# Categorical Aspects of Parameter Learning

ERATO MMSD, Tokyo
Oct. 24, 2018

Bart Jacobs — Radboud University
bart@cs.ru.nl





## Outline







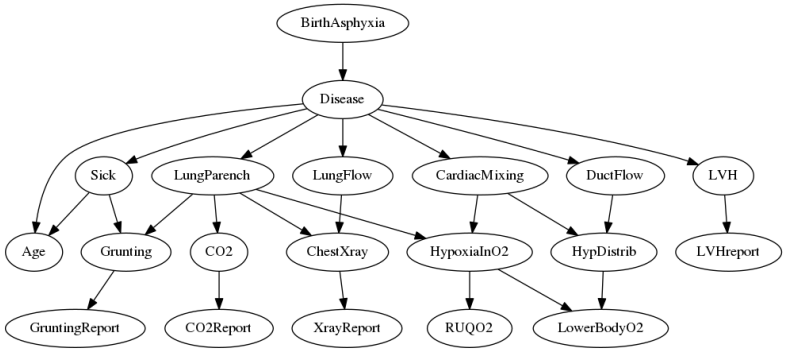
Categorical Aspects of

## Where we are, so far





## Example medical Bayesian network




- Such networks are used for inference: given this-and-this evidence, the likelihood of that becomes ...
- far-reaching decisions can be taken, based on such outcomes.





## Starting point

**Basic question**: how do we get "accurate" Bayesian networks?

- Extraction of knowledge from experts is cumbersome
- Learning from data is the current paradigm.

There are **two forms of learning**:

(1) structure learning: what is the graph?

(2) parameter learning: given the graph structure, what are the numbers in Conditional Probability Tables (really: channels)?

We only look at the second point.





## Multisets

- A multiset is a set in which elements may occur multiple times
- Convenient notation as formal sum $\sum_i n_i|\, x_i \,\rangle$, where $n_i \in \mathbb{N}$ is the multiplicity of element $x_i \in X$
- Equivalently,

$$\mathcal{M}(X) := \{\varphi \colon X \to \mathbb{N} \mid \mathrm{supp}(\varphi) \text{ is finite}\}$$

  where $\mathrm{supp}(\varphi) = \{x \in X \mid \varphi(x) \neq 0\}$.
- This is functorial: for $f \colon X \to Y$ one gets $\mathcal{M}(f) \colon \mathcal{M}(X) \to \mathcal{M}(Y)$ with:

$$\mathcal{M}(f)\Big(\textstyle\sum_i n_i|\, x_i \,\rangle\Big) := \sum_x n_i|\, f(x_i) \,\rangle.$$

- This $\mathcal{M}\colon$ **Sets** $\to$ **Sets** is in fact a monad.
- We use subsets: We use special subsets of multisets $\sum_i n_i|\, x_i \,\rangle$
  - $\mathcal{M}_*(X)$ for non-empty multisets, with $n := \sum_i n_i > 0$
  - $\mathcal{M}_{\circledast}(X)$ for multisets with full support, equal to $X$, so each $n_i > 0$.





## Distributions

- A (finite, discrete probability) distribution is a convex combination of elements
- Now written as formal convex sum $\sum_i r_i|\, x_i \,\rangle$, where $r_i \in [0,1]$ is the multiplicity of element $x_i \in X$, where $\sum_i r_i = 1$.
- Again a monad $\mathcal{D}\colon$ **Sets** $\to$ **Sets**
- $\mathcal{D}_{\circledast}(X)$ contains distributions with 'full' support





## Where we are, so far





## The frequentist learning map

**Theorem**

*There is a natural transformation, via normalisation:*

$$\mathcal{M}_*(X) \xrightarrow{\text{flrn}} \mathcal{D}(X) \quad \textit{via} \quad \text{flrn}\Big(\textstyle\sum_i n_i|\, x_i\,\rangle\Big) := \displaystyle\sum_i \frac{n_i}{\sum n_j}|\, x_i\,\rangle$$

**Proof**

$$\begin{aligned}(\text{flrn} \circ \mathcal{M}(f))\big(\textstyle\sum_i n_i|\, x_i\,\rangle\big) &= \text{flrn}\big(\textstyle\sum_i n_i|\, f(x_i)\,\rangle\big)\\ &= \textstyle\sum_i \frac{n_i}{\sum n_j}|\, f(x_i)\,\rangle\\ &= \mathcal{D}(f)\big(\textstyle\sum_i \frac{n_i}{\sum n_j}|\, x_i\,\rangle\big)\\ &= \big(\mathcal{D}(f) \circ \text{flrn}\big)\big(\textstyle\sum_i n_i|\, x_i\,\rangle\big).\end{aligned}$$

