Indicator Frameworks

November 10, 2017

Abstract

We develop a diagrammatic tool for constructing correlations between random variables, called an abstract indicator framework. Abstract indicator frameworks are modeled off operational (key performance) indicator frameworks as they are used in city planning and project governance, and give a rigorous, statistically-motivated process for constructing operational indicator frameworks.

1 Introduction

Performance indicators are measurements that enable operators or planners to understand how well a system achieves a desired outcome. In most traditional performance management systems, individual key performance indicators gauge the current level of a particular measure and determine whether it lies in an acceptable range or an unacceptable range.

However, most performance management systems do not provide frameworks to understand the relationship between individual indicators, especially how well they collectively describe the ability of the system to achieve the desired outcomes. Describing these relationships allow us to evaluate the relative value of new indicators in this ensemble. With new sensor and human-generated data from smart cities and the industrial Internet of Things (IoT), evaluating the relative value of indicators allows practitioners to optimize the investment on new performance data. This evaluation will be beneficial both to the developers and the consumers of the technology behind this new data, estimating the projected value that can be then contextualized with development risks. [Clarify. Why?]

We take as our starting point a diagrams of simple correlations, as below:

This diagram correlates the measurement variables of an air pollution monitoring system with other, partially-observable variables like traffic composition, mechanical turbulence, and the presence of sunlight. It presents an intuitive and apparently useful description of a system at large. We would like to clarify the meaning of this diagram, and of others like it, by giving its interconnections a precise mathematical meaning. Clarifying the meaning of the diagram will not only make it more useful; it will allow us to connect this local, correlation-based picture of a system with other local pictures, as well as with more sophisticated scientific models of the world. [Practical promise: an easy, plug and play approach to adding models.]

In this paper, we develop *abstract indicator frameworks*, a diagrammatic tool for constructing causally-linked sets of random variables and their correlations. Abstract indicator frameworks are modeled off operational (key performance) indicator frameworks, especially as they are used in city planning and project governance. Such operational indicator frameworks have three main uses: (1) to communicate quantitative information and strategic priorities to a wide audience, (2) to enable policy reactions to data, especially in the optimization of processes, and (3) to restrict attention to a set of 'relevant' indicators—thus discarding the information from many other, 'non-relevant' indicators.

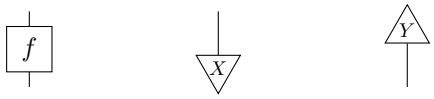
In city planning, there are several strategy-setting frameworks for constructing operational indicator frameworks, from balanced scorecards [**?**] to SMART [**?**] to more specialized urban planning frameworks; in such frameworks, the indicators are often designed by mayors, chief strategy officers, and sizable expert committees in tandem with new projects, new policies, and new processes. Even assuming that the participants adhere to a framework, the process of choosing indicators is often ad hoc, the results do not account for statistical relationships between the indicators, and the generated data is hard to translate across localities.

We propose an alternative. Instead of constructing operational indicator frameworks expensively and internally, meaning indicator-by-indicator, we can specify them abstractly and externally, by means of their causal and statistical relationships to other, already-extant sets of indicators. Our approach is especially suited to situations where heterogeneous data is distributed across many projects and many localities.

For example, cities are often interested in understanding the second-order impacts of specific projects, e.g. the impacts on health, crime, and jobs of a smart shuttle system. Assuming the existence of a local indicator framework for the shuttle system, and the existence of a top-level indicator framework representing broad priorities such as health, crime, jobs, and so on, then we can construct a mediating indicator set whose indicators satisfy certain statistical and causal relationships generated by the given indicator frameworks; these mediating indicators represent the second-order impacts of the local project to the city's other priorities.

This motivates the definition of abstract indicator frameworks, which we define as the objects of a certain "category of diagrams of random variables", Ind. We will apply *category theory*, originally developed to relate and analyze topological spaces, as an efficient language for relating and analyzing the causal and statistical aspects of indicator frameworks. To illustrate the mathematics, we will develop a running example of how a mayor can implement the approach.

