The Foerder Institute for Economic Research
at Tel Aviv University

The Eitan Berglas School of Economics

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

מכון למחקר כלכלי על שם
ד"ר ישעיהו פורדר
על יד אוניברסיטת תל אביב

בית-הספר לכלכלה ע"ש איתן ברגלס

עמותה רשומה

Cheating with (Recursive) Models

Kfir Eliaz, Ran Spiegler and Yair Weiss

Working Paper No. 8-2020

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

Cheating with (Recursive) Models*

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

May 15, 2020

Abstract

To what extent can misspecified subjective models distort correlations? We study an "analyst" who utilizes models that take the form of a recursive system of linear regression equations. The analyst fits each equation to an objective empirical distribution. We characterize the maximal pairwise correlation that the analyst's model can predict given a generic objective covariance matrix, subject to the constraint that the estimated model does not distort the mean and variance of individual variables. We show that as the number of variables in the model grows, the estimated pairwise correlation can become arbitrarily large, regardless of the objective correlation.

[*] Financial support from ERC Advanced Investigator grant no. 692995 is gratefully acknowledges. Eliaz and Spiegler thank Briq and the Economics Department at Columbia University for their generous hosptiality 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 preoccupies decision makers and scientific researchers, for purposes of diagnostication, prediction or causal inference. In some cases, agents are 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 origins. Belief in a model allows us to extrapolate from incomplete 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 - often in the form of memorable "narratives".

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). We pose 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 (which, to our knowledge, this paper is the first to pose). First, when analyzing the behavior of decision makers with misspecified models, we would like to know how large their errors can get. Our paper thus introduces worst-case analysis into the literature on decision making under misspecified models.[1] Second, politicians use false narratives (which can be regarded as misspecified causal models) to exaggerate the perceived impact of policies and attribute credit/blame for social outcomes (Eliaz and Spiegler (2018)). Our exercise helps quantifying the extent to which they can do so. Third, 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".[2] 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 decision maker with a

[1] See Piccione and Rubinstein (2003), Jehiel (2005), Esponda and Pouzo (2016) for a few examples. Spiegler (2019) reviews this literature through the prism of causal models.

[2] This idea is related to the notion of "competing models" in Montiel Olea et al. (2018).

particular agenda (e.g. a study demonstrating that taxation has an adverse affect on economic growth would be of use for a policy maker with a tax-cutting agenda). Alternatively, they may have staked their reputation on a claim that strongly links two variables. Or they may want to make a splash with a strong finding. Such incentives may lead to (possibly subconscious) self-serving model selection.

For expositional purposes, we will focus on the latter, "bad researcher" metaphor throughout the paper. In our model, an *analyst* wishes to demonstrate to a lay audience that x and y are strongly associated (we will bounce between the diagnostic, predictive and causal interpretations of this association). The analyst has statistical data about the joint behavior of many variables, including x and y. His method is to propose a model, fit it to the data and use the estimated model to compute the correlation between x and y. 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.

We assume that the analyst is restricted to use *recursive linear-regression models* ("recursive models" henceforth). A model in this class consists of a list of linear-regression equations, such that an explanatory variable in one equation cannot appear as a dependent variable in another equation down the list. We assume that the recursive model includes the variables x and y, as well as a selection of up to $n-2$ additional variables. Thus, the total number of variables in the analyst's model is n, which is a natural measure of the model's complexity. Each equation is estimated via Ordinary Least Squares (OLS).

We examine this class of models for a number of reasons. First, recursive models are a convenient object of study because they would be familiar to readers from their elementary econometrics studies (as a special case of simultaneous equations). Second, outside economics, recursive models are often known as Gaussian Bayesian Networks. The structure of a recursive model can be described by a directed acyclic graph (DAG). At least since

Pearl (2009), DAGs are the leading formalization of causal models. Indeed, when x appears exclusively as an explanatory variable in the system of equations while y appears exclusively as a dependent variable, the recursive model intuitively charts a causal explanation that pits x as a primary cause of y, such that the estimated correlation between x and y can be interpreted as an estimated causal effect of x on y. Our analysis characterizes the maximal false effect that a misspecified causal model can generate in Gaussian environments.

Finally, in their simplest forms (which we will encounter in Sections 1.1 and 5), recursive models are regularly employed by empirical researchers. But rather than viewing our analysis in terms of its direct applicability for practitioners, we invite the reader to think about it more abstractly, as a *model* of how researchers (and other types of agents) misuse models. From this point of view, the class of recursive models is a stylized representation of an aspect of empirical research, not a literal description of its current practice. Our approach is akin to the way economists capture imperfect competition with abstract models like Cournot/Bertrand competition or repeated games - except that the objects of our abstraction are researchers rather than firms. By way of analogy, just as the large theoretical literature on Bayesian persuasion (e.g. Dworczak and Martini (2019)) examines the extent to which agents' beliefs can be manipulated by prior commitment to an information structure, our exercise examines the extent to which beliefs can be manipulated by misspecified models.

1.1 An Example: Fishing for Surrogate Markers

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 estimates the following three-variable recursive model:

$$\begin{aligned} x_1 &= \varepsilon_1 \\ x_2 &= \beta_1 x_1 + \varepsilon_2 \\ x_3 &= \beta_2 x_2 + \varepsilon_3 \end{aligned} \tag{1}$$

where x_1, x_2, x_3 are normalized to have zero mean and unit variance. The analyst assumes that all the ε_k's are mutually uncorrelated, and also that for every $k > 1$ and $j < k$, ε_k is uncorrelated with x_j (for $j = k - 1$, this is mechanically implied by the OLS method).

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 (1) 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 misinterpreting biomarker data.

Let ρ_{ij} denote the correlation between x_i and x_j according to the *true* data-generating process. 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 procedure and its underlying assumptions, is $\hat{\rho}_{13} = \rho_{12}\rho_{23}$. 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 come at the expense of raising ρ_{23}. Formally, consider the true correlation matrix:

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

By definition, this matrix is positive semi-definite. 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 $\frac{1}{2}$, hence this is the maximal false correlation that the model can generate. This bound is tight: It can be 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 recursive 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 "fishing for markers" 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 give a concrete demonstration for the latter claim, consider Figure 1, which is extracted from a database compiled by the World Health Organization and collected by Reshef et al. (2011).[3] All the variables are taken from this database. The figure displays the maximal $\hat{\rho}_{13}$ correlation that the model (1) can generate for two fixed pairs of variables x_1 and x_3 with ρ_{13} close to zero, when the auxiliary variable is selected from a pool whose size is given by the horizontal axis (variables were added to the pool in a particular order). When the analyst can choose x_2 from only ten possible auxiliary variables, the estimated correlation between x_1 and x_3 he can generate with (1) is still modest. In contrast, once the pool size is in the hundreds, the

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

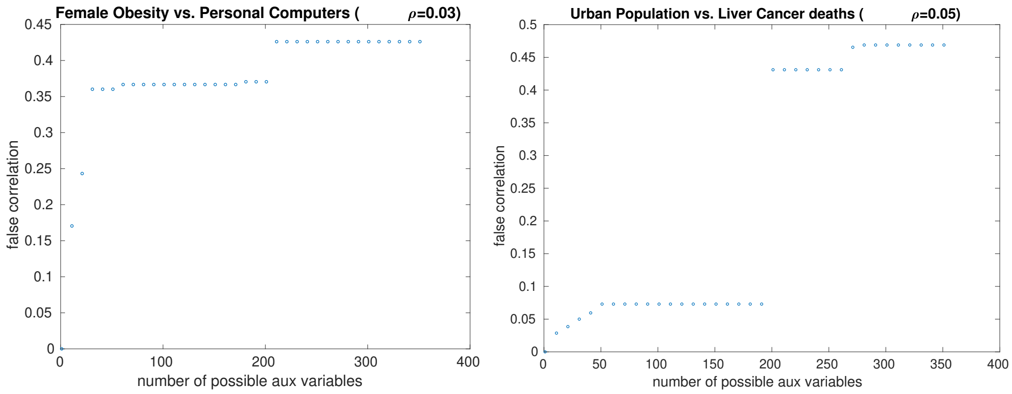

Figure 1: False correlation in a recursive model with one auxiliary variable, as a function of the number of possible auxiliary variables the researcher can choose from. All variables and their correlations are taken from a database compiled by the World Health Organization. Even though the true correlation is close to zero in both cases, as the number of possible auxilary variable increases, the estimated correlation rises yet never exceeds 0.5.

maximal estimated correlation approaches the upper bound (slightly above $\frac{1}{2}$).