**Aside:** flrn is not a map of monads


## Example, part I: medicine and blood pressure

| | no medicine | medicine 1 | medicine 2 | totals |
|---|---|---|---|---|
| **high** | 10 | 35 | 25 | 70 |
| **low** | 5 | 10 | 15 | 30 |
| **totals** | 15 | 45 | 40 | 100 |

- Use sets for medicine and blood pressure $B = \{H, L\},\ M = \{0, 1, 2\}$
- Write table as multiset $\tau \in \mathcal{M}(B \times M)$

  $\tau = 10|\, H, 0\,\rangle + 35|\, H, 1\,\rangle + 25|\, H, 2\,\rangle + 5|\, L, 0\,\rangle + 10|\, L, 1\,\rangle + 15|\, L, 2\,\rangle$
- Total column and row are first and second marginal:

$$\begin{aligned}\mathcal{M}(\pi_1)(\varphi) &= 70|\, H\,\rangle + 30|\, L\,\rangle\\ \mathcal{M}(\pi_2)(\varphi) &= 15|\, 0\,\rangle + 45|\, 1\,\rangle + 40|\, 2\,\rangle\end{aligned}$$


## Example, part II

We 'learn' probabilities from the table by normalisation:

$$\begin{aligned}\omega &:= \text{flrn}(\varphi)\\ &= 0.1|\, H, 0\,\rangle + 0.35|\, H, 1\,\rangle + 0.25|\, H, 2\,\rangle + 0.05|\, L, 0\,\rangle + 0.1|\, L, 1\,\rangle + 0.15|\, L, 2\,\rangle\end{aligned}$$

- By naturality, it does not matter how one learns from the 'totals': the marginals of $\omega$ can also be learned from these side column/row
- That is:

$$\mathcal{D}(\pi_i)(\omega) = \mathcal{D}(\pi_i)\big(\text{flrn}(\tau)\big) = \text{flrn}\big(\mathcal{M}(\pi_i)(\tau)\big)$$


## Frequentist learning and disintegration

Write $n = \{1, 2 \ldots, n\}$

- For **multisets** there is an isomorphism:

  $\mathcal{M}(n \times m) \xrightarrow[\cong]{\text{row}} \mathcal{M}(m)^n$ via $\text{row}\big(\sum_{ij} k_{ij}|\, ij\,\rangle\big)(i) = \sum_j k_{ij}|\, j\,\rangle$
- For **distribution** there is disintegration

  $\mathcal{D}(n \times m) \xrightarrow{\text{dis}} \mathcal{D}(m)^n$ via $\text{dis}\big(\sum_{ij} r_{ij}|\, ij\,\rangle\big)(i) = \sum_j \frac{r_{ij}}{\sum_i r_{ij}}|\, j\,\rangle$

**Theorem**

*Learning commutes with disintegrations, as in:*

$$\begin{array}{ccc}\mathcal{M}_*(n \times m) & \xrightarrow{\text{row}} & \mathcal{M}_*(m)^n\\ {\scriptstyle\text{flrn}}\downarrow & & \downarrow{\scriptstyle\text{flrn}^n}\\ \mathcal{D}(n \times m) & \xrightarrow{\text{dis}} & \mathcal{D}(m)^n\end{array}$$


## Example, part III

Recall the table:

| | no medicine | medicine 1 | medicine 2 | totals |
|---|---|---|---|---|
| **high** | 10 | 35 | 25 | 70 |
| **low** | 5 | 10 | 15 | 30 |
| **totals** | 15 | 45 | 40 | 100 |

- By the theorem we can learn row-by-row, or disintegrate the joint distribution learned from the entire table
- The extracted channel $c \colon B \to \mathcal{D}(H)$ is thus:

$$c(H) = \tfrac{1}{7}|0\rangle + \tfrac{1}{2}|1\rangle + \tfrac{5}{14}|2\rangle$$
$$c(L) = \tfrac{1}{6}|0\rangle + \tfrac{1}{3}|1\rangle + \tfrac{1}{2}|2\rangle.$$

They give the conditional proabilities $P(M \mid B)$



## Learning Bayesian network parameters

- Assume you already know the graph $G$ of the Bayesian network
- and you have a table of data — for all the nodes of the graph — in a multi-dimensional table, in $\mathcal{M}(X_1 \times \cdots \times X_n)$
- Then you can:
  (1) "reshape" the table so that it fits the graph $G$, via appropriate marginalisations, and row (or column) extractions
  (2) learn the Conditional Probability Tables locally, from this reshaped table
  (3) extend $G$ to a Bayesian network.