Let $\mathcal{X}$ stand for an indicator set. Abstract indicator frameworks have a notion of *process* that transforms one indicator set into another, a notion of *state* that represents the process of picking a specific indicator in $\mathcal{X}$, and a notion of *effect* that represents the process of "measuring" or computing the correlation with respect to a specific indicator in $\mathcal{X}$. Processes, states, and effects are represented, respectively:



These are needed to capture the operations of composing processes in sequence, called *composition*, and combining them in parallel, called *tensoring*. The composition $g \circ f$ (first f, then g) and tensor $f \otimes g$ are represented as:

We call any formalism with a notion of composition and tensoring a *process theory*. The semantics of process theories and their diagrams are governed by the theory of monoidal categories, which is surveyed in [**?**]. The goal of the paper is to specify an appropriate symmetric monoidal category, Rand, representing the appropriate operations on random variables, after which we can define a causal model as a strong monoidal functor from a causal theory into Rand. These causal models—essentially, diagrams in Rand—will be the promised abstract indicator frameworks. In Section 2, we will consider a preliminary version of Rand along with some of the possible alternatives. In Section 3, we will give the statistical justification for our choice of Rand, review the notion of a causal model from [**?**], and then give the full definition of abstract indicator frameworks.

2 Background

As mentioned above, there are a variety of approaches to choosing indicator frameworks as part of the process of strategic priorities. Of the many specialized approaches to choosing indicator frameworks in various fields, Niemejer and de Groot [**?**] have suggested a similar methodology for choosing environmental indicator sets based on explicit causal networks of environmental forces and societal response; while their methodology is still largely qualitative rather

than formal or statistical, their paper handily illustrates how (diagrams of) causal models can facilitate the selection of relevant indicator sets. In statistics, Horvath [**?**] also takes a compositional approach to correlation by focusing on weighted correlation networks, which represent random variables by nodes in a graph and edges between variables by a soft threshold on their correlation. These correlation networks have proved useful for analyzing high-dimensional data sets, especially gene expression data.

In Bayesian statistics and machine learning, a variety of generic graphical approaches, called graphical models, have been developed to model the conditional (in)dependence of multivariate random variables; the joint distribution over all the random variables in a graphical model is the product of their conditional distributions. Among the most familiar examples of graphical models are (directed) Bayesian networks and (undirected) Markov networks. The upshot is that complex questions about joint distributions of many interrelated variables can be answered in terms of the topology of the graph. We mention these graphical models, and especially Bayesian networks (and by extension stochastic matrices), since they ground many existing approaches to integrating causality with probability, including that of Stoch and FinStoch.

[Other things: structural equation modeling and construct validity, ref. Petter 2007 or Churchill 1979, where constructs are structural and measurement models used in the study of systems, e.g. as in structural equation modeling. Also, stress-condition-response models, which are mentioned in Niemejer as well.]

[TO ADD: Pearl's discussion of causality and causal models using diagrams, methods of doing observational studies by eliminating interventions = surgeries on equations via diagrams. Simpson's paradox (see Fong's example).]

[EXAMPLE, part 1?]

Even within the constraints of a process theory, there are still a number of diagrammatic approaches to probability. In this section, we will go over three examples: the traditional Hilbert space interpretation of random variables, the original category of probabilistic mappings suggested by Lawvere [**?**], and the diagrammatic approach of Coecke and Spekkens [**?**] to Lawvere's work. We also briefly discuss graphical models such as those surveyed in [**?**], which are the most obvious applications of [**?**] and [**?**], e.g. see [**?**].

The traditional approach, which we call Rand, uses the fact that real-valued random variables over some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ form a Hilbert space H where the inner product $\langle X, Y \rangle$ is just the covariance $\mathbb{E}(XY)$. Assuming that we restrict ourselves to standard variables with zero mean and unit variance, the covariance equals the correlation, and we can represent both by the inner product in Rand. This inner product can be represented by a process diagram, namely as the composition of a state and an effect in Rand:

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y) = \langle X, Y \rangle =$$

As we will see in the next section, Rand is actually already very close to what we want; the problem is that the obvious categorical interpretation of Rand does not give a natural way of analyzing the data of "intermediate" correlations.

