CONDENSATION

SAM EISENSTAT

1. Introduction

We are concerned here with what it is to *understand* a scientific theory—as opposed, for example, to merely being able to use it to make certain predictions. The leading idea here will be that in order to understand a theory, we should seek to organize it in terms of certain concepts. This asks a different kind of question than statistical methods, living in the "context of justification" ask, but nonetheless we begin to approach it using the mathematical methods of probability theory. In particular, we here model concepts as latent variables, though we also believe that some of the ideas here should generalize to concepts that cannot be thus modeled.

The structuring of scientific theories has long been a theme of methodological and philosophical reflection. Lewis [4] proposes that we should understand the world in terms of natural properties. We are more skeptical about the natural; like Dennett [1], we want our ability to use concepts to be the criterion for their "reality". Dennett's approach leads in the direction of algorithmic information theory, as he discusses. While some simple models of concepts have been made using information theory (algorithmic or otherwise), we believe that none of these faces up to challenge of understanding a complicated theory in terms of many concepts. Nevertheless, we do not believe that information theory is powerless here—we seek to begin the necessary work. Mathematically, our methods have some resemblance to those of Wentworth and Lorell [7], who are interested in similar problems.

2. Set-up

In this section, we will introduce the central concepts of latent variable models. We intend for latent variable models to organize the structure of random variable models by positing additional latent variables, which cannot necessarily be defined from the given random variables. However, our definition of latent variable models will be rather weak. We ask that the given variables can be recovered from the latent variable, but we don't ask that this serve any organizing role, we don't ask that we attain an enlightening perspective on the given variables. So, we supplement this using various scoring functions. A latent variable model that gets a good score may more likely help us understand the underlying random variable model.

Now, we can proceed with our objects of study.

Definition 2.1. A *random variable model* is a standard Borel probability space Ω together with a countable family of random variables $X_i \colon \Omega \to R_i$ valued in standard Borel spaces.

We will consider many information-theoretic quantities—entropies, mutual informations, and so on—in relation to random variable models. These quantities will always be defined when the probability space and the family $(X_i)_{i \in I}$ are finite, but

not necessarily otherwise. Hence, in some cases some analytic care will be taken, avoiding the unnecessary exclusion of some infinitary cases from the domain of the theory.

We will also define morphisms of random variable models. We'll give a bit of the idea first. A probability-preserving map of probability spaces $\pi\colon \Omega \to \Lambda$ forgets distinctions. We can think of Ω as an extension of Λ. In other words, we can say that Ω has all the measurable sets $\pi^{-1}E$ corresponding to measurable sets E in Λ, but it can also have other measurable sets. So, a morphism of random variable models will be similar, but we also account for the random variables named by our index set. In particular, we correspondingly let the random variables of the source model make more distinctions than those of the target model. Now, we'll say all this more precisely.

Definition 2.2. A *morphism* of random variable models has the form

$$\left(\pi, \iota, (f_j)_{j\in J}\right) : \left(\Omega, (X_i)_{i\in I}\right) \to \left(\Lambda, (Y_j)_{j\in J}\right), \tag{2.1}$$

where $\pi\colon \Omega \to \Lambda$ is a probability-preserving map, ι is a function $J \to I$, and f_j is a measurable function from the range of $X_{\iota(j)}$ to the range of Y_j. We require, for all $j \in J$, that $Y_j = f_j\left(X_{\iota(j)}\right)$ almost everywhere on Ω. Note that, under the hypotheses stated, the condition $Y_j = f_j\left(X_{\iota(j)}\right)$ defines a measurable set, since $X_{\iota(j)}$, Y_j, and f_j are measurable functions, and the diagonal of $\mathcal{R}Y_j \times \mathcal{R}Y_j$ is a measurable set (by the standard Borel property).

Making the pullback explicit, we can write this as $\pi^*Y_j = f_j\left(X_{\iota(j)}\right)$. Also, note that if J is countable, then this is equivalent to the condition that on a set of full measure in Ω, we have $Y_j = f_j\left(X_{\iota(j)}\right)$ for all $j \in J$.

