"Data Monkeys": A Procedural Model of Extrapolation from Partial Statistics*

Ran Spiegler†

January 12, 2017

Abstract

I present a behavioral model of a "data analyst" who extrapolates a fully specified probability distribution over observable variables from a collection of statistical datasets that cover partially overlapping sets of variables. The analyst employs an iterative extrapolation procedure, whose individual rounds are akin to the stochastic-regression method of imputing missing data. Users of the procedure's output fail to distinguish between raw and imputed data, and it functions as their practical belief. I characterize the ways in which this belief distorts the correlation structure of the underlying data generating process - focusing on cases in which the distortion can be described as the imposition of a causal model (represented by a directed acyclic graph over observable variables) on the true distribution.

*Earlier versions of this paper were circulated under the titles "Bayesian Networks and Missing-Data Imputation" and "On the Limited Feedback Foundation of Boundedly Rational Expectations". The paper has benefitted from ESRC grant no. ES/L003031/1 and ERC grant no. 692995. I am grateful to Noga Alon, Yair Antler, Simon Byrne, Philip Dawid, Andrew Ellis, Erik Eyster, Kfir Eliaz, Ehud Lehrer, Itay Saporta-Ecksten, Heidi Thysen, numerous seminar participants, the editor and three referees for useful comments.

†Tel Aviv University, University College London and Centre for Macroeconomics. URL: http://www.tau.ac.il/~rani. E-mail: rani@post.tau.ac.il.

"Data Monkey: One who spends the majority of their time running data and creating useless PowerPoint slides to please the upper echelons of management." (Urban Dictionary, http://www.urbandictionary.com)

1 Introduction

Members of modern organizations are often required to process and present statistical data. Conventional economic models assume that agents in such situations act as impeccable statisticians. In reality, the typical analyst will not reach the heights of statistical sophistication that characterize, say, an academic econometrician. He will often perform statistical procedures without putting much thought to them, without understanding the implicit assumptions behind them, and without internalizing their implications for the validity of various inferences.

Even when the analyst does know what he is doing, he faces pressure to present data in an easily digestible format and underplay its noisiness. As a result, his final report may shroud the underlying data limitations and data-processing methods. For instance, Silver (2012) criticizes economic forecasters' tendency to present point estimates without providing confidence intervals. Within organizations, the pressure to avoid technical details in communication with "upper echelons of management" is known in management folklore as one of "Putt's Laws": "Technical analyses have no value above the midmanagement level" (see Putt (2006, p. 109)). Finally, the analyst's successors may be unaware of "how the data sausage was made" (the analyst himself may later forget this), due to imperfect organizational memory.

The upshot in all these situations is that users of the processed data are likely to take it at face value, without accounting for how the bottom line was reached. I refer to a data analyst who communicates processed data without imparting the underlying data limitations and data-processing methods to

his audience as a "*data monkey*", adapting the motto's colloquial term.

One example of this general phenomenon involves extrapolation from "non-rectangular" databases. Analysts regularly confront datasets with missing values, or multiple datasets that cover different sets of variables. Turning them into presentable output requires the analyst to adopt methods for handling missing data. However, these methods often remain hidden. In the world of academic research, this problem has been documented in various areas, including medicine (Burton and Altman (2004), Wood et al. (2004), Mackinnon (2010)), education studies (Manly and Wells (2015)) and economics (Kaplan and Schulhofer-Wohl (2012), Meyer et al. (2014)). In particular, Kaplan and Schulhofer-Wohl describe how an undocumented change in the US Census Bureau's method of imputing missing data in the CPS had led economists and demographers to identify a spurious trend in geographic mobility. They speculate that a certain break in the data may be due to an unknown aspect of the data-processing methods employed by the Census Bureau. I am not aware of systematic evidence about this phenomenon outside academia, but causal observation suggests that it is at least as high.

This paper presents a model of an analyst who employs a natural procedure for extrapolating a probability distribution from partial statistics. Users of the procedure's output (including possibly the analyst's future self) take it at face value and it becomes their practical belief. I take this attitude of users as given, and do not try to derive it from a more basic model. I characterize the ways in which this belief distorts the correlation structure of the underlying data generating process. My perspective in this paper is descriptive rather than normative. I do not deal with how processing, reporting and consuming statistical data *should* be done, but rather with (a stylized model of) how they *are* done, and I am interested in the belief errors that result from the failure to distinguish between raw and imputed data.

1.1 An "Oligopoly" Example

To introduce the main idea, consider the following scenario. A fresh business graduate has just landed a job as a junior analyst in a consulting firm. The analyst is ordered to write a report about an oligopolistic industry. He is not told who is going to use this report or for what purpose. The analyst gathers data about three variables: the product price (denoted y) and the production quantities chosen by producers 1 and 2 (denoted x_1 and x_2). Specifically, he manages to get hold of two proprietary datasets. Each dataset $i = 1, 2$, belonging to producer i, consists of a large number of historical realizations of the product price and producer i's quantity. The datasets cover different, non-overlapping time periods.

Having collected the data, the analyst wishes to prepare a file that subsequent users (himself included) can readily process. As a first step, he merges the two datasets into a single spreadsheet, which is schematically illustrated by the following table:

x_1	x_2	y
$+$	$-$	$+$
$-$	$+$	$+$

Each row in this table represents a block of observations originating from one of the datasets; a "+" ("−") sign in a cell indicates that the value of the relevant variable is recorded (missing) in the relevant spreadsheet fields.

Because the original datasets cover different sets of variables, the merged spreadsheet is "non-rectangular" - i.e., it contains missing values. The analyst wants to fill the missing cells, in order to produce a "rectangular" spreadsheet that is amenable to rudimentary statistical analysis that can be conveyed to users in the form of plot diagrams, tables, simple regressions, etc. The rub is that subsequent users of the analyst's report may treat the processed spreadsheet as if it purely consisted of raw data, whereas in fact it mixes raw and imputed values. The frequencies of (x_1, x_2, y) in the rectangularized spreadsheet will serve as a practical estimate of the joint distribution

over prices and quantities in the industry, and this de-facto belief may systematically distort the true underlying data generating process.

Our analyst employs the following method for filling the missing cells in his spreadsheet. When the value of x_1 is missing in some row, he relies on the observed realization of y in the row, combined with the joint distribution over (x_1, y) given by the first underlying dataset, to impute a value for x_1 in this row. Specifically, he draws this value from the first dataset's distribution over x_1 conditional on y. Likewise, when the value of x_2 is missing in some row, the analyst draws its imputed value from the second dataset's distribution over x_2 conditional on the realization of y in the row.

This extrapolation method is intuitive: the analyst fills the missing values of x_1 and x_2 according to his best available evidence (namely, how these variables are correlated with y), without invoking any explicit prior theory about the joint distribution of x_1 and x_2, for which he has no evidence. The method also has professional credentials: in fact, it is a variant on a familiar imputation method known as "stochastic conditional imputation" or "stochastic regression" (Little and Rubin (2002, Ch. 4), Gelman and Hill (2006, Ch. 25)). Thus, the difficulties are not with the procedure *per se*, but with the danger that users will take its output at face value.

To see these difficulties, let us derive the distribution over (x_1, x_2, y) in the rectangularized spreadsheet. First, we need to specify the data generating process. Suppose that each observation in each of the original datasets was independently drawn from an objective joint distribution p over (x_1, x_2, y). Assume further that the two datasets are arbitrarily large. Therefore, the first dataset enables the analyst to learn the true marginal distribution $p(x_1, y)$, whereas the second dataset enables him to learn the true marginal distribution $p(x_2, y)$. As a result, in the block of observations that originates from the first dataset, frequencies of (x_1, x_2, y) in the rectangularized spreadsheet are given by $q_1(x_1, x_2, y) = p(x_1, y)p(x_2 \mid y)$. Similarly, in the block that originates from the second dataset, these frequencies are given by

$q_2(x_1, x_2, y) = p(x_2, y)p(x_1 \mid y)$. The overall frequencies in the final spreadsheet, denoted q, are thus given by a *weighted average* of q_1 and q_2, where the weights match the blocks' relative size. However, observe that by the basic rules of conditional probability,

$$q_1(x_1, x_2, y) = q_2(x_1, x_2, y) = p(y)p(x_1 \mid y)p(x_2 \mid y) \tag{1}$$

Thus, both blocks in the rectangularized spreadsheet exhibit the same frequencies, such that q will be given by the R.H.S of (1), independently of the original datasets' relative size. It is evident from this expression that q satisfies the conditional-independence property $x_1 \perp x_2 \mid y$. Indeed, both q_1 and q_2 satisfy it. This property is a direct consequence of the analyst's extrapolation method; it may be violated by the objective distribution p itself. When users of the analyst's report ignore the extrapolation method behind it and treat it as raw data, they will misperceive the true correlation structure of p and potentially make economically significant errors.

For instance, a user may notice that q can be expressed by the R.H.S of (1). This formula suggests that the price is an exogenous variable and that producers are independent "price takers". This is a *causal* story that can be summarized by the *directed acyclic graph* (DAG) $x_1 \leftarrow y \rightarrow x_2$ - i.e., the price is a primary cause, whereas quantities are conditionally independent consequences. However, this "price taking" account is empirically groundless - it is merely an artifact of the data limitations and the imputation procedure. Alternatively, suppose that the report's user does not reach the insight that q is consistent with a causal model. Instead, he directly measures the correlation between x_1 and x_2 induced by q. It is easy to construct objective distributions p for which x_1 and x_2 are statistically independent and nevertheless deemed correlated by q. Upon noticing this correlation, the user may suspect that producers coordinate their activities, in violation of anti-trust regulations. However, this is a false interpretation of the data: it is due to the fact that the imputation procedure makes these variables appear

independent *conditionally* on y.