Rand is closely related to older work on the categorical foundations of probability initiated by Lawvere in [**?**] and developed in Giry [**?**]:

Definition 2.1. The category Stoch of stochastic processes is defined by the following data:

1. objects are measurable spaces (A, Σ_A) of sets A with a σ-algebra Σ_A
2. morphisms $P : (A, \Sigma_A) \to (B, \Sigma_B)$ are stochastic kernels, i.e. functions $P : A \times \Sigma_B \to [0, 1]$ that assign to (a, σ_B) the probability of σ_B given a, denoted $P(\sigma_B|a)$
3. composition $Q \circ P : A \times \Sigma_C \to [0, 1]$ of $P : (A, \Sigma_A) \to (B, \Sigma_B)$ and $Q : (B, \Sigma_B) \to (C, \Sigma_C)$ is defined by

$$(Q \circ P)(\sigma_C|a) = \int_{b \in B} Q(\sigma_C|b) dP_a,$$

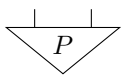
   i.e. marginalization over B

As suggested by the notation, morphisms in Stoch represent probability measures on an event/outcome space (A, Σ_A). If we restrict to the subcategory FinStoch whose objects are *finite* measurable spaces—since the outcomes are finite, one can imagine these spaces as sets of natural numbers $\{1, ..., n\}$—then we can think of stochastic kernels are stochastic matrices, i.e. matrices whose column entries sum to 1. Taking $1 = (\{*\}, \Sigma_*)$ as the monoidal unit, we can see that a probability distribution in FinStoch is just a vector $P : 1 \to (A, \Sigma_A)$ whose entries are the probabilities of all the possible atomic outcomes in A. The usual tensor product described in Stoch is the functor $\otimes$: Stoch $\times$ Stoch $\to$ Stoch that assigns to two probability distributions $P : 1 \to (A, \Sigma_A)$, $Q : 1 \to (B, \Sigma_B)$ their product measure, i.e. the map $PQ : 1 \to (A \times B, \Sigma_A \otimes \Sigma_B)$ s.t. $PQ(*, (a, b)) = P(*, a)Q(*, b)$.

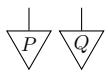
Using the language of symmetric monoidal categories,, Coecke and Spekkens [**?**] give a graphical calculus for FinStoch and use it to elaborate Bayesian reasoning (in particular, a diagrammatic representation of Bayes' rule). As above, objects of FinStoch are natural numbers, morphisms from m to n are $n \times m$ stochastic matrices, composition is matrix product, and the monoidal product is the matrix tensor product. States are probability distributions over the set $1, ..., n$:

$: 1 \to A = (p_1, ..., p_n)$ such that $\sum_{j=1}^{n} p_j = 1$

A *joint state* is a state over the composite object $1, ..., mn$, of the form



A joint state is "uncorrelated"—one should be careful, since this is correlation between probability distributions, not between random variables per se—when it can be decomposed into a tensor product, and perfectly correlated when it can be represented as a delta function. Uncorrelated and perfectly correlated joint states are depicted, respectively:



$: 1 \to A \otimes B = (p_1 q_1, ..., p_n q_m)$

$: I \to A \otimes A = (\delta_{i,i'} \in \{1, ..., n\}.$

A perfectly anti-correlated joint state in FinStoch is just a cup with a NOT-gate attached to one end. More generally, any correlation can be obtained by attaching a suitable box to one of the ends of the cup.

There are two major problems with this graphical formalism, and with Stoch and FinStoch in general. First and foremost, there is not a very convenient way of talking about *random variables*. Technically, a real-valued random variable is given by the diagram below, of a measurable function $X : (\Omega, \Sigma_\Omega) \to (\mathbb{R}, \Sigma_\mathbb{R})$ which takes possible outcomes in Ω to their numerical representations in $\mathbb{R}$ (technically, X is not a function but a stochastic matrix whose columns represent point probabilities), a probability measure $P : 1 \to (\Omega, \Sigma_\Omega)$ on the outcomes in Ω, and finally the pushforward $X(P)$ of P along X, which represents the probability distribution of the random variable X.

$$\begin{array}{ccc} 1 & \xrightarrow{P} & (\Omega, \Sigma_\Omega) \\ & \searrow^{X(P)} & \downarrow X \\ & & (\mathbb{R}, \Sigma_\mathbb{R}) \end{array}$$

Besides being difficult to work with, the point of view of this paper is that probability measures are not random variables nor are they sufficient replacements; a probability measure is something probability theorists invented in order to talk about random variables. While useful in the context of Bayesian networks, which can be articulated primarily in terms of stochastic processes, probability measures are less visible in cases driven by data and by correlational arguments.