Our aim here is to understand random variable models by means of auxiliary random variables, which we'll call latent variables. In particular, in a random variable model $\left(\Omega, (X_i)_{i\in I}\right)$, we want the random variables X_i to be functions of certain latent variables. We won't necessarily define these latent variables on Ω; instead we might need an extension of the probability space. This leads to a definition.

Definition 2.3. A *latent variable model* for a random variable model $\left(\Omega, (X_i)_{i\in I}\right)$ is an ordered pair consisting of a random variable model $\left(\Lambda, (Y_A)_{A\in\mathcal{P}^+I}\right)$ and a probability-preserving map $\pi\colon \Lambda \to \Omega$ subject to certain conditions. The random variables $(Y_A)_{A\in\mathcal{P}^+I}$ are indexed by the nonempty powerset of I, and we require that at most countably many of them are nontrivial, in the sense that they have range other than a one-point space. Further, for each random variable X_i, we require that the pullback π^*X_i be almost everywhere a function of the random variables Y_A such that $A \subseteq I$ and $A \ni i$. In other words, π^*X_i is almost everywhere equal to $f_i\left(Y_A\colon A \subseteq I, A \ni i\right)$ for some measurable function f_i from the product of the ranges of the random variables $(Y_A)_{A\subseteq I, A\ni i}$ to the range of X_i. We call the variables $(Y_A)_{A\in\mathcal{P}^+I}$ *latent variables*.

Note that the map π is not necessarily a morphism of random variable models here, since each random variable X_i may depend nontrivially on multiple latent variables.

Definition 2.4. Let $\mathcal{M}$ be a random variable model, with random variables $(X_i)_{i\in I}$, and $\mathcal{L}$ an associated latent variable model with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$. The

simple score of $\mathcal{L}$ at $A \subseteq I$ is

$$\sigma_{\mathcal{L}}(A) = \sum_{\substack{B \in \mathcal{P}^+ I \\ B \cap A \neq \emptyset}} H(Y_B), \tag{2.2}$$

where H is the Shannon entropy, and correspondingly, the *conditioned score* of $\mathcal{L}$ at A is

$$\chi_{\mathcal{L}}(A) = \sum_{B \cap A \neq \emptyset} H\left(Y_B \mid (Y_C)_{C \supsetneq B}\right). \tag{2.3}$$

In particular, we have defined these scores for $A = \emptyset$ for convenience, but they are always zero in this case. These scores are in some sense local, measuring the complexity of the latent model as it pertains to the product of the random variables $(X_i)_{i \in A}$.

Definition 2.5. Let $\mathcal{M}$ be a random variable model with variables $(X_i)_{i \in I}$ and $\mathcal{L}$ an associated latent variable model with variables $(Y_A)_{A \in \mathcal{P}^+ I}$. We will write X and Y with certain subscripts other than elements, respectively, of I and $\mathcal{P}^+ I$ to denote certain products of the random variables in these families. If $A \subseteq I$, we write X_A to denote the product random variable $(X_i)_{i \in A}$. Similarly, for any $\mathcal{F} \subseteq \mathcal{P}^+ I$, we write $Y_{\mathcal{F}}$ to denote the product random variable $(Y_A)_{A \in \mathcal{F}}$. We also define the following notations:

$$Y_{\cap A} = \left(Y_B \colon B \in \mathcal{P}^+ I, B \cap A \neq \emptyset\right) \tag{2.4}$$

$$Y_{\supseteq A} = \left(Y_B \colon B \in \mathcal{P}^+ I, A \subseteq B\right) \tag{2.5}$$

$$Y_{\supsetneq A} = \left(Y_B \colon B \in \mathcal{P}^+ I, A \subsetneq B\right) \tag{2.6}$$

$$Y_{\ni i} = \left(Y_B \colon B \in \mathcal{P}^+ I, i \in B\right). \tag{2.7}$$

In particular, note that all these random variables are valued in standard Borel spaces, since note that for any $i \in I$,

$$Y_{\ni i} = Y_{\cap \{i\}} = Y_{\supseteq \{i\}}. \tag{2.8}$$

3. Perfect condensation

In order to understand our scoring functions, we will ask some questions broadly following two directions of inquiry. First, what is a "good" score? When is a score good enough that we should be interested in a latent variable model that attains that score? Second, what can we conclude about the structure of a latent variable model that gets a good score? In the sequel, we assume that all latent variable models are such that the simple and conditional scores at every set of indices is finite. First, we note that we do indeed have uninteresting latent variable models with bad scores.