Finally, suppose that at some later point in time, our analyst receives a third dataset that documents the joint distribution of y and x_3, the production quantity of a third producer (without giving any information that relates realizations of x_3 to simultaneous realizations of x_1 or x_2). By then, the analyst has forgotten the origins of his earlier report, having left no record of its preparation. Like earlier users of that report, he treats q as raw data. The following is a schematic illustration of the spreadsheet that is produced by merging the earlier processed database and the new dataset (the star symbol indicates imputed values - although this is a distinction that the analyst himself does not make):

x_1	x_2	y	x_3
$+$	$*$	$+$	$-$
$*$	$+$	$+$	$-$
$-$	$-$	$+$	$+$

The analyst proceeds to extrapolate a new joint distribution from this database, using the same "stochastic conditional imputation" method. When the value of x_3 is missing in some row, he relies on the observed realization of y in the row, combined with the distribution q over (x_1, x_2, y), to impute a value for x_1 in this row. And when the value of (x_1, x_2) is missing in some row, he relies on the observed realization of y in the row, combined with the distribution $p(y, x_3)$ provided by the new dataset, to impute a value for (x_1, x_2) in this row. The overall frequencies in the first and second blocks of the newly rectangularized spreadsheet, denoted r_1 and r_2, are given by

$$\begin{aligned} r_1(x_1, x_2, y, x_3) &= q(x_1, x_2, y)p(x_3 \mid y) \\ r_2(x_1, x_2, y, x_3) &= p(y, x_3)q(x_1, x_2 \mid y) \end{aligned}$$

Plugging (1) for q, we see that r_1 and r_2 coincide and can be written as

follows:

$$r(x_1, x_2, y, x_3) = p(y)p(x_1 \mid y)p(x_2 \mid y)p(x_3 \mid y)$$

As before, this formula for r suggests a causal story that can be described by a DAG:

$$\begin{array}{ccccc} x_1 & \leftarrow & y & \rightarrow & x_2 \\ & & \downarrow & & \\ & & x_3 & & \end{array}$$

This is essentially the same causal story as before: all three producers are "price takers" whose quantities are conditionally independent consequences of the exogenous product price y. Thus, the basic extrapolation procedure can be iterated as new datasets continue to arrive, and its output can be consistent with increasingly elaborate causal models.

1.2 Preview of the Model and the Main Results

In Section 2, I define the notion of an ordered database as a sequence of variable sets for which the joint distribution p is given. In the oligopoly example, the ordered database consisted of the datasets {product price, producer 1's quantity}, {product price, producer 2's quantity} and {product price, producer 3's quantity}. I formalize the iterative extrapolation procedure, which takes the ordered database as an input and produces a fully specified probability distribution as an output. The main question is how this output distorts the correlation structure of the true distribution p.

In Section 1.1, we saw how the procedure's output distorts the objective distribution p as if it imposes a causal model on it. I generalize this idea and define the notion of a "DAG representation" of systematic belief distortion - following Spiegler (2015), which itself drew on the Bayesian-networks literature (Cowell et al. (1999), Pearl (2009)). In Section 3, I apply the iterative extrapolation procedure to a few economically motivated examples of ordered databases, and I examine whether the output has a (possibly "mixed") DAG

causal interpretation.

Section 4 is the analytical heart of this paper. I introduce a combinatorial property of ordered databases - known in the Bayesian-networks literature as the "running intersection property" - which requires that the intersection of every dataset with the union of its predecessors is contained in one of them. The "oligopoly" example satisfies this property. It turns out that the running intersection property ensures that the iterative extrapolation procedure's output has an essentially unique DAG representation. (The result is in fact stronger, as it is based on a weaker, analogous property of *unordered* databases.) Moreover, it forces the DAG to be "perfect" - i.e., it has the property that if two variables are perceived as direct causes of some third variable, they must be directly linked (the DAGs in the "oligopoly" example are perfect because they vacuously satisfy the antecedent). Conversely, I show that if a given collection of datasets cannot be ordered with a running intersection, the procedure's output lacks a DAG representation - i.e., for every DAG we can find p for which the procedure's output does not distort p according to the DAG.

As a corollary of these two results, I obtain the paper's main result: A DAG representation can be justified as the procedure's output if and only if the DAG is perfect. Perfect DAGs are special in the sense that the direction of their links cannot be identified from observational data: for any link in the DAG, we can find an observationally equivalent DAG that inverts it. The lesson is that the only DAG representations that can be extrapolated from ordered databases (via the iterative procedure) are ones whose causal interpretation is vacuous. Armed with this result, I comment on whether familiar examples of misspecified subjective models (involving phenomena such as correlation neglect) can be justified as the outcome of procedural extrapolation from partial statistics. In later parts, I examine two variations of the model: an alternative extrapolation method based on the maximal-entropy principle, and extrapolation from selective datasets.

2 Datasets and Extrapolation

Let $x_1, ..., x_n$ be a collection of observable random *variables,* where $n \geq 2$ and x_k takes values in the set X_k. Denote $X = X_1 \times \cdots \times X_n$. Let N be the set of variable *indices*. For most purposes, it will be simplest to set $N = \{1, ..., n\}$. However, in examples, it is often useful to notate indices such that their relation to the variables is more transparent. For every $S \subseteq N$, denote $x_S = (x_k)_{k \in S}$ and $X_S = \times_{k \in S} X_k$. Let $p \in \Delta(X)$ be an *objective probability distribution.* I use $p^S \in \Delta(X_S)$ to denote the marginal of p over X_S.

An *analyst* obtains partial statistical data regarding p, in the form of a sequence of m datasets, enumerated $1, ..., m$. The interpretation is that the datasets gradually become available to the analyst. The k^{th} dataset consists of infinitely many observations of a subset of variables $S_k \subset N$. Each of these observations is a random independent draw from p. Therefore, the k^{th} dataset enables the analyst to learn the true marginal p^{S_k}. The sequence $(S_1, ..., S_m)$ is referred to as an *ordered database* and denoted $\bar{\mathcal{S}}$, whereas the set $\{S_1, ..., S_m\}$ is referred to as an *unordered database* and denoted $\mathcal{S}$. For every $k = 1, ..., m$, denote $B_k = \cup_{j \leq k} S_k$. Assume $B_m = N$ - i.e., $\mathcal{S}$ is a cover of N. This reflects the definition of $x_1, ..., x_n$ as a collection of *observable* variables. I also assume that no two subsets $S, S' \in \mathcal{S}$ contain one another; this assumption is made purely because it simplifies notation at certain points.

At first glance, my formulation of databases does not fit the "spreadsheet" metaphor. However, the extrapolation procedure that I present below can be defined for a more elaborate formulation that matches the metaphor more closely. Indeed, when stated in such terms, the procedure's interpretation as a method of imputing missing values becomes manifest. However, because the results in this paper are the same under either formalism, I opted for the simpler version. For this reason, I will tend to avoid the term "imputation" in the sequel and use the more neutral term "extrapolation" instead.

2.1 An Iterative Extrapolation Procedure

The analyst extends the collection of marginals $(p^{S_1}, ..., p^{S_m})$ into a fully specified probability distribution over X, according to the following iterative procedure, which consists of m rounds. The output of each round is a provisional belief $q^k \in \Delta(X_{B_k})$, defined inductively as follows:

- For $k = 1$, $q^1 = p^{S_1}$.
- For every subsequent round $k = 2, ..., m$, define two auxiliary distributions over X_{B_k}:

$$\begin{aligned} q_1^k(x_{B_k}) &= q^{k-1}(x_{B_{k-1}}) p(x_{S_k - B_{k-1}} \mid x_{S_k \cap B_{k-1}}) \\ q_2^k(x_{B_k}) &= p(x_{S_k}) q^{k-1}(x_{B_{k-1} - S_k} \mid x_{S_k \cap B_{k-1}}) \end{aligned} \tag{2}$$

  and let

$$q^k = \alpha^k \cdot q_1^k + (1 - \alpha^k) \cdot q_2^k \tag{3}$$

  where $\alpha^k \in (0, 1)$ is some constant.

The distribution q^m is the procedure's final output, which I also denote by $f(\bar{S}, \alpha, p)$.

This procedure is a straightforward iteration of the basic "stochastic conditional imputation" method described in Section 1.1. In round 1, the analyst only has access to the dataset that covers the set of variables S_1, and this enables him to learn p^{S_1}. In round 2, he receives an additional dataset that covers S_2, and he employs the "stochastic conditional imputation" method to extrapolate the learned marginals p^{S_1} and p^{S_2} into a distribution $q^2 \in \Delta(X_{S_1 \cup S_2})$. By the end of round $k - 1$, having confronted the partial (ordered) database $(S_1, ..., S_{k-1})$, the analyst has tentatively extended the marginals $p^{S_1}, ..., p^{S_{k-1}}$ into a probability distribution $q^{k-1} \in \Delta(X_{B_{k-1}})$. He subsequently treats q^{k-1} *as if it were "raw" data* - although (for every

$k > 2$) it is partly a consequence of earlier extrapolations. And when he receives the k^{th} dataset, he once again employs "stochastic conditional imputation" to extrapolate q^{k-1} and p^{S_k} into a distribution $q^k \in \Delta(X_{B_{k-1} \cup S_k})$. The procedure is terminated at the end of round m, such that the belief q^m is the procedure's final output.

Each round k in the iterative extrapolation procedure (referred to as **IEP** henceforth) involves *two simultaneous extrapolations*, given by the auxiliary distributions q_1^k and q_2^k. The coefficient α^k is the weight of q_1^k in determining the provisional output of round k. These weights may reflect some intuitive perception of the quality of different data sources, the importance of the variables they cover, or the datasets' relative size (although they are all assumed to contain infinitely many observations, this is a mere approximation for large finite datasets).

The latter interpretation fits the "spreadsheet" metaphor. Let $\sigma_k > 0$ represent the size of the k^{th} dataset. When the analyst performs "stochastic conditional imputation" in round k, this method produces $\alpha^k = 1 - \sigma_k / \sum_{j \leq k} \sigma_j$. The reason is that rectangularizing the spreadsheet blocks that cover B_{k-1} and S_k into a single block that covers $B_k = B_{k-1} \cup S_k$ involves two steps: using p^{S_k} to impute missing values of $x_{S_k - B_{k-1}}$ in the first block, and using q^{k-1} to impute missing values of $x_{B_{k-1} - S_k}$ in the second block. The two steps yield q_1^k and q_2^k, respectively. In the rectangularized block, the weights that q_1^k and q_2^k get depend on the relative size of B_{k-1} and S_k.

