Probabilistic morphisms and Bayesian statistical models

Hông Vân Lê
Institute of Mathematics, CAS, Praha
partly based on a joint work with Jürgen Jost, Duc Hoang Luu and Tat Dat Tran

Categorical Probability and Statistics,
Ottawa June 2020

OUTLINE

1) Probabilistic morphisms.

2) Revisiting Dirichlet measures using probabilistic morphisms.

3) Revisiting posterior distribution.

4) Bayesian statistical models and diffeological statistical models.

1. Probabilistic morphisms

• 1962: Lawvere proposed a categorical approach to probability theory. His new concepts: probabilistic mappings (morphisms), the canonical σ-algebra on the space $\mathcal{P}(\mathcal{X})$ of probability measures on a measurable space $\mathcal{X}$.

• $\mathcal{F}_s(\mathcal{X})$ - the set of simple functions on $\mathcal{X}$.
• $\mathcal{S}(\mathcal{X})$ ($\mathcal{M}(\mathcal{X})$, $\mathcal{P}(\mathcal{X})$) - the set of signed finite measures (resp. nonnegative finite and probability measures on $\mathcal{X}$).

• $I : \mathcal{F}_s(\mathcal{X}) \to \mathcal{S}^*(\mathcal{X})$, $f \mapsto I_f$, $I_f(\mu) := \int_{\mathcal{X}} f d\mu$.
• Σ_w - the smallest σ-algebra on $\mathcal{S}(\mathcal{X})$ (resp. on $\mathcal{M}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$) s.t. I_f is measurable for all $f \in \mathcal{F}_s(\mathcal{X})$.

• For a topological space $\mathcal{X}$ we consider the Borel σ-algebra $\mathcal{B}(\mathcal{X})$, the set $C_b(\mathcal{X})$ of bounded continuous functions on $\mathcal{X}$, and the smallest topology τ_v on $\mathcal{S}(\mathcal{X})$ (resp. $\mathcal{M}(\mathcal{X})$, $\mathcal{P}(\mathcal{X})$) s.t. for any $f \in C_b(\mathcal{X})$ the map $I_f : (\mathcal{S}(\mathcal{X}), \tau_v) \to \mathbf{R}$ is continuous. If $\mathcal{X}$ is separable and metrizable then the Borel σ-algebra on $\mathcal{M}(\mathcal{X})$ generated by τ_v coincides with Σ_w.

Proposition 1.(JLLT2019) (1) Assume that $\Sigma_{\mathcal{X}}$ has a countable generating algebra $\mathcal{A}_{\mathcal{X}}$. Then $\mathcal{M}(\mathcal{X})$ is a measurable subset of $\mathcal{S}(\mathcal{X})$, and $\mathcal{P}(\mathcal{X})$, $\mathcal{M}^*(\mathcal{X}) := \mathcal{M}(\mathcal{X})\backslash\{0\}$ are measurable subsets of $\mathcal{M}(\mathcal{X})$.

(2) $\mathfrak{a} : (\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}), \Sigma_w \otimes \Sigma_w) \to (\mathcal{M}(\mathcal{X}), \Sigma_w)$ $(\mu, \nu) \mapsto \mu + \nu$, is a measurable map. If $\mathcal{X}$ is a topological space, then the addition $\mathfrak{a}$ is τ_v-continuous.

• A probabilistic morphism $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ from a measurable space $\mathcal{X}$ to a measurable space $\mathcal{Y}$ is an arrow associated to a measurable mapping $\overline{T} : \mathcal{X} \to (\mathcal{P}(\mathcal{Y}), \Sigma_w)$. We say that T is generated by $\overline{T}$.

• $Id_{\mathcal{P}} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ generates $ev : \mathcal{P}(\mathcal{X}) \rightsquigarrow \mathcal{X}$, so $\overline{ev} := Id_{\mathcal{P}}$.

2. The map $\delta : \mathcal{X} \to (\mathcal{P}(\mathcal{X}), \Sigma_w)$, $x \mapsto \delta_x$, is measurable. Hence any measurable mapping $\kappa : \mathcal{X} \to \mathcal{Y}$ defines a probabilistic morphism, also denoted by κ, with $\overline{\kappa} := \delta \circ \kappa$.

$T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ induces $S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$:

$$S_*(T)(\mu)(B) := \int_{\mathcal{X}} \overline{T}(x)(B) d\mu(x).$$

- $M_* := S_*(T)|_{\mathcal{M}(\mathcal{X})}$, $P_* := S(T)|_{\mathcal{P}(\mathcal{X})}$.