The second problem is that there is not a good interpretation of *effect*, i.e. of "measuring" or computing something with respect to a specific state. An effect in FinStoch

is defined to be a morphism $X^\dagger : (A, \Sigma_A) \to 1$, i.e. a function $X^\dagger : A \times \Sigma_* \to [0, 1]$. The problem is that 1 is terminal in Stoch: there is only one possible morphism from any object to 1 due to the constraint on morphisms of being a probability measure. In particular, for any (A, Σ_A) and for all $a \in A$, the unique map $X^\dagger : (A, \Sigma_A) \to 1$ is given by $X^\dagger(*|a) = 1$ and $X^\dagger(\varnothing|a) = 0$. In other words, any 'measurement' of a state (i.e. a probability distribution) in FinStoch and Stoch simply kills the state.

3 Indicator Frameworks

In this section, we define the category Ind of abstract indicator frameworks and give an example, with diagrams, of a concrete indicator framework.

Before giving the definition of the category Ind of abstract indicator frameworks, we will go through some of the statistical justification. Suppose that we have a correlation between random variables X and Y and another one between Y and Z. What can we say about the correlation between X and Z? One obvious guess would be

$$\mathrm{Cor}(X, Z) = \mathrm{Cor}(X, Y)\mathrm{Cor}(Y, Z). \tag{1}$$

Of course we know that Equation **??** is, in general, false.[1] But it is, under certain conditions, still the best guess.

[EXAMPLE, part 2.]

The following result is a standard exercise in statistics.

Lemma 3.1. If $a = \mathrm{Cor}(X, Y)$ and $b = \mathrm{Cor}(Y, Z)$, then

$$\mathrm{Cor}(X, Z) \geq ab - \sqrt{1 - a^2}\sqrt{1 - b^2} \tag{2}$$

$$\mathrm{Cor}(X, Z)\rangle \leq ab + \sqrt{1 - a^2}\sqrt{1 - b^2} \tag{3}$$

Proof. WLOG, assume that A, B, C are standard variables with zero mean and unit variance, since the correlation is invariant under changes to mean and variance. We can write $X = aY + E_{Y,X}$ and $Z = bY + E_{Y,Z}$ where, by construction, $E_{Y,X}, E_{Y,Z}$ are random variables uncorrelated with Y.

Then $\langle X, Z\rangle = \mathrm{Cor}(X, Z) = \langle aY + E_{Y,X}, bY + E_{Y,Z}\rangle = ab + \langle E_{Y,X}, E_{Y,Z}\rangle$, and we can use the Cauchy-Schwarz inequality to bound $\langle E_{Y,X}, E_{Y,Z}\rangle$ from above and from below, giving the lemma. □

The lemma tells us that there is a range of possible values, centered around $\mathrm{Cor}(X, Y)\mathrm{Cor}(Y, Z)$, for the composite correlation; unfortunately, in practice that range can so large as to be useless. In such a situation, we may ask what is the obstruction, given $\mathrm{Cor}(X, Y)$ and $\mathrm{Cor}(Y, Z)$, to knowing the canonical or 'true' correlation of their composite, and whether we can reduce or get around that obstruction. Reading the proof of the lemma, we know the obstruction is just the correlation $\langle E_{Y,X}, E_{Y,Z}\rangle$; that is, if $\langle E_{Y,X}, E_{Y,Z}\rangle$ were 0, our guess would be valid.

