A graphical approach to measuring smart cities

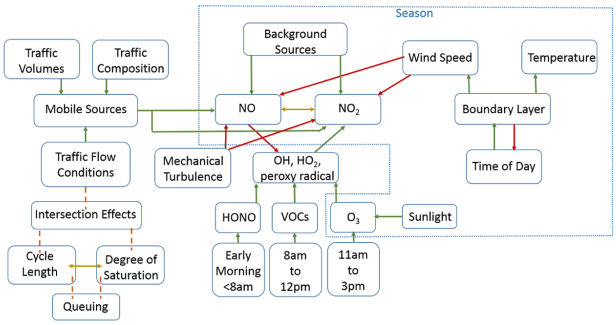
Joshua Z. Tan, Abhishek Dubey, and Sokwoo Rhee

Suggestions from Ed: make sure to motivate the math before even mentioning "hybrid indicator frameworks" as a solution. There are things called "process indicators" in systems engineering, which are a bit different from KPIs. We need a completeness theorem stating that the diagrammatic calculus for indicator frameworks actually models all possible indicator frameworks and correlations.

Publication strategy: use this paper to focus on formalizing just indicator frameworks*without mentioning the hybrid part*, with a diagrammatic calculus, and publish in the CPS week workshop. Create an application-focused paper, again without the 'hybrid' part, that connects this workshop paper with the CPS Framework, and publish that in Ed's special issue. Finally, develop in late spring a full hybrid indicator framework paper.

1. INTRODUCTION

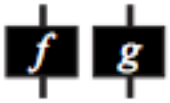
Suppose we take as our starting point a simple diagram of correlations between the system variables of a complex system, as below.



The diagram correlates the measurement variables of an air pollution monitoring system with other, partially-observable variables like traffic composition, mechanical turbulence, 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 the diagram a precise mathematical semantics. Clarifying the meaning of the diagram will not only make it more useful; it will allow us to connect this simple, non-causal picture of the system with more sophisticated *causal* models.

In this essay, we develop a diagrammatic calculus for correlations between random variables. Broadly, we expect our diagrammatic calculus to have a notion of *process* (i.e. correlations), *state* (i.e. random variables), and *effect* (i.e. random variables). These are needed to capture the operations

of composing correlations in sequence, called *composition*, and combining them in parallel, called *tensoring*. Composition and tensoring are to be represented, respectively, as below:



The semantics of such diagrams are governed by the theory of monoidal categories, which is surveyed in [5].

There are three reasons for our focus on correlations. First, our choice is motivated by practical considerations in the use of (key performance) indicator frameworks, especially as they are used in city planning and project governance. Such indicator frameworks present 'relevant' system variables to a human decision-maker, but the determination of 'relevant' becomes difficult as the system increases in size and complexity. Since indicator frameworks are so structurally simple, one of the few structures we can define on them is correlation. Second, correlations are practical; in many settings involving multivariate data, correlations are the analyst's workhorse. In observational studies of complex systems, often the best we can do is derive correlations, since observational data "from nature" rarely support conditional relations of the sort sought in laboratory experiments. Finally, part of the aim of this paper is to clearly distinguish

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

causal from non-causal models, and correlations seem like a core feature of any non-causal model of multivariate data.

NB: Correlations are rarely composed in practice because we can usually compute the composite correlation directly from the data. If we lack the data, then we can almost always do some estimation of parameters and then use the learned model (e.g. a Kalman filter or dynamical Bayes network) to estimate the correlation. However, the use of the learned model typically incurs a causal framework.

2. BACKGROUND

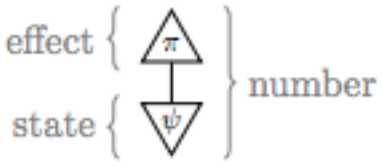
Even within the constraints of a process theory, there are still a number of diagrammatic approaches to correlation and covariance. In this section, we will go over three examples: the traditional Hilbert space interpretation of random variables, the Bayesian approach of Coecke and Spekkens [1], and graphical models such as those surveyed in [3].

The most straightforward approach, which we call Rand, uses the fact that real-valued, standard random variables over some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ form an (infinite-dimensional) Hilbert space $\mathcal{H}$ where the inner product $\langle X, Y\rangle$ is just the covariance/correlation $\mathbb{E}(XY)$. The objects of the category Rand are subspaces H of $\mathcal{H}$, morphisms are unitary matrices of the appropriate dimensions, composition $g \circ f : H \to H'$ is given by matrix multiplication, and the tensor is the matrix tensor product with unit $\mathbb{R}$. A state of H is a normalized vector representing a specific random variable (qua probability distribution) in the subspace H.