The IEP is entirely non-parametric and invariant to the variables' meaning. This is an attractive feature when X lacks intrinsic structure. In contrast, when variables get real values and the analyst has prior reasons to hypothesize, say, a linear relation among variables, it would be plausible to incorporate this hypothesis into the extrapolation process (by literally using a linear regression in the construction of q_1^k and q_2^k). Thus, while the procedure's generality makes it widely applicable, it also calls for adjustments in certain applications.

2.2 DAG Representations

I now formally introduce the notion of a DAG representation, which is a class of functions that systematically distort the correlation structure of objective probability distributions as if they fit them to a causal model. The IEP's output in the "oligopoly" example of Section 1.1 belongs to this class.

Let (N, R) be a directed graph, where N (the set of variable indices) is the set of nodes and R is the set of directed links. (In some cases, as in the oligopoly example, I abuse notation and take the variable labels themselves to be the nodes, in order to make the DAG's meaning more transparent.) I use the notations jRi and $j \to i$ interchangeably. The graph is *acyclic* if it does not contain any directed path from a node to itself. From now on, I identify R itself with the DAG. For every $i \in N$, denote $R(i) = \{j \in N \mid jRi\}$.

Fix a DAG R. For every objective distribution $p \in \Delta(X)$, define

$$p_R(x) = \prod_{i \in N} p(x_i \mid x_{R(i)}) \tag{4}$$

The distribution p_R is said to *factorize* p *according to* R. For instance, when $R : 1 \to 2 \to 3 \leftarrow 4$, $p_R(x) = p(x_1)p(x_4)p(x_2 \mid x_1)p(x_3 \mid x_2, x_4)$.

A DAG and the set of distributions that it factorizes constitute a *Bayesian network*. In what follows, I refer to p_R as a DAG representation. Its interpretation here and in Spiegler (2015) differs from existing interpretations in the Statistics and Artificial-Intelligence literature (e.g., see Cowell et al. (1999)), in that p is viewed as a *true* "steady state" distribution, such that p_R *systematically distorts* an objective distribution into a subjective belief. Following Pearl (2009), a DAG can be interpreted as a causal model, such that the link $i \to j$ means that x_i is considered to be an immediate cause of x_j. From this point of view, $R(i)$ represents the set of immediate causes of the variable x_i.

Different DAGs can be equivalent in terms of the distributions they factorize.

Definition 1 (Equivalent DAGs) *Two DAGs R and Q are* ***equivalent*** *if $p_R = p_Q$ for every $p \in \Delta(X)$.*

To take the simplest example, the DAGs $1 \rightarrow 2$ and $2 \rightarrow 1$ are equivalent, since $p(x_1)p(x_2 \mid x_1) = p(x_2)p(x_1 \mid x_2)$. Likewise, all fully connected DAGs are equivalent: in this case, the factorization formula (4) reduces to a textbook chain rule.

Frydenberg (1990) and Verma and Pearl (1991) provided a complete characterization of the equivalence relation. The *skeleton* of R, denoted $\tilde{R}$, is its non-directed version - that is, $i\tilde{R}j$ if iRj or jRi. The *v-structure* of a DAG R is the set of all triples of nodes i, j, k such that iRk, jRk, $i \not R j$ and $j \not R i$.

Proposition 1 (Verma and Pearl (1991)) *Two DAGs R and Q are equivalent if and only if they have the same skeleton and the same v-structure.*

For instance, $1 \rightarrow 3 \leftarrow 2$ and $1 \rightarrow 3 \rightarrow 2$ have identical skeletons but different v-structures. Therefore, these DAGs are not equivalent: there exist distributions that can be factorized by one DAG but not by the other. In contrast, $1 \rightarrow 3 \rightarrow 2$ and $1 \leftarrow 3 \leftarrow 2$ are equivalent because they have the same skeleton and the same (vacuous) v-structure. As will see in the next section, two DAGs can be equivalent in terms of Definition 1, and yet differ in terms of the plausibility of their causal interpretation.

3 Examples

The following examples illustrate the IEP and examine whether its output has a DAG representation. By this, I mean that the IEP's output q^m is equal to the formula (4) for some DAG R and for *every* possible objective distribution p (or, in some cases, for every p in some restricted domain).

Example 3.1: An "availability bias"

Let $m = n - 1$. The unordered database is $\mathcal{S} = \{\{1, k+1\}\}_{k=1,...,m}$. This specification fits the "oligopoly" example of Section 1.1. An alternative story is that each observation in a dataset consists of various characteristics of some individual. A basic demographic characteristic such as age or gender - summarized by x_1 - is available in every dataset. Other characteristics - summarized by $x_2, ..., x_n$ - are domain-specific (e.g., tax or health records) and appear in one dataset only. Each dataset is a large sample drawn from an arbitrarily larger population, such that the probability that the same individual appears in multiple samples is negligible.

For any ordering of $\mathcal{S}$, the output of the IEP is

$$q^m(x_1, ..., x_n) = p(x_1) \cdot \prod_{i=2}^{n} p(x_i \mid x_1)$$

The proof is by simple induction. Without loss of generality, order the database as follows: $\bar{\mathcal{S}} = (\{1, 2\}, ..., \{1, m\})$. Suppose that the provisional output of round $k \geq 2$ is

$$q^{k-1}(x_1, ..., x_k) = p(x_1) \cdot \prod_{i=2}^{k} p(x_i \mid x_1)$$

For $k = 2$, this can be established exactly as in Section 1.1. In round $k + 1$, the auxiliary distribution q_1^{k+1} is by definition

$$q_1^{k+1}(x_1, ..., x_{k+1}) = q^k(x_1, ..., x_k) \cdot p(x_{k+1} \mid x_1) = p(x_1) \cdot \prod_{i=2}^{k+1} p(x_i \mid x_1)$$

The auxiliary distribution q_2^{k+1} is

$$\begin{aligned} q_2^{k+1}(x_1, ..., x_{k+1}) &= p(x_1, x_{k+1}) \cdot q_1^k(x_1, ..., x_k \mid x_1) = p(x_1, x_{k+1}) \cdot \prod_{i=2}^{k} p(x_i \mid x_1) \\ &= p(x_1)p(x_{k+1} \mid x_1) \cdot \prod_{i=2}^{k} p(x_i \mid x_1) = q_1^{k+1}(x_1, ..., x_{k+1}) \end{aligned}$$

hence $q^{k+1} = q_1^{k+1}$, which completes the proof.

The output q^m is a DAG representation, where the DAG R consists of all links $1 \to k$, $k = 2, ..., m$. It is consistent with a causal story: the individual's basic demographic characteristic is a primary cause and the other characteristics are conditionally independent consequences. This causal interpretation suffers from an "*availability bias*": One characteristic ends up appearing like as a cause of the others, only because it happens to be available in every dataset. Moreover, by Proposition 1, every other DAG in the equivalence class of R reverses *exactly* one link (otherwise, the DAG's v-structure would not be preserved), such that x_1 is perceived as the cause of all other variables save one. I will comment on the generality of this availability bias in Section 4.1.

Example 3.2: "Education"
Our tireless analyst now performs a consulting job for a higher-education institution. He gets access to three datasets that cover (in total) four individual characteristics: number of years of schooling (denoted s), outcome of a childhood intelligence test (denoted c), father's number of years of schooling (denoted f) and wage earnings in adulthood (denoted w). Let δ_y denote the index of any variable $y = s, c, f, w$. The ordered database is $\bar{\mathcal{S}} = (\{\delta_f, \delta_c\}, \{\delta_c, \delta_s\}, \{\delta_s, \delta_w\})$. The interpretation is that datasets arrive gradually over time, and each dataset only covers recent variables. For instance, records about individuals' wage earnings are likely to indicate whether they have a college degree, but less likely to specify their childhood test per-

formance. Here is a schematic illustration of the "spreadsheet" that merges the datasets:

$$\begin{array}{cccc} f & c & s & w \\ + & + & - & - \\ - & + & + & - \\ - & - & + & + \end{array}$$

Let us execute the IEP. In round 1, $q^1(f,c) = p(f,c)$. Round 2 works exactly as in Example 1.1, such that $q^2(f,c,s)$ can be written as

$$q^2(f,c,s) = p(c)p(f \mid c)p(s \mid c) = p(f)p(c \mid f)p(s \mid c) = p(s)p(c \mid s)p(f \mid c)$$

Here is what our metaphorical spreadsheet looks like at the end of round 2:

$$\begin{array}{cccc} f & c & s & w \\ + & + & * & - \\ * & + & + & - \\ - & - & + & + \end{array}$$

Let us turn to the final round, by writing down explicit formulas for the auxiliary distributions:

$$q_1^3(f,c,s,w) = q^2(f,c,s)p(w \mid s) = p(f)p(c \mid s)p(s \mid c)p(w \mid s)$$

and

$$q_2^3 = p(s,w)q^2(f,c \mid s) = p(s)p(w \mid s)p(c \mid s)p(f \mid c) = q^2(f,c,s)p(w \mid s)$$

The two distributions thus coincide, such that the procedure's final output can be written as

$$q^3(f,c,s,w) = p(f)p(c \mid f)p(s \mid c)p(w \mid s) \tag{5}$$

As in previous examples, the exact values of the coefficients (α^k) are irrelevant for the final output.

Formula (5) is a DAG representation, where the DAG $f \to c \to s \to w$ can be interpreted as a causal chain: an individual's paternal education causes his childhood test performance, which in turn causes his schooling, which in turn causes his adult earnings. This causal interpretation is intuitive, because it tracks the chronological order of the variables' realizations. However, it may be false - e.g., when all four variables have a common unobserved cause. In this case, the conditional-independence properties of q^3 (such as $s \perp f \mid c$) will not be satisfied in reality. A user of the analyst's output may falsely infer that when he controls for children's test scores, he can predict their future school performance independently of their father's education.