One may also derive the above result from the definition of the partial correlation of X and Z, fixing Y. Recall that the partial correlation $\rho_{XZ \cdot Y}$ is defined

[1] Correlations are rarely composed in practice because (1) the computation is usually false and (2) because we can usually compute the composite correlation directly from the data. It is only when we lack the data (which is quite often in studies of complex systems) that we use a causal model to infer the correlation. Unfortunately, causal models are often invoked in the process of imposing a learned model such as a Kalman filter or a dynamical Bayesian network, which will often conflate the statistical and causal contributions.

as the correlation between the residuals of X and of Z, fixing Y. In terms of their component correlations,

$$\rho_{XZ\cdot Y} = \frac{\mathrm{Cor}(X,Z) - \mathrm{Cor}(X,Y)\mathrm{Cor}(Y,Z)}{\sqrt{1-\mathrm{Cor}(X,Y)^2}\sqrt{1-\mathrm{Cor}(Y,Z)^2}}.$$

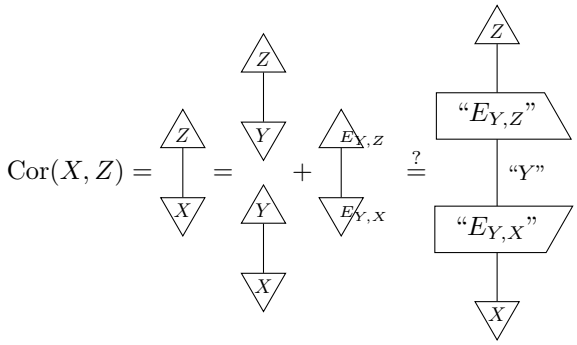
Thus

$$\langle E_{Y,X}, E_{Y,Z}\rangle = \rho_{XZ\cdot Y}\sqrt{1-\mathrm{Cor}(X,Y)^2}\sqrt{1-\mathrm{Cor}(Y,Z)^2}.$$

In other words, Equation **??** is correct just when the partial correlation $\rho_{XZ\cdot Y} = 0$, when $\mathrm{Cor}(X,Y) = 1$ or -1 (i.e. X and Y are linear functions of each other), or when $\mathrm{Cor}(Y,Z) = 1$ or -1. This allows us to produce another guess:

$$\text{“Cor}(X,Z)\text{”} = Y \text{ s.t. } \rho_{XZ\cdot Y} = 0 \tag{4}$$

Explicitly, $E_{Y,X}$ measures the nonlinear component of the relation between X and Y. But one may also think of it as a measure of the 'noise' or 'error' between X and Y, at least as it concerns the correlation. The idea of Equation **??** is that, if we are lucky in choosing Y, then the noise factors $E_{Y,X}$ and $E_{Y,Z}$ will "cancel out" to produce the true correlation $\mathrm{Cor}(X,Z) = \mathrm{Cor}(X,Y)\mathrm{Cor}(Y,Z)$. Heuristically, we can represent this process as below:



That is, the correlation between X and Z can be computed by applying a transformation "$E_{Y,X}$", representing some sort of structured noise factor, then applying a transformation "$E_{Y,Z}$" that cancels out the noise introduced by "$E_{Y,X}$".

We can formalize this intuition. Recall that real-valued, square-integrable random variables over a given probability space form a Hilbert space $L^2(\Omega, \Sigma, \mathbb{P})$ whose inner product is just the covariance.

Definition 3.2. The category of random variables, Rand, is defined by the following data:

1. objects are finite-dimensional Hilbert spaces

$$\mathcal{X} = L^2(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$$

of square-integrable random variables (under the equivalence relation $X_1 \sim X_2$ if $\mathbb{P}_{\mathcal{X}}(X_1 = X_2) = 1$) with inner product $\langle X, Y\rangle = E(XY)$, defined over probability spaces $(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$, with an associated basis $\mathcal{B}_{\mathcal{X}} =$

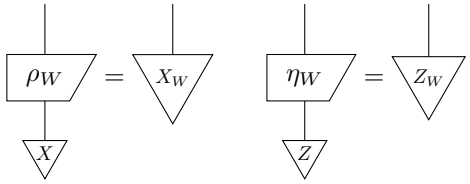
$\{X_1, X_2, ..., X_n\} \cup \mathbf{1}$, where $\mathbf{1}$ is the random variable with constant value 1.

2. morphisms $F : \mathcal{X} \to \mathcal{Y}$ are bounded linear operators
3. identity $1 : \mathcal{X} \to \mathcal{X}$ is the identity matrix.
4. the composition is the usual composition of bounded linear operators
5. the tensor product of $\mathcal{X}$ and $\mathcal{Y}$ is the pushout over their joint support in $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

Lemma 3.3. Rand is a symmetric monoidal category with the above tensor.