Example 3.1. Let $\mathcal{M} = \left(\Omega, (X_i)_{i \in I}\right)$ be a random variable model. Consider the latent variable models $\mathcal{L}_1$ and $\mathcal{L}_2$ associated with $\mathcal{M}$, defined as follows. First, $\mathcal{L}_1$ and $\mathcal{L}_2$ have the same underlying probability space as $\mathcal{M}$, that is,

$$\mathcal{L}_1 = \left(\left(\Omega, (Y_A)_{A \in \mathcal{P}^+ I}\right), \mathrm{id}_\Omega\right) \qquad \mathcal{L}_2 = \left(\left(\Omega, (Z_A)_{A \in \mathcal{P}^+ I}\right), \mathrm{id}_\Omega\right) \tag{3.1}$$

for some families Y and Z of random variables. We will set $Y_{\{i\}} = X_i$ for $i \in I$. For $A \in \mathcal{P}^+ I$ with $|A| \neq 1$, let Y_A be constant. Next, let $Z_I = X_I$, and let Z_A be

constant for $A \subsetneq I$. As previously stated, we have assumed that all the following quantities are defined, so we have

$$\sigma_{\mathcal{L}_1}(A) = \sum_{B \cap A \neq \emptyset} H(Y_B) = \sum_{i \in A} H(X_i) \tag{3.2}$$

$$\chi_{\mathcal{L}_1}(A) = \sum_{B \cap A \neq \emptyset} H(Y_B \mid Y_{\supsetneq B}) = \sum_{i \in A} H(X_i) \tag{3.3}$$

$$\sigma_{\mathcal{L}_2}(A) = \sum_{B \cap A \neq \emptyset} H(Z_B) = H(Z_I) = H(X_I) \tag{3.4}$$

$$\chi_{\mathcal{L}_2}(A) = \sum_{B \cap A \neq \emptyset} H(Z_B \mid Z_{\supsetneq B}) = H(X_I). \tag{3.5}$$

Since we didn't use anything about the structure of $\mathcal{M}$ to produce these latent variable models, we expect that they don't tell us much about $\mathcal{M}$, at least in the typical case. So, these should usually be "bad" scores. If we want to produce even worse scores, we could add more entropy to the latent variables in a way that is irrelevant to determining the variables X_i.

Now, we can establish some easy lower bounds on the simple and conditioned scores.

Proposition 3.2. *Let $(\Omega, (X_i)_{i \in I})$ be a random variable model and $\mathcal{L}$ an associated latent variable model with latent variables $(Y_A)_{A \in \mathcal{P}^+ I}$. Then, for any $A \subseteq I$, we have*

$$\sigma_{\mathcal{L}}(A) \geq \chi_{\mathcal{L}}(A) \geq H(Y_{\cap A}) \geq H(X_A). \tag{3.6}$$

This motivates a definition of perfect condensation.

Definition 3.3. A latent variable model $\mathcal{L}$ *perfectly condenses* a random variable model $\mathcal{M} = (\Omega, (X_i)_{i \in I})$ if $\chi_{\mathcal{L}}(A) = H(X_A)$ for all $A \subseteq I$. Further, $\mathcal{L}$ *simply-perfectly condenses* $\mathcal{M}$ if $\sigma_{\mathcal{L}}(A) = H(X_A)$ for all $A \subseteq I$.