Comment: Order effects
Unlike Example 3.1, here the IEP's output is sensitive to the order in which datasets appear. To see why, let $\bar{\mathcal{S}} = (\{\delta_f, \delta_c\}, \{\delta_s, \delta_w\}, \{\delta_c, \delta_s\})$. The provisional output of round 2 is $q^2(f, c, s, w) = p(f, c)p(s, w)$. In the final round, we have $q_1^3 = q^2$ and $q_2^3(f, c, s, w) = p(c, s)q^2(f, w \mid c, s)$. The procedure's final output is

$$q^3(f, c, s, w) = q^2(f, w \mid c, s) \left[\alpha^3 q^2(c, s) + (1 - \alpha^3)p(c, s)\right]$$

Consider an objective distribution p under which each of the variables f and w is independently distributed, whereas the variables c and s are mutually correlated. Then, $q^2(f, c, s, w) = p(f)p(c)p(s)p(w)$. Therefore,

$$q^3(f, c, s, w) = p(f)p(w) \left[\alpha^3 p(c)p(s) + (1 - \alpha^3)p(c, s)\right]$$

This expression underestimates the objective correlation between c and s (to an extent given by α^3). By comparison, under the same assumptions on p, expression (5) would be reduced to $p(f)p(w)p(c, s)$, which fully accounts for

the objective correlation between c and s.

Comment: The causal interpretation of q^m

the interpretation that I pursue in this paper is that the analyst lacks any prior theoretical prejudice: his objective is to obtain a rectangular database that is amenable to simple, presentable statistical analysis, and he employs an intuitive extrapolation procedure toward that end. It is not essential to this interpretation that a user of the procedure's output notices that it exhibits conditional-independence patterns that suggest a causal mechanism.

Suppose, however, that the user *does* notice that q^m is consistent with a DAG R. This may suggest a causal interpretation to him. And indeed, in cases like Example 3.2, the causal story *is* intuitive because it matches the variables' chronological ordering. This may reassure the user of the validity of the analyst's report, exacerbating his lack of interest in the methods behind it. However, recall the equivalence relation over DAGs. In Example 3.2, the procedure's output could be equivalently written as $q^3(f, c, s, w) = p(w)p(s \mid w)p(c \mid s)p(f \mid c)$, an expression that manifestly factorizes p according to the DAG $w \to s \to c \to f$. This DAG entirely reverses the chronological ordering and therefore makes no sense as a causal chain.

In other cases, q^m factorizes p according to some DAG R, and yet *no* DAG in the equivalence class of R would make sense as a causal model. For instance, in the context of the "education" story of Example 3.2, suppose that the database is as in Example 3.1, consisting of three datasets that record the correlation of w with each of the other variables f, c, s. Then, although q^3 has a DAG representation, the DAG necessarily regards w as an immediate cause of at least two other variables (per our discussion at the end of Example 3.1). This causal interpretation is absurd, because w is the *last* in the chronological order of the variables' realizations. In situations like these, when the user notices that q^m is consistent with an implausible causal story, he may be impelled to probe into the methods behind q^m, and thus become a more sophisticated user of processed statistical data.

Example 3.3: "Partial cursedness"
Let $N = \{1, 2, 3\}$, where x_1 represents the action of an uninformed player in a simultaneous-move game, whereas x_2 and x_3 represent the information and action of the player's opponent, respectively. In this story, we could identify the analyst with player 1 (or rather his strategic advisor). Consider the ordered database $\bar{\mathcal{S}} = (\{1, 2\}, \{1, 3\}, \{2, 3\})$. The first two rounds of the IEP are the same as in the previous examples. The provisional output of round 2 can be written as $q^2(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)$. Turning to the final round, note that $q_1^3 = q^2$, whereas

$$q_2^3(x_1, x_2, x_3) = p(x_2, x_3)q^2(x_1 \mid x_2, x_3) = q^2(x_1, x_2, x_3)\frac{p(x_2, x_3)}{q^2(x_2, x_3)}$$

This expression does not coincide with q_1^3. The final output is

$$q^3(x_1, x_2, x_3) = q^2(x_1, x_2, x_3)\left[\alpha^3 + (1 - \alpha^3)\frac{p(x_2, x_3)}{q^2(x_2, x_3)}\right] \tag{6}$$

This is not a DAG representation, and cannot be rewritten as such. However, the above simultaneous-game story suggests that we should restrict the domain of permissible objective distributions p to those for which $x_1 \perp x_2, x_3$ - equivalently, those that are consistent with the DAG R^* : $1 \qquad 3 \leftarrow 2$. The domain restriction implies $q^2(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$, and therefore $q^2(x_2, x_3) = p(x_2)p(x_3)$. The formula (6) is then simplified into

$$q^3(x_1, x_2, x_3) = p(x_1)p(x_2)\left[(1 - \alpha^3)p(x_3 \mid x_2) + \alpha^3 p(x_3)\right] \tag{7}$$

This is an example of what Spiegler (2015) calls a "mixed-DAG representation" - namely, a convex combination of two DAG representations - assigning weight $1 - \alpha^3$ to R^* and weight α^3 to the empty DAG. It matches what Eyster and Rabin (2015) call "partial cursedness", where α^3 measures player 1's "degree of cursedness" - i.e., the extent to which he neglects the correlation between the opponent's information and action. In this sense,

the IEP provides a foundation for partial cursedness. However, this foundation crucially relies on the assumption that $\{2,3\}$ is the last dataset in the ordered database; any other ordering would have let to an output that factorizes any objective distribution in the restricted domain according to R^*. Also, the foundation does not extend to other restricted domains of p with a simultaneous-game motivation. For instance, when x_2 represents an uninformed opponent's action and x_3 represents the game's outcome, the natural domain restriction is $x_1 \perp x_2$, and then q^3 loses the partial-cursedness structure.

4 General Analysis

In this section I characterize databases for which the IEP's output has a DAG representation, and the class of DAG representations that can emerge as outputs of the IEP. A few preliminaries are in order before I can state the main results. First, let us introduce a few properties of databases.

Definition 2 *An ordered database $\bar{\mathcal{S}} = (S_1, ..., S_m)$ satisfies the **running intersection property** (**RIP**) if for every $k = 2, ..., m$, $S_k \cap (\cup_{i<k} S_i) \subseteq S_j$ for some $j < k$.*

Definition 3 *An unordered database $\mathcal{S}$ satisfies **RIP*** if there exists an ordering $\bar{\mathcal{S}}$ of $\mathcal{S}$ that satisfies RIP.*

Definition 4 *An ordered database $(S_1, ..., S_m)$ is **maximally overlapping** if $|S_k \cap (\cup_{j<k} S_j)| \geq |S_i \cap (\cup_{j<k} S_j)|$ for every $k = 2, ..., m-1$ and $i = k+1, ..., m$.*

RIP requires that the intersection between any set along the sequence $\bar{\mathcal{S}}$ and the union of its predecessors is weakly contained in one of them. This

combinatorial property is familiar from the Bayesian-network literature - see below. Although it lacks an a-priori appealing economic interpretation, it happens to hold in a number of realistic situations (including Examples 3.1 and 3.2).

RIP* requires that the collection $\mathcal{S}$ can be ordered with a running intersection. It holds trivially for $m = 2$. To illustrate the definition for $m = 3$, let $\mathcal{S} = \{\{1,3,5\},\{2,4,5\},\{1,2,5\}\}$. If we order $\mathcal{S}$ as $(\{1,3,5\},\{2,4,5\},\{1,2,5\})$, RIP is violated because $\{1,2,5\} \cap (\{1,3,5\} \cup \{2,4,5\}) = \{1,2,5\}$ is not contained in any of the first two sets in the sequence. In contrast, the sequence $(\{1,3,5\},\{1,2,5\},\{2,4,5\})$ satisfies RIP because $\{2,4,5\} \cap (\{1,3,5\} \cup \{1,2,5\}) = \{2,5\} \subset \{1,2,5\}$. Therefore, $\{\{1,3,5\},\{1,2,5\},\{2,4,5\}\}$ satisfies RIP*. Conversely, the database $\{\{1,3\},\{1,2\},\{2,3\}\}$ violates RIP*; its members cannot be ordered in a way that satisfies RIP.

The maximal-overlap property requires that S_k has at least as many variables in common with previous observed datasets as any subsequent dataset. This property makes particular sense when, for instance, variables are realized according to some chronological order, and as new datasets arrive over time, they tend to cover recent variables. Example 3.2 fits this interpretation. The property can also reflect the analyst's own initiative: in his attempts to broaden the database, he deliberately seeks datasets that maximally overlap prior datasets, because he wishes to extrapolate as little as possible.

The following result, proved by Noga Alon, links the three properties.

Lemma 1 (Alon (2016)) *Suppose that the database $\mathcal{S}$ satisfies RIP*. Then, every maximally overlapping ordering of $\mathcal{S}$ satisfies RIP.*

This lemma provides a simple tool for checking whether $\mathcal{S}$ satisfies RIP*: order this collection according to any maximally overlapping sequence, and check whether the sequence satisfies RIP.

Perfect DAGs

Let us turn from properties of databases to a property of DAGs. A subset of nodes $C \subseteq N$ is a *clique* in R if $i\tilde{R}j$ for every distinct $i, j \in C$. A clique is *maximal* if it is not a strict subset of another clique. A clique C is *ancestral* if $R(i) \subset C$ for every $i \in C$.

Definition 5 (perfect DAGs) *A DAG R is perfect if whenever iRk and jRk, it is the case that $i\tilde{R}j$.*

Equivalently, R is perfect if $R(i)$ is a clique for every $i \in N$.[1] To illustrate the definition, $3 \leftarrow 1 \rightarrow 2 \rightarrow 4$ is perfect, whereas $3 \leftarrow 1 \rightarrow 2 \leftarrow 4$ is imperfect. When R is perfect, I refer to p_R as a *perfect-DAG* representation.

Remark 1 *Two perfect DAGs are equivalent if and only if they have the same set of cliques. In particular, we can set any one of these cliques to be ancestral w.l.o.g. (This is a direct implication of Proposition 1.)*

What is the meaning of perfection in light of the causal interpretation of DAGs? By definition, all the postulated causes of a variable in a perfect DAG are also presumed to have direct causal links among them. However, Remark 1 implies that the causal interpretation of perfect DAGs is spurious in the following sense: every causal link postulated by the DAG is reversed in some equivalent DAG.

Perfect DAGs are related to RIP*, via the following result, which is familiar in the Bayesian-networks literature.