Theorem 1. Let T be a probabilistic morphism.
- $S_*(T)$ is a linear bounded map wrt $\|\cdot\|_{TV}$ [Chentsov1972].
- M_* and P_* are faithful functors [JLLT2019].
- If $\nu \ll \mu \in \mathcal{M}^*(\mathcal{X})$ then $M_*(T)(\nu) \ll M_*(T)(\mu)$ [MS1966].

• Lawvere introduced the σ-algebra Σ_w and defined ev in terms of Markov kernels.

• Giry proved that the triple $(P_*, \delta, ev_{\mathcal{P}})$ is a (Giry's) monad.

• Chentsov called the category of Markov kernels the statistical category and their morphisms Markov morphisms. Chentsov also showed that $S_*(T)$ sends a probability measure to a probability measure.

2. Rivisiting Dirichlet measures using probabilistic morphisms

- $\Omega_k := \{\omega_1, \cdots, \omega_k\}$. Then

$(\mathcal{S}(\Omega_k), \tau_v) \sim \mathbf{R}^k$, $(\mathcal{M}(\Omega_k), \tau_v) \sim \mathbf{R}^n_{\geq 0}$, $(\mathcal{P}(\Omega_k), \tau_v) \sim \Delta_k := \{(x_1, \cdots, x_k) \in \mathbf{R}^k_{\geq 0} | \sum_{i=1}^k x_i = 1\}$.

- For $\alpha \in \mathcal{M}^*(\Omega_k)$ set

- $\Omega(\alpha) := \{\omega_j \in \Omega_k | \alpha(\omega_j) \neq 0\}$.

- $\pi_\alpha : \Omega(\alpha) \to \Omega_k$ - the natural inclusion.

- $l(\alpha) := \#\Omega(\alpha)$.

• The Dirichlet distribution $Dir(\alpha) \in \mathcal{P}^2(\Omega_k)$ - $P_*^2(\pi_\alpha) Dir(\alpha|_{\Omega(\alpha)})$, where $Dir(\alpha|_{\Omega(\alpha)})$ is the classical $(l(\alpha)-1)$-dimensional Dirichlet distribution on $\Delta_{l(\alpha)}$

$$Dir(x_{i_1}, \cdots, x_{i_{l(\alpha)}}) = \frac{\Gamma(\alpha_{i_1} + \cdots + \alpha_{i_{l(\alpha)}})}{\Gamma(\alpha_{i_1}) \cdots \Gamma(\alpha_{i_{l(\alpha)}})} \prod_{j=1}^{l(\alpha)} x_{i_j}^{\alpha_{i_j}-1}.$$

- $\mathcal{X}$ - a measurable space, $\alpha \in \mathcal{M}^*(\mathcal{X})$.

- $\mathcal{D}(\alpha) \in \mathcal{P}^2(\mathcal{X})$ is called a Dirichlet measure on $\mathcal{P}(\mathcal{X})$ parameterized by α, if for all surjective measurable mappings $\pi_k : \mathcal{X} \to \Omega_k$ we have

$$P^2_*(\pi_k)(\mathcal{D}(\alpha)) =$$

$$Dir(\alpha(\pi_k^{-1}(\omega_1), \cdots, \alpha(\pi_k^{-1}(\omega_k)) \in \mathcal{P}^2(\Omega_k),$$

where $Dir(\alpha_1, \cdots, \alpha_k)$ is the Dirichlet distribution with the parameter $(\alpha_1, \cdots, \alpha_k)$ on Δ_k. If $\mathcal{D}(\alpha)$ is defined for all $\alpha \in \mathcal{M}^*$ we shall call $\mathcal{D}$ a Dirichlet map.

Recall that $M_* : \mathcal{X} \mapsto \mathcal{M}^*(\mathcal{X})$ and $P_*^2 : \mathcal{X} \mapsto \mathcal{P}^2(\mathcal{X})$ are functors from the category of measurable spaces with measurable mappings.

Theorem 4. (JLLT2019) For any measurable space $\mathcal{X}$ there exists a measurable mapping $\mathcal{D} : \mathcal{M}^*(\mathcal{X}) \to \mathcal{P}^2(\mathcal{X})$ such that $\mathcal{D}(\alpha)$ is a Dirichlet measure parameterized by α. Moreover $\mathcal{D} : M_* \to P_*^2$ is a natural transformation.

3. Revisiting posterior distribution

• Bayesian statistical model:= $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$, where $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$ is a measurable mapping and $\mu_\Theta \in \mathcal{P}(\Theta)$ - a prior measure.

Example. Assume that (M, μ_M) is a smooth finite dimensional manifold endowed with a volume element μ_M and $\mathbf{p} : M \to (\mathcal{P}(\mathcal{X}), d_H)$ is a continuous map, for instance, if $\mathbf{p}$ is a smooth map, i.e., the composition $i \circ \mathbf{p} : M \to \mathcal{S}(\mathcal{X})_{TV}$ is a smooth map, then $(M, \mu_M, \mathbf{p}, \mathcal{X})$ is a Bayesian statistical model.