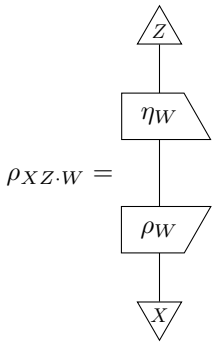
Example 3.4. Suppose that the transportation department buys a new bus and designates an indicator, X, that counts the number of riders on the bus per day. Elsewhere, the education department tracks an indicator, Z, that counts the number of students per day who are absent from class across the whole city. Assume that X and Z live in indicator sets $\mathcal{X}$ and $\mathcal{Z}$.

First, we "integrate" the data by computing $\mathcal{X} \otimes \mathcal{Z}$, so that the correlation is computed only on days for which X, Z both have data. We compute the correlation: then the correlation may be very small, or conversely it may be absurdly high, especially if there is some confounding variable correlated with both X and Z, e.g. an economic boom. Suppose the federal government tracks a separate variable, W, on aggregate economic performance per quarter. The first step is to get rid of the influence of W, i.e. compute the residuals X_W, Z_W of X, Z resulting from their linear regression with W. Assuming that X and Z live in indicator sets $\mathcal{X}$ and $\mathcal{Z}$ respectively, and that W lives in both $\mathcal{X}$ and $\mathcal{Z}$, we can represent computing the residuals as applying transformations ρ_W, η_W on $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$, respectively:



For variables X, Y, let us denote the linear regression of X with respect to Y by $X|Y$. ρ_W and η_W are indeed morphisms in Rand, i.e. bounded linear operators, since the residual of a linear regression can be written in the form $X_W = X - X|W = X - (a\mathbf{1} + bW)$, where a, b are constants. (Technically, everything above happens in the "larger" Hilbert space $\mathcal{X} \otimes \mathcal{Z}$; ρ_W, η_W are

projections from this larger space.) Then by definition, we have



As shorthand, we will sometimes refer to the space of random variables $\mathcal{X}$ as a set of variables or indicators; in such cases, we always mean the basis set of random variables, $B_{\mathcal{X}}$.

It will help to think of random variables as representing column vectors or "dimensions" of data in a table of such data, where row vectors in that table represent particular data points. The correlation between two column vectors is just their sample correlation. This has several benefits: it makes the inner product (correlation) and tensor product (entity resolution) very concrete, it is what a data analyst actually looks at, and it highlights the restrictions and challenges imposed by the presence and absence of data. In fact, we can define a category Data explicitly in such terms:

Definition 3.5. The category of $\mathbb{R}$-valued data tables, Data, is defined by the following data:

1. objects $\mathcal{X} = (\mathcal{X}, \Omega_{\mathcal{X}}, \mathbb{I}_{\mathcal{X}})$ of Data are $m \times n$ tables of $\mathbb{R}$-valued data vectors whose rows are assigned an index key given by $\mathbb{I}_{\mathcal{X}} : \Omega_{\mathcal{X}} \to \mathbb{R}$ and whose columns, $B_{\mathcal{X}} = \{X_1, ..., X_n\}$, represent indicators
2. morphisms $f : \mathcal{X} \to \mathcal{Y}$ are linear transformations of the column values of $\mathcal{X}$ by vector addition (of other columns in $\mathcal{X}$) and scalar multiplication
3. the identity $1 : \mathcal{X} \to \mathcal{X}$ is the identity matrix
4. the composition is just the matrix product
5. the tensor product of $\mathcal{X} \otimes \mathcal{Y}$ is the integrated table of their data values over a table of linkages, $S \subset \Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

[EXAMPLE, part 3. Show some actual tables with actual data.]

Suppose we are working in a 3-dimensional Hilbert space $\mathcal{X}$ with a basis of random variables $B_{\mathcal{X}} = \{X, Y, Z\}$. In this basis, the random variable $E_{Y,X}$ is just the vector $X - \langle X, Y \rangle Y$ (and similarly with $E_{Z,X}$), but the problem with this space... is that there is no problem! In Rand, having a basis in X, Y, Z corresponds to the observation, in Data, that we already have the tabular data

we need to compute $\langle X, Z \rangle$ directly. But in many situations of interest in a complex, open system, e.g. in computing the second-order impacts of local and/or technical projects, such broad-based data is difficult to obtain.

