Cheating with (Recursive) Models*

Kfir Eliaz†, Ran Spiegler‡ and Yair Weiss§

September 16, 2019

Abstract

To what extent can misspecified models generate false estimated correlations? We focus on models that take the form of a recursive system of linear regression equations. Each equation is fitted to minimize the sum of squared errors against an arbitrarily large sample. We characterize the maximal estimated correlation between a given pair of variables that this procedure can generate for a generic objective covariance matrix, provided 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 false correlation can become arbitrarily close to one, even for a pair of objectively uncorrelated variables.

*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, Martin Weidner and the audience at the Warwick Economic Theory workshop 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

Consider the following situation. A "researcher" wishes to demonstrate to an audience that two variables, x and y, are strongly related. Direct evidence about the correlation between these variables is hard to come by. However, the researcher has access to data about the correlation of x and y with other variables. He therefore constructs a *model* that involves x, y and a selection of auxiliary variables. He fits this model to a large sample and uses the estimated model to predict the correlation between x and y. The researcher is unable (or unwilling) to tamper with the data. However, he is free to choose the auxiliary variables and how they operate in the model. To what extent does this degree of freedom enable the researcher to attain his objective?

This question is motivated by a number of real-life situations. First, academic researchers often serve as consultants to policy makers or activist groups in pursuit of a particular agenda. E.g., consider an economist consulting a policy maker who pursues a tax-cutting agenda and would benefit from an academic study showing a strong quantitative relation between tax cuts and economic growth. Second, the researcher may be wedded to a particular stand regarding the relation between x and y, either because he has a pet theory or because he staked his reputation on this claim in the past. Finally, the researcher may be interested in publishing a particular controversial/counterintuitive result and may stop exploring alternative model specifications once he obtains the desired result.

To address our question, we restrict the class of models that the researcher can employ to be *recursive linear-regression models*. A model in this widely used family 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. Each equation is estimated via Ordinary Least Squares (OLS) against an arbitrarily large sample. The following quote lucidly summarizes two attractions of recursive models:

> "A system of equations is recursive rather than simultaneous if there is unidirectional dependency among the endogenous variables such that, for given values of exogenous variables, values for

> the endogenous variables can be determined sequentially rather than jointly. Due to the ease with which they can often be estimated and the temptation to interpret them in terms of causal chains, recursive systems were the earliest equation systems to be used in empirical work in the social sciences."[1]

The causal interpretation of recursive models is particularly resonant. If x only appears as an explanatory variable in the system of equations while y only appears 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 legitimately interpreted as an estimated *causal* effect of x on y.

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 researcher's model is n, which is one measure of the model's complexity. In this environment, we pose the following question: Given the complexity bound n and the objective correlation r between x and y, what is the maximal estimated correlation that the researcher's model can generate in an arbitrarily large (and unbiased) sample?

A three-variable example

To illustrate the problem, suppose that the researcher 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 all have zero mean and unit variance. The researcher assumes that ε_j is uncorrelated x_k for every $j, k = 1, 2, 3$ (for $j = k$, this is mechanically implied by the OLD method). Let ρ_{ij} denote the correlation between x_i and x_j according to the *true* data-generating process.

[1]The quote is taken from the International Encyclopedia of the Social Sciences, https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/recursive-models.

Suppose that x_1 and x_3 are objectively uncorrelated - i.e. $r = \rho_{13} = 0$. The estimated correlation between these variables according to the model (1) is $\hat{\rho}_{13} = \beta_1/\beta_2$. The researcher's procedure and its underlying assumptions imply

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

It is easy to see from this expression how the model can generate spurious estimated correlation between x_1 and x_3, even though none exists in reality. All the researcher has to do is select a variable x_2 that is positively correlated with both x_1 and x_3, such that $\rho_{12} \cdot \rho_{23} > 0$.