Example 3.4. Let I be an index set, and consider any random variable model $\mathcal{L} = (\Omega, (Y_A)_{A \in \mathcal{P}^+ I})$, indexed by the nonempty power set of I, such that the variables Y_A are jointly independent. We will construct a random variable model $\mathcal{M} = (\Omega, (X_i)_{i \in I})$ such that $\mathcal{M}$ is perfectly condensed by $\mathcal{L}$, as related to $\mathcal{M}$ via the identity map id_Ω. For each $i \in I$, we define X_i to be the product random variable

$$X_i = Y_{\ni i} = (Y_A \colon i \in A \subseteq I). \tag{3.7}$$

Now, for any set $A \subseteq I$, we have

$$\begin{aligned} H(X_A) &= H(X_i \colon i \in A) = H(Y_B \colon B \cap A \neq \emptyset) \\ &= \sum_{B \cap A \neq \emptyset} H(Y_B), \end{aligned} \tag{3.8}$$

using the independence assumption in the last step. This is just the simple score, so $\mathcal{L}$ simply-perfectly condenses $\mathcal{M}$.

We can say quite a lot about a latent variable model if we know that it is a perfect condensation or a simple-perfect condensation.

Lemma 3.5. *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i\in I}$, let $\mathcal{L}$ be an associated latent variable model with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$. Then, the following are equivalent.*

(1) *For all $A \in \mathcal{P}^+I$, we have $H(Y_{\cap A}) = H(X_A)$.*
(2) *For all $i \in I$ and $A \in \mathcal{P}I$ such that $i \in A$, there is some measurable function $f^i_A \colon \mathcal{R}X_i \to \mathcal{R}Y_A$ such that $Y_A = f^i_A(X_i)$ almost everywhere.*

Corollary 3.6. *Let $\mathcal{M}$ be a random variable model with random variables $(X_i)_{i\in I}$, let $\mathcal{L}$ be an associated latent variable model with latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ that perfectly condenses $\mathcal{M}$. Then, whenever we have $i \in A \in \mathcal{P}I$, there is some measurable function $f^i_A \colon \mathcal{R}X_i \to \mathcal{R}Y_A$ such that $Y_A = f^i_A(X_i)$ almost everywhere.*

We can also express the conclusion of this corollary in terms of an equivalence.

Proposition 3.7. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model and $\mathcal{L} = \left(\left(\Lambda, (Y_A)_{A\in\mathcal{P}^+I}\right), \pi\right)$ an associated latent variable model. Then, the following are equivalent.*

(1) For all $i \in I$ and $A \in \mathcal{P}^+I$ such that $i \in A$, there is some measurable function $f^i_A \colon \mathcal{R}X_i \to \mathcal{R}Y_A$ such that $Y_A = f^i_A(X_i)$ almost everywhere.
(2) $\left(\Lambda, (X_i)_{i\in I}\right)$ and $\left(\Lambda, \left(Y_{\cap\{i\}}\right)_{i\in I}\right)$ are equivalent as random variable models, via an equivalence of the form $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (g_i)_{i\in I}\right)$ and $\left(\mathrm{id}_\Lambda, \mathrm{id}_I, (h_i)_{i\in I}\right)$ for some families of functions g and h.

Corollary 3.6 tells us something about perfect, and hence simply-perfect, condensations. By imposing further conditions, we can define stronger properties, which will give us equivalences.

Theorem 3.8. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model and $\mathcal{L} = \left(\left(\Lambda, (Y_A)_{A\in\mathcal{P}^+I}\right), \pi\right)$ an associated latent variable model. The following are equivalent.*

(A1) *$\mathcal{L}$ is a simple-perfect condensation of $\mathcal{M}$.*
(A2) *For all $i \in I$ and $A \in \mathcal{P}^+I$ such that $i \in A$, the latent variable Y_A is a function of X_i. Further, the latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ are jointly independent.*
(A3) *$\mathcal{L}$ is a perfect condensation of $\mathcal{M}$ and the latent variables $(Y_A)_{A\in\mathcal{P}^+I}$ are jointly independent.*

Further, the following are also equivalent:

(B1) *$\mathcal{L}$ is a perfect condensation of $\mathcal{M}$.*
(B2) *For all $i \in I$ and $A \in \mathcal{P}^+I$ such that $i \in A$, the latent variable Y_A is a function of X_i. Further, the latent variables obey the following independence condition, which is a form of the Markov condition from the theory of Bayesian networks. For any $A \in \mathcal{P}^+I$, let $\mathcal{F} \subseteq \mathcal{P}^+I$ be the family of all $B \in \mathcal{P}^+I$ such that B is incomparable in the inclusion order to A, i.e. B is neither a subset nor a superset of A. Then, the random variables Y_A and $Y_{\mathcal{F}}$ are independent conditional on $Y_{\supsetneq A}$.*

This theorem looks like an analogue of Lemma 3.5, strengthening the condition that $H(Y_{\cap A}) = H(X_A)$. This condition is fairly different from perfect condensation in other ways though. Recall the latent variable model $\mathcal{L}_1$ from Example 3.1, in which $Y_{\{i\}} = X_i$ and Y_A is constant for all other A. Here, the condition $H(Y_{\cap A}) =$

$H(X_A)$ is satisfied for every set A. We were able to construct such a latent variable model for any given random variable model—we could for example construct a random variable model $\mathcal{M}$ with random variables X and an associated perfect condensation with many nontrivial latents Y, using Theorem 3.8, and then $\mathcal{M}$ would admit a very different random variable model as in Example 3.1 with latents Z, and both these latent variable models would satisfy the same condition:

$$H(Y_{\cap A}) = H(Z_{\cap A}) = H(X_A) \tag{3.9}$$

for all subsets A of the index set.

By contrast, the condition of perfect condensation is much more rigid. Given a random variable model $\mathcal{M}$ and associated latent variable models $\mathcal{L}_1$ and $\mathcal{L}_2$, we want to say that $\mathcal{L}_1$ and $\mathcal{L}_2$ are essentially the same. It would be straightforward to express this by asserting the existence of an equivalence between $\mathcal{L}_1$ and $\mathcal{L}_2$ satisfying certain properties. Under some hypotheses, this will hold, but mostly we will be concerned with something somewhat weaker. One aspect of this is that it may be that the underlying measure spaces of our two latent variable models—call them Λ_1 and Λ_2—differ in a way that does not interact with the random variables of interest. Maybe different points of Λ_1 can always be distinguished by some latent variable, but Λ_2 is the product of Λ_1 by the unit interval equipped with Lebesgue measure, for example. In order to regard such a difference as inessential, we should be willing to extend our latent variable models by arbitrary morphisms. That is, we should be satisfied with studying latent variable models $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$, together with morphisms $\widetilde{\mathcal{L}_k} \to \mathcal{L}_k$ for each k, and an equivalence between $\widetilde{\mathcal{L}_1}$ and $\widetilde{\mathcal{L}_2}$.

Definition 3.9. Let Ω, Λ_1, and Λ_2 be probability spaces, and $\pi_k \colon \Lambda_k \to \Omega$ probability preserving maps for $k \in \{1,2\}$. A *amalgamation* of Λ_1 and Λ_2 given Ω is a probability space Λ_0 together with probability-preserving maps $\rho_k \colon \Lambda_0 \to \Lambda_k$ such that the diagram

$$\begin{array}{ccc} \Lambda_0 & \xrightarrow{\rho_1} & \Lambda_1 \\ \downarrow{\scriptstyle \rho_2} & & \downarrow{\scriptstyle \pi_1} \\ \Lambda_2 & \xrightarrow{\pi_2} & \Omega \end{array} \tag{3.10}$$

of probability-preserving maps commutes.

Lemma 3.10. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model, and let $\mathcal{L}_1 = \left(\left(\Lambda_1, (Y_A)_{A\in\mathcal{P}^+I}\right), \pi_1\right)$ and $\mathcal{L}_2 = \left(\left(\Lambda_2, (Z_A)_{A\in\mathcal{P}^+I}\right), \pi_2\right)$ be latent variable models associated with $\mathcal{M}$. Then, there is a standard Borel probability space Λ_0 with maps $\rho_k \colon \Lambda_0 \to \Lambda_k$ for $k \in \{1,2\}$, which is an amalgamation of Λ_1 and Λ_2 given Ω. Further, the structures*