So we lack data. But to take just one example, in a database setting there are ways to reason about "missing data", e.g. database nulls, especially when that data is the subject of a data migration or integration, as described in [**?**]. In particular, one can impose a set of algebraic equations that each null value must satisfy, where the equations are given by a diagram of database schema mappings. More generally, almost every diagram in a category articulates a set of constraints on the objects of that category.

Example 3.6. Recall our earlier example, where X stands for the number of riders on a particular bus in a city, and Z stands for the number of absent students across a city. Suppose that we have already controlled X and Z for economic performance (i.e. W) along with any number of other confounding variables, and that we have found (or suspect) a small but significant correlation between X and Z. We are now interested in understanding *how* X correlates with Z.

There may be a variety of possible explanations for why this correlation exists: maybe dropping the price of a ticket (thus promoting more bus ridership) allows more students to go to school, or perhaps additional bus ridership decreases traffic, which gives harried parents more time to track their truant children. Without choosing any one explanation, we can represent the statistical properties of a set of mediating, "explanatory" variables $\mathcal{Y}$ by a 'sum' of the possible explanations:

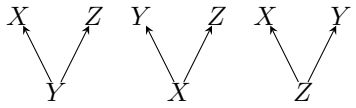
$$\mathrm{Cor}(X, Z) = \sum_{Y \in \mathcal{Y}} \left(\begin{array}{c} Z \\ \eta_Y \\ \mathcal{Y} \\ \rho_Y \\ X \end{array} \right) \tag{5}$$

The equation above succinctly represents a set of constraints that we can impose on the intermediate framework $\mathcal{Y}$, and motivates the following definition.

[REWORK? Introduce adjustment problem / Simpson's paradox first, and then define a confounding framework as "the set of confounding variables you should measure and marginalize over?"

Definition 3.7. A *mediating framework* between two spaces of random variables $\mathcal{X}, \mathcal{Z}$ is a space of random variables $\mathcal{Y}$ such that Equation **??** is satisfied for all variables $X \in \mathcal{X}, Z \in \mathcal{Z}$.

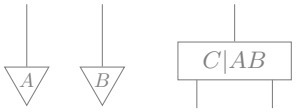
Rand, Data, and the notion of mediating framekwork supply the basic statistical foundation for a theory of indicator frameworks. We would now like to incorporate a causal foundation. There are several reasons for doing so. First: many statistical arguments, e.g. partial correlation, actually rely on an implicit choice of causal model—see discussions related to confounding and mediating variables by Pearl [**?**], among others. For example, given three random variables X, Y, Z; the partial correlations $\rho_{XZ \cdot Y}$, $\rho_{XY \cdot Z}$, or $\rho_{YZ \cdot X}$ are all equally valid; which one we take as 'true' depends on which of the following causal structures we believe is true:



Second, many operational indicator frameworks are constructed based on experts' causal models of the indicators, e.g. as in [**?**] or as in the air pollution diagram shown at the very beginning of this paper. Any story of indicator frameworks would be incomplete without mentioning causation. And third, the pattern developed here for causation will be useful later, when we want to incorporate not only causal models but arbitrary scientific models (such as those in Bayesian networks) into our indicator frameworks.

We recall the definition of causal theory from Fong [**?**], as a certain symmetric monoidal category induced from a directed acyclic graph (i.e. the causal structure), such as any of the three graphs above. Without going into the details, the idea is that given such a causal structure, we can specify a symmetric monoidal category whose objects are collections of the letters $\{X, Y, Z\}$, and whose morphisms are generated by the counit (representing 'deletion') and comultiplication (representing 'copying'), depicted respectively by

and by a set of causal mechanisms generated from the causal structure, $[A] : \varnothing \to A$, $[B] : \varnothing \to B$, and $[C|AB] : AB \to C$, depicted as