But how large can the false estimated correlation $\hat{\rho}_{13}$ be? Intuitively, since x_1 and x_3 are objectively uncorrelated, if we choose x_2 such that it is highly correlated with x_1, the lower its correlation with x_3 will be. In other words, increasing ρ_{12} will come at the expense of decreasing ρ_{23}. Formally, consider the correlation matrix induced by the true joint distribution over the three variables:

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

Since this is a correlation matrix, it must be positive semi-definite by definition. This implies $(\rho_{12})^2 + (\rho_{23})^2 \leq 1$. The maximal value of $\rho_{12} \cdot \rho_{23}$ subject to this constraint is $\frac{1}{2}$, and therefore this is the maximal false correlation that the above recursive 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 $x_2 = \frac{1}{2}(x_1 + x_3)$. Thus, while a misspecified recursive model can generate spurious estimated correlation between objectively independent variables, there is a limit to how far it can go along these lines.

Our interest in the upper bound on $\hat{\rho}_{13}$ is not purely mathematical. We have in mind situations in which the researcher can select x_2 from a *large pool* of potential auxiliary variable. In the current age of "big data", researchers have access to data about a huge number of covariates. As a result, they have considerable freedom when deciding which variables to incorporate into their models. This helps our researcher generate a false correlation that

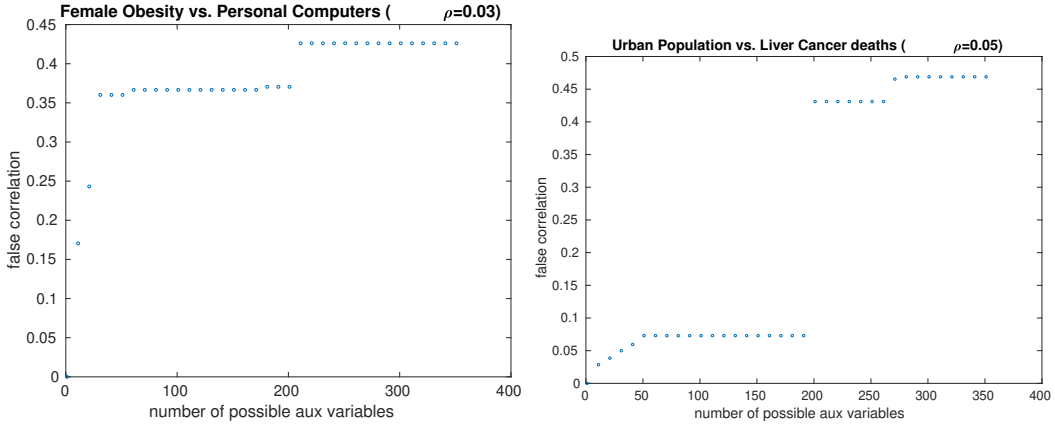

Figure 1: False correlation in a recursive model with one auxiliary variable, as a function of the number of possible auxiliary variables the researcher can choose from. All variables and their correlations are taken from a database compiled by the World Health Organization. Even though the true correlation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

approaches the theoretical upper bound.

To make this claim concrete, consider Figure 1, which is extracted from a database compiled by the World Health Organization and collected by Reshef et al. (2011).[2] All the variables are taken from this database. The figure displays the average 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 (the average is taken over all pools of this size). When the researcher can choose x_2 from only ten possible auxiliary variables, the estimated correlation between x and y he can generate with (1) is still modest. In contrast, once the size of the pool of variables from which x_2 is opportunistically selected is in the hundreds, the estimated correlation that (1) can generate approaches the upper bound of $\frac{1}{2}$.

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 per 100,000 men. The true correlation between

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

these variables is 0.05. If the researcher selects x_2 to be coal consumption (measured in tonnes oil equivalent), 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 an intuitive causal mechanism: Urban population causes liver cancer deaths via its effect on coal consumption.

Review of the results

We present our formal model in Section 2 and pose our main problem: What is the largest estimated correlation between x_1 and x_n that a recursive, n-variable linear-regression model can generate? We impose one constraint on this maximization problem: While the estimated model is allowed to distort correlations among variables, it is constrained to produce correct estimates of the mean and variance of individual variables. This constraint can be viewed as a minimal test of model misspecification, which is easy to implement. We relax this constraint for models that consist of a single non-degenerate regression equation in Section 4, and use this case to shed light on the researcher's opportunistic use of "*bad controls*".