For a specific variable that gets us near the theoretical upper bound, consider the figure's R.H.S, where x_1 represents urban population and x_3 represents liver cancer deaths. The true correlation between these variables is 0.05. If the analyst selects x_2 to be coal consumption, the estimated correlation between x_1 and x_3 is 0.43, far above the objective value. 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. This numerical illustration shows that the upper bound on $\hat{\rho}_{13}$ may have real-life relevance when researchers have substantial freedom to "fish for markers"

1.2 Preview of the Analysis

We present our model in Section 2 and pose the main question: What is the largest estimated correlation $\hat{\rho}_{1n}$ that an n-variable recursive model can generate? We impose one constraint: *The estimated model cannot distort individual variables' mean and variance.* We interpret this constraint as an elementary “misspecification test” that an unsophisticated decision maker can implement, since monitoring individual variables (unlike their comovement) is relatively straightforward.[4]

In Section 3, we state our main result. For a generic objective correlation matrix with $\rho_{1n} = r$, the maximal $\hat{\rho}_{1n}$ that an n-variable recursive model can generate, subject to the correct-variance constraint, is

$$\left(\cos \left(\frac{\arccos r}{n-1} \right) \right)^{n-1} \tag{2}$$

The upper bound is tight. It is attained by the simplest connected n-variable model: For every $k = 2, ..., n$, x_k is regressed on x_{k-1} only. This model is represented graphically by the chain $x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_n$. The variables $x_2, ..., x_{n-1}$ are all deterministic linear functions of x_1 and x_n.

Formula (2) reproduces the value $\hat{\rho}_{13} = \frac{1}{2}$ from Section 1.1 when $r = 0$, and it is strictly increasing in n. When $n \rightarrow \infty$, the expression converges to 1 for *any* r. That is, regardless of the objective correlation between x_1 and x_n, a sufficiently large recursive model can generate an arbitrarily large estimated correlation. Furthermore, if the analyst only uses discretion when selecting the variables and then applies a sparsity-seeking model-discovery tool like the Chow-Liu algorithm (Chow and Liu (1968)), the algorithm will “discover” the above chain.

The detailed proof of our main result is presented in Section 4. It relies on the graphical representation of recursive models and employs tools from the Bayesian-networks literature (Cowell et al. (1999), Koller and Friedman (2009)).

In Section 5 we relax the correct-variance constraint. Our motivation for doing so is the use of recursive models for estimating the *causal* effect of x_1 on x_n, when their observed correlation is claimed to be contaminated by confounding. We analyze the special case of models that consist of a single non-degenerate regression equation. We show that the maximal causal effect (standardized to be comparable to Pearson correlation) is $1/\sqrt{2 - r^2}$. The model and variables that implement this bound are an instance of “*bad*

[4] The model (1) in the example of Section 1.1 satisfies the constraint.

controls".

Finally, in Section 6, we present partial analysis of our question for a non-Gaussian Bayesian network that involves uniformly distributed *binary* variables. Compared with the Gaussian case, the chain model $x_1 \to x_2 \to \cdots \to x_n$ is more limited in its ability to produce false correlations. For example, when $\rho_{1n} = 0$, the maximal estimated correlation given this model converges to $1/e$ (from below) when $n \to \infty$.

2 The Model

Let p be an objective probability measure over n variables, $x_1, ..., x_n$. We interpret p as a "true" underlying distribution, or as an empirical distribution given by a finite random sample. For every $A \subset \{1, ..., n\}$, denote $x_A = (x_i)_{i \in A}$. Assume that the marginal of p on each of these variables has *zero mean and unit variance.* This will entail no loss of generality for our purposes. 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 covariance matrix that characterizes p is therefore (ρ_{ij}).

An analyst estimates a recursive model that involves these variables. This model consists of a system of linear-regression equations. For every $k = 1, ..., n$, the k^{th} equation takes the form

$$x_k = \sum_{j \in R(k)} \beta_{jk} x_j + \varepsilon_k$$

where:

- $R(k) \subseteq \{1, .., k-1\}$. This restriction captures the model's recursive structure: An explanatory variable in one equation cannot appear as a dependent variable in a later equation.

- In the k^{th} equation, the β_{jk}'s are parameters to be estimated against p. The analyst assumes that each ε_k has zero mean and that it is uncorrelated with all other ε_j's, as well as with $(x_j)_{j<k}$. The vector of

coefficients $\beta_k = (\beta_{jk})_{j \in R(k)}$ is selected to minimize the mean squared error of the k^{th} regression equation, which gives the standard Ordinary Least Squares estimate:

$$\beta_k = \rho_{k,R(k)} \left(\rho_{R(k),R(k)} \right)^{-1} \tag{3}$$

where $\rho_{k,R(k)}$ denotes the row of correlations between x_k and each of the explanatory variables x_j, $j \in R(k)$; and $\rho_{R(k),R(k)}$ denotes the submatrix of the correlations among the explanatory variables.

We refer to such a system of regression equations as an *n-variable recursive model.* The function R effectively defines a directed acyclic graph (DAG) over the set of nodes $\{1, ..., n\}$, such that a link $i \to j$ exists whenever $i \in R(j)$. We will make use of the DAG representation in the proof of our main result, as well as in Section 6. DAGs are often interpreted as causal models (see Pearl (2009)). We will occasionally invoke the causal interpretation. However, in this section and the next, we will not explicitly address the question of causal inference, and rather emphasize the model's diagnostic and predictive functions.

Note that in the analyst's model, the partial ordering given by R is consistent with the natural enumeration of variables (i.e., $i \in R(j)$ implies $i < j$). In particular, the equation for x_1 has no explanatory variables, and x_n is not an explanatory variable in any equation. This restriction is made for notational convenience; relaxing it would not change our results. However, it has the additional advantage that the causal interpretation of the model-estimated correlation between x_1 and x_n is sensible. Indeed, it is legitimate according to Pearl's (2009) rules for causal inference based on DAG-represented models. We will develop this point further in Section 5.

The analyst's assumption that ε_k is uncorrelated with $x_{R(k)}$ is redundant, because it is an automatic consequence of his OLS procedure for estimating β_k. In contrast, his assumption that ε_k is uncorrelated with *all other* x_j, $j < k$, is the basis for how he combines the individually estimated equations into a joint estimated distribution. It is fundamentally a conditional-independence assumption - namely, that x_k is independent of $(x_j)_{j\in\{1,..,k-1\}-R(k)}$ conditional on $x_{R(k)}$. This assumption will typically be false - indeed, it is what makes his model misspecified and what enables the analyst to "cheat" with his model.

