Information topology and probabilistic graphical models

Juan Pablo Vigneaux
IMJ-PRG - Université Paris 7

August 8, 2017

Outline

Minimal background

Shannon (1948): information content of a random variable $X : \Omega \to \{x_1, ..., x_n\}$ is

$$S_1[X](P) = -\sum_{k=0}^{n} \mathbb{P}(X = x_i) \log_2 \mathbb{P}(X = x_i) = \mathbb{E}_P \left\{ \log \left(\frac{1}{P(X)} \right) \right\}. \qquad (1)$$

where $\mathbb{P}$ is a probability on Ω. The function S_1 is called entropy.

Minimal background

Shannon (1948): information content of a random variable $X : \Omega \to \{x_1, ..., x_n\}$ is

$$S_1[X](P) = -\sum_{k=0}^{n} \mathbb{P}(X = x_i) \log_2 \mathbb{P}(X = x_i) = \mathbb{E}_P \left\{ \log \left(\frac{1}{P(X)} \right) \right\}. \tag{1}$$

where $\mathbb{P}$ is a probability on Ω. The function S_1 is called entropy.

Information is related to **uncertainty**.

1. If $P(X = x_i) = 1$ for certain i, then $S_1[X] = 0$.
2. Uniform distribution on $\{x_1, ..., x_n\}$ implies $S_1[X]$ maximal.

Minimal background

Shannon (1948): information content of a random variable $X : \Omega \to \{x_1, ..., x_n\}$ is

$$S_1[X](P) = -\sum_{k=0}^{n} \mathbb{P}(X = x_i) \log_2 \mathbb{P}(X = x_i) = \mathbb{E}_P \left\{ \log \left(\frac{1}{P(X)} \right) \right\}. \quad (1)$$

where $\mathbb{P}$ is a probability on Ω. The function S_1 is called entropy.

Information is related to **uncertainty**.

1. If $P(X = x_i) = 1$ for certain i, then $S_1[X] = 0$.
2. Uniform distribution on $\{x_1, ..., x_n\}$ implies $S_1[X]$ maximal.

Shannon recognized an important relation:

$$S_1[X, Y] = S_1[X] + S_1[Y|X].$$

Where $S_1[Y|X] = \mathbb{E}_P(S_1[Y](P|_X))$. (It looks like $0 = X.f[Y] - f[XY] + f[X]$.)

Havrda - Charvát (1967), Tsallis (1988): α-entropy, defined as

$$S_\alpha[X](P) = c_\alpha \left(\sum_{k=1}^{n} P(X = x_k)^\alpha - 1 \right), \tag{2}$$

where $\alpha \in (0, \infty) \setminus \{1\}$ and c_α is some constant. Typical choices give $S_\alpha[X] \to S_1[X]$ when $\alpha \to 1$.

Generalized functions: α-entropies

Havrda - Charvát (1967), Tsallis (1988): α-entropy, defined as

$$S_\alpha[X](P) = c_\alpha \left(\sum_{k=1}^{n} P(X = x_k)^\alpha - 1 \right), \tag{2}$$

where $\alpha \in (0, \infty) \setminus \{1\}$ and c_α is some constant. Typical choices give $S_\alpha[X] \to S_1[X]$ when $\alpha \to 1$.

It satisfies a similar cocycle relation (stay tuned).

Our purpose

To introduce a generalized notion of *statistical space*.
Entropies appear as cocycles for a suitably defined cohomology on this space: information cohomology.

Our purpose

To introduce a generalized notion of *statistical space*.
Entropies appear as cocycles for a suitably defined cohomology on this space: information cohomology.

'A theory of "shape" for information.'

Our purpose

To introduce a generalized notion of *statistical space*.
Entropies appear as cocycles for a suitably defined cohomology on this space: information cohomology.

'A theory of "shape" for information.'

We will see that this space is the natural ground for some probabilistic problems with geometrical flavour.

Outline

Topoi

For any "good" space X, the category $\mathrm{Sh}(X)$ contains basically the same topological information as X and should be seen as X disguised as a topos (Moerdijk).

Grothendieck, Verdier,... (SGA IV): generalized the definition of sheaves to allow cases not covered by usual topology.

Our topos: presheaves on a small category $\mathcal{S}$ such that...

Observables

Given a set Ω, introduce the category of "finite observables" $\mathcal{O}(\Omega)$:

- objects: finite partitions.
- arrows: $X \to Y$ if X refines Y.

Definition

An **information structure** $\mathcal{S}$ is a full subcategory of $\mathcal{O}(\Omega)$ such that

1. $\mathrm{Ob}(\mathcal{S})$ contains $\mathbf{1} := \{\Omega\}$.
2. If $Y \leftarrow X \to Z$ in $\mathcal{S}$, then $Y \times Z$ (the product in $\mathcal{O}(\Omega)$) is in $\mathrm{Ob}(\mathcal{S})$.

Dictionary

$$X : \Omega \to E_X = \{x_1, ..., x_n\} \text{ r.v.} \quad \leftrightarrow \quad \text{Partition } \{X = x_i\}_{i=1}^n$$

$$(Y, Z) : \Omega \to E_Y \times E_Z \quad \leftrightarrow \quad \text{Product of partitions } Y \times Z$$

$$\omega \mapsto (Y(\omega), Z(\omega))$$

Structure sheaf

For each $X \in \mathrm{Ob}(\mathcal{S})$, the set

$$\mathcal{S}_X := \{\, Y \mid X \to Y \,\}$$

has the structure of a monoid, with binary law $(Y, Z) \mapsto YZ := Y \times Z$.

Definition

The **structure sheaf** $\mathcal{A}$ is the sheaf of algebras $X \mapsto \mathcal{A}_X := \mathbb{R}[\mathcal{S}_X]$ (the monoid algebra).

Family of examples: Graphical models

$[n] := \{1, ..., n\}$
Set $\Omega = \prod_{i \in [n]} E_i$ and let X_i be the projection on the i-th component.

For each $I \subset [n]$, $X_I := \prod_{i \in I} X_i$. Then the abstract simplex $\Delta([n])$ can be seen as an information structure.

Family of examples: Graphical models

$[n] := \{1, ..., n\}$
Set $\Omega = \prod_{i \in [n]} E_i$ and let X_i be the projection on the i-th component.

For each $I \subset [n]$, $X_I := \prod_{i \in I} X_i$. Then the abstract simplex $\Delta([n])$ can be seen as an information structure.

Any simplicial subcomplex K of $\Delta([n])$ also defines a structure $\mathcal{S}_K$.

Family of examples: Graphical models

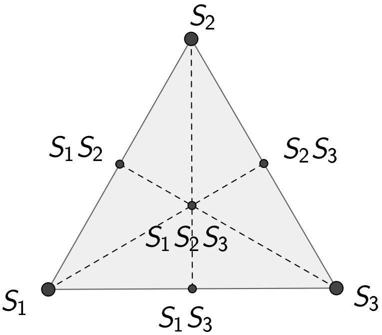
$[n] := \{1, ..., n\}$
Set $\Omega = \prod_{i \in [n]} E_i$ and let X_i be the projection on the i-th component.

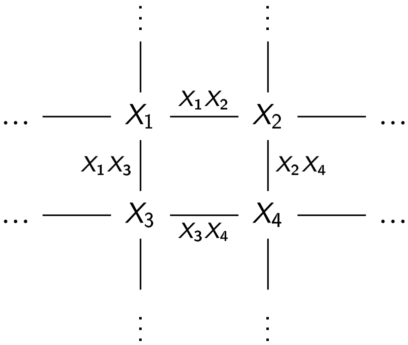
For each $I \subset [n]$, $X_I := \prod_{i \in I} X_i$. Then the abstract simplex $\Delta([n])$ can be seen as an information structure.

Any simplicial subcomplex K of $\Delta([n])$ also defines a structure $\mathcal{S}_K$.

Examples:

- Ising model: Fix a lattice $\Lambda \subset \mathbb{Z}^d$. If each X_i represent the "spin" on a site of Λ, and we consider $X_i X_j$ each time that $(i,j) \in \Lambda$, we obtain an information structure $\mathcal{S}_K$ where $K = \Lambda$ (1-dim. complex).
- More general: Factor graphs.


S_2
$S_1 S_2$
$S_2 S_3$
$S_1 S_2 S_3$
S_1
S_3
$S_1 S_3$


X1
X2
X3
X4
X1X2
X1X3
X2X4
X3X4

Probabilities

For each $X \in \mathrm{Ob}(\mathscr{S})$, set

$$\Delta(X) = \left\{ p : X \to [0,1] \,\middle|\, \sum_{x \in X} p(x) = 1 \right\}. \tag{3}$$

Note that $\Delta(X) \cong \{(p_0, \ldots, p_n) \in \mathbb{R}^n \,|\, p_i \geq 0$ and $\sum p_i = 1\}$ whenever $E_X = \{x_0, \ldots, x_n\}$.

For each $X \in \mathrm{Ob}(\mathcal{S})$, set

$$\Delta(X) = \left\{ p : X \to [0,1] \,\Big|\, \sum_{x \in X} p(x) = 1 \right\}. \tag{3}$$

Note that $\Delta(X) \cong \{ (p_0, \dots, p_n) \in \mathbb{R}^n \,|\, p_i \geq 0 \text{ and } \sum p_i = 1 \}$ whenever $E_X = \{x_0, \dots, x_n\}$. To each arrow of refinement $\pi : X \to Y$, it corresponds an operation called marginalization, $\Delta(\pi) \equiv \pi_* : Q_X \to Q_Y$, given by

$$\forall y \in Y, \quad \pi_* p(y) = \sum_{x \in \pi^{-1}(y)} p(x). \tag{4}$$

Notation: Y_* instead of π_* if π clear from context.

Definition

A **probability functor** $Q : \mathcal{S} \to \mathcal{S}ets$ associates:

- To each $X \in \text{Ob}(\mathcal{S})$ a subset Q_X of $\Delta(X)$ stable under conditioning by variables $Y \in \mathcal{S}_X$;
- To each $\pi : X \to Y$, the map $Q(\pi) = \Delta(\pi)|_{Q_X}$.

Simplicial families are always stable under conditioning.

An important presheaf

For each X, define

$$F(Q_X) := \{f : Q_X \to \mathbb{R} \mid f \text{ measurable}\}. \tag{5}$$

For each arrow $\pi : X \to Y$, there is a map $\pi^* : F(Q_Y) \to F(Q_X)$, $f \mapsto f \circ \pi_*$.

For each $\alpha > 0$, define an action of $\mathcal{A}_X$ on $F(Q_X)$ such that

$$\forall Y \in \mathcal{S}_X, \quad (Y.f)(P) = \sum_{y \in E_Y} P(Y = y)^\alpha f(P|_{Y=y}). \tag{6}$$

The corresponding $\mathcal{A}_X$-module is denoted $F_\alpha(Q_X)$.

	Differential	Algebraic	Statistical
Space	Open(M) for M manifold	Spec(R) for ring R	Information structure $\mathcal{S}$
Structure sheaf	$C^0(U)$	R_p	$\mathcal{A}_X$
Sheaves of modules	$\Omega^k(U)$	...	$F_\alpha(Q_X)$
Action	$C^0(U)$ acts on $\Omega^k(U)$ by multiplication		$\mathcal{A}_X$ acts on $F_\alpha(Q_X)$ by conditioning.

Outline

Information cohomology: Definition

The information topos is the ringed topos $(\mathrm{PSh}(\mathcal{S}), \mathcal{A})$. We are particularly interested in:
$\mathcal{Mod}(\mathcal{A})$ = category of presheaves $M : \mathcal{S} \to \mathcal{Ab}$ such that, for all $X \in \mathrm{Ob}(\mathcal{S})$, $\mathcal{A}_X$ acts on $M(X)$, functorially.

Information cohomology: Definition

The information topos is the ringed topos $(\mathrm{PSh}(\mathcal{S}), \mathcal{A})$. We are particularly interested in:
$\mathcal{Mod}(\mathcal{A})$ = category of presheaves $M : \mathcal{S} \to \mathcal{Ab}$ such that, for all $X \in \mathrm{Ob}(\mathcal{S})$, $\mathcal{A}_X$ acts on $M(X)$, functorially.
This is an abelian category (kernels and cokernels are computed over each X).

Information cohomology: Definition

The information topos is the ringed topos $(\mathrm{PSh}(\mathcal{S}), \mathcal{A})$. We are particularly interested in:
$\mathcal{Mod}(\mathcal{A})$ = category of presheaves $M : \mathcal{S} \to \mathcal{Ab}$ such that, for all $X \in \mathrm{Ob}(\mathcal{S})$, $\mathcal{A}_X$ acts on $M(X)$, functorially.
This is an abelian category (kernels and cokernels are computed over each X).

Let $\mathbb{R}_{\mathcal{S}}$ be the presheaf that associates to each $X \in \mathrm{Ob}(\mathcal{S})$ the abelian group $\mathbb{R}$ with trivial $\mathcal{A}_X$ action (for $s \in \mathcal{S}_X$ and $r \in \mathbb{R}$, take $s \cdot r = r$).

Information cohomology: Definition

The information topos is the ringed topos $(\mathrm{PSh}(\mathcal{S}), \mathcal{A})$. We are particularly interested in:
$\mathcal{Mod}(\mathcal{A})$ = category of presheaves $M : \mathcal{S} \to \mathcal{Ab}$ such that, for all $X \in \mathrm{Ob}(\mathcal{S})$, $\mathcal{A}_X$ acts on $M(X)$, functorially.
This is an abelian category (kernels and cokernels are computed over each X).

Let $\mathbb{R}_{\mathcal{S}}$ be the presheaf that associates to each $X \in \mathrm{Ob}(\mathcal{S})$ the abelian group $\mathbb{R}$ with trivial $\mathcal{A}_X$ action (for $s \in \mathcal{S}_X$ and $r \in \mathbb{R}$, take $s \cdot r = r$).
The functor $\mathrm{Hom}(\mathbb{R}_{\mathcal{S}}, -)$ is left exact, its associated derived functors are $\mathrm{Ext}^n(A, -)$, for $n \geq 0$.

Definition

The **information cohomology** groups with coefficients in F are

$$H^n(\mathcal{S}, F) := \mathrm{Ext}^n(\mathbb{R}_{\mathcal{S}}, F).$$

The existence of a projective resolution $0 \leftarrow \mathbb{R}_{\mathcal{S}} \leftarrow B_\bullet$ (via the bar construction) implies that information cohomology can be computed as the cohomology of the differential complex $(\mathrm{Hom}(B_\bullet, F), \delta)$.

In particular,

$$H^1(\mathcal{S}, F_\alpha(Q)) = \ker\{\delta : \mathrm{Hom}(B_1, F_\alpha(Q)) \to \mathrm{Hom}(B_2, F_\alpha(Q))\}. \tag{7}$$

(Since $\mathrm{im}\,\delta_0$ is trivial.)
$B_1(X) := \mathcal{A}_X$-module freely generated by the symbols $[Y]$ with $Y \in S_X$, and

The existence of a projective resolution $0 \leftarrow \mathbb{R}_{\mathcal{S}} \leftarrow B_\bullet$ (via the bar construction) implies that information cohomology can be computed as the cohomology of the differential complex $(\mathrm{Hom}(B_\bullet, F), \delta)$.

In particular,

$$H^1(\mathcal{S}, F_\alpha(Q)) = \ker\{\delta : \mathrm{Hom}(B_1, F_\alpha(Q)) \to \mathrm{Hom}(B_2, F_\alpha(Q))\}. \tag{7}$$

(Since $\mathrm{im}\,\delta_0$ is trivial.)
$B_1(X) := \mathcal{A}_X$-module freely generated by the symbols $[Y]$ with $Y \in S_X$, and

$$(\delta f)[X|Y] = X.f[Y] - f[XY] + f[X]. \tag{8}$$

Locality

Let $f \in \mathrm{Hom}(B_1, F_\alpha(Q))$. Whenever $X \to Y \to Z$,

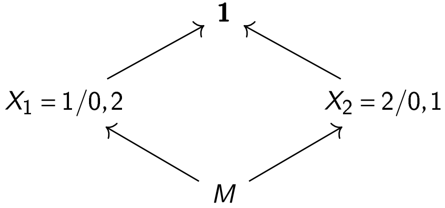
$$\begin{array}{ccccc} [Z] & & B_1(Y) \xrightarrow{f_Y} F_\alpha(Q_Y) & & f_Y[Z] \\ \downarrow & & \downarrow \qquad\qquad \downarrow & & \downarrow \\ [Z] & & B_1(X) \xrightarrow{f_X} F_\alpha(Q_X) & & f_X[Z] = \pi^* f_Y[Z] \end{array}$$

Equivalently:

$$f_X[Z](P_X) = f_Z[Z](Z_* P_X). \tag{9}$$

Particular case

Let $\Omega = \{0, 1, 2\}$. Structure



A cocycle is given by three local functions:

$$f[X_1](p_0, p_1, p_2) = f[X_1](p_1, p_0 + p_2)$$
$$f[X_2](p_0, p_1, p_2) = f[X_2](p_2, p_0 + p_2)$$
$$f[X_1X_2](p_0, p_1, p_2)$$

They verify

$$0 = X.f[Y] - f[XY] + f[X]$$
$$0 = Y.f[X] - f[XY] + f[Y]$$

Particular case (continued)

The equations above give

$$X.f[Y] - f[Y] = Y.f[X] - f[X]. \tag{10}$$

Therefore, we look for two functions f_1, f_2 that verify: for all $(p_0, p_1, p_2) \in \Delta^2$,

$$\begin{aligned}(1-p_2)^\alpha f_1\left(\frac{p_0}{1-p_2}, \frac{p_1}{1-p_2}\right) - f_1(1-p_1, p_1) \qquad (11)\\ = (1-p_1)^\alpha f_2\left(\frac{p_0}{1-p_1}, \frac{p_2}{1-p_1}\right) - f_2(1-p_2, p_2).\end{aligned}$$

Solutions: multiples of the corresponding entropy

$$s_\alpha(x) = \begin{cases} -x\log x - (1-x)\log(1-x) & \text{if } \alpha = 1 \\ \frac{1}{1-\alpha}(x^\alpha + (1-x)^\alpha - 1) & \text{otherwise} \end{cases}$$

Theorem

Under appropriate non-degeneracy hypotheses

$$H^1(\mathcal{S}, F_\alpha(Q)) \cong \prod_{c \in H_0^{CW}(\mathcal{S}^*)} \mathbb{R} \cdot S_\alpha^{(c)}$$

where c *represents a connected component of* $\mathcal{S}^* = \mathcal{S} \setminus \mathbf{1}$ *and*

$$S_\alpha^{(c)}[X] = \begin{cases} S_\alpha[X] & \text{if } X \in \mathrm{Ob}(c) \\ 0 & \text{if } X \notin \mathrm{Ob}(c) \end{cases}$$

Extension of algebras

$\mathrm{Ext}^n(\mathbb{R}_{\mathcal{S}}, F_\alpha(Q))$ encodes the different ways of constructing an **split** extension of algebras

$$0 \to F_\alpha(Q) \to E \to \mathcal{A} \to 0. \tag{12}$$

In fact, these are always congruent to

$$0 \to F_\alpha(Q) \to F_\alpha \rtimes \mathcal{A} \to \mathcal{A} \to 0. \tag{13}$$

and the splitting is given by a morphism of presheaves algebras $d : \mathcal{A} \to F_\alpha \rtimes \mathcal{A}$ such that

$$(d[Y], Y) \bullet (d[X], X) = (d[YX], YX) \Leftrightarrow (d[Y] + Y.d[X], YX) = (d[YX], YX). \tag{14}$$

Only solution: the entropy.

This transformation of products (successive measurements) into a sum (information) correspond to the third axiom of Shannon's characterization (Theorem 2 in his paper) and we see now that it suffices in generic situations.

But why are we interested in this property?

$X : \Omega \to E_X = \{x_0, x_1, ..., x_s\}$ r.v. and $\pi : X \to Y$.
Take $N \in \mathbb{N}$ and $N(0), ..., N(s)$ such that $\sum N(i) = N$; define $\nu(i) = N(i)/N$.

From $(x_0 + \cdots + x_s)^N = (\sum_{y_i}(x_{j_1(i)} + \cdots + x_{j_k(i)}))^N$, we deduce that

$$\log\binom{N}{N(0), ..., N(s)} = \log\binom{N}{NY_*\nu(0), ..., NY_*\nu(t)} + \sum_{i=0}^{t} \log\binom{NY_*\nu(i)}{N(j_1(i)), ..., N(j_k(i))} \tag{15}$$

If we take the limit $N \to \infty$ imposing $N(i)/N \to \mu_i$, we obtain

$$S(\mu) = S(Y_*\mu) + \sum_{i=0}^{t} Y_*\mu(i) S(\mu|_{Y=y_i}). \tag{16}$$

Outline

Frustration

A problem from statistical mechanics:
Consider the Ising model $\mathcal{S}_\Lambda$; over each edge $e = \{i, j\}$ of Λ, prescribe the law $p_e \equiv p_{i,j} \in \Delta(X_e)$ (local interactions). These must be compatible i.e. for each segment

$$i \text{ ------ } j \text{ ------ } k$$

in Λ, they must verify

$$(X_j)_* p_{i,j} = (X_j)_* p_{j,k}. \tag{17}$$

Equivalently: p must be a section of $\Delta : \mathcal{S}_\Lambda \to s\mathcal{S}ets$.

Problem: when does it exist $q \in \Delta(X_\Lambda)$ such that, for all $e \in \Lambda$,

$$(X_e)_* q = p_e\,? \tag{18}$$

P. Baudot and D. Bennequin, *The homological nature of entropy*, Entropy, 17 (2015), pp. 3253–3318.

I. Csiszár, *Axiomatic characterizations of information measures*, Entropy, 10 (2008), pp. 261–273.

M. Gromov, *In a search for a structure, part 1: On entropy*, Preprint available at http://www.ihes.fr/gromov, (2012).

J. Havrda and F. Charvát, *Quantification method of classification processes. concept of structural a-entropy*, Kybernetika, 3 (1967), pp. 30–35.

C. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, 27 (1948), pp. 379–423, 623–656.