- $\Pi_{\mathcal{X}} : (\Theta \times \mathcal{X}, \Sigma_\Theta \otimes \Sigma_{\mathcal{X}}) \to (\mathcal{X}, \Sigma_{\mathcal{X}})$.

- A family of probability measures $\mu_{\Theta|\mathcal{X}}(\cdot|x) \in \mathcal{P}(\Theta)$, $x \in \mathcal{X}$, is called a family of posterior distributions of μ_Θ after seeing the data x if for all $B \in \Sigma_\Theta$ we have

$$\mu_{\Theta|\mathcal{X}}(B|x) = \frac{d(\Pi_{\mathcal{X}})_*(1_{B\times\mathcal{X}}\mu)}{d(\Pi_{\mathcal{X}})_*\mu}(x) \in L^1(\mathcal{X}, \mu_{\mathcal{X}}),$$

Proposition 2. (JLLT2019) Given a Bayesian statistical model $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ a family of posterior distributions $\mu_{\Theta|\mathcal{X}}(\cdot|x)$ exists if and only if the statistical models (Θ, μ_Θ) and $(\mathcal{X}, \mu_\mathcal{X})$ are equivalent, i.e. $\mu_\mathcal{X} = \mathbf{p}_*(\mu_\Theta) \in \mathcal{P}(\mathcal{X})$, where $\mu_\mathcal{X}$ is the marginal distribution of $\mathcal{X}$.

• A Souslin space is a Hausdorff space admitting a surjective continuous mapping from a complete metrizable space. In particular, every Polish space (a complete separable metrizable space) is a Souslin space, and more general, every standard Borel space (a measurable space admitting a bijective, bimeasurable correspondence with a Borel subset of a Polish space) is a Souslin space.

Theorem 3. Let $\mathcal{X} \subset (M, g)$ with the induced metric and $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ - a Bayesian statistical model, where Θ is a Souslin space. Let $D_r(x)$ denote the open ball of radius r centered at $x \in \mathcal{X}$. Then there exist a measurable subset $S \subset \mathcal{X}$ of zero $\mu_\mathcal{X}$-measure, and family of posterior distributions $\mu_{\Theta|\mathcal{X}}(\cdot|x)$ on Θ after seeing data $x \in \mathcal{X}$ such that

$$\mu_{\Theta|\mathcal{X}}(B|x) = \lim_{r \to 0} \frac{\int_B \mathbf{p}_\theta(D_r(x)) d\mu_\Theta}{\int_\Theta \mathbf{p}_\theta(D_r(x)) d\mu_\Theta} \qquad (1)$$

for any $B \in \mathcal{B}(\Theta) = \Sigma_\Theta$, and for any $x \in \mathcal{X} \setminus S$. For $x \in S$ we have $\mu_{\Theta|\mathcal{X}}(B|x) := 0$ for any $B \in \Sigma_\Theta$.

Corollary 2. Assume the conditions of Theorem 2. Given a point $x_0 \in \mathcal{X} \setminus S$ assume that $p_\theta(x_0) = 0$ for all $\theta \in \Theta$. If the condition for differentiation w.r.t. r at 0 under the integral $\int_C \mathbf{p}_\theta(D_r(x_0))d\mu_\Theta$ holds for $C \in \mathcal{U}_\Theta \cup \{\Theta\}$, then for any $B \in \mathcal{U}_\Theta$ we have

$$\mu_{\Theta|\mathcal{X}}(B|x_0) = \frac{\int_B \frac{d}{dr}|_{r=0}\mathbf{p}_\theta(D_r(x_0))d\mu_\Theta}{\int_\Theta \frac{d}{dr}|_{r=0}\mathbf{p}_\theta(D_r(x_0))d\mu_\Theta}, \quad (2)$$

if the dominator in the RHS of (2) does not vanish.

4. Bayesian statistical models and diffeological statistical models

- A statistical model is a subset $P_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$.
- A parameterized statistical model is a triple $(\Theta, \mathbf{p}, \mathcal{X})$, $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$ (classical definition).
- A parameterized statistical model is a triple $(M, \mathcal{X}, \mathbf{p})$ where M is a Banach manifold, $\mathcal{X}$ is a measurable space, and $i \circ \mathbf{p} : M \xrightarrow{\mathbf{p}} \mathcal{P}(\mathcal{X}) \xrightarrow{i} \mathcal{S}(\mathcal{X})$ is a C^1-map [AJLS, 2015-2017].

The last concept of parameterized statistical models aims to give a framework for applying differential geometric methods to statistics.