Under this assumption, the analyst proceeds to estimate the correlation between x_1 and x_n, according to the following recursive procedure. Consider the n^{th} equation. For every $j \in R(n) - \{1\}$, replace x_j with the R.H.S of the j^{th} equation. This produces a new equation for x_n, with a different set of explanatory variables. Repeat the substitution for each one of these variables (except x_1) until the only remaining explanatory variable is x_1. The procedure's final output is the equation

$$x_n = \alpha x_1 + \sum_{j=1}^{n} \gamma_j \varepsilon_j$$

The coefficients $\alpha, \gamma_1, ..., \gamma_n$ are combinations of β parameters (which were obtained by OLS estimation of the individual equations). Likewise, the distribution of each error term ε_j is taken from the estimated j^{th} equation.[5]

The analyst uses this equation to estimate the variance of x_n and its covariance with x_1, implementing the normalization that all the x_k's have zero mean and unit variance, and his (partly erroneous) assumption that the ε_k's have zero covariance with x_1 and among themselves:

$$\widehat{Var}(x_n) = \alpha^2 \cdot Var(x_1) + \sum_{j=1}^{n} (\gamma_j)^2 Var(\varepsilon_j) = \alpha^2 + \sum_{j=1}^{n} (\gamma_j)^2 Var(\varepsilon_j)$$

and

$$\widehat{Cov}(x_1, x_n) = E\left[x_1 \left(\alpha x_1 + \sum_{j=1}^{n} \gamma_j \varepsilon_j\right)\right] = \alpha \cdot Var(x_1) + \sum_{j=1}^{n} \gamma_j E(x_1 \varepsilon_j) = \alpha$$

[5] E.g., suppose $x_2 = \beta_{12}x_1 + \varepsilon_2$ and $x_3 = \beta_{13}x_1 + \beta_{23}x_2 + \varepsilon_3$. Then, $\alpha = \beta_{23}\beta_{21} + \beta_{13}$, $\gamma_2 = \beta_{23}$ and $\gamma_3 = 1$.

Therefore, the estimated coefficient of correlation between x_1 and x_n is

$$\hat{\rho}_{1n} = \frac{\widehat{Cov}(x_1, x_n)}{\sqrt{Var(x_1)\widehat{Var}(x_n)}} = \frac{\alpha}{\sqrt{\widehat{Var}(x_n)}} \tag{4}$$

In the Introduction, we discussed why this method for estimating $\hat{\rho}_{1n}$ can be "legitimized". First, it may be impossible to directly compute ρ_{1n}, as in the "surrogate marker" example of Section 1.1. Second, suppose the data-generating process is a multivariate normal distribution that obeys the conditional-independence properties that the analyst's recursive model assumes (and that the DAG given by R represents). Suppose further that p is an empirical distribution given by a finite sample of independent draws from the underlying distribution. Then, $\hat{\rho}_{1n}$ is the maximum-likelihood estimate of the true correlation between x_1 and x_n. Of course, this "legitimacy" is ill-founded when the model is misspecified.[6]

We assume that the analyst faces the constraint that the estimated mean and variance of all individual variables must be correct. 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. 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 recursive models can distort pairwise correlations without distorting marginals.

[6] This is particularly relevant in high-dimensional settings. As Cai et. al. (2016, p. 3) note: "The standard and most natural estimator, the sample covariance matrix, performs poorly and can lead to invalid conclusions in the high-dimensional settings...To overcome the difficulty due to the high dimensionality, structural assumptions are needed in order to estimate the covariance or precision matrix consistently."

The requirement that the estimated means of individual variables are undistorted has no bite: The OLS procedure for individual equations satisfies it. Thus, the constraint is reduced to the requirement that

$$\widehat{Var}(x_k) = 1$$

for all k (we can calculate this estimated variance for every $k > 1$, using the same recursive procedure we applied to x_n). This reduces (4) to

$$\hat{\rho}_{1n} = \alpha$$

Our objective is to examine how large this expression can be, given n and a generic objective covariance matrix.

Motivations for using recursive models

In Section 1.1, we gave an example for a primary motivation for using recursive models, namely extrapolating from partial data. Another is *covariance estimation*. Suppose a researcher obtains a finite sample of several variables. He believes the data-generating process obeys a recursive model, and wishes to learn the covariance matrix of this process. Accordingly, he estimates the model's parameters from all available data and uses the estimated parameters to compute the covariance matrix. Given the model, this procedure can be more reliable than computing the covariance matrix directly from the sample distribution. Indeed, it can be a consequence of maximum-likelihood estimation (e.g. see Cai et al. (2016)).[7]

Recursive models are also relevant because they satisfy a demand for mechanisms and stories, thanks to their underlying DAG structure. Suppose we are interested in the causal effect of one variable on another. Even when we can reliably measure it (e.g. by a randomized experiment), we often tend to be dissatisfied with mere measurement and look for mechanisms that transmit the causal effect. This contributes to our subjective sense that we understand the effect, and also suggests opportunities for future interven-

[7] Gaussian Bayesian Networks serve covariance estimation in diverse contexts: brain imaging data (Steimer et al. (2015)), call center data (Rajaratnam et al. (2007)), wireless communication (Wiesel and Hero (2011)), etc.

tions.[8] A DAG is a way of articulating such a mechanism. In other contexts, causal models package observed correlations into "*narratives*". Since people are typically more comfortable with stories than with statistical tables or correlation matrices as a means of organizing empirical regularities, causal models help remembering and communicating these regularities. In Section 5 we will discuss another role of recursive models - namely, eliciting causal effects from observational data in the presence of confounding.

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 covariance matrix* (ρ_{ij}) *satisfying* $\rho_{1n} = r$*, if the estimated recursive model satisfies* $\widehat{Var}(x_k) = 1$ *for all* k*, then the estimated correlation between* x_1 *and* x_n *satisfies*

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

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

(i) A recursive model defined by $R(k) = \{k-1\}$ *for every* $k = 2, ..., n$*.*
(ii) A multivariate Gaussian distribution satisfying, for every $k = 1, ..., n$*:*

$$x_k = s_1 \cos((k-1)\theta_{r,n}) + s_2 \sin((k-1)\theta_{r,n}) \tag{5}$$

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

[8] E.g., even when we know a genetic cause of a phenotype, a mechanism that involves other variables can inspire future medical interventions, since the gene itself is an unlikely locus of intervention.

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 adding a variable to the false correlation that the analyst's model can produce decays quickly. However, when $n \to \infty$, the upper bound converges to one. This is the case for any value of r. That is, even if the objective empirical correlation between x_1 and x_n is strongly negative, a sufficiently large model can produce a large positive estimated correlation.

The recursive model that attains the upper bound has a simple structure. Its DAG representation is a single chain

$$1 \to 2 \to \cdots \to n$$

Intuitively, this is the simplest connected n-node DAG, in terms of the number of links and size of $R(k)$. The distribution over the auxiliary variables $x_2, ..., x_n$ 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.

Variable selection vs. model selection

Suppose the analyst selects the variables $x_2, ..., x_{n-1}$ 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 $x_2, ..., x_{n-1}$ are selected as in (5), 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 recursive models whose representative DAG satisfies a property called *perfection* (see Cowell et al. (1999)). A DAG given by R is perfect if for every $k > 1$, if $i, j \in R(k)$ and $i < j$, then $i \in R(j)$. The chain $1 \rightarrow 2 \rightarrow \cdots \rightarrow n$ 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 recursive models.