Remark 2 (Cowell et al. (1999, p. 54)) *The set of maximal cliques in a perfect DAG satisfies RIP*.*

[1] Note that by the definition of $R(i)$, it is a clique if and only if $R(i) \cup \{i\}$ is a clique.

Which marginals get distorted by a DAG representation?
The DAG representation p_R generally distorts the objective distribution p: unless R is fully connected, there exists an objective distribution p for which $p_R \neq p$. However, certain marginal distributions are not distorted by p_R. The following proposition, which will be useful in the proof of the main result, characterizes these cases. The proof is relegated to Appendix I.

Proposition 2 *Let R be a DAG and let $C \subseteq N$. Then, $p_R(x_C) = p(x_C)$ for every p if and only if C is an ancestral clique in some DAG in the equivalence class of R.*

Thus, the marginal distribution over X_C induced by p_R never distorts the true marginal if C is an ancestral clique in R, or in some DAG that is equivalent to R. Note that if R is perfect, then by Remark 1, $p_R(x_C) = p(x_C)$ for every p and *every* clique C in R.

The intuition for Proposition 2 can be conveyed through the causal interpretation of DAGs. Suppose that C consists of a single node i. When i is ancestral, it represents a "primary cause". The belief distortions that arise from a misspecified DAG concern variables that are either independent of x_i or (possibly indirect) effects of x_i. These distortions are irrelevant for calculating the marginal of p_R over x_i. In contrast, suppose that i is not ancestral in any DAG in the equivalence class of R. Then, there must be two other variables x_j, x_k, deemed independent by R, which function as (possibly indirect) causes of x_i. This failure to account for the full dependencies among the causes of x_i can lead to distortion of the marginal distribution over x_i.

4.1 The Main Results

We can now state the first main result. If an unordered database $\mathcal{S}$ satisfies RIP*, then for any maximally overlapping ordering of $\mathcal{S}$, the IEP will generate an output that can be written as a perfect-DAG representation. Moreover,

the DAG is *essentially unique*, because its set of maximal cliques is $\mathcal{S}$ (and by Remark 1, all perfect DAGs with the same set of cliques are equivalent).

Theorem 1 *Suppose that $\mathcal{S}$ satisfies RIP*. Let R be any perfect DAG whose set of maximal cliques is $\mathcal{S}$. Then, $f(\bar{\mathcal{S}}, \alpha, p) = p_R$ for every maximally overlapping ordering $\bar{\mathcal{S}}$ of $\mathcal{S}$, collection of coefficients $\alpha = (\alpha^k)_{k=1,...,m}$ and objective distribution $p \in \Delta(X)$.*

Proof. I will show that for every p and every $k = 1, ..., m$, the belief $q^k \in \Delta(X_{B_k})$ has a perfect-DAG representation, where the DAG R^k is defined over B_k and its set of maximal cliques is $\{S_1, ..., S_k\}$. The proof is by induction on k. Let $k = 1$. By definition, $q^1 = p^{S_1}$, which is trivially a perfect-DAG representation (the DAG R^1 is fully connected over $B_1 = S_1$).

Now consider $k \in \{2, ..., m\}$. By assumption, $\mathcal{S} = \{S_1, ..., S_m\}$ satisfies RIP* and $\bar{\mathcal{S}} =(S_1, ..., S_m)$ is some maximally overlapping ordering of $\mathcal{S}$. By Lemma 1, $(S_1, ..., S_m)$ satisfies RIP. This immediately implies that $(S_1, ..., S_k)$ satisfies RIP, too. By assumption, $\mathcal{S}$ does not include sets that contain one another. Therefore, $S_k - B_{k-1}$ and $B_{k-1} - S_k$ are both non-empty. The auxiliary beliefs q_1^k and q_2^k over $B_k = B_{k-1} \cup S_k$ are given by (2).

Consider the expression for q_1^k. The inductive hypothesis is that q^{k-1} has a perfect-DAG representation, where the DAG R^{k-1} is defined over B_{k-1}, and its set of maximal cliques is $\{S_1, ..., S_{k-1}\}$. By RIP, $S_k \cap B_{k-1}$ is weakly contained in one of the sets $S_1, ..., S_{k-1}$. Extend R^{k-1} to a DAG R^k over B_k, by adding a link $i \to j$ for every $i \in S_k \cap B_{k-1}$ and $j \in S_k - B_{k-1}$, as well as adding directed links among all nodes in $S_k - B_{k-1}$ without destroying acyclicity. The DAG R^k is perfect and its set of maximal cliques is $\{S_1, ..., S_k\}$. Thus, q_1^k is a perfect-DAG representation, where the DAG is R^k.

It remains to show that q_2^k coincides with q_1^k, such that by (3), $q^k = q_1^k$ for any α^k. If $S_k \cap B_{k-1} = \varnothing$, this is self-evident. Now suppose $S_k \cap B_{k-1} \neq \varnothing$.

Note that q_1^k and q_2^k can be written as

$$\begin{aligned} q_1^k(x_{B_k}) &= p(x_{S_k})q^{k-1}(x_{B_{k-1}}) \cdot \frac{1}{p(x_{S_k \cap B_{k-1}})} \\ q_2^k(x_{B_k}) &= p(x_{S_k})q^{k-1}(x_{B_{k-1}}) \cdot \frac{1}{q^{k-1}(x_{S_k \cap B_{k-1}})} \end{aligned} \tag{8}$$

Since q^{k-1} is a perfect-DAG representation - where the DAG is R^{k-1}, and $S_k \cap B_{k-1}$ is a clique in R^{k-1} - Remark 1 implies that w.l.o.g it is an ancestral clique. Proposition 2 then implies that $q^{k-1}(x_{S_k \cap B_{k-1}}) = p(x_{S_k \cap B_{k-1}})$. Therefore, q_2^k coincides with q_1^k. ■

The following result is a converse to Theorem 1, which shows that RIP* is *necessary* for the IEP's output to have a DAG representation.

Theorem 2 *Suppose that $\mathcal{S}$ violates RIP*. Then, for every DAG R and every ordering $\bar{\mathcal{S}}$ of $\mathcal{S}$, there exists an objective distribution $p \in \Delta(X)$ such that $f(\bar{\mathcal{S}}, \alpha, p) \neq p_R$ for any collection of coefficients α.*

Proof. Suppose that $\mathcal{S}$ violates RIP*. Then, any ordering of $\mathcal{S}$ violates RIP. Note that this means $m \geq 3$. Let $k > 2$ be the earliest round for which $S_k \cap B_{k-1}$ is *not* weakly contained in any of the sets $S_1, ..., S_{k-1}$. By the proof of Theorem 1, $q^{k-1} \in \Delta(X_{B_{k-1}})$ has a perfect-DAG representation, where the perfect DAG R^{k-1}, defined over B_{k-1}, is characterized by the set of maximal cliques $\{S_1, ..., S_{k-1}\}$. It follows that $S_k \cap B_{k-1}$ is not a clique in R^{k-1}. By Proposition 2, there exist distributions p for which $q^{k-1}(x_{S_k \cap B_{k-1}}) \neq p(x_{S_k \cap B_{k-1}})$. Therefore, by (8), there exists p for which q_1^k and q_2^k do not coincide. I now construct a family of such distributions.

Since $S_k \cap B_{k-1}$ is not a clique in R^{k-1}, it must contain two nodes, denoted w.l.o.g 1 and 2, that are not linked under R^{k-1}. Let p be an arbitrary objective distribution for which x_i is independently distributed for every

$i \neq 1, 2$, whereas x_1 and x_2 are mutually correlated. Then,

$$q^{k-1}(x_{B_{k-1}}) = \prod_{i \in B_{k-1}} p(x_i)$$

It follows that q_1^k and q_2^k can be written as

$$\begin{aligned} q_1^k(x_{B_k}) &= p(x_1)p(x_2) \cdot \prod_{i \in B_k - \{1,2\}} p(x_i) \\ q_2^k(x_{B_k}) &= p(x_1)p(x_2 \mid x_1) \cdot \prod_{i \in B_k - \{1,2\}} p(x_i) \end{aligned}$$

such that

$$q^k(x_{B_k}) = \left(\prod_{i \in B_k - \{2\}} p(x_i) \right) \cdot \left[\alpha^k \cdot p(x_2) + (1 - \alpha^k) \cdot p(x_2 \mid x_1) \right]$$

Since all the variables $i \neq 1, 2$ are independently distributed under p and $\alpha^k \in (0, 1)$ for every k, the continuation of the IEP will eventually produce a final belief of the form

$$q^m(x) = \left(\prod_{i \neq 2} p(x_i) \right) \cdot [\beta \cdot p(x_2) + (1 - \beta) \cdot p(x_2 \mid x_1)]$$

where $\beta \in (0, 1)$ is some combination of $\alpha^k, ..., \alpha^m$. Since $p(x_2 \mid x_1) \neq p(x_2)$ for some x_1, x_2, q^{m-1} does not have a DAG representation. ■

The key to understanding Theorems 1 and 2 is whether the provisional distribution q^k (at any round k in the IEP) distorts the true marginal over $X_{S_k \cap B_{k-1}}$. When the ordered database $(S_1, ..., S_k)$ satisfies RIP, $S_k \cap B_{k-1}$ is weakly contained in some dataset S_i, $i \leq k - 1$. This means that the i^{th} dataset contains complete raw information about the marginal of p over $S_k \cap B_{k-1}$. The proof-by-induction of Theorem 1 employs the properties of perfect DAGs to ensure that this information is not lost or distorted by the

time we reach round k, and therefore the marginal of q^{k-1} over $X_{S_k \cap B_{k-1}}$ does not contradict the true marginal of p over this set, which is given by the dataset S_k.

In contrast, when the ordered database violates RIP, there will be a round k for which $S_k \cap B_{k-1}$ is not weakly contained in any dataset S_i, $i \leq k-1$. This means that the marginal of q^{k-1} over $X_{S_k \cap B_{k-1}}$ is not exclusively based on raw data, and thus inevitably involves extrapolation, potentially missing correlations among variables in $S_k \cap B_{k-1}$. As a result, the marginal of q^{k-1} over $X_{S_k \cap B_{k-1}}$ may contradict the true marginal of p over this set, which is given by the dataset S_k. In other words, the k^{th} dataset will disagree with the provisional output of round $k-1$ over the correlation structure of the variables in $S_k \cap B_{k-1}$. This disagreement will persist until the procedure's very end, such that its output will lack the coherent correlation structure that characterizes a DAG representation.