In Section 3, we derive the following result. For a generic true $n \times n$ covariance matrix with $\rho_{1n} = r$, the maximal estimated correlation $\hat{\rho}_{1n}$ that a recursive model can generate subject to preserving the mean and variance of individual variables is

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

The upper bound given by (2) is tight. Specifically, it is attained by the simplest recursive model that involves n variables: For every $k = 2, ..., n$, x_k is regressed on x_{k-1} only. This model is represented graphically by the chain $x_1 \to x_2 \to \cdots \to x_n$, and each of the variables $x_2, ..., x_{n-1}$ are deterministic linear functions of x_1 and x_n. This chain has an intuitive causal interpretation, which enables the researcher to present $\hat{\rho}_{1n}$ as an estimated causal effect of x_1 on x_n.

Formula (2) reproduces the value $\hat{\rho}_{13} = \frac{1}{2}$ that we derived in our illustrative example, and it is strictly increasing in n. When $n \to \infty$, the expression converges to 1. That is, regardless of the true correlation between x_1 and

x_3, a sufficiently large recursive model can generate an arbitrarily large estimated correlation. The lesson is that when the researcher is free to select his model and the variables that inhabit it, he can deliver any conclusion about the effect of one variable on another - unless we impose constraints on his procedure, such as bounds on the complexity of his model or additional misspecification tests.

The formula (2) has a simple geometric interpretation, which also betrays the construction of the recursive model and objective distribution that implement the upper bound. Take the angle that represents the objective correlation r between x_1 and x_n; divide it into $n-1$ equal sub-angles; this sub-angle represents the correlation between adjacent variables along the above causal chain; the product of these correlations produces the estimated correlation between x_1 and x_n.

The detailed proof of our main result is presented in Appendix I. It relies on the graphical representation of recursive models using concepts and tools from the Bayesian-networks literature (Cowell et al. (1999), Koller and Friedman (2009)). In Appendix II, we present partial analysis of our question for a different class of models involving *binary* variables.

Related literature

A number of works in Economics have explicitly modeled researcher 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 (2006) 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.

Finally, the researcher in our model need not be a scientist - he could be a politician or a pundit. Under this interpretation, constructing a model and fitting it to data is not an explicit, formal affair. Rather, it involves spinning a "narrative" about the effect of policy on consequences and using casual empirical evidence to substantiate it (see Eliaz and Spiegler (2018) for a model of political beliefs that is based on this idea). From this point of view, our exercise in this paper explores the extent to which false narratives can exaggerate the effect of policy.

2 The Model

Let p be an objective probability measure over n variables, $x_1, ..., x_n$. 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.* Since we are interested in the correlation *coefficient* between a pair of variables, this entails no loss of generality. We use ρ_{ij} to denote the 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}).

A researcher estimates a recursive linear-regression 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 an infinitely large sample drawn from p, and the ε_k's are assumed to be zero mean and uncorrelated with any of the variables $x_1, ..., x_n$. Specifically, the β_{jk}'s are 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 $\beta_k = (\beta_{jk})_{j \in R(k)}$, $\rho_{k,R(k)}$ 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 $i \in R(j)$. DAGs are often interpreted as causal models (see Pearl (2009)). We will make use of the DAG representation in the appendices. We will also allude to the recursive model's causal interpretation, but without addressing explicitly the question of causal inference. In other words, the researcher in our model engages in a problem of fitting a model to data. While the causal interpretation of his model may add to its appeal and be useful for "selling" its conclusions to an audience, he does not engage in an explicit procedure for drawing causal inference from his model.

Note that in the researcher's model, x_1 is not a dependent variable and x_n is not an explanatory variable in the system of equations, such that the partial ordering given by R is consistent with the natural enumeration of variables. This restriction is made mainly for notational simplicity; 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 sound causal inference based on DAG-represented models.