Proposition 1 *Consider a recursive model represented by a perfect DAG. For every objective covariance matrix* (ρ_{ij}) *satisfying* $\rho_{1n} = r$, $\hat{\rho}_{1n} \leq (\cos\theta_{r,n})^{n-1}$.

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

General outline of the proof
The proof of Theorem 1 proceeds in three major steps. First, the constraint that the estimated model preserves the variance of individual variables for a generic objective distribution reduces the class of candidate recursive models to those that can be represented by perfect DAGs.

In the second step, we use the tool of *junction trees* in the Bayesian-networks literature (Cowell et al. (1999)) to perform a further reduction in the class of relevant recursive models. Consider a recursive model represented by a non-chain perfect DAG. We show that the analyst can generate the same $\hat{\rho}_{1n}$ with another objective distribution and a recursive model that takes the form of a chain $1 \rightarrow \cdots \rightarrow n$ of length n at most.

To illustrate this argument, consider the following recursive model with

$n = 4$:

$$\begin{aligned} x_1 &= \varepsilon_1 \\ x_2 &= \beta_{12}x_1 + \varepsilon_2 \\ x_3 &= \beta_{13}x_1 + \beta_{23}x_2 + \varepsilon_3 \\ x_4 &= \beta_{24}x_2 + \beta_{34}x_3 + \varepsilon_4 \end{aligned}$$

This recursive model has the following DAG representation:

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

Because x_4 depends on x_2 and x_3 only through their linear combination $\beta_{24}x_2 + \beta_{34}x_3$, we can replace (x_2, x_3) with a scalar variable x_5, such that the recursive model becomes

$$\begin{aligned} x_1 &= \varepsilon_1 \\ x_5 &= \beta'_{15}x_1 + \varepsilon'_5 \\ x_4 &= \beta'_{54}x_5 + \varepsilon_4 \end{aligned}$$

This model is represented by the DAG $1 \rightarrow 5 \rightarrow 3$, which is a simple chain with fewer nodes than the original DAG.

This means that in order to calculate the upper bound on $\hat{\rho}_{1n}$, we can restrict attention to the chain model. But in this case, the analyst's objective function has a simple explicit form:

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

Thus, in the third step, we derive the upper bound by finding the objective correlation matrix (ρ_{ij}) that maximizes the R.H.S of this formula, subject to the constraint that $\rho_{1n} = r$ and that the matrix is positive semi-definite (the defining property of covariance matrices). The solution to this problem has a simple geometric interpretation.

4 Proof of Theorem 1

4.1 Preliminaries: Bayesian Networks

The proof relies on concepts and tools from the Bayesian-network literature (Cowell et al. (1999), Koller and Friedman (2009)). Therefore, we introduce a few definitions that will serve us in the proof.

A DAG 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 $N = \{1, ..., n\}$. With some abuse of notation, $R(i)$ is the set of nodes j for which the DAG includes a link $j \to i$. Recall that (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 $i, j \in C$, iRj or jRi. 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 $i \in N$.

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 $i \in R(j)$. In line with our definition of recursive models in Section 2, we assume that 1 is ancestral and n is terminal. It is also easy to verify that we can restrict attention to DAGs in which n is the *only* terminal node - otherwise, we can remove the other terminal nodes from the DAG, without changing $\hat{p}(x_n \mid x_1)$. We will take these restrictions for granted henceforth.

The analyst's procedure for estimating $\hat{\rho}_{1n}$, as described in Section 2, has an equivalent description in the language of Bayesian networks, which we now describe.

Because the analyst estimates a linear model, it is *as if* he believes that the underlying distribution p is *multivariate normal*, where the estimated k^{th} equation is a complete description of the conditional distribution $(p(x_k \mid x_{R(k)}))$. Therefore, from now, we will proceed as if p were indeed a standardized multivariate normal with covariance matrix (ρ_{ij}), such that the

k^{th} regression equation corresponds to measuring the correct distribution of x_k conditional on $x_{R(k)}$. This is helpful expositionally and entails no loss of generality.

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

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

We say that p is consistent with G if $p_G = p$.

By Koller and Friedman (2009, Ch. 7), when p is multivariate normal, p_G is reduced to the estimated joint distribution as described in Section 2. In particular, we can use p_G to calculate the estimated marginal of x_k for any k:

$$p_G(x_k) = \int_{(x_j)_{j<k}} \prod_{j \leq k} p(x_j \mid x_{R(j)})$$

Likewise, the induced estimated distribution of x_n conditional on x_1 is

$$p_G(x_n \mid x_1) = \int_{x_2,...,x_{n-1}} \prod_{k \in K} p(x_k \mid x_{R(k)}) \tag{6}$$

This conditional distribution, together with the marginals $p_G(x_1)$ and $p_G(x_n)$, induce the estimated correlation coefficient $\hat{\rho}_{1n}$ given by (4).

Because we take p to be multivariate normal, the constraint that p_G does not distort the mean and variance of individual variables is equivalent to the requirement that the estimated marginal distribution $(p_G(x_k))$ coincides with the objective marginal distribution of x_k. This constraint necessarily holds if G is perfect. Furthermore, when G is perfect, $p_G(x_C) \equiv p(x_C)$ for every clique C in G (see Spiegler (2017)).

4.2 The Proof

Our first step is to establish that for generic p, perfection is necessary for the correct-marginal constraint.

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

Proof. Recall that we list the variables $x_1, ..., x_n$ such that $R(i) \subseteq \{1, ..., i-1\}$ for every i. Consider the lowest k for which $R(k)$ is not a clique. This means that there exist two nodes $h, l \in R(k)$ that are unlinked in G, whereas for every $k' < k$ and every $h', l' \in R(k')$, h' and l' are linked in G.

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

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

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

The equation for x_k is

$$x_k = \sum_{i \in R(k)} \beta_{ik} x_i + \varepsilon_k \tag{7}$$

Let β denote the vector $(\beta_{ik})_{i \in R(k)}$. Let A denote the correlation sub-matrix

$(\rho_{ij})_{i,j\in R(k)}$ that fully characterizes the objective joint distribution $(p(x_{R(k)}))$. Then, the objective variance of x_k can be written as

$$Var(x_k) = 1 = \beta^T A\beta + \sigma^2 \tag{8}$$

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

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

$$Var_G(x_k) = \beta^T C\beta + \sigma^2 \tag{9}$$

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

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

$$\beta^T A\beta = \beta^T C\beta \tag{10}$$

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

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

$$\Sigma_{\{h\}\times\{l\}} = \Sigma_{\{h\}\times M}\Sigma^{-1}_{M\times M}\Sigma_{M\times\{l\}} \tag{11}$$

Note that equation (11) 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 (11).

The L.H.S of (11) is simply $\hat{\rho}_{hl}$. The R.H.S of (11) is induced by $p_G(x_1, ..., x_{k-1})$. As noted earlier, this distribution is pinned down by G and the entries in $(\rho_{ij})_{i,j=1,...,k-1}$ except for ρ_{hl}. That is, if we are not informed of ρ_{hl} but we are informed of all the other entries in $(\rho_{ij})_{i,j=1,...,k-1}$, we are able to pin down the R.H.S of (11).