## Background: frequentist learning as *maximal* likelihood

- Consider $[0,1]^{\mathcal{D}(X)}$, the set of predicates on $\mathcal{D}(X)$
  - it's a commutative monoid, via $\mathbf{1}$ and & (pointwise mult.)
  - there is a 'point validity' map $\mathrm{pv} \colon X \to [0,1]^{\mathcal{D}(X)}$
  - It is $\mathrm{pv}(x)(\omega) = \omega \models \mathbf{1}_x = \omega(x)$
- By freeness we get $\overline{\mathrm{pv}} \colon \mathcal{M}(X) \to [0,1]^{\mathcal{D}(X)}$
  - it sends data $\sum_i n_i |x_i\rangle$ to the predicate $\omega \mapsto \prod_i \omega(x_i)^{n_i}$
- **Theorem.** For fixed $\varphi \in \mathcal{M}(X)$, the learned distribution $\mathrm{flrn}(\varphi) \in \mathcal{D}(X)$ gives the maximum for the predicate:

$$\mathcal{D}(X) \xrightarrow{\overline{\mathrm{pv}}(\varphi)} [0,1]$$

**Thus**: $\mathrm{flrn}(\varphi) \in \mathcal{D}(X)$ best fits the data $\varphi \in \mathcal{M}(X)$



## Disadvantages of frequentist learning

- Variance is not taken into account: you learn the same coin bias from
  - 3 heads out of 5 throws
  - 3000 heads out of 5000 throws
- Prior knowledge, if any, is not taken into account

## Where we are, so far





## The Giry monad $\mathcal{G}$

- Defined in general, on a measurable space $X = (X, \Sigma)$ as:

$$\mathcal{G}(X) \coloneqq \{\omega \colon \Sigma \to [0,1] \mid \omega \text{ is a probability measure}\}$$

- In practice $\omega \in \mathcal{G}(X)$ is often given by a pdf $f \colon X \to \mathbb{R}_{\geq 0}$ as:

$$\omega(M) = \int_M f(\vec{x}) d\vec{x}$$

where $X \subseteq \mathbb{R}^n$. This suffices here.





## Bayesian learning

- Instead of *calculating* a distribution in $\mathcal{D}(X)$, we now look at distributions over $\mathcal{D}(X)$
- We restrict to finite sets, actually to $n = \{1, \ldots, n\}$
  - then $\mathcal{D}(n) \hookrightarrow \mathbb{R}^n_{\geq 0}$
  - subset of $(x_1, \ldots, x_n)$ with $\sum_i x_i = 1$
  - actually we only use $x_i > 0$, for all $i$, so $\mathcal{D}_{\circledast}(n)$
- Data $\sum_i k_i | \, i \, \rangle \in \mathcal{M}_{\circledast}(n)$, with each $k_i > 0$
  - $\mathcal{M}(h) \colon \mathcal{M}_{\circledast}(n) \to \mathcal{M}_{\circledast}(m)$ is only well-defined for surjective $h \colon n \to m$





## Dirichlet

The aim is to get:

$$\begin{array}{ccc} \mathcal{M}_{\circledast}(n) & \xRightarrow{\quad \mathrm{Dir}_n \quad} & \mathcal{G}\big(\mathcal{D}_{\circledast}(n)\big) \\ \vec{k} & \longmapsto & \Big(M \mapsto \displaystyle\int_M d_n(\vec{k})(\vec{x}) \, \mathrm{d}\vec{x}\Big) \end{array}$$

where $d_n(\vec{k}) \colon \mathcal{D}_{\circledast}(n) \to \mathbb{R}_{\geq 0}$ is the Dirichlet pdf:

$$d_n(k_1, \ldots, k_n)(x_1, \ldots, x_n) \coloneqq \frac{\Gamma(\sum_i k_i)}{\prod_i \Gamma(k_i)} \cdot \prod_i x_i^{k_i - 1}.$$

with $\Gamma(k) = (k-1)!$





## Naturality of Dirichlet

**Theorem**

Dir: $\mathcal{M}_{\circledast} \Longrightarrow \mathcal{GD}_{\circledast}$ *is natural in* $n$, *for surjective functions*

**Proof** There is a known aggregation property of dirichlet:

$$\begin{aligned} & d_{n-1}(k_1 + k_2, k_3, \ldots, k_n)(x_2, x_3, \ldots, x_n) \\ &= \int_{y \in (0, x_2)} d_n(k_1, k_2, k_3, \ldots, k_n)(y, x_2 - y, x_3, \ldots, x_n)\, \mathrm{d}y. \end{aligned}$$

It can be "pimped" to: for surjective $h\colon n \twoheadrightarrow m$,

$$d_m\Big(\mathcal{M}_{\circledast}(h)(\vec{k})\Big)(\vec{x}) = \int_{\vec{y} \in \mathcal{D}_{\circledast}(h)^{-1}(\vec{x})} d_n(\vec{k})(\vec{y})\, \mathrm{d}\vec{y}.$$

This gives naturality of Dir: $\mathcal{M}_{\circledast} \Longrightarrow \mathcal{GD}_{\circledast}$.



## Compositionality and locality properties

- The (output) type $\mathcal{GD}$ of the dirichlet map makes compositionality more difficult
  - in order to learn from a multi-dimensional table in separate steps
- How this is done in practice is not so clear.



## Conjugate prior and updating

- Calculating (eg. conditioning) with complicated distributions like Dirichlet is a pain
- Instead, one wants to calculate via the "hyper" parameters $\vec{k}$
  - this is the essence of conjugate prior properties
  - for categorical description: **arXiv:1707.00269**
- Here we give a logical formulation



## Predicates, for relating frequentist & Bayesian

- For $\omega \in \mathcal{D}(n)$ and $p \in [0,1]^n$ we have validity $\omega \models p := \sum_x \omega(x) \cdot p(x) \in [0,1]$
- Thus $\widehat{p} := (-) \models p$ is a predicate $\mathcal{D}(n) \to [0,1]$
  - hence we can take $\sigma \models \widehat{p}$ for $\sigma \in \mathcal{G}(\mathcal{D}(n))$

**Theorem**

*For* $p \in [0,1]^n$ *the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{M}_{\circledast}(n) & \xrightarrow{\mathsf{flrn}_n} & \mathcal{D}_{\circledast}(n) \\ {\scriptstyle \mathsf{Dir}_n}\downarrow & & \downarrow{\scriptstyle (-) \models p} \\ \mathcal{G}(\mathcal{D}_{\circledast}(n)) & \xrightarrow[(-) \models \widehat{p}]{} & [0,1] \end{array}$$

## Incremental Bayesian learning

- Suppose we have data $\vec{k} \in \mathcal{M}_{\circledast}(n)$ and associated $\mathrm{Dir}(\vec{k})$
- We then find one new data item, at position $|\, i \,\rangle$
  - we can change the data to $\vec{k} +\!\!+ i = (k_1, \ldots, k_i + 1, \ldots, k_n)$, and then apply Dir again
  - We can also update $\mathrm{Dir}(\vec{k})$ with the singleton predicate $\mathbf{1}_i$ (actually with $\widehat{\mathbf{1}_i}$)

**Theorem**

*The last two bullets give the same outcome:*

$$\mathrm{Dir}(\vec{k} +\!\!+ i) = \mathrm{Dir}(\vec{k})\big|_{\widehat{\mathbf{1}_i}}$$





## Where we are, so far



iCIS | Digital Security
Radboud University

## Concluding remarks

- There is remarkably much categorical structure in learning

$$\mathcal{M} \xRightarrow{\text{frequentist}} \mathcal{D} \qquad \text{and} \qquad \mathcal{M} \xRightarrow{\text{Bayesian}} \mathcal{GD}.$$

- This work uncovers some basic/deep relations between
  (1) multiset $\mathcal{M}$ on the one hand
  (2) distribution $\mathcal{D}$ and Giry $\mathcal{G}$ on the other hand
- Further applications lay ahead
  - latent Dirichlet allocation for document classification
  - ...
- For more details, see paper: **arXiv:1810.05814**



iCIS | Digital Security
Radboud University