Example 3.2 demonstrated the role of the maximal-overlap property. The unordered database satisfied RIP*, and yet we saw that it could be ordered in a way that violates the maximal-overlap property. This ordering gave rise to an output that lacked a DAG representation. However, unlike RIP*, the maximal-overlap property is not necessary for the IEP's output to have a DAG representation. For instance, the ordered database $(\{1,2,3\},\{3,4\},\{2,3,5\})$ fails the maximal-overlap property but satisfies RIP, and therefore the IEP's output in this case has a DAG representation.

The availability bias illustrated in Example 3.1 is a general feature of the model, in the following sense. Consider a database that satisfies RIP*. Suppose that the database records the correlation between some variable x_i and many other variables, but does not record the correlation among the latter. Then, for any DAG representation of the IEP's output, x_i will appear above the other variables (with the exception of at most one of them) in the DAG's causal hierarchy.

The following corollaries of Theorems 1 and 2 examine the relation between DAG representations and the IEP from a different perspective. Rather than taking the database as primitive and checking whether it leads to a DAG representation, we can take the representation as given and ask whether it can be justified as the procedure's output for a suitably specified database.

Corollary 1 *Suppose that R is a perfect DAG. Let $\mathcal{S}$ be an unordered database consisting of the maximal cliques of R. Then, for every maximally overlapping ordering $\bar{\mathcal{S}}$ of $\mathcal{S}$ and any collection of coefficients α, $f(\bar{\mathcal{S}}, \alpha, p) = p_R$ for all objective distributions $p \in \Delta(X)$.*

Corollary 2 *Suppose that R is an imperfect DAG. Then, for every ordered database $\bar{\mathcal{S}}$, there exists an objective distribution $p \in \Delta(X)$ such that $f(\bar{\mathcal{S}}, \alpha, p) \neq p_R$ for any collection of coefficients α.*

Corollary 1 is an immediate implication of Remark 2 and Theorem 1. Corollary 2 is a consequence of Proposition 1, which implies that imperfect and perfect DAGs can never belong to the same equivalence class. By Theorem 1, if $\mathcal{S}$ satisfies RIP*, the procedure's output has a perfect-DAG representation. And by Theorem 2, if $\mathcal{S}$ violates RIP*, the procedure's output lacks a DAG representation altogether. Therefore, no database can generate an imperfect-DAG representation.

4.2 Which Causal Models can be Extrapolated?

Corollaries 1 and 2 have special significance when we consider the causal interpretation of DAGs. Recall that perfect DAGs do not postulate identifiable causal links (in the sense that if iRj, there exists an equivalent DAG R' such that $jR'i$). In contrast, every imperfect DAG R contains at least two identifiable causal links that belong to the graph's v-structure: these links remain unreversed in any DAG in the equivalence class of R. Corollaries 1 and 2

imply that only causal models that make unidentifiable causal assumptions can be extrapolated (via the IEP) from partial statistics. This is consistent with the familiar motto "correlation does not imply causation": the analyst's dataset contains purely observational data; the extrapolation method that the analyst employs does not create meaningful beliefs about causality out of thin air.

The two corollaries enable us to shed light on whether natural classes of subjective causal models can be justified as the outcome of procedural extrapolation from partial statistics. The following are a few examples.

Fixed-lag causal models

Consider a decision maker whose subjective belief q over X distorts the objective distribution p by treating each variable as a stochastic function of its L immediate predecessors:

$$q(x_1, ..., x_n) = p(x_1, ..., x_L) \prod_{i=L+1}^{n} p(x_i \mid x_{i-L}, ..., x_{i-1}) \tag{9}$$

For instance, when $L = 1$, $q(x_1, ..., x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1})$. This specification fits environments in which variables have a natural chronological ordering, and where subjective perception of the variables' history dependence is "coarse" or "truncated" (as in Piccione and Rubinstein (2003)).

For any $L < n$, the ordered database $\bar{\mathcal{S}} = (\{k, ..., k+L\})_{k=1,...,n-L}$ satisfies RIP and generates (9) as the IEP's output. Thus, a fixed-lag subjective causal model can be justified (via the IEP) as the outcome of extrapolation from datasets with memory of fixed length. Example 2.2 illustrates this result for $n = 4$, $L = 1$.

Correlation neglect

Let $n = 3$, and consider the DAG $R : \theta \rightarrow z \leftarrow a$, where θ represents a state of Nature that affects the value of an object, a represents the bidding behavior of a player in some trading mechanism, and z represents the mech-

anism's outcome. Then, $p_R(\theta, a, z) = p(\theta)p(a)p(z \mid \theta, a)$. The DAG R can be interpreted as a subjective model that postulates the independence between the player's bidding behavior and the object's value. If p violates this independence property - because in reality the player conditions his behavior on some unobserved signal of θ - we have $p_R \neq p$, and thus p_R exhibits "correlation neglect" (as in Eyster and Rabin (2005) or Jehiel and Koessler (2008)).

Is it possible for an analyst who has access to partial statistics about θ, a, z to extrapolate the correlation structure given by R? Because R is imperfect, Corollary 2 implies that no database will generate an output that factorizes every p according to R. The intuition for this impossibility is simple. In order to estimate the term $p(z \mid \theta, a)$, the analyst must have access to joint observations of all three variables. But this would also enable him to grasp whatever correlation exists between θ and a, whereas p_R treats them as mutually independent.

Thus, the particular form of correlation neglect captured by p_R cannot be justified by our notion of extrapolation from partial statistics. Of course, this failure is specific to the present example; other forms of correlation neglect are consistent with the IEP (see Example 3.3).

It is easy to see from this expression why it cannot be obtained by the IEP.

Observability structures

Let $N = \{1, 2, 3, 4\}$. Each variable x_i represents the action of a different player in an extensive game. Each of the following two DAGs represents a game form in which each player moves once and the order of moves is fixed. A link $i \to j$ means that j always observes i's move. DAGs thus represent what Eyster and Rabin (2014) refer to (in the context of a social-learning

model) as "observability structures".

The question is whether an outside observer could extrapolate a belief in these observability structures from some partial statistical data. The left-hand DAG is imperfect ($R(4) = \{2, 3\}$, and yet 2 and 3 are not linked). Therefore, no database can generate an output that factorizes every p according to this graph. In contrast, the right-hand DAG is perfect, and it will be extrapolated from any ordering of $\mathcal{S} = \{\{1, 2, 3\}, \{1, 4\}\}$.

What is the broader significance of these examples? When a subjective causal model is consistent with the IEP, we can tell an "as if" story about the model's origin: Agents do not necessary have an explicit prior theory regarding the correlation structure of their environment; instead, their belief is based on procedural extrapolation from partial statistics (possibly performed by another agent - our eponymous "data monkey"); and the belief only *appears* as if the agent were trying to impose an explicit causal model on the true distribution. This argument is in the spirit of axiomatic decision theory: DAG representations are a tractable formula that captures a range of systematic belief distortions and possesses a natural (causal) interpretation; it is instructive to know whether they have a plausible "origin story".

5 Relation to the MaxEnt Problem

The IEP is a "behaviorally motivated" procedure for extending partial statistics into a fully-specified probability distribution over X. An alternative, more "normatively motivated" extrapolation method is based on the criterion of *maximal entropy*. Consider the unordered database $\mathcal{S}$, which enables the analyst to learn the marginals $(p^S)_{S \in \mathcal{S}}$. The analyst's problem is to find

a probability distribution $q \in \Delta(X)$ that maximizes entropy subject to the constraint that for every $S \in \mathcal{S}$, the marginal of q over X_S is p^S. A more general version of this problem was originally stated by Jaynes (1957) and has been studied in the Machine Learning literature, where it is known as the MaxEnt problem.

The maximal-entropy criterion generalizes the "principle of insufficient reason" (recall that unconstrained entropy maximization yields the uniform distribution). The idea behind it is that the analyst wishes to be "maximally agnostic" about the aspects of the distribution he has not learned, while being entirely consistent with the aspects he has learned. For instance, suppose that the analyst only manages to learn the marginal distributions over all individual variables - i.e., $\mathcal{S} = \{\{1\}, ..., \{n\}\}$. Then, the maximal-entropy extension of these marginals is $p(x_1) \cdots p(x_n)$.[2]

The following is an existing result, reformulated to suit our present purposes.

Proposition 3 (Hajek et al. (1992)) *Suppose $\mathcal{S}$ satisfies RIP*. Then, the maximal-extension entropy of the marginals $(p^S)_{S \in \mathcal{S}}$ is p_R, where R is any perfect DAG whose set of maximal cliques is $\mathcal{S}$.*

This result establishes a connection between the IEP and the MaxEnt problem: the former can be viewed as an algorithm for implementing the latter whenever $\mathcal{S}$ satisfies RIP*. Recall that RIP* holds automatically for $m = 2$. Thus, the basic extrapolation method employed in the first round of the IEP always implements the maximal-entropy principle. Indeed, I motivated this method by the idea that the analyst wishes to use all available evidence without making any active assumption about things for which he has no evidence. This is essentially the maximal-entropy principle; and the proposition

[2] For a more recent study of information-theoretic methods of extrapolation from limited data, see Miller and Liu (2002).

formalizes the connection for $m = 2$. When $m > 2$, the unqualified equivalence between the IEP and the maximal-entropy principle breaks down; it holds only when $\mathcal{S}$ satisfies RIP*.

6 Related Literature

This paper draws on two literatures in statistics: graphical models and statistical inference with missing data. Both links were explained earlier in the paper. This section focuses on the paper's connection to literature *within* economics. Recent years have seen intensified interest in equilibrium models with "boundedly rational expectations", in which agents' subjective beliefs systematically distort the correlation structure of the steady-state distribution. The distortions take various forms: "coarse" beliefs that neglect correlations (Piccione and Rubinstein (2003), Jehiel (2005), Koessler and Jehiel (2008), Mullainathan et al. (2008), Eyster and Piccione (2013)); failure to realize how action-consequence correlation would change if off-equilibrium actions were played (Esponda (2008)); belief in spurious correlations due to naive extrapolation from small samples (Osborne and Rubinstein (1998)); or attributing fluctuations in a certain variable to the wrong cause (Eyster and Rabin (2005), Ettinger and Jehiel (2010)).