Now, when we draw the objective correlation submatrix $(\rho_{ij})_{i,j=1,...,k-1}$ at random, we can think of it as a two-stage lottery. In the first stage, all the entries in this submatrix except ρ_{hl} are drawn. In the second stage, ρ_{hl} is drawn. The only constraint in each stage of the lottery is that $(\rho_{ij})_{i,j=1,...,k-1}$ has to be positive-semi-definite and have 1's on the diagonal. Fix the outcome of the first stage of this lottery. Then, it pins down the R.H.S of (11). In the lottery's second stage, there is (for a generic outcome of the lottery's first stage) a continuum of values that ρ_{hl} could take for which $(\rho_{ij})_{i,j=1,...,k-1}$ will be positive-semi-definite. However, there is only value of ρ_{hl} that will coincide with the value of $\hat{\rho}_{hl}$ that is given by the equation (11). We have thus established that $A \neq C$ for generic $(\rho_{ij})_{i,j=1,...,k-1}$.

Recall once again that we can regards β as a parameter of p that is independent of A (and therefore of C as well), because A describes $(p(x_{R(k)}))$ whereas β, σ^2 characterize $(p(x_k \mid x_{R(k)}))$. Then, since we can assume $A \neq C$, (10) 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}). ■

Corollary 1 *For almost every (ρ_{ij}), if a DAG G satisfies $E_G(x_k) = 0$ and $Var_G(x_k) = 1$ for all $k = 1, ..., n$, 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 $\{1, ..., n\}$ 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. ■

The next step is based on the following definition.

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 $1 \to \cdots \to n$. 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 exists 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 false correlation induced by G' on ρ' is exactly the same as the false correlation induced by G on ρ: $cov_{G'}(x_1, x_n) = cov_G(x_1, x_n)$.*

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 arbitrary 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 $C_1, ..., C_K$ 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 $1 \to C_1 \to \cdots \to C_K \to n$, as follows:

$$p_G(x_1, x_{C_1}, ..., x_{C_K}, x_n) = p(x_1) \prod_{k=1}^{K} p(x_{C_k} | x_{C_{k-1}}) p(x_n | x_{C_K}) \tag{12}$$

where $C_0 = \{1\}$. The length of this linear DAG is $K + 2 \leq n$. While this factorization formula superficially completes 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 $cov_G(x_1, x_n)$.

Recall that we can regard p as a multivariate normal distribution without loss of generality. Furthermore, under such a distribution and 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 $p_G(x_1, x_{C_1}, ..., x_{C_K}, x_n)$

via the following recursion:

$$\begin{aligned} x_1 &\sim N(0,1) \\ x_{C_1} &= A_1 x_1 + \eta_1 \\ &\vdots \\ x_{C_k} &= A_k x_{C_{k-1}} + \eta_k \\ &\vdots \\ x_{C_K} &= A_K x_{C_{K-1}} + \eta_K \\ x_n &= A_{K+1} x_{C_K} + \eta_n \end{aligned} \tag{13}$$

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 recursive 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 (with $k > 1$) is conditionally independent of $x_1, ..., x_{k-1}$ given $x_{R(k)}$. Since the junction-tree factorization (12) 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 $x_1, ..., x_{C_{k-2}}$. 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 (13), we obtain a recursion for z:

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

Given that p is taken to be multivariate normal, the equation for z_k measures the objective conditional distribution $(p_G(z_k \mid z_{k-1}))$. Since p_G does not distort the objective distribution over cliques, $(p_G(z_k \mid z_{k-1}))$ coincides with $(p(z_k \mid z_{k-1}))$. This means that an analyst who fits a recursive 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 $z_1, ..., z_k$ (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. ■

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

$$\hat{\rho}_{1n} = \prod_{k=1}^{n-1} \rho_{k,k+1} \tag{14}$$

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:

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

Note that the positive semi-definiteness 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 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:

$$\rho^*_{1n} = \max_{\substack{a_i^T a_i=1 \text{ for all } i \\ a_1^T a_n=r}} \prod_{i=1}^{n-1} a_i a_{i+1}^T \tag{15}$$

Denote $\alpha = \arccos r$. Since the solution to (15) 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 $a_2, ..., a_{n-1}$.

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

$$\|u\|^2 + \|v\|^2 = 1 \tag{16}$$

The terms in the objective function (15) 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 (16) 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 a_k 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* $k = 2, ..., n-1$, 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,

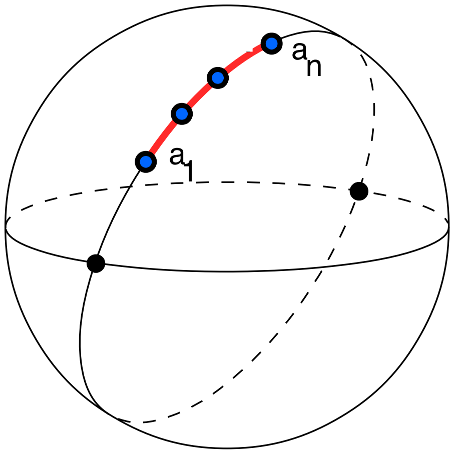

Figure 2: Geometric intuition for the proof.

equally spaced between a_1 and a_n. The explicit formulas for these vectors are given by (5).

The formula for the upper bound has a simple geometric interpretation (illustrated by Figure 2). 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 average distance 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 Single-Equation Models: Causal Inference and Bad Controls

Analysts often propose models that take the form of a *single* linear-regression equation, consisting of a dependent variable x_n (where $n > 2$), an explanatory variable x_1 and $n-2$ "*control*" variables $x_2, ..., x_{n-1}$. Using the language of Section 2, this corresponds to the specification $R(k) = \emptyset$ for all $k = 1, ..., n-1$ and $R(n) = \{1, ..., n-1\}$. That is, the only non-degenerate equation is the one for x_n, hence the term "single-equation model".

Using the graphical representation, the single-equation model corresponds to a DAG in which $x_1, ..., x_{n-1}$ are all ancestral nodes that send links into x_n. Since this DAG is imperfect, Lemma 1 implies that for almost all objective covariance matrices, the estimated variance of x_n according to the single-equation model will differ from its objective value. (All the other variables are ancestral, and therefore their marginals are not distorted (see Spiegler (2017).) Accordingly, we relax the correct-variance constraint and look for the maximal estimated correlation $\hat{\rho}_{1n}$ that such models can generate.

There is a clear rationale for focusing on this class of imperfect DAGs and relaxing the correct-variance constraint. So far, we have been fairly agnostic about the meaning of the estimated correlation $\hat{\rho}_{1n}$. One interpretation is that the analyst wants to persuade the audience that x_1 is a diagnostic or predictive signal of x_n. From this point of view, there is no need to attach any causal meaning to $\hat{\rho}_{1n}$. Nevertheless, we have already noted that when x_1 is not a dependent variable in any equation (i.e., it is an ancestral node in the DAG given by R), $\hat{\rho}_{1n}$ also has a natural interpretation as the estimated causal effect of x_1 on x_n.

However, as mentioned in Section 2, one reason for using a model when estimating the causal effect of one variable on another is that the observed correlation between the variables is not a reliable estimate of this effect, due to *confounding*. In this vein, suppose the analyst proposes a recursive model R^* postulating that the correlation between x_1 and x_n is confounded by another variable x_k. Unlike the specification of Section 2, x_1 is a dependent variable in some equation. The DAG representation of this phenomenon is as follows: Both x_1 and x_n are descendants of some other variable x_k, and there exists a directed path from x_k to x_n that does not pass through x_1. A special case of such a DAG R^* is the following: $x_2, ..., x_{n-1}$ are all ancestral nodes; and each of them sends direct links into x_1 as well as x_n. When our analyst claims that this model R^* explains the objective distribution p, he implies that the empirical correlation ρ_{1n} does not measure the average causal effect of x_1 on x_n.