With this setup, the correlation between two standard random variables is just the inner product in $(\Omega, \mathcal{F}, \mathbb{P})$, i.e.

$$\langle X, Y\rangle := \mathbb{E}(XY)/(\sigma_X \sigma_Y) = \mathbb{E}(XY).$$

It is depicted:

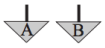


[1] gives a similar graphical calculus for Bayesian inference. Restricting to standard probability, objects of the category Bayes are natural numbers, morphisms from m to n are $n \times m$ stochastic matrices, composition is matrix product, and the monoidal product is the matrix tensor product. Normalized states are probability distributions over the set $1, ..., n$:


$$: \mathrm{I} \to A = (p_1, p_2, \ldots, p_n) \text{ such that } \sum_{j=1}^{n} p_j = 1.$$

A joint state is a normalized state over the composite object $1, ..., mn$. A joint state is uncorrelated when it can be decomposed into a tensor product, and perfectly correlated when it can be represented as a delta function. In Bayes, uncorrelated and perfectly correlated joint states are depicted, respectively:

The work of [1] is itself related to older work on the categorical foundations of probability initiated by Lawvere in


$$: \mathrm{I} \to A \otimes B = (p_i q_j | i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}),$$


$$: I \to A \otimes A = \mathbf{e}^T = (\delta_{i,i'} | i, i' \in \{1, \ldots, n\}).$$

[4]. In that paper, the category of probabilistic mappings has objects Borel spaces (X, Σ_X) and morphisms Markov kernels; a Markov kernel $T : (X, \Sigma_X) \to (Y, \Sigma_Y)$ represents a function $T : \sigma_Y \times X \to [0, 1]$ that assigns to each $x \in X$ and each measurable set $B \in B_Y$ the probability of B given x, denoted $T(B|x)$; in this way, each random variable can be represented as a morphism between Borel spaces (technically, the probability measure of T), and the composition of random variables corresponds to marginalization.

Finally, in Bayesian statistics and machine learning, a variety of more 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. The upshot is that complex questions about joint distributions of many interrelated variables can be answered in terms of the topology of the graph.

[Need more explanation and figures.]

Bayes and similar formalisms give a semantics for probabilistic reasoning, often with a specific causal model. Causal models are useful for interpreting the results of an experiment. It is not useful, generally speaking, when the data do not provide evidence for causation, such as in purely observational studies of complex systems. We want to carefully distinguish the causal aspects of our models from their non-causal aspects, and the first step is to develop a coherent framework for *non-causal* reasoning.

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

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

Of course we know that (1) is, in general, false. But it is, under certain conditions, still the best guess. The following result is an exercise in [**?**].

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

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

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

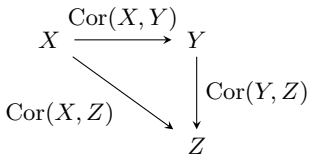
PROOF. WLOG, assume that A, B, C are standard variables with zero mean and unit variance, since the correlation is invariant under changes to mean and variance. We can write $X = aY + O_Y^X$ and $Z = bY + O_Y^Z$ where, by assumption, O_Y^X, O_Y^Z are random variables uncorrelated with Y.

Then $\langle X, Z\rangle = \mathrm{Cor}(X, Z) = \langle aY + O_Y^X, bY + O_Y^Z\rangle = ab + \langle O_Y^X, O_Y^Z\rangle$.

We can use the Cauchy-Schwarz inequality to bound $\langle O_Y^X, O_Y^Z\rangle$ 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. 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. Reading the proof of the lemma, we know this is just the correlation $\langle O_Y^X, O_Y^Z\rangle$. Intuitively, since $\langle O_Y^X, Y\rangle = \langle O_Y^Z, Y\rangle = 0$, they give 'all' the extra information relevant to $\mathrm{Cor}(X, Z)$.

Now, another way of telling the same story above is to say that we are interested in how correlations between two variables, X and Z, may 'factor through' an intermediate variable, Y, in such a way that the diagram below commutes.



In this paper, we have a more specific way of thinking about this diagram: if X and Z are fixed, how can we design a random variable Y, or a set of random variables $\{Y_i\}$, such that X and Z are maximally correlated?