Given a causal theory, i.e. a symmetric monoidal category, we can define a model of that causal theory $\mathcal{C}$ in Rand as a strong monoidal functor $F : \mathcal{C} \to$ Rand. To specify such a functor, it suffices to define its behavior on every atomic variable and every generating map in $\mathcal{C}$, i.e. the counit, comultiplication, and causal mechanisms of $\mathcal{C}$, since the values of the functor on the rest of $\mathcal{C}$ is specified up to isomorphism by the definition of a strong monoidal functor. For example, if A is an atomic causal variable of the causal theory $\mathcal{C}$, then F sends A to a one-dimensional Hilbert space, e.g. one with basis set $\{X\}$. On tensor products

of atomic causal variables, $F(A \otimes B)$ gives the tensor product $F(A) \otimes F(B)$, i.e. the space of random variables with basis set $\{F(A), F(B)\}$ and probabilities inherited from the product measure. On morphisms, $F([A]) : F(*) \to F(A)$ is just the single random variable in $F(A)$, and a causal mechanism $[C|AB]$ becomes a linear operator $F([C|AB]) : F(A \otimes B) \to F(C)$. Note that diagrams in the causal theory do not, typically, give rise to diagrams of the same shape in Rand. For example, a confounding variable Y with causal structure $X \leftarrow Y \to Z$ will typically generate the diagram corresponding to $\rho_{XZ \cdot Y}$. In general, each strong monoidal functor $\mathcal{C} \to$ Rand converts a causal theory into a certain "package" of related indicator sets, where the operational indicator framework is represented by the terminal leaves of the causal theory. Picking the appropriate functor constitutes an optimization problem.

We can now state the definition of Ind.

Definition 3.8. The category Ind of abstract indicator frameworks is defined by the following data:

1. an object I of Ind is a strong symmetric monoidal functor $\mathcal{C} \to$ Rand from a causal theory $\mathcal{C}$ to the category of random variables.

2. a morphism η between abstract indicator frameworks is a natural transformation of strong symmetric monoidal functors

In other words, an object of Ind represents a diagram in Rand, whose nodes are indicator sets and whose edges have been organized to represent the various relationships between indicator sets. One may directly compare Ind with the category of stochastic causal models in [**?**], which are generalizations of Bayesian networks.

4 Conclusion

In this paper, we sought to give a rigorous mathematical alternative to the traditional, indicator-by-indicator process of constructing indicator frameworks, especially in city planning and project governance. We proposed that indicator frameworks could be defined (and optimized) by means of their relationships to other indicator frameworks. These relationships were probabilistic as well as causal. Therefore, we sought to develop a semantics for the problem of constructing indicator frameworks that could accommodate both probabilistic and causal modes of reasoning.

We examined several options for the semantics of probability, including Stoch [**?**], FinStoch, and their corresponding diagrammatic representations [**?**]. After reflecting on the practical necessities of data analysis, we decided to base our construction on a category more directly in terms of random variables and correlations, and defined the symmetric monoidal category Rand of (spaces of) random variables. We then introduced the idea of a causal model from [**?**], and used this to motivate the definition of the category Ind of abstract indicator frameworks as models of a causal theory in Rand.

We then used Ind as the setting for an optimization problem: how to construct a mediating indicator framework that best explains the relationship between a given set of indicators, such as those of a specialized project in a city, and another set of indicators, such as headline indicators of broad interest to the public. Such a mediating framework can be used to answer the question, "what are the secondary impacts of my project?"

[Given a heterogeneous set of models, what, in some sense, is the "total prediction" of that set of models? Given the total prediction, can we then define a consistent, heterogeneous "total prediction error"? Draw inspiration from Friston 2009 (Trends in Cog Sci).

The ultimate goal of this line of research, following the proposal for S&CC, is develop a consistent notion of what it means to have a "status quo". So not only a total prediction, but also a sense of how the status quo prediction changes.]

This paper is the subject of ongoing research; future versions will address applications to more complicated, real-world examples in city administration, as well as the integration of other mathematical models beyond causal ones into the framework. Other additional future work include the possibility of studying the constraints on data-supported applications introduced by constraints on their underlying indicators, an analysis of how to translate an "ontological model" of the event space into constraints on the data, as well as closer examinations of the phenomenon of tensoring or 'gluing' data sets and the possible obstacles such gluings may introduce to producing a consistent global picture of the complex system.

We would like to thank Bob Coecke, Bilin Guvenc, Levent Guvenc, Derek Loftis, and Ed Griffor for helpful conversations in the writing of this paper.

Disclaimer

Certain commercial products may be identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.