According to Pearl (2009), the procedure for eliciting the causal effect of x_1 on x_n in the presence of confounding is as follows. Modify the DAG $G = (N, R^*)$ by removing all the links that go into x_1, thus producing the *imperfect* DAG $G' = (N, R)$, where R is *precisely* the function we introduced at the beginning of this section. The DAG G' coincides with G except that x_1 is an ancestral node. The estimated causal effect of x_1 on x_n according to the original model G is then calculated according to $p_{G'}$. For example, suppose the analyst proposes a model represented by the DAG

$$\begin{tikzcd} & x_2 \arrow[dl] \arrow[d] \\ x_1 \arrow[r] & x_3 \end{tikzcd} \tag{17}$$

Then, in order to estimate the causal effect of x_1 on x_3, the analyst would compute $\hat{\rho}_{1n}$ under the following auxiliary DAG,

$$\begin{tikzcd} & x_2 \arrow[d] \\ x_1 \arrow[r] & x_3 \end{tikzcd}$$

The quantity $\hat{\rho}_{1n}$ summarizes this causal effect - standardized according to the estimated variance of x_n. There is no reason to demand this estimated variance to coincide with the empirically observed variance, because G' does not purport to describe the empirical distribution; rather, it intends to describe a *counterfactual* distribution in which x_1 is truly randomized instead of being affected by the variables $x_2, ..., x_{n-1}$. Thus, the combination of a model that corresponds to an imperfect DAG and the relaxation of the correct-variance constraint is a consequence of the analyst's claim that he is trying to elicit a *causal* effect from an empirical distribution in the presence of confounding effects.

Proposition 2 *A single-equation model* $x_n = \sum_{i=1}^{n-1} \beta_i x_i + \varepsilon$ *can generate an estimated coefficient* $\hat{\rho}_{1,n}$ *of at most*

$$\frac{1}{\sqrt{2 - r^2}}$$

This bound is tight, and can be implemented with $n = 3$*, such that* $x_2 = \delta \cdot x_1 + \sqrt{1 - \delta^2} \cdot x_3$*, where* $\delta = \sqrt{1/(1 + r^2)}$*.*[9]

Proof. Because $x_2, ..., x_{n-1}$ are Gaussian without loss of generality, we can replace their linear combination $(\sum_{i=2}^{n-1} \beta_i x_i)/(\sum_{i=2}^{n-1} \beta_i)$ (where the β_i's are determined by the objective p) by a single Gaussian variable z that has mean zero, but its variance need not be one. Its objective distribution conditional on x_1, x_n can be written as a linear equation $z = \delta x_1 + \gamma x_n + \eta$, where η is uncorrelated with x_1, x_n and has variance σ^2. Since x_1 and x_n are standardized normal variables, the objective variance of z is

$$Var(z) = \delta^2 + \gamma^2 + 2\delta\gamma r + \sigma^2$$

The analyst's model can be written as

$$x_n = \frac{1}{\gamma} z - \frac{\delta}{\gamma} x_1 - \frac{1}{\gamma} \eta \tag{18}$$

Our objective is to find the values of δ, γ and σ that maximize

$$\hat{\rho}_{1,n} = \frac{\hat{E}(x_1, x_n)}{\sqrt{\widehat{Var}(x_n)\widehat{Var}(x_1)}}$$

[9] When $r = 0$, the upper bound is approximated arbitrarily well, by setting δ slightly below 1.

Because x_1 and x_n are standardized normal, it follows from (18) that $\hat{E}(x_1, x_n) = -\delta/\gamma$. The analyst's model does not distort the variance of x_1. The reason is that the node that represents x_1 in the DAG representation of the model is ancestral. By Spiegler (2017), the estimated model does not distort the marginals of such variables. Therefore, $\widehat{Var}(x_1) = 1$. And since the analyst's model regards z, x_1 and η as independent,

$$\begin{aligned} \widehat{Var}(x_n) &= \left(\frac{1}{\gamma}\right)^2 Var(z) + \left(\frac{\delta}{\gamma}\right)^2 + \left(\frac{\sigma}{\gamma}\right)^2 \\ &= \left(\frac{1}{\gamma}\right)^2 \left(\delta^2 + \gamma^2 + 2\delta\gamma r + \sigma^2\right) + \left(\frac{\delta}{\gamma}\right)^2 + \left(\frac{\sigma}{\gamma}\right)^2 \end{aligned}$$

It is clear from this expression that in order to maximize $\hat{\rho}_{1,n}$, we should set $\sigma = 0$. It follows that

$$\hat{\rho}_{1,n} = -\frac{\frac{\delta}{\gamma}}{\sqrt{1 + 2(\frac{\delta}{\gamma})^2 + 2r\frac{\delta}{\gamma}}}$$

This attains a maximum of $1/\sqrt{2 - r^2}$ at $\delta/\gamma = -1/r$. Since without loss of generality we can set $\gamma = \sqrt{1 - \delta^2}$ such that $z \sim N(0, 1)$, the upper bound is attained at

$$\begin{aligned} \delta &= \sqrt{\frac{1}{1 + r^2}} \\ \gamma &= -\sqrt{\frac{r^2}{1 + r^2}} \end{aligned}$$

When $r = 0$, we approximate the upper bound arbitrarily well, by setting $\delta \approx 1$. ■

Thus, to magnify the estimated causal effect of x_1 on x_3, the analyst would select a single "control" variable x_2 that is a deterministic linear combination of x_1 and x_3 (with a positive weight on one variable and a negative weight on the other). This collinearity inflates the estimated variance of x_3 to $\widehat{Var}(x_3) = (2/r^2) - 1$. This is more than compensated by the increase in the estimated covariance between x_1 and x_3.

The analyst justifies the use of the "control" variable x_2 by the original model R^*, in which x_2 acts as a confounder of the causal effect of x_1 on x_3. In other words, the original model R^* is represented by the DAG depicted in (17). Suppose the true causal model that underlies the objective distribution is different: x_2 is a descendant of x_1 or x_3 rather than their parent. Then, adding x_2 to the regression means that the analyst controls for a "post-treatment" variable (where x_1 is viewed as the treatment). In other words, x_2 is a "bad control" (Angrist and Pischke (2008), p. 64).[10]

6 Uniform Binary Variables

The general Bayesian-network formulation of our problem in Section 4.1 is equivalent to its presentation in Section 2 when the objective distribution p is multivariate normal. Once we depart from Gaussian environments, the general problem is largely open. In this section, we present partial analysis of a specific non-Gaussian setting. Suppose the variables $x_1, ..., x_n$ all take values in $\{-1, 1\}$, and restrict attention to the class of objective distributions p whose marginal on each variable is uniform - i.e., $p(x_i = 1) = \frac{1}{2}$ for every $i = 1, ..., n$. As in our main model, fix the correlation between x_1 and x_n to be r - that is,

$$\rho_{1n} = p(x_n = 1 \mid x_1 = 1) - p(x_n = 1 \mid x_1 = -1) = r$$

The question of finding the distribution p (in the above restricted domain) and the DAG G that maximize the induced $\hat{\rho}_{in}$ subject to $p_G(x_n = 1) = \frac{1}{2}$ is generally open. However, when we fix G to be the linear DAG

$$1 \rightarrow 2 \rightarrow \cdots \rightarrow n$$

we are able to find the maximal $\hat{\rho}_{1n}$. It makes sense to consider this specific DAG, because it proved to be the one most conducive to generating false

[10] See also http://causality.cs.ucla.edu/blog/index.php/2019/08/14/a-crash-course-in-good-and-bad-control/.

correlations in the case of linear-regression models.

