Diagonalization of Dependencies

Towards Trivariate Mutual Information

Challenges in Generalizing MI

Inclusion-Exclusion —

Venn Diagrams Mislead — My intuition is that mutual information in 3 variables is so slippery to define-with-complete-consensus due to a failure of additivity-of-information. An analogous situation occurs in the dimension theory of vector spaces: Note that if A, B, C are subspaces of a finite dimensional vector space, then

$$\dim(A + B) = \dim(A) + \dim(B) - \dim(A \cap B)$$

$$\dim(A \cap B) = \dim(A) + \dim(B) - \dim(A + B)$$

These look like inclusion-exclusion in a venn diagram. Are the following then true?

$$\dim(A{+}B{+}C) = \dim(A){+}\dim(B){+}\dim(C){-}\dim(A{\cap}B){-}\dim(B{\cap}C){-}\dim(C{\cap}A){+}\dim(A{\cap}B{\cap}C)$$

$$\dim(A{\cap}B{\cap}C) = \dim(A){+}\dim(B){+}\dim(C){-}\dim(A{+}B){-}\dim(B{+}C){-}\dim(C{+}A){+}\dim(A{+}B{+}C)$$

No! That's what we would expect from a venn diagram, but it is wrong. Likewise, venn models MI badly. For these reasons, I think $\mathrm{MI}(X;Y;Z)$ is not a real number but instead something that remembers more geometric structure (much as a lattice of subspaces generated by sums and intersections of given subspaces carries more data than just the dimensions of the pure sums).

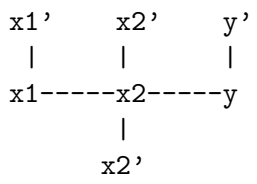
Fix a vector space with basis u, v, x. Let subspaces A, B, C be spanned by $\{u\}, \{v\}, \{u{+}v\}$. Let subspaces A', B', C' be spanned by $\{u\}, \{v\}, \{x\}$. Each of these spaces has dimension 1. The pairwise and triple intersections within each triplet all have dimension 0. And yet the two situations are not isomorphic, since $C \subseteq A{+}B$ but $C' \subsetneq A'{+}B'$. Another example is that (A, B, C) have the same intersection dimensions as (B, C, A), even though C is distinguished.

Distributivity —

Geometry —

Intersections, Complements, Cohomology

Geometry



We're studying a perturbation of a system exactly described by a markov structure (shoe horizontal edges far from nearly-independent). One interesting task in this setting is to predict y based on x2, then fine-tune based on x1. This relates to self-supervised learning for representation learning; pretraining. We want to perturb in "p(y—x2) vs p(y—x2,x1)"; perhaps the right analogue to the original story is for us to assume access to two-point but not three-point marginals.

Start with understanding for one simple SVD in nearly markov how to interpret that SVD.

Now, *what are analogues of the original story's singular vectors?* The bounded-computation framework suggests that we maximize something like

$$\mathrm{MI}'(x1'; y' \mid x2) \qquad \text{or maybe} \qquad \mathrm{MI}'(x1'; y' \mid x2')$$

The thing to do, though, is to define MI′! For instance, might we condition on a "mediating" variable just by taking an expectation over the MI′s of its conditionals, as follows?

$$\mathrm{MI}'(X; Z \mid M) = \mathbb{E}_{m\sim\mu_M} \sup_{c \text{ bilinear}} \mathbb{E}_{x,z\sim\mu_{X\times Z|M}(\cdot|m)}\left[c(x,z)\right] - \log \mathbb{E}_{x,z\sim\mu_{X|M}(\cdot|m)\times\mu_{Z|M}(\cdot|m)}\left[\exp c(x,z)\right]$$

Or should c instead somehow know about m as well?

Relatedly, if we condition on $x2'$ then is there a way we would like to incorporate $x2'$'s linear structure? And if we condition instead on $x2$, is there a decomposition theorem reducing to old work.

In deep learning, we implicitly characterize a concept-to-be-learned by specifying how it should behave in composition with other concepts. That is, *we build our tools by using them.* This Grothendieckian framework sheds light on many global architectural themes of deep learning: autoencoders, cycle consistency losses, attention layers, , siamese networks, contrastive losses. A key example is self-supervised representation learning, wherein we learn to represent data by attempting to predict some of its "parts" or "aspects" from others.

In these notes we elaborate on this theme in the case of weak dependencies.

In what sense is the left figure a better featurization of our data than the right?

Universal Features

He had bought a large map representing the sea,
Without the least vestige of land:
And the crew were much pleased when they found it to be
A map they could all understand.
— The Hunting of the Snark

We summarize the (2019 theory) of Shao-Lun Huang, Anuran Makur, Gregory W. Wornell, and Lizhong Zheng.

MI between nearly joint-uniform variables

Suppose X, Z have uniform marginals and are nearly independent —

$$p(x', z') = \frac{\exp(\epsilon\pi(x', z'))}{|X \times Z|}$$

— in that their mutual information $\mathrm{MI}(X;Z)$ is $\ll \epsilon^2$. Here, $\pi(\cdot,\cdot) \ll 1$. Write $M = |X \times Z|$.

BIRD'S EYE STORY — We wish to featurize X, Z into k-dimensional variables X', Z' so as to preserve as much shared information as possible:

$$X' - X - Z - Z' \qquad \mathrm{MI}'(X';Z') \lesssim \mathrm{MI}(X;Z)$$

Now, our notion MI' of shared information ought not literally be mutual information; otherwise, for positive k almost all embeddings of X, Z into $\mathbb{R}^k$ would trivially preserve information. Still, there is a sense in which some embeddings of X, Z are "more aligned" or "more co-informative" than others. We'll focus on the notion of *accessible co-informativity*: that is, on a measure of how related variables appear despite bounded computation to compare them. Variational bounds provide a convenient way to model bounded computation:

$$\mathrm{MI}(X,Z) = \sup_c \mathbb{E}_{x,z\sim\mu_{X\times Z}}\left[c(x,z)\right] - \log \mathbb{E}_{x,z\sim\mu_X\times\mu_Z}\left[\exp c(x,z)\right]$$

with the supremum achieved at $c_{x,z} = \log(p_{x,z}/(p_x p_z)) = \epsilon\pi_{x,z}$. Let's define MI' to be the same thing except the function c over which we vary must be **bilinear**.

What does that MI objective look for c near the unconstrained optimum $c = \epsilon\pi$? Well, for $c = \epsilon(\gamma + \pi)$ we have to second order in ϵ and terms constant in γ:

$$\begin{aligned}
\cdots &= \sum_{x,z} \epsilon\gamma_{x,z}\cdot\exp(\epsilon\pi_{x,z})/M - \log\left(1 + \sum_{x,z}(\exp(\epsilon\gamma_{x,z})-1)\cdot\exp(\epsilon\pi_{x,z})/M\right) \\
&\in C - \frac{1}{2}\sum_{x,z}(\epsilon\gamma_{x,z})^2\cdot\exp(\epsilon\pi_{x,z})/M + \frac{1}{2}\left(\sum_{x,z}\epsilon\gamma_{x,z}\cdot\exp(\epsilon\pi_{x,z})/M\right)^2 + o(\epsilon^2) \\
&= C - \frac{\epsilon^2}{2}\mathrm{Var}_{x,z\sim\mu_{X\times Z}}[\gamma_{x,z}] + o(\epsilon^2)
\end{aligned}$$

Neither the true nor approximate MI objectives care whether we translate γ vertically. So let's take γ to have mean zero with respect to the true joint. The above says that for constrained c (so long as the feasible set gets very close to the true optimum), we want a constrained c that minimizes a least-squares error $\mathrm{Var}_{x,z\sim\mu_{X\times Z}}[\epsilon\pi - c]$ — wow!

Let's consider c bilinear in x', z' (the latter featurized per f, g):

$$c_{x,z} = \epsilon\sum_k f_x^k g_z^k$$

Momentarily neglecting centering and the difference between variance with respect to joint vs independent marginals, we see that the SVD minimizes that least-squares error!

Desired Result: *To leading order in ϵ, the optimal features f, g are proportional to the top k non-trivial singular vectors of $\exp(\epsilon\pi)$ (when SVD and optimization are uniquely soluble).*

TECHNICAL DETAILS —

This result justifies the truncation developed in HMWZ 2019 (§IV.C).

COMPARISON TO HMWZ — That paper characterizes its truncation as a solution to five optimization problems (§III.B; §V.C thru §V.F), each of which : (TODO: verify!)

- relies on a norm introduced manually rather than derived from KL (Ky-fan, nuclear,) or

	§III.B	§V.C	§V.D	§V.E	§V.F	ours
norm	Ky-fan				nuclear	information
constraint				feature noise		computation
	Ky-fan					

- supposes, besides near-independence of X, Z, an asymptotically extreme feature noise

By contrast, in our analysis the universal features fall out from a linearity constraint (having to do with pure linearity rather than norms).

adaptation to non-uniform marginals

principal components

We review principal component analysis.

PYTHAGOREAS AND GAUSS —

SIDETRACK: CAUSAL GRASSMANIANS — Pearl's do-calculus –¿ fit a nilpotent matrix to data?! PCA:correlation :: ?:causation

```
(I + T) nxn
T^n = 0
 FED
(1...
( 1..
(  1.
(   1)
```

TODO: in nearly markov case, notion of "orthogonality" is different... offdiagonals ... perhaps a triangularization?? what does this look like?

—

FISCHER METRIC —

greg's analysis of weak pairwise dependencies

Nearly Markov

Independence is a heady draft...
— Maya Angelou

```
x1'      y'
 |       |
x1 - x2- y
     |
    x2'
```

In nearly markov structure, predict y based on x2; and fine-tune based on x1 — relate to self-supervised learning for representation learning; pretrainig. (Assume have two-point but not three-point marginals). PRIORIZZE THIS PARAGRAPH perturb in "p(y—x2) vs p(y—x2,x1)" — what is analogue of singular vectors/values here?

Generalized Self-Supervised

Sample Complexity