A common justification for models of this kind is that agents receive *partial feedback* as they try to learn statistical regularities in their environment, and therefore their subjective beliefs distort the correlation structure of the steady-state distribution. In some cases (e.g. Osborne and Rubinstein (1998), Jehiel (2005)), this idea is formally built into the definition of equilibrium. In other cases the "limited feedback" justification is informal and outside the model. The exercise in this paper can be viewed as a novel formalization of this justification and an exploration of its limits. Note that the maximal-entropy aspect of the IEP is akin to the "Occam's Razor" aspect of the notion of analogy-based expectations in Jehiel (2005), which assumes

that the agent imposes an analogy partition on the set of possible contingencies and requires the agent's belief to be measurable with respect to that partition.

The link between partial data and non-rational expectations was studied from an explicitly *dynamic* perspective in two recent papers. Esponda and Pouzo (2016a) propose a general game-theoretic framework, in which each player has a "subjective model", which is a set of stochastic mappings from his action a to a primitive set of payoff-relevant consequences y he observes during his learning process. The feedback is limited because in the true model, other "latent" variables may affect the action-consequence mapping. Esponda and Pouzo define an equilibrium concept in which each player best-replies to a subjective distribution (of y conditional on a) that is closest in his subjective model to the true equilibrium distribution. Distance is measured by a weighted version of Kullback-Leibler divergence. Esponda and Pouzo justify this equilibrium concept as the steady-state of a Bayesian learning process.

Schwartzstein (2014) studies a dynamic model in which an analyst tries to predict a variable y as a function of two variables x and z. At every period, he observes the realizations of y and x. In contrast, he pays attention to the realization of z only if his belief at the beginning of the period is that z is sufficiently predictive of y. When the analyst chooses not to observe z, he imputes a constant value. Schwartzstein examines the long-run belief that emerges from this learning process, and in particular the analyst's failure to perceive correlations among the three variables.

Finally, the broad notion of decision makers as "imperfect statisticians" is far from original and has many precedents in the literature - too many to cite here. What is arguably new about this paper is the emphasis on the processing of explicitly statistical data, and the distortions that arise when users of processed data take it at face value and do not internalize the processing methods.

7 Conclusion

Economists have two prevailing images of how agents form beliefs. At one extreme, we have the conventional view of the economic agent as an infallible (Bayesian) statistician. At the other extreme, we have the image promoted by behavioral economics, which emphasizes the role of intuition in belief formation. These two pictures have implications for the kind of *data* that could be involved in the agent's reasoning. By definition, the conventional picture is compatible with *any* data - from espresso-machine gossip to large statistical tables. In contrast, intuitive judgments are more naturally associated with the former kind. It makes sense to think about intuitive probability judgments in response to a rumor, an anecdote or a terse statement about probabilities; it makes weaker sense to think in such terms when we describe reasoning about data that are arrive in the form of large spreadsheets. If we wish to depart from the view of the economic agent as a supreme statistician, a different metaphor is needed for such situations.

In this paper I offered the image of a "data monkey", to describe an economic agent who faces partial statistical data and subjects it to some method of extrapolation in order to produce a digestible output. Such an agent may have an imperfect understanding of his own methods, or he may fail to impart such understanding to his audience. This belief-formation error can be viewed as an organizational, data-saturated analogue to the phenomenon of "base rate neglect" that has been observed in the context of intuitive probability judgments (Tversky and Kahneman (1974)).[3] Of course, the "data monkey" image is not restricted to the extrapolation problem; developing other "behaviorally motivated" models of how people reason about large statistical datasets is an exciting direction for future research.

[3]The idea that agents over-interpret posterior beliefs because they neglect the prior theories that shaped them is related to other psychological phenomena, such as confirmatory bias (see Rabin and Schrag (1999)).

References

[1] Alon, N. (2016), "Problems and Results in Extermal Combinatorics - III," Journal of Combinatorics 7, 233–256.

[2] Burton, A. and D. Altman (2004), "Missing Covariate Data within Cancer Prognostic Studies: A Review of Current Reporting and Proposed Guidelines," British Journal of Cancer 91, 4-8.

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

[4] Esponda, I. (2008), "Behavioral Equilibrium in Economies with Adverse Selection," *The American Economic Review,* 98, 1269-1291.

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

[6] Esponda, I. and D. Pouzo (2016b), "Conditional Retrospective Voting in Large Elections," *American Economic Journal: Microeconomics*, forthcoming.

[7] Ettinger, D. and P. Jehiel (2010), "A Theory of Deception," *American Economic Journal: Microeconomics* 2, 1-20.

[8] Eyster, E. and M. Piccione (2013), "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions," *Econometrica,* 81, 1483-1506.

[9] Eyster, E. and M. Rabin (2005), "Cursed Equilibrium," *Econometrica,* 73, 1623-1672.

[10] Eyster, E. and M. Rabin (2014), "Extensive imitation is irrational and harmful," *Quarterly Journal of Economics*, 129, 1861-1898.

[11] Frydenberg, M. (1990), "The Chain Graph Markov Property," *Scandinavian Journal of Statistics* 17, 333-353.

[12] Gelman, A. and J. Hill (2006), *Data Analysis using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

[13] Hajek, P., T. Havranek and R. Jirousek (1992), *Uncertain Information Processing in Expert Systems*, CRC Press.

[14] Jaynes, E. T. (1957), "Information Theory and Statistical Mechanics," *Physical Review,* 106, 620-630.

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

[16] Jehiel, P. (2015), "Investment Strategy and Selection Bias: An equilibrium Perspective on Overconfidence," mimeo.

[17] Jehiel, P. and F. Koessler (2008), "Revisiting Games of Incomplete Information with Analogy-Based Expectations," *Games and Economic Behavior,* 62, 533-557.

[18] Kaplan, G. and S. Schulhofer-Wohl (2012), "Interstate Migration has Fallen Less than You Think: Consequences of Hot Deck Imputation in the Current Population Survey," *Demography* 49, 1061-1074.

[19] Little, R. and D. Rubin (2002), *Statistical Analysis with Missing Data,* Wiley, New Jersey.

[20] Mackinnon, A. (2010), "The Use and Reporting of Multiple Imputation in Medical Research – A Review," Journal of Internal Medicine 268, 586-593.

[21] Manly, C. and R. Wells (2015), "Reporting the Use of Multiple Imputation for Missing Data in Higher Education Research," Research in Higher Education 56, 397-409.

[22] Meyer, B., R. Goerge and N. Mittag (2014), "Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation," University of Chicago, mimeo.

[23] Miller, D. and W. Liu (2002), "On the Recovery of Joint Distributions from Limited Information," *Journal of Econometrics* 107, 259-274.

[24] Mullainathan, S. J. Schwartzstein and A. Shleifer (2008), "Coarse Thinking and Persuasion," *Quarterly Journal of Economics* 123, 577-619.

[25] Osborne, M. and A. Rubinstein (1998), "Games with Procedurally Rational Players," *American Economic Review,* 88, 834-849.

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

[27] 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.

[28] Rabin, M. and J. Schrag (1999), "First Impressions Matter: A Model of Confirmatory Bias," Quarterly Journal of Economics 114, 37-82.

[29] Schwartzstein, J. (2014), "Selective Attention and Learning," *Journal of European Economic Association,* 12, 1423-1452.

[30] Silver, N. (2012), *The Signal and the Noise: Why so Many Predictions Fail - but Some Don't*, Penguin.

[31] Sloman, S. (2009), Causal Models: How People Think about the World and its Alternatives, Oxford University Press.

[32] Spiegler, R. (2015), "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics* 131, 1243-1290.

[33] Tversky, A. and D. Kahneman (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, 1124–1131.

[34] Verma, T. and J. Pearl (1991), "Equivalence and Synthesis of Causal Models," *Uncertainty in Artificial Intelligence,* 6, 255-268.

[35] Wood, A., I. White and S. Thompson (2004), "Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals," Clinical Trials 1, 368-376.

Appendix I: Proof of Proposition 2

For convenience, label the variables in C by $1, ..., m$. Let us write down the explicit expression for $p_R(x_C)$:

$$\begin{aligned} p_R(x_C) &= \sum_{x'_{m+1},...,x'_n} p_R(x_1, ..., x_m, x'_{m+1}, ..., x'_n) \qquad (10) \\ &= \sum_{x'_{m+1},...,x'_n} \prod_{i \in C} p(x_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) \prod_{i \notin C} p(x'_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) \end{aligned}$$

(i). Assume C is an ancestral clique in R. Then,

$$\prod_{i \in C} p(x_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) = p(x_1) \prod_{i=2}^{m} p(x_i \mid x_1, ..., x_{m-1}) = p(x_C)$$

Expression (10) can thus be written as

$$p(x_C) \sum_{x'_{m+1},...,x'_n} \left(\prod_{i=m+1}^{n} p(x'_i \mid x_{R(i) \cap C}, x'_{R(i)-C}) \right) = p(x_C)$$

Therefore, $p_{R'}(x_C) = p(x_C)$ for every R' that is equivalent to R.
(ii). Let us distinguish between two cases.

Case 1: C is not a clique in R (and therefore also not a clique in any DAG that is equivalent to R). Then, C contains two variables, labeled w.l.o.g 1 and 2, such that $1 \not R 2$ and $2 \not R 1$. Consider an objective distribution p, for which every x_i, $i > 2$, is distributed independently, whereas x_1 and x_2 are mutually correlated. Then, expression (10) is simplified into

$$\prod_{i=1}^{m} p(x_i) \sum_{x'_{m+1},...,x'_n} \prod_{i=m+1}^{n} p(x'_i) = \prod_{i=1}^{m} p(x_i)$$

whereas the objective distribution can be written as

$$p(x_C) = p(x_1)p(x_2 \mid x_1) \prod_{i=3}^{m} p(x_i)$$

The two expressions are different because x_2 and x_1 are not independent.