Given the DAG G and the objective distribution p, the correlation between x_i and x_j that is induced by p_G is

$$\hat{\rho}_{ij} = p_G(x_j = 1 \mid x_i = 1) - p_G(x_j = 1 \mid x_i = -1)$$

Let $j > i$. Given the structure of the linear DAG, we can write

$$p_G(x_j \mid x_i) = \sum_{x_{i+1},\dots x_{j-1}} p(x_{i+1} \mid x_i) p(x_{i+2} \mid x_{i+1}) \cdots p(x_j \mid x_{j-1}) \tag{19}$$

In particular,

$$\begin{aligned} p_G(x_n \quad \mid \quad x_1) &= \sum_{x_2,\dots x_{n-1}} p(x_2 \mid x_1) p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1}) \qquad (20) \\ &= \sum_{x_2} p(x_2 \mid x_1) p_G(x_n \mid x_2) \end{aligned}$$

Note that $p_G(x_n \mid x_2)$ has the same expression that we would have if we dealt with a linear DAG of length $n-1$, in which 2 is the ancestral node: $2 \to \cdots \to n$. Applying algebraic manipulation to (20), we obtain that $\hat{\rho}_{1n}$ is equal to

$$\left[p(x_2 = 1 \mid x_1 = 1) - p(x_2 = 1 \mid x_1 = -1)\right] \left[p_G(x_n = 1 \mid x_2 = 1) - p_G(x_n = 1 \mid x_2 = -1)\right]$$

and therefore

$$\hat{\rho}_{1n} = \rho_{12} \cdot \hat{\rho}_{2n}$$

This formula will enable us to apply an inductive proof to our result.

We can now derive an upper bound on $\hat{\rho}_{in}$ for the environment of this section - i.e., the estimated model is a linear DAG, and the objective distribution has uniform marginals over binary variables.

Proposition 3 *For every n,*

$$\hat{\rho}_{1n} \leq \left(1 - \frac{1-r}{n-1}\right)^{n-1}$$

Proof. The proof is by induction on n. Let $n = 2$. Then, $p_G(x_2 \mid x_1) = p(x_2 \mid x_1)$, and therefore $\hat{\rho}_{12} = r$, which confirms the formula.

Suppose that the claim holds for some $n = k \geq 2$. Now let $n = k + 1$. Consider the distribution of x_2 conditional on x_1, x_n. Denote $\alpha_{x_1,x_n} = p(x_2 = 1 \mid x_1, x_n)$. We wish to derive a relation between ρ_{12} and ρ_{2n}. Denote

$$q = \frac{1+r}{2} = p(x_n = 1 \mid x_1 = 1) = p(x_n = -1 \mid x_1 = -1)$$

Then,

$$\begin{aligned} p(x_2 &= 1 \mid x_1 = 1) = p(x_n = 1 \mid x_1 = 1) \cdot \alpha_{1,1} + p(x_n = -1 \mid x_1 = 1) \cdot \alpha_{1,-1} \\ &= q\alpha_{1,1} + (1-q)\alpha_{1,-1} \end{aligned}$$

Likewise,

$$\begin{aligned} p(x_2 &= 1 \mid x_1 = 0) = p(x_n = 1 \mid x_1 = 0) \cdot \alpha_{-1.1} + p(x_n = 0 \mid x_1 = 0) \cdot \alpha_{-1,-1} \\ &= q\alpha_{-1,-1} + (1-q)\alpha_{-1,1} \end{aligned}$$

The objective correlation between x_1 and x_2 is thus

$$\rho_{12} = q(\alpha_{1,1} - \alpha_{-1,-1}) + (1-q)(\alpha_{1,-1} - \alpha_{-1,1}) \tag{21}$$

Let us now turn to the joint distribution of x_n and x_2. Because the marginals on both x_2 and x_n are uniform, $p(x_n \mid x_2) = p(x_2 \mid x_n)$. Therefore, we can obtain ρ_{2n} in the same manner that we obtained ρ_{12}:

$$\rho_{2n} = q(\alpha_{1,1} - \alpha_{-1,-1}) + (1-q)(\alpha_{-1,1} - \alpha_{1,-1}) \tag{22}$$

We have thus established a relation between ρ_{12} and ρ_{2n}.

Recall that $\hat{\rho}_{2n}$ is the estimated correlation between x_2 and x_n under the objective distribution p and the linear DAG $2 \to \cdots \to n$, when $p(x_2 =$

$x_n) = \tilde{q}$. Therefore, by the inductive step,

$$\begin{aligned}\hat{\rho}_{1n} &= \rho_{12} \cdot \hat{\rho}_{2n} \qquad (23)\\ &\leq [q(\alpha_{1,1} - \alpha_{-1,-1}) + (1-q)(\alpha_{1,-1} - \alpha_{-1,1})] \cdot \left(1 - \frac{1-\rho_{2n}}{k-1}\right)^{k-1}\end{aligned}$$

Both ρ_{12} and ρ_{2n} increase in $\alpha_{1,1}$ and decrease in $\alpha_{-1,-1}$, such that we can set $\alpha_{1,1} = 1$ and $\alpha_{-1,-1} = 0$ without lowering the R.H.S of (23). This enables us to write

$$\rho_{12} = q + (1-q)(\alpha_{1,-1} - \alpha_{-1,1})$$

such that

$$\rho_{2n} = 1 + r - \rho_{12}$$

Therefore, we can transform (23) into

$$\hat{\rho}_{1n} \leq \max_{\rho_{12}} \quad \rho_{12} \cdot \left(1 - \frac{\rho_{12} - r}{k-1}\right)^{k-1}$$

The R.H.S is a straightforward maximization problem. Performing a logarithmic transformation and writing down the first-order condition, we obtain

$$\rho_{12}^{*} = 1 - \frac{1-r}{k}$$

and

$$\left(1 - \frac{\rho_{12}^{*} - r}{k-1}\right)^{k-1} = \left(1 - \frac{1-r}{k}\right)^{k-1}$$

such that

$$\hat{\rho}_{1n} \leq \left(1 - \frac{1-r}{k}\right)^{k}$$

which completes the proof. ■

How does this upper bound compare with the Gaussian case? For illustration, let $r = 0$. It is easy to see that for $n = 3$, we obtain $\hat{\rho}_{13} = \frac{1}{3}$, which is below the value of $\frac{1}{2}$ we were able to obtain in the Gaussian case. And as $n \to \infty$, $\hat{\rho}_{1n} \to 1/e$. That is, unlike the Gaussian case, the maximal estimated correlation that the linear DAG can generate is bounded (far) away from one.

The upper bound obtained in this result is tight. The following is one way to implement it. For the case $r = 0$, take the exact same Gaussian distribution over $x_1, ..., x_n$ that we used to implement the upper bound in Theorem 1, and now define the variable $y_k = sign(x_k)$ for each $k = 1, ..., n$. Clearly, each $y_k \in \{-1, 1\}$ and $p(y_k = 1) = p(y_k = -1) = \frac{1}{2}$ since each x_k has zero mean. To find the correlations between different y_k variables, we use the following lemma.

Lemma 3 *Let w_1, w_2 be two unit vectors in R^2 and let z be a multivariate Gaussian with zero mean and unit covariance. Then,*

$$E(sign(w_1^T z) sign(w_2^T z)) = 1 - \frac{2\theta}{\pi}$$

where θ is the angle between the two vectors.