Now, suppose we had two variables, X and Z, and we wanted to 'tensor' them in a way that captures the same extra information relevant to $\mathrm{Cor}(X, Z)$, given any random variable Y or set of random variables, possibly correlated with each other.

I.e. so that $\mathrm{Cor}(X \otimes_Y Z, Y) = \langle O_Y^X, O_Y^Z\rangle$.

Perfectly-correlated variables duplicate knowledge; the more correlated two variables are, the more information is lost when they are combined.

which is perfectly correlated with another variable, it does not add much to our knowledge of the world. perfectly correlated variables which then correlate do not offer much more information about

Now suppose we have two correlations, $\mathrm{Cor}(X, Y)$ and $\mathrm{Cor}(U, V)$, "parallel" to each other in the sense.

The natural product between random variables should be a tensor product.

Suppose we had some initial notion of how "significant" certain random variables was, relative to other variables.

Of course, we didn't have to derive this result to guess

Now suppose that we have multiple correlations $\mathrm{Cor}(X, Y_i)$ and $\mathrm{Cor}(Y_i, Z)$. Can we say anything more about the correlation between X and Z?

for which we have correlations between X and Z.

$$\langle X, Z\rangle = \frac{1}{n}\sum_{i\in B}^{n}\langle X, Y_i\rangle\langle Y_i, Z\rangle.$$

This assumes, however, that all the correlations are equally significant in supplying the correlation. Moreover, it fails to square with our intuition

Definition 1 *The category* Ind *of indicator frameworks takes objects as finite-dimensional Hilbert spaces, morphisms from V^m to V^n as an $n \times m$ matrix of correlation coefficients, and composition as the correlation matrix generated by the following equation:*

Lemma 3.2 Ind *is a symmetric monoidal category with tensor $\otimes$ and monoidal unit $\mathbb{C}$.*

Proof. First, we can define states: a state. □

Given this monoidal structure, a state is a [0,1]-valued vector in that Hilbert space, which is interpreted as a "trustworthiness" or "prior" vector on each of the indicators. An effect is interpreted as a "priority" vector on a (typically smaller) set of indicators.

In essence, the idea is to treat indicators as something closer to *features* of a given optimization problem, rather than as random variables.

4. APPLICATIONS

In later work, we will use Ind to give a precise definition of "hybrid indicator frameworks" that relate non-causal information with causal models such as those found in [1] or [2]. For now, even without mentioning causal models, we can set up interesting questions from statistics and analyze them using Ind.

TO DO: build an example optimization.

5. CONCLUSION

6. MISC. NOTES

Correlation theory wishlist

1. Cost heuristic on (sets of) indicators, such that imposing a state and an effect creates an optimization problem. Recall: *For a given policy or project, what is the 'right' set of indicators to measure it? Subtext: "holistic"?*

2. A super-indicator is a "conceptual" indicator, whose correlation is not set down in terms of its data, and which is supposed to represent the priorities of the Mayor, and it should be definable, at least partly, in terms of the particular projects and policies related to it.

   "Super" indicators, e.g. quality-of-life, that represent high-level goals disconnected from reality or data, as joint indicators with special properties. A super-indicator (or any joint indicator) like quality-of-life should be at least partially defined in terms of the projects and policies which can affect it.

3. Statistical significance of the correlation. E.g. how can we check that we have enough data points, or represent the fact that we don't have enough?

4. Correlation of residuals?

5. "Knowledge about the indicators" can be formalized through functors from other categories to Ind.

Process theory wishlist:

1. Measurements of state

2. Correlations between indicators

3. General processes in the underlying complex system

Arbitrary system variables in the complex system do not have an assigned meaning of measurement (in other ways, we're not collecting data on them)... but that doesn't mean they don't matter. Similarly [but in what way?], there is no underlying reality or total state of the complex system; the state may not be measurable, but that doesn't mean it doesn't matter. We should be able to put them into larger equations that govern the whole.

7. REFERENCES

[1] B. Coecke and R. Spekkens. Picturing classical and quantum bayesian inference. *Synthese*, 186(3):651–696, June 2012.

[2] B. Fong. Causal theories: A categorical perspective on bayesian networks. Preprint, April 2013.

[3] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[4] F. W. Lawvere. The category of probabilistic mappings. Unpublished seminar notes.

[5] P. Selinger. A survey of graphical languages for monoidal categories. Preprint, August 2009.