• Assume that (M, μ_M) is a smooth finite dimensional manifold endowed with a volume element μ_M and $(M, \mathcal{X}, \mathbf{p})$ is a parameterized statistical model. Then it has been shown in [JLLT2019] that $(M, \mu_M, \mathbf{p}, \mathcal{X})$ is a Bayesian statistical model, i.e., the map $\mathbf{p} : (M, \mathcal{B}(M)) \to (\mathcal{P}(\mathcal{X}), \Sigma_w)$ is measurable.

• In the definition of Bayesian statistical models $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ we don't assume that Θ has a manifold structure and therefore $(\Theta, \mathbf{p}, \mathcal{X})$ is not a (AJLS) parameterized statistical model and $\mathbf{p}(\Theta)$ is simply a statistical model.

• A statistical model is a pair $(\mathcal{X}, P_{\mathcal{X}})$ where $\mathcal{X}$ is a measurable space and $P_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$. The category of statistical models consists of statistical models as its objects whose morphisms $\varphi : (\mathcal{X}, P_{\mathcal{X}}) \rightsquigarrow (\mathcal{Y}, P_{\mathcal{Y}})$ are probabilistic morphisms $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ such that $T_*(P_{\mathcal{X}}) \subset P_{\mathcal{Y}}$. A morphism $T : (\mathcal{X}, P_{\mathcal{X}}) \rightsquigarrow (\mathcal{X}, P_{\mathcal{X}})$ will be called a unit, if $T_* : P_{\mathcal{X}} \to P_{\mathcal{X}}$ is the identity.

Two statistical models $(\mathcal{X}, P_{\mathcal{X}})$ and $(\mathcal{Y}, P_{\mathcal{Y}})$ are called equivalent, if there exist morphisms $T_1 : (\mathcal{X}, P_{\mathcal{X}}) \rightsquigarrow (\mathcal{Y}, P_{\mathcal{Y}})$ and $T_2 : (\mathcal{Y}, P_{\mathcal{Y}}) \rightsquigarrow (\mathcal{X}, P_{\mathcal{X}})$ such that $T_1 \circ T_2$ and $T_2 \circ T_1$ are units. In this case T_1 and T_2 will be called equivalences.

A morphism $T : (\mathcal{X}, P_{\mathcal{X}}) \rightsquigarrow (\mathcal{Y}, P_{\mathcal{Y}})$ will be called sufficient if there exists $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ s.t. for all $\mu \in P_{\mathcal{X}}$ and $h \in L(\mathcal{X})$ we have

$$T_*(h\mu) = \underline{\mathbf{p}}^*(h) T_*(\mu)$$

$$\iff \underline{\mathbf{p}}^*(h) = \frac{dT_*(h\mu)}{dT_*(\mu)} \in L^1(\mathcal{Y}, T_*(\mu)).$$

In this case we shall call $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ a probabilistic morphism sufficient for $\mathcal{P}_{\mathcal{X}}$ and we shall call the measurable mapping $\mathbf{p} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ defining the probabilistic morphism $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ a conditional mapping for T.

Definition For $k \in \mathbf{N}^+ \cup \infty$ and a nonempty set X, a C^k-diffeology of X is a set $\mathcal{D}$ of mappings $\mathbf{p} : U \to X$, where U is an open domain in $\mathbf{R}^n$, and n runs over nonnegative integers, such that the three following axioms are satisfied.

D1. Covering. The set $\mathcal{D}$ contains the constant mappings $\mathbf{x} : r \mapsto x$, defined on $\mathbf{R}^n$, for all $x \in X$ and for all $n \in \mathbf{N}$.

D2. Locality. Let $\mathbf{p} : U \to X$ be a mapping. If for every point $r \in U$ there exists an open

neighborhood V of r such that $\mathbf{p}_{|V}$ belongs to $\mathcal{D}$ then the map $\mathbf{p}$ belongs to $\mathcal{D}$.

D3. Smooth compatibility. For every element $\mathbf{p} : U \to X$ of $\mathcal{D}$, for every real domain V, for every $\psi \in C^k(V, U)$, $\mathbf{p} \circ \psi$ belongs to $\mathcal{D}$.

A C^k-diffeological space is a nonempty set equipped with a C^k-diffeology $\mathcal{D}$. Elements $\mathbf{p} : U \to X$ of $\mathcal{D}$ will be called C^k-maps from U to X.

A statistical model $P_{\mathcal{X}}$ endowed with a C^k-diffeology $\mathcal{D}_{\mathcal{X}}$ will be called a C^k-diffeological

statistical model, if for any map $\mathbf{p} : U \to P_{\mathcal{X}}$ in $\mathcal{D}_{\mathcal{X}}$ the composition $i \circ \mathbf{p} : U \to \mathcal{S}(\mathcal{X})$ is a C^k-map (Le2020).

• C^k-diffeological statistical models are invariