think of our analyst as effectively choosing the variables $x_2, \ldots, x_{n-1}$ 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 versus Model Selection.—Suppose the analyst selects the variables $x_2, \ldots, x_{n-1}$ 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 $\left|R(k)\right| \leq 1$. Then, it is easy to show that when $r > 0$ and $x_2, \ldots, x_{n-1}$ are selected as in (3), the Chow-Liu algorithm will select the model given by the chain $1 \to 2 \to \cdots \to n$. 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 $1 \to 2 \to \cdots \to n$ 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* $\hat{\rho}_{1n} \leq \left(\cos\theta_{r,n}\right)^{n-1}$ *for every objective correlation matrix* $\left(\rho_{ij}\right)$ *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.

III. 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 invokes genericity. To give a rough intuition for this step, consider the imperfect DAG $G: 1 \rightarrow 3 \leftarrow 2$. UMC requires $\mathrm{var}_G(x_3) = 1$. To satisfy this condition, it can be shown that $\hat{\rho}_{12}$ should coincide with ρ_{12}. But since G assumes $x_1 \perp x_2$, $\hat{\rho}_{12} = 0$. The resulting equation $\rho_{12} = 0$ is violated by generic (ρ_{ij}).

Second, we 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 that form cliques with single "mega-nodes." The Gaussianity assumption means that each term $p(x_k | x_{R(k)})$ in p_G is given by a linear regression equation, such that $x_{R(k)}$ affects x_k only via a scalar variable that is a linear combination of the variables in $R(k)$. This enables us to mimic the "mega-nodes" by scalar Gaussian variables.

Thus, in order to calculate the upper bound on $\hat{\rho}_{1n}$, we can restrict attention to chain models involving univariate normal variables. 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 semidefinite (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.

A. *The First Step: From DAGs to Perfect DAGs*

We first establish that for generic Gaussian p, perfection is necessary for UMC. Because p is Gaussian, p_G is Gaussian as well (Koller and Friedman 2009, chapter 7). Therefore, the marginal of p_G over x_k is given entirely by its mean and variance, denoted $E_G(x_k)$ and $\mathrm{var}_G(x_k)$.

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

PROOF:

For notational convenience, renumber the variables $x_1, \ldots, x_n$ such that $R(i) \subseteq \{1, \ldots, i-1\}$ for every i. (This is legitimate because at this stage of our proof, x_1 and x_n lack a fixed meaning yet.) Consider the lowest k for which $R(k)$ is not a clique. This means that there exists two nodes $h, l \in R(k)$, $h < l$, such that h and l are unlinked in G, whereas for every $k' < k$ and every $h', l' \in R(k')$, h' and l' are linked in G.

Because p is Gaussian, the conditional distribution $\left(p\left(x_k \mid x_{R(k)}\right)\right.$ is given by a linear regression equation:

$$x_k = \sum_{i \in R(k)} \beta_{ik} x_i + \varepsilon_k. \tag{4}$$

Denote $\beta = \left(\beta_{ik}\right)_{i \in R(k)}$. Let A denote the correlation submatrix $\left(\rho_{ij}\right)_{i,j \in R(k)}$ that fully characterizes $\left(p\left(x_{R(k)}\right)\right)$. Then,

$$\operatorname{var}\left(x_k\right) = 1 = \beta^T A \beta + \sigma^2, \tag{5}$$

where $\sigma^2 = \operatorname{var}\left(\varepsilon_k\right)$. In contrast, the estimated variance of x_k, denoted $\operatorname{var}_G\left(x_k\right)$, obeys the equation

$$\operatorname{var}_G\left(x_k\right) = \beta^T A_G \beta + \sigma^2, \tag{6}$$

where A_G denotes the correlation submatrix $\left(\hat{\rho}_{ij}\right)_{i,j \in R(k)}$ that characterizes $\left(p_G\left(x_{R(k)}\right)\right)$. 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. UMC requires $\operatorname{var}_G\left(x_k\right) = 1$. This implies the equation

$$\beta^T A \beta = \beta^T A_G \beta. \tag{7}$$

We now wish to show that this equation fails for generic $\left(\rho_{ij}\right)_{i,j=1,\ldots,k-1}$. The proof is based on a standard result in the Bayesian-network literature: if a Gaussian distribution p with correlation matrix ρ is consistent with a DAG G (that is, $p_G = p$), then $\rho_{kl}^{-1} = 0$ for any k, l such that x_k and x_l are independent conditional on all the other variables under p_G (Koller and Friedman 2009, pp. 69, 251). Thus, if we choose the elements of ρ at random, then almost surely $\rho_{kl}^{-1} \neq 0$, and so for almost any ρ, $p_G \neq p$.

Now, consider the subgraph of G restricted to the nodes $1, \ldots, l$. By definition, this subgraph is perfect, l is a terminal node, and there is no link between h and l. Applying the rules of d-separation (see Koller and Friedman 2009, pp. 69–72), $x_h \perp x_l \mid x_{\{1,\ldots,l-1\}-\{h\}}$ under p_G. Thus, the (h,l) entry in the inverse of the covariance matrix of $p_G\left(x_1, \ldots, x_l\right)$ must be *exactly* zero. Hence, for generic ρ, $A \neq A_G$.

When we draw the objective correlation submatrix $\left(\rho_{ij}\right)_{i,j=1,\ldots,k}$ at random, we can think of it as a two-stage lottery. In the first stage, we draw the correlation submatrix over $x_1, \ldots, x_{k-1}$. In the second stage, we draw the vector β that defines the correlation between x_k and the preceding $k-1$ variables. Since $A_G \neq A$ for generic ρ, (7) is a nontautological quadratic equation of β (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 $\operatorname{var}_G\left(x_k\right) = 1$ is violated by almost every $\left(\rho_{ij}\right)$. ∎

COROLLARY 1: *For almost every* $\left(\rho_{ij}\right)$, *if a DAG* G *satisfies* $E_G\left(x_k\right) = 0$ *and* $\operatorname{var}_G\left(x_k\right) = 1$ *for all* $k = 1, \ldots, n$, *then* G *is perfect.*

PROOF:

By Lemma 1, 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 $\{1,\ldots,n\}$ 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. ∎

B. *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 $1 \rightarrow \cdots \rightarrow n$. Every linear DAG is perfect by definition.

LEMMA 2: *For every Gaussian distribution with correlation matrix ρ and nonlinear 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):

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

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—that is, $1, n \notin C$ 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 $C_1, \ldots, C_K$ are maximal, it follows that every C_k along the sequence must introduce at least one new element $i \notin \bigcup_{j<k} C_j$ (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 *nonlinear*, 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 $1 \rightarrow C_1 \rightarrow \cdots \rightarrow C_K \rightarrow n$, as follows:

$$p_G\left(x_1, x_{C_1}, \ldots, x_{C_K}, x_n\right) = p\left(x_1\right) \prod_{k=1}^{K} p\left(x_{C_k} \mid x_{C_{k-1}}\right) p\left(x_n \mid x_{C_K}\right), \tag{8}$$

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 (that is, univariate) normal variables without changing $\mathrm{cov}_G\left(x_1, x_n\right)$.

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 $p_G\left(x_1, x_{C_1}, \ldots, x_{C_K}, x_n\right)$ via the following recursion:

$$\begin{aligned}
x_1 &\sim N(0,1), \\
x_{C_1} &= A_1 x_1 + \eta_1, \\
&\vdots \\
x_{C_k} &= A_k x_{C_{k-1}} + \eta_k, \\
&\vdots \\
x_{C_K} &= A_K x_{C_{K-1}} + \eta_K, \\
x_n &= A_{K+1} x_{C_K} + \eta_n,
\end{aligned} \tag{9}$$

where each equation describes an objective conditional distribution—in particular, the equation for x_{C_k} describes $\left(p\left(x_{C_k} \mid x_{C_{k-1}}\right)\right)$. The matrices A_k are functions of the vectors β_i in the original model. The η_k terms are all zero mean and uncorrelated with the explanatory variables $x_{C_{k-1}}$, such that $E\left(x_{C_k} \mid x_{C_{k-1}}\right) = A_k x_{C_{k-1}}$.

Furthermore, according to p_G (that is, the analyst's estimated model), each x_k is conditionally independent of $x_1, \ldots, x_{k-1}$ given $x_{R(k)}$. (Here we use the earlier result that we can enumerate the variables such that $R(k) \subseteq \{1, \ldots, k-1\}$ for every k.) Since the junction-tree factorization (8) represents exactly the same distribution p_G, this means that every η_k is uncorrelated with all other η_j terms as well as with $x_1, \ldots, x_{C_{k-2}}$. Therefore,

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

Since p_G preserves the marginals of individual variables, $\mathrm{var}_G(x_k) = 1$ for all k. In particular $\mathrm{var}_G(x_1) = \mathrm{var}_G(x_n) = 1$. 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 in the recursion (9), 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 $\big(p_G(z_k \,|\, z_{k-1})\big)$. Since p_G does not distort the objective distribution over cliques, $\big(p_G(z_k \,|\, z_{k-1})\big)$ coincides with $\big(p(z_k \,|\, z_{k-1})\big)$. This means that an analyst who fits a model given by the linear DAG $G' : x_1 \to z_1 \to \cdots \to z_K \to x_n$ 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 terms as well as with $z_1, \ldots, z_k$ (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, $\text{var}_{G'}(x_n) \;=\; 1$, 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. ∎

C. *The Final Step: Solving the Reduced Problem*

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

$$\hat{\rho}_{1n} \;=\; \prod_{i=1}^{n-1} \rho_{i,i+1}. \tag{10}$$

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 semidefinite, all diagonal elements are equal to one, and entry $(1, n)$ is equal to r:

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

The positive semidefiniteness constraint is what makes the problem nontrivial. 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 semidefiniteness. Since positive semidefiniteness can be rephrased as the requirement that $(\rho_{ij}) \;=\; AA^T$ for some matrix A, we can rewrite the constrained maximization problem as follows:

$$\max_{\substack{a_i^T a_i=1 \text{ for all } i \\ a_1^T a_n=r}} \prod_{i=1}^{n-1} a_i a_{i+1}^T. \tag{11}$$

Denote $\alpha \;=\; \arccos r$. Since the solution to (11) is invariant to a rotation of all vectors a_i, we can set

$$a_1 \;=\; e_1,$$

$$a_n \;=\; e_1 \cos\alpha + e_2 \sin\alpha,$$

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 $a_2, \ldots, a_{n-1}$.