(3.11)

$$\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, (\rho_1^* Y_A)_{A\in\mathcal{P}^+I}\right), \pi_1 \circ \rho_1\right) \qquad \widetilde{\mathcal{L}_2} = \left(\left(\Lambda_0, (\rho_2^* Z_A)_{A\in\mathcal{P}^+I}\right), \pi_2 \circ \rho_2\right)$$

are latent variable models associated with $\mathcal{M}$, and

$$\boldsymbol{\rho}_1 = \left(\rho_1, \mathrm{id}_{\mathcal{P}^+I}, (\mathrm{id}_{\mathcal{R}Y_A})_{A\in\mathcal{P}^+I}\right) \colon \widetilde{\mathcal{L}_1} \to \mathcal{L}_1 \tag{3.12}$$

$$\boldsymbol{\rho}_2 = \left(\rho_2, \mathrm{id}_{\mathcal{P}^+I}, (\mathrm{id}_{\mathcal{R}Z_A})_{A\in\mathcal{P}^+I}\right) \colon \widetilde{\mathcal{L}_2} \to \mathcal{L}_2 \tag{3.13}$$

are morphisms of random variable models.

Theorem 3.11. *Let $\mathcal{M} = \left(\Omega, (X_i)_{i\in I}\right)$ be a random variable model, and suppose that $\mathcal{L}_1 = \left(\left(\Lambda_1, (Y_A)_{A\in\mathcal{P}^+I}\right), \pi_1\right)$ and $\mathcal{L}_2 = \left(\left(\Lambda_2, (Z_A)_{A\in\mathcal{P}^+I}\right), \pi_2\right)$ are both perfect condensations of $\mathcal{M}$. Then, we can put the latent variables Y and Z in correspondence in the following sense. There is a standard Borel probability space Λ_0 and maps $\rho_1 \colon \Lambda_0 \to \Lambda_1$ and $\rho_2 \colon \Lambda_0 \to \Lambda_2$ forming an amalgamation of Λ_1 and Λ_2 given Ω, latent variable models*

(3.14)
$$\widetilde{\mathcal{L}_1} = \left(\left(\Lambda_0, \left(\widetilde{Y_A}\right)_{A\in\mathcal{P}^+I}\right), \pi_1 \circ \rho_1\right) \qquad \widetilde{\mathcal{L}_2} = \left(\left(\Lambda_0, \left(\widetilde{Z_A}\right)_{A\in\mathcal{P}^+I}\right), \pi_2 \circ \rho_2\right),$$

and morphisms

(3.15)
$$\boldsymbol{\rho}_1 \colon \widetilde{\mathcal{L}_1} \to \mathcal{L}_1 \qquad \boldsymbol{\rho}_2 \colon \widetilde{\mathcal{L}_2} \to \mathcal{L}_2$$

such that for all $A \in \mathcal{P}^+I$, the random variable $\widetilde{Y_A}$ is a function of $\widetilde{Z_{\supseteq A}}$ almost everywhere, and reciprocally $\widetilde{Z_A}$ is a function of $\widetilde{Y_{\supseteq A}}$ almost everywhere.

References

Dennett, Daniel C. (1991). "Real Patterns". In: *Journal of Philosophy* 88.1, pp. 27–51.

Giry, Michèle (1982). "A categorical approach to probability theory". In: *Categorical Aspects of Topology and Analysis*, pp. 68–85.

Kallenberg, Olav (2021). *Foundations of Modern Probability*. Vol. 99. Probability Theory and Stochastic Modelling. Cham: Springer International Publishing.

Lewis, David K. (1983). "New Work for a Theory of Universals". In: *Australasian Journal of Philosophy* 61.4, pp. 343–377.

Mackey, George W. (1957). "Borel structure in groups and their duals". In: *Transactions of the American Mathematical Society* 85.1, pp. 134–165.

Quine, Willard Van Orman (Jan. 25, 2013). *Word and Object*. The MIT Press.

Wentworth, John and David Lorell (Mar. 20, 2024). "Natural Latents: The Concepts". In: URL: `https://www.alignmentforum.org/posts/mMEbfooQzMwJERAJJ/natural-latents-the-concepts`.