The recursive structure allows the researcher to estimate each regression equation separately. We take the limit case in which the researcher's sample is infinitely large. Because the researcher 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.

The estimated joint distribution over $x_1, ..., x_n$ is thus

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

We can use $\hat{p}$ to calculate the estimated marginal of x_k for any k:

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

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

$$\hat{p}(x_n \mid x_1) = \int_{x_2,...,x_{n-1}} \prod_{k \in K} p(x_k \mid x_{R(k)}) \tag{4}$$

This conditional distribution induces an estimated correlation coefficient $\hat{\rho}_{1n}$. Our objective is to understand how large $\hat{\rho}_{1n}$ can be under the constraint that $(\hat{p}(x_k))$ coincides with the objective marginal distribution of x_k for generic p. Because we can regard p as multivariate normal without loss of generality, the constraint is reduced to the requirement that the estimated mean and variance of individual variables are correct. But we can go even further. As shown in Spiegler (2019), when p is multivariate normal, the estimated mean of any individual variable is unbiased. Therefore, our constraint can

be boiled down to the requirement that the estimated variance of x_k is equal to 1 for all k.

Interpretation of the researcher's procedure

In our model, the researcher relies on a structural model to generate an estimate of the correlation between x_1 and x_n. Our primary motivation for this assumption is that it seems to approximate what researchers often do when they interact with an audience: They fit a model to data and use the estimated model to generate predictions, and the process by which they select which model to estimate is to some extent hidden from their audience. Researchers may sometimes (perhaps unwittingly) search for a model that helps them deliver strong results in a certain direction.

However, this description raises the question of *why* the researcher relies on a structural model in the first place. In particular, why does he use a model to estimate the correlation between x_1 and x_n, rather than estimating it *directly*? One answer may be that direct evidence on this correlation is unavailable (as, for example, in the case of long-term health effects of nutritional choices). In this case, the researcher *must* use a model to extrapolate an estimate of ρ_{1n} from observed data.

Another answer is that researchers use models as simplified representations of a complex reality, which they can consult for *multiple* conditional-estimation tasks: Estimating the effect of x_1 on x_n is only one of these tasks. This is illustrated in the following quote: "The economy is an extremely complicated mechanism, and every macroeconomic model is a vast simplification of reality... the large scale of FRB/US [*a general equilibrium models employed by the Federal Reserve Bank - the authors*] is an advantage in that it can perform a wide variety of computational 'what if' experiments."[3] From this point of view, our analysis concerns the maximal distortion of pairwise correlations that such models can produce.

[3]This quote is taken from a speech by Stanley Fisher: See https://www.federalreserve.gov/newsevents/speech/fischer20170211a.htm.

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 true covariance matrix (ρ_{ij}) satisfying $\rho_{1n} = r$, if the estimated recursive model satisfies $\hat{Var}(x_k) = 1$ 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 standard normal variables with correlation r.

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 researcher'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 in $[-1, 1)$. That is, even if the true correlation between x_1 and x_n is strongly negative, a sufficiently large model can produce a large positive 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 DAG with n nodes (it has the smallest number of links among this class of DAGs, and it has no junctions). The distribution over the auxiliary variables $x_2, ..., x_n$ in the construction of the upper bound also has a simple structure: 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.

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 generic p reduces the class of candidate recursive models, to those that can be represented by *perfect* DAGs. Since perfect DAGs preserve marginals of individual variables for every objective distribution, the theorem is actually stronger: If the recursive model is represented by a perfect DAG, the upper bound on $\hat{\rho}_{1n}$ holds for *any* objective covariance matrix.

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, and recall that we can assume without loss of generality that p is multivariate normal. We show that the researcher can generate the same $\hat{\rho}_{1n}$ with another Gaussian distribution p' and a recursive model that takes the form of a simple chain $1 \to \cdots \to n$. Furthermore, this chain will involve fewer variables than the original model.