Proof. This follows from the fact that the product $sign(w_1^T z) sign(w_2^T z)$ is equal to 1 whenever z is on the same side of the two hyperplanes defined by w_1 and w_2, and -1 otherwise. Since the Gaussian distribution of z is circularly symmetric, the probability that z lies on the same side of the two hyperplanes depends only on the angle between them. ■

Returning to the definition of the Gaussian distribution over $x_1, ..., x_n$ that we used to implement the upper bound in Theorem 1, we see that in the case of $r = 0$, the angle between w_1 and w_n will be $\frac{\pi}{2}$, so that by the above lemma, y_1 and y_n will be uncorrelated. At the same time, the angle between any w_k and w_{k-1} is by construction $\frac{\pi}{2}\frac{1}{n-1}$ because the vectors were chosen at equal angles along the great circle. Substituting this angle into the lemma, we obtain that the correlation between y_k and y_{k-1} is $1 - \frac{1}{n-1}$.

For the case where $r \neq 0$, the same argument holds, except that we need to choose the original vectors w_1, w_n so that the correlation between y_1 and y_n will be r (these will not be the same vectors that give a correlation of r between the Gaussian variables x_1 and x_n) and then choose the rest of the vectors at equal angles along the great circle. By applying the lemma again, we obtain that the angle between y_k and y_{k-1} is $1 - \frac{1-r}{n-1}$, which again attains the upper bound.

This method of implementing the upper bound also explains why false correlations are harder to generate in the uniform binary case, compared with the case of linear-regression models. The variable y_k is a coarsening of the original Gaussian variable x_k. It is well-known that when we coarsen Gaussian variables, we weaken their mutual correlation. Therefore, the correlation between any consecutive variables y_k, y_{k+1} in the construction for the uniform binary case is lower than the corresponding correlation in the Gaussian case. As a result, the maximal correlation that the model generates is also lower.

The obvious open question is whether the restriction to linear DAGs entails a loss of generality. We conjecture that in the case of uniform binary variables, a non-linear perfect DAG can generate larger estimated correlations for sufficiently large n.

7 Conclusion

This paper performed a worst-case analysis of estimated correlations in misspecified recursive models. We showed that within this class, the worst case can be very bad indeed, growing quickly in the number of model variables.

Within our framework, several questions are left open. First, we do not know whether Theorem 1 would continue to hold if we replaced the quantifier "for almost every p" with "for every p". Second, we do not know how much bite the undistorted-variance constraint has outside the single-equation case. Third, we lack complete characterizations for recursive models outside the linear-regression family (beyond our partial characterization for models that involve uniform binary variables in Section 6). Finally, it would be interesting to devise a small collection of misspecification or robustness tests that would be effective in trading off the belief errors that result from misspecified models and the benefits of using such models.

Related literature

There is a huge literature on misspecified models in various branches of Economics and Statistics, which is too vast to survey in detail here. A few recent references can serve as entry points for the interested reader: Esponda and Pouzo (2016), Bonhomme and Weidner (2018) and Molavi (2019).

The "analyst" story that motivated our exercise brings to mind the phenomenon of researcher bias. A few works in Economics have explicitly modeled this bias and its implications for statistical inference. (Of course, there is a larger literature on how econometricians should cope with researcher/publication bias, but here we only describe exercises that contain explicit models of the researcher's behavior.) Leamer (1974) suggests a method of discounting evidence when linear regression models are constructed after some data have been partially analyzed. Lovell (1983) considers a researcher who chooses k out of n independent variables as explanatory variables in a single regression with the aim of maximizing the coefficient of correlation between the chosen variables and the dependent variable. He argues that a regression coefficient that appears to be significant at the α level should be regarded as significant at only the $1-(1-\alpha)^{n/k}$ level. Glaeser (2008) suggests a way of correcting for this form of data mining in the coefficient estimate.

More recently, Di-Tillio, Ottaviani and Sorensen (2017,2019) characterize data distributions for which strategic sample selection (e.g., selecting the k highest observations out of n) benefits an evaluator who must take an action after observing the selected sample realizations. Finally, Spiess (2018) proposes a mechanism-design framework to align the preferences of the researcher with that of "society": A social planner first chooses a menu of possible *estimators*, the investigator chooses an estimator from this set, and the estimator is then applied to the sampled observations.

References

[1] Angrist, J. and J. Pischke (2008), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

[2] Bonhomme, S. and M. Weidner (2018), Minimizing Sensitivity to Model Misspecification, arXiv preprint arXiv:1807.02161.

[3] Cai, T., Ren, Z., and Zhou, H. H. (2016). Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation, Electronic Journal of Statistics, 10, 1-59.

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

[5] Cowell, R., P. Dawid, S. Lauritzen and D. Spiegelhalter (1999), Probabilistic Networks and Expert Systems, Springer, London.

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

[7] Dworczak, P. ad G. Martini (2019), The Simple Economics of Optimal Persuasion, Journal of Political Economy 127, 1993-2048.

[8] Eliaz, K. and R. Spiegler (2018), A Model of Competing Narratives, mimeo.

[9] Esponda, I. and D. Pouzo (2016), Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models, Econometrica 84, 1093-1130.

[10] Fleming, T. and D. DeMets (1996), Surrogate End Points in Clinical Trials: Are we Being Misled?, Annals of internal medicine 125, 605-613.

[11] Glaeser, E. (2008), Researcher Incentives and Empirical Methods, in The Foundations of Positive and Normative Economics (Andrew Caplin and Andrew Schotter, eds.), Oxford: Oxford University, 300-319.

[12] Jehiel, P. (2005), Analogy-Based Expectation Equilibrium, Journal of Economic Theory, 123, 81-104.

[13] Katz, R. (2004), Biomarkers and Surrogate Markers: An FDA Perspective, NeuroRx 1, 189-195.

[14] Leamer, E. (1974), False Models and Post-Data Model Construction, Journal of the American Statistical Association, 69(345), 122-131.

[15] Molavi, P. (2019), Macroeconomics with Learning and Misspecification: A General Theory and Applications, mimeo.

[16] Lovell, M. (1983), Data Mining, The Review of Economics and Statistics, 65(1), 1-12.

[17] Di Tillio, A., M. Ottaviani, and P. Sorensen (2017), Persuasion Bias in Science: Can Economics Help? Economic Journal 127, F266–F304.

[18] Di Tillio, A., M. Ottaviani, and P. Sorensen (2019), Strategic Selection Bias, Working Paper.

[19] Pearl, J. (2009), Causality: Models, Reasoning and Inference, Cambridge University Press, Cambridge.

[20] Koller, D. and N. Friedman. (2009). Probabilistic Graphical Models: Principles and Techniques, MIT Press, Cambridge MA.

[21] Monteal Olea, J., P. Ortoleva, M. Pai and A. Prat (2018), Competing Models, mimeo.

[22] 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.

[23] Rajaratnam, B., H. Massam and C. M. Carvalho (2007), Flexble Covariance Estimation in Graphical Gaussian Models, Annals Statistics 36, 2818-2849.

[24] 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.

[25] Spiegler, R. (2017), "Data Monkeys": A Procedural Model of Extrapolation From Partial Statistics, Review of Economic Studies 84, 1818-1841.

[26] Spiegler, R. (2019), Behavioral Implications of Causal Misperceptions, forthcoming in Annual Review of Economics.

[27] Spiess, J. (2018), Optimal Estimation when Researcher and Social Preferences are Misaligned, Working Paper.

[28] Steimer, A., F. Zubler and K. Schindler (2015), Chow-Liu Trees are Sufficient Predictive Models for Reproducing Key Features of Functional Networks of Periictal EEG Time-Series, NeuroImage 118, 520-537.

[29] Wiesel, A. and A. Hero (2011), Distributed Covariance Estimation in Gaussian Graphical Models, IEEE Transactions on Signal Processing 60, 211-220.