Case 2: C is a clique which is not ancestral in any DAG in the equivalence class of R. Suppose that for every node $j \in C$, j has no "unmarried parents" - i.e., if there exist nodes k, k' such that kRj and $k'Rj$, then kRk' or $k'Rk$. In addition, if there is a directed path from some $i \notin C$ to j, then i has no unmarried parents either. Transform R into another DAG R' by inverting every link along every such path. The DAGs R and R' share the same skeleton and v-structure. By Proposition 1, they are equivalent. By construction, C is an ancestral clique in DAG that is equivalent to R, a contradiction.

It follows that R has the following structure. First, there exist three distinct nodes, denoted w.l.o.g $1, 2, 3$, such that $1, 2 \notin C$, $1R3$, $2R3$, $1 \not R 2$ and $2 \not R 1$. Second, there is a directed path from 3 to some node $s \in C$, $s \geq 3$. For convenience, denote the path by $(3, 4, ..., s)$ - i.e., the immediate predecessor of any $j > 3$ along the path is $j - 1$. It is possible that $s = 3$, in which case the path is degenerate. W.l.o.g, we can assume that $i \notin C$ for every $i \neq s$ along this path (otherwise, we can take s to be the lowest-numbered node that belongs to C along the path).

Consider any p which is consistent with a DAG R^* that has the following structure: first, $1R^*2R^*3$ and $1R^*3$; second, for every $j \in \{4, ..., s\}$, $R^*(j) = \{j-1\}$; and $R^*(j) = \varnothing$ for every $j \notin \{2, ..., s\}$. (Note that the latter property means that every x_j, $j \notin \{2, ..., s\}$ is independently distributed. Then,

$$p(x) = p_{R^*}(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdot \prod_{i=4}^{s} p(x_i \mid x_{i-1}) \cdot \prod_{j>s} p(x_j)$$

In contrast,

$$p_R(x) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2) \cdot \prod_{i=4}^{s} p(x_i \mid x_{i-1}) \cdot \prod_{j>s} p(x_j)$$

By definition, every $i = 4, ..., s-1$ does not belong to C. Denote

$$q(x') = p(x'_1)p(x'_3 \mid x'_1, x'_2) \left(\prod_{i=4}^{s-1} p(x'_i \mid x'_{i-1}) \right) p(x_s \mid x'_{s-1})$$

Therefore,

$$\begin{aligned} p(x_C) &= \prod_{j \in C - \{s\}} p(x_j) \sum_{x'} p(x'_2 \mid x'_1) q(x') \\ p_R(x_C) &= \prod_{j \in C - \{s\}} p(x_j) \sum_{x'_1, ..., x'_{s-1}} p(x'_2) q(x') \end{aligned}$$

It is easy to see from these expressions that we can find a distribution p which is consistent with R^* such that $p_R(x_C) \neq p(x_C)$ for some x.

Appendix II: Selective Datasets

An assumption that runs throughout the paper is that the process that generates missing data is independent of variables' realizations. This is what enables the analyst to learn the true marginal of p over X_S from the dataset that covers the set of variables S. However, this independence property is naturally violated in many contexts. For example, data about a politician's quality is more likely to arrive when he is elected for office. Likewise, data about the value of an investment is more likely to arrive when the investment is taken. Because the decision to elect a politician or make an investment is typically based on information that may be correlated with the variable in question (the politician's quality, the investment's value), we cannot assume that the process that generates data is independent of the process that generates the relevant variables' realizations.

A number of recent works (e.g., Esponda (2008), Jehiel (2015), Esponda and Pouzo (2016b)) have analyzed models in which agents naively extrapolate their equilibrium beliefs from endogenously selective samples. In this Appendix, I use a basic version of Jehiel's (2015) model to illustrate how the current formalism can be adapted to the case of selective datasets, thus providing a new perspective into this interesting class of models.

In Jehiel's model, there are three variables: a represents an investor's decision whether to invest in a project; θ is the actual value of the project; and t is the investor's private information regarding the project's quality prior to taking an action. The variable a takes two values, 0 (not investing) and 1 (investing). The objective distribution p is a long-run joint distribution over these three variables. A realization of this distribution corresponds to an episode in which an individual investor faced an investment opportunity and had some private information regarding its value. Given the variables' interpretation, it is natural to assume that every distribution in the relevant domain satisfies the conditional independence property $a \perp \theta \mid t$.

An analyst wishes to understand empirical regularities in this environ-

ment. His database consists of two datasets. The first dataset is an infinitely large collection of independent joint observations of θ and t. This enables him to learn $p(\theta, t)$. The second dataset consists of infinitely many joint observations of θ and a. However, each of these observations satisfies $a = 1$ - that is, the dataset only records the value of an investment when it is taken; there are no observations of an investment's counterfactual value when it is not taken. The interpretation is that each observation in this dataset describes the outcome of a particular investment decision, without recording the private information that lay behind it. This dataset is clearly selective; and it can be described by a distribution $\hat{p}$ over (a, θ) that is defined as follows: $\hat{p}(a = 1) = 1$, and $\hat{p}(\theta \mid a = 1) = p(\theta \mid a = 1)$.

Suppose that the analyst applies the IEP to this database. Because $m = 2$, any ordering of the database is maximally overlapping. The procedure's final output is

$$q^2(a, \theta, t) = \alpha^2 \cdot p(\theta, t)\hat{p}(a \mid \theta) + (1 - \alpha^2) \cdot \hat{p}(a, \theta)p(t \mid \theta)$$

Because there are no observations of $a = 0$, $q^2(a = 0, \theta, t) = 0$ for every θ, t. For $a = 1$, the expression for q^2 simplifies into

$$\begin{aligned} q^2(a &= 1, \theta, t) = \alpha^2 \cdot p(\theta, t) + (1 - \alpha^2) \cdot p(\theta \mid a = 1)p(t \mid \theta) \\ &= p(t \mid \theta) \left[\alpha^2 \cdot p(\theta) + (1 - \alpha^2) \cdot p(\theta \mid a = 1)\right] \end{aligned}$$

Suppose that the analyst now makes the output of his research available to a potential new investor who receives a particular private signal t. The investor knows that not investing leads to a sure payoff of 0, independently of θ or t. He relies on q^2 for estimating the expected payoff from investing conditional on his signal. In other words, he will invest if and only if

$$\sum_{\theta} q^2(\theta \mid a = 1, t)\theta > 0$$

In the $\alpha^2 \to 1$ limit, the conditional distribution $q^2(\theta \mid a = 1, t)$ converges to $p(\theta \mid t)$. This is the quantity a rational investor should calculate, because p satisfies $\theta \perp a \mid t$. In this case, the new investor will choose to invest if and only if it is rational to do so.

In contrast, in the $\alpha^2 \to 0$ limit, $q^2(\theta \mid a = 1, t)$ converges to

$$\frac{p(\theta \mid a = 1)p(t \mid \theta)}{\sum_{\theta'} p(\theta' \mid a = 1)p(t \mid \theta')}$$

Given our assumptions on p, the term $p(\theta \mid a = 1)p(t \mid \theta)$ can be rewritten as

$$p(t)p(\theta \mid t) \sum_{t'} p(t' \mid \theta)p(a = 1 \mid t')$$

Therefore, the new investor will choose to invest if and only if

$$\frac{\sum_\theta p(t)p(\theta \mid t) \left(\sum_{t'} p(t' \mid \theta)p(a = 1 \mid t')\right) \theta}{\sum_\theta p(t)p(\theta \mid t) \left(\sum_{t'} p(t' \mid \theta)p(a = 1 \mid t')\right)} > 0 \tag{11}$$

In Jehiel (2015), this inequality is precisely the criterion that defines investors' subjectively optimal behavior. The distribution p is in equilibrium when an investor whose private signal is t chooses $a = 1$ if and only if the inequality holds. This notion of equilibrium is meant to capture the idea that when investors try to evaluate the consequences of an investment decision, they naively extrapolate from a large sample of prior investments, without taking into account that such a sample exhibits positive selection. The same criterion emerges from our procedure of extrapolating from partial statistics, when the dataset that records joint observations of a and θ gets arbitrarily large weight.

Surprisingly, the investment criterion given by (11) would also emerge if we assumed that the objective distribution is induced by a population of investors whose subjective belief over (θ, t, a) is given by a perfect-DAG representation, where the DAG is $R : t \leftarrow \theta \to a$. Corollary 1 thus implies that

a database that consists of *non-selective* datasets covering (θ, t) and (θ, a) would have led to the same prediction. That is, the selectiveness assumption is inessential for the model's prediction in the $\alpha^2 \to 0$ limit.

To derive this result, note that if $p(a \mid t) > 0$, then the investor should find a to be subjectively optimal given t - i.e., a maximizes

$$\sum_{\theta} p_R(\theta \mid t, a) \cdot \theta a$$

where θa is the investor's payoff from his decision. Note that

$$\begin{aligned} p_R(t, a) &= \sum_{\theta} p(\theta) p(t \mid \theta) p(a \mid \theta) \\ &= \sum_{\theta} p(\theta) p(t \mid \theta) \sum_{t'} p(t' \mid \theta) p(a \mid t') \end{aligned}$$

Therefore, if $p(a \mid t) > 0$, then a maximizes

$$\frac{\sum_{\theta} p(\theta) p(t \mid \theta) \left(\sum_{t'} p(t' \mid \theta) p(a \mid t')\right) \theta a}{\sum_{\theta} p(\theta) p(t \mid \theta) \left(\sum_{t'} p(t' \mid \theta) p(a \mid t')\right)}$$

It follows that the DM's subjective evaluation of $a = 0$ is zero, and his evaluation of $a = 1$ coincides with the L.H.S of (11).

This coincidence is due to the strong structure of the selective dataset. In Jehiel's investment problem (as in Esponda's (2008) leading bilateral trade example), there are two actions: one action generates a fixed payoff and no observations, whereas another action generates observations and uncertain payoffs. As a result, although the IEP's output q^2 does not take the form of a DAG representation, the *behavior* it induces in the $\alpha^2 \to 0$ limit is as if the investor's subjective belief had a perfect-DAG representation.