Now consider some $k = 2, \ldots, n-1$. 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

$$\| u \|^2 + \| v \|^2 = 1. \tag{12}$$

The terms in the objective function (11) that depend on a_k are simply $\left(a_{k-1}^T u\right)\left(a_{k+1}^T u\right)$. 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 $\left(a_{k-1}^T u\right)\left(a_{k+1}^T u\right)$ will be replaced by

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

By (12) 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, $\alpha + \beta = \gamma$. Fixing a_j for all $j \neq k$ and applying a logarithmic transformation to the objective function, the optimal α must maximize $\log\cos(\alpha) + \log\cos(\gamma - \alpha)$. 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* $k = 2, \ldots, n-1$, 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. 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 $\hat{\rho}_{1n} = \left(\cos\theta_{r,n}\right)^{n-1}$. ∎

Comment: The Set of Attainable $\hat{\rho}_{1n}$.—Throughout the paper, we focused on the *maximal* level that $\hat{\rho}_{1n}$ can get, given r, n. Does it follow that *any* value of $\hat{\rho}_{1n}$ between r and the upper bound $\left(\cos\theta_{r,n}\right)^{n-1}$? The final step in our proof confirms that the answer is affirmative. Consider our geometric construction, and gradually

shift one of the interior points toward one of its adjacent points. By continuity of the objective function in (11), at some stage (before the two points perfectly coincide), $\hat{\rho}_{1n}$ will coincide with the upper bound on $\hat{\rho}_{1,n-1}$, which is strictly lower than the upper bound on $\hat{\rho}_{1n}$. Thus, using the chain model $1 \rightarrow \cdots \rightarrow n$, the analyst can select the intermediate variables $x_2, \ldots, x_{n-1}$ to attain any $\hat{\rho}_{1n} \in \left[r, \left(\cos\theta_{r,n}\right)^{n-1}\right]$.

IV. Conclusion

Many real-life decisions by policymakers, firms, or individuals are guided by models that estimate (informally or explicitly) correlations 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 pairwise correlations. This paper showed that within the class of models represented by Gaussian Bayesian networks, things can get bad indeed. When the model includes a moderately large number of variables, it can lead decision-makers to conclude 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," which disqualifies a model if it distorts the marginal distribution of some variable.

Our analysis raises an important open 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 usually wrong) models?

REFERENCES

Caron, Richard, and Tim Traynor. 2005. "The Zero Set of a Polynomial." WSMR Report 05-02. Chow, C., and C. Liu. 1968. "Approximating Discrete Probability Distributions with Dependence Trees." *IEEE Transactions on Information Theory* 14 (3): 462–67.

Cowell, Robert G., A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.

Eliaz, Kfir, and Ran Spiegler. 2020. "A Model of Competing Narratives." *American Economic Review* 110 (12): 3786–816.

Eliaz, Kfir, Rani Spiegler, and Yair Weiss. 2019. "Cheating with (Recursive) Models." https://ssrn.com/abstract=3486251.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica* 84 (3): 1093–1130.

Eyster, Erik, and Michele Piccione. 2013. "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions." *Econometrica* 81 (4): 1483–506.

Jehiel, Philippe. 2005. "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory* 123 (2): 81–104.

Jehiel, Philippe, and Frédéric Koessler. 2008. "Revisiting Games of Incomplete Information with Analogy-Based Expectations." *Games and Economic Behavior* 62 (2): 533–57.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press.

Mailath, George, and Larry Samuelson. 2020. "Learning under Diverse World Views: Model Based Inference." *American Economic Review* 110 (5): 1464–501.

Monteal Olea, Jose Luis, Pietro Ortoleva, Mallesh M. Pai, and Andrea Prat. 2018. "Competing Models." Unpublished.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference.* Cambridge: Cambridge University Press.

Piccione, Michele, and Ariel Rubinstein. 2003. "Modeling the Economic Interaction of Agents with Diverse Abilities to Recognize Equilibrium Patterns." *Journal of the European Economic Association* 1 (1): 212–23.

Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. 2011. "Detecting Novel Associations in Large Data Sets." *Science* 334 (6062): 1518–24.

Sloman, Steven. 2005. *Causal Models: How People Think about the World and Its Alternatives*. Oxford: Oxford University Press.

Spiegler, Ran. 2017. "'Data Monkeys': A Procedural Model of Extrapolation from Partial Statistics." *Review of Economic Studies* 84 (4): 1818–41.

Spiegler, Ran. 2020. "Behavioral Implications of Causal Misperceptions." *Annual Review of Economics* 12: 81–106.

VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.