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 researcher's ob-

jective 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 is tantamount to finding a correlation matrix that maximizes the R.H.S of this formula, subject to the constraints that $\rho_{1n} = r$ and that the matrix is positive semi-definite (which is the property that defines the class of covariance matrices). We solve this problem using the method of Lagrange multipliers and show that our solution is a global maximum of the problem.

4 Single-Equation Models

Researchers 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 $x_2, ..., x_{n-1}$. In terms of the model 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 in Appendix I implies that it distorts the marginal of x_n for generic p.[4] That is, for almost all objective covariance matrices, the estimated variance of x_n according to the single-equation model will differ from its true value. However, given the particular interest in this class of models, we relax the correct-marginal constraint in this section and look for the maximal false correlation that such models can generate. For expositional convenience, we focus on the case of $r = 0$.

Proposition 1 *Let $r = 0$. Then, a single-equation linear-regression 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*

[4]All the other variables are represented by ancestral nodes, and therefore their marginals are not distorted (see Spiegler (2017)).

$1/\sqrt{2}$. *This bound is tight, and can be approximated arbitrarily well with* $n = 3$ *such that* $x_2 = \alpha x_1 + \sqrt{1-\alpha^2}x_3$, *where* $\alpha \approx -1$.

Proof. See Appendix III. ■

Thus, to magnify the false correlation between x_1 and x_n, the researcher would select the "control" variables such that a certain linear combination of them has strong negative correlation with x_1. That is, the researcher will prefer his regression model to exhibit multicollinearity. This inflates the estimated variance of x_n but increases the estimated correlation of this variable with x_1.

Note that when the bound on the model's complexity n is sufficiently large, the single-equation model is outperformed by the multi-equation chain model, which has the added benefit that it does not distort the variance of individual variables.

5 Conclusion

This paper addressed the problem of researcher bias, when the researcher's only tool is model selection. We showed that when the researcher makes use of recursive linear models, this tool is very powerful: Provided that the researcher is allowed to select a moderate number of variables from a large pool, he can produce a very large estimated correlation between two variables of interest. Furthermore, the structure of his model allows him to interpret this correlation as a causal effect. This is true even if the two variables are objectively independent, or if their correlation is in the opposite direction. Imposing a bound on the model's complexity (measured by its number of auxiliary variables) is an important constraint on the researcher. However, the value of this bound decays quickly, as even with one or two auxiliary variables the researcher can greatly distort objective correlations.

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 correct-marginal constraint has in models that consist of more

than one equation. Third, we lack complete characterizations for recursive models outside the linear-regression family (see our partial characterization for models that involve binary variables in Appendix II). Finally, it would be interesting to devise a sparse collection of misspecification or robustness tests that would restrain our opportunistic researcher.

Taking a broader perspective into the last question, our exercise suggests a novel approach to the study of biased estimates due to misspecified models in Statistics and Econometric Theory. Under this approach, the researcher who employs a structural model for statistical or causal analysis is viewed as a player in a game with his audience. Researcher bias implies a conflict of interests between the two parties. This bias means that the researcher's model selection is opportunistic. The question is which strategies the audience can play (in terms of robustness or misspecification tests it can demand) in order to mitigate errors due to researcher bias, without rejecting too many valuable models.

6 Appendix I: Proof of Theorem 1

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 $j \to i$. A DAG 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.

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.

Given an objective distribution p over $x_1, ..., x_n$ and a DAG G, define

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

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

As mentioned in Section 2, we can assume without loss of generality that p is multivariate normal with covariance matrix (ρ_{ij}). We will make use of this convenience throughout the proof.

When G is perfect, $p_G(x_C) \equiv p(x_C)$ for every clique C in G (see Spiegler (2017)).

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 true 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 true 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 true value of ρ_{hl}. We will make use of this observation toward the end of this proof.

The equation for x_k is

$$x_k = \sum_{i \in R(k)} \beta_{ik} x_i + \varepsilon_k \tag{6}$$

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

$$Var(x_k) = 1 = \beta^T A \beta + \sigma^2 \tag{7}$$

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

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

$$Var_G(x_k) = \beta^T C \beta + \sigma^2 \tag{8}$$

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 true joint distributed 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

$$\beta^T A \beta = \beta^T C \beta \tag{9}$$

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),

$$\Sigma_{\{h\} \times \{l\}} = \Sigma_{\{h\} \times M} \Sigma^{-1}_{M \times M} \Sigma_{M \times \{l\}} \tag{10}$$

Note that equation (10) 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 (10).

The L.H.S of (10) is simply $\hat{\rho}_{hl}$. The R.H.S of (10) 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 (10).

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 (10). 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 (10). 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 $A \neq C$, (9) 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 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 strictly 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 prefect 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. Such a tree graph is known as the "junction tree" corresponding to G and we can write (Koller and Friedman (2009, p. 363)):

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

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. The conditional distribution $p(C_i|C_{r(i)})$ can be written

$$p(x_{C_i}|x_{C_{r(i)}}) = \frac{p(x_{C_i})p(x_{C_{r(i)}})}{p(x_{S_{ir(i)}})}$$

where $S_{ir(i)} = C_i \cap C_{r(i)}$ is known as a "separator".

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. Note that the length of this path is at most $n-1$. This can be seen by using the *running intersection property* of the junction tree (Koller and Friedman (2009, p. 348)): If $i \in C_k, C_j$ for some $1 \leq k < j \leq K$, then $i \in C_h$ for every $h = k+1, ..., j-1$. And since the cliques $C_1, ..., C_K$ are maximal, it follows that every C_k along the sequence must introduce at least one element $i \notin \cup_{j<k} C_j$. As a result, it must be the case that $K \leq n-1$. Furthermore, if G is not linear, the inequality is strict.

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 over a suitably defined set of

nodes:

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

where $C_0 = \{1\}$.

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. Therefore, all the conditional distributions it induces are Gaussian, and thus we can write $p_G(x_1, x_{C_1}, ..., x_{C_{K-1}}, x_n)$ via the following recursion:

$$\begin{aligned} x_1 &= N(0,1) \\ x_{C_1} &= A_1 x_1 + \varepsilon_1 \\ &\vdots \\ x_{C_k} &= A_{k-1} x_{C_{k-1}} + \varepsilon_k \\ &\vdots \\ x_n &= A_K x_{C_K} + \varepsilon_n \end{aligned} \tag{11}$$

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, and the ε_k's are all independently distributed with zero mean, such that $E(x_{C_k}|x_{C_{k-1}}) = A_k x_{C_{k-1}}$. Therefore,

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

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

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

Step 2: Defining a new distribution over scalar variables
For every k, define the variable

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

Plugging the recursion (11), we obtain a recursion for z:

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

This means that a researcher who fits a recursive model given by the linear DAG $G' : x_1 \to z_2 \to \cdots \to z_{K-1} \to x_n$ will obtain the following estimated model, where all the ε_k's are independent variables:

$$\begin{aligned} x_1 &= N(0,1) \\ z_2 &= \alpha_1 A_1 x_1 + \varepsilon_2 \\ &\vdots \\ z_{k+1} &= z_k + \alpha_{k+1} \varepsilon_{k+1} \\ &\vdots \\ x_n &= z_{K-1} + \varepsilon_n \end{aligned}$$

Therefore, $E_{G'}(x1, xn)$ is given by:

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

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

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

We have thus reduced our problem to finding the largest $\hat{\rho}_{1n}$ that can be

attained by a linear DAG $G : 1 \to \cdots \to n$ of length n at most. ■

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

$$\hat{\rho}_{1n} = \prod_{k=1}^{n-1} \rho_{k,k+1} \tag{12}$$

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-definitenss. 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:

$$\rho^*_{1n} = \max_{\substack{a_i^T a_i=1 \text{ for all } i \\ a_1^T a_n=r}} \prod_{i=1}^{n-1} a_i a_{i+1}^T \tag{13}$$

Denote $\alpha = \arccos r$. Since the solution to (13) 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.

Perform a logarithmic transformation of the objective function and write

down the Lagrangian for the problem:

$$L = \sum_{i=1}^{n-1} \log(a_i a_{i+1}^T) + \sum_i \lambda_i (a_i a_i^T - 1)$$

The first-order condition of L with respect to a_i gives:

$$\frac{a_{i-1}}{a_i a_{i-1}^T} + \frac{a_{i+1}}{a_i a_{i+1}^T} = \lambda_i a_i \tag{14}$$

This means that there exist scalars α, β such that $a_i = \alpha a_{i-1} + \beta a_{i+1}$ where a_i, a_{i-1}, a_{i+1} are vectors. This means that any local extremum must satisfy that a_k be a linear combination of a_{k-1} and a_{k+1}. In other words, a_k must lie on the plane defined by the origin and a_{k-1}, a_{k+1}.

Now suppose that a_{k-1} and a_{k+1} are fixed and we wish to maximize the objective with respect to a_k. 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, $\alpha + \beta = \gamma$. Thus, when we are optimizing with respect to a_k, we maximize

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

Differentiating this expression with respect to α and setting the derivative to zero, we obtain $\alpha = \beta = \frac{\gamma}{2}$. Since this must hold for *any* k, 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 (5).

To summarize this argument, equation (14) is a necessary condition for a local extremum, implying that any local extremum $(a_2, ..., a_{n-1})$ must lie on a great circle connecting a_1 and a_n. Since we showed that the maximum among these local extrema satisfies the equi-distance property, we have established that this is a global maximum of the entire problem.

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

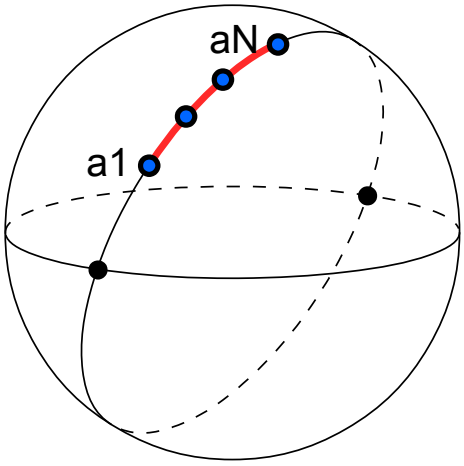

Figure 2: Geometric intuition for the proof.

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 harmonic 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.

7 Appendix II: Uniform Binary Variables

Suppose now that 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 $p_G(x_n) = \frac{1}{2}$ is generally open. However, when we fix G to be the linear DAG

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

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 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

$$p_G(x_j \mid x_i) = \sum_{x_{i+1},\ldots x_{j-1}} p(x_{i+1} \mid x_i)p(x_{i+2} \mid x_{i+1}) \cdots p(x_j \mid x_{j-1}) \tag{15}$$

In particular,

$$\begin{aligned} p_G(x_n \mid x_1) &= \sum_{x_2,\ldots x_{n-1}} p(x_2 \mid x_1)p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1}) \\ &= \sum_{x_2} p(x_2 \mid x_1)p_G(x_n \mid x_2) \end{aligned} \tag{16}$$

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: $2 \to \cdots \to n$. This observation will enable us to apply an inductive proof to our result.

Lemma 3 *For every p,*

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

Proof. Applying simple algebraic manipulation of (16), $\hat{\rho}_{1n}$ is equal to

$$\begin{aligned} & [p(x_2 = 1 \mid x_1 = 1) - p(x_2 = 1 \mid x_1 = -1)]\,[p_G(x_n = 1 \mid x_2 = 1) - p_G(x_n = 1 \mid x_2 = -1)] \\ = \;& \rho_{12} \cdot \hat{\rho}_{2n} \end{aligned}$$

■

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

Proposition 2 *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 $\hat{\rho}_{in} = r$, 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,

$$\begin{aligned} p(x_2 &= 1 \mid x_1 = 1) = p(x_n = 1 \mid x_1 = 1) \cdot \alpha_{11} + p(x_n = -1 \mid x_1 = 1) \cdot \alpha_{10} \\ &= q\alpha_{11} + (1-q)\alpha_{10} \end{aligned}$$

Likewise,

$$\begin{aligned} p(x_2 &= 1 \mid x_1 = 0) = p(x_n = 1 \mid x_1 = 0) \cdot \alpha_{01} + p(x_n = 0 \mid x_1 = 0) \cdot \alpha_{00} \\ &= q\alpha_{00} + (1-q)\alpha_{01} \end{aligned}$$

The objective correlation between x_1 and x_2 is thus

$$\rho_{12} = q(\alpha_{11} - \alpha_{00}) + (1-q)(\alpha_{10} - \alpha_{01}) \tag{17}$$

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}:

$$\rho_{2n} = q(\alpha_{11} - \alpha_{00}) + (1-q)(\alpha_{01} - \alpha_{10}) \tag{18}$$

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

Recall that $\hat{\rho}_{2n}$ is the expression we would have for the linear DAG $2 \to \cdots \to n$ when $p(x_2 = x_n) = \tilde{q}$. Therefore, by the inductive step,

$$\begin{aligned} \hat{\rho}_{1n} &= \rho_{12} \cdot \hat{\rho}_{2n} \\ &\leq [q(\alpha_{11} - \alpha_{00}) + (1-q)(\alpha_{10} - \alpha_{01})] \cdot \left(1 - \frac{1-\rho_{2n}}{k-1}\right)^{k-1} \end{aligned} \tag{19}$$

Both ρ_{12} and ρ_{2n} increase in α_{11} and decrease in α_{00}, such that we can set $\alpha_{11} = 1$ and $\alpha_{00} = 0$ without lowering the R.H.S of (19). This enables us to write

$$\rho_{12} = q + (1-q)(\alpha_{10} - \alpha_{01})$$

such that

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

Therefore, we can transform (19) 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 $r = 0$. Then, 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 false 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 4 *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 in a loss of generality. We conjecture that in the case of uniform binary variables, a non-linear perfect DAG can generate larger false correlations for sufficiently large n.

8 Appendix III: Proof of Proposition 1

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 $z = \alpha x_1 + \gamma x_n + \eta$. Since all variables on the R.H.S of this equation are independent (and since x_1 and x_n are standardized normal variables), it follows that the objective variance of z is

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

The researcher's model can now be written as

$$x_n = \frac{1}{\gamma} z - \frac{\alpha}{\gamma} x_1 - \frac{1}{\gamma} \eta \tag{20}$$

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

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

Because x_1 and x_n are independent, standardized normal, $\hat{E}(x_1, x_n) = -\alpha/\gamma$. The researcher's model does not distort the variance of x_1.[5] Therefore, $\hat{Var}(x_1)$. And since the researcher's model regards z, x_1 and η as independent,

$$\hat{Var}(x_n) = \left(\frac{1}{\gamma}\right)^2 Var(z) + \left(\frac{\alpha}{\gamma}\right)^2 + \left(\frac{\sigma}{\gamma}\right)^2 = \left(\frac{1}{\gamma}\right)^2 \left(\alpha^2 + \gamma^2 + \sigma^2\right) + \left(\frac{\alpha}{\gamma}\right)^2 + \left(\frac{\sigma}{\gamma}\right)^2$$

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

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

which is increasing in $|\alpha/\gamma|$ and attains an upper bound of $1/\sqrt{2}$ when $\alpha/\gamma \to -\infty$.

Note that since without loss of generality we can set $\gamma = \sqrt{1 - \alpha^2}$ such

[5]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.

that $z \sim N(0,1)$. Therefore, the upper bound is approximated to an arbitrarily fine degree when we set $\alpha \to -1$ such that $\gamma \to 0$. As a result, the estimated variance of x_n diverges.

References

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

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

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

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

[5] 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.

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

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

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

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

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

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

[12] 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.

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

[14] Spiegler, R. (2019), Can Agents with Causal Misperceptions be Systematically Fooled?, *Journal of the European Economic Association*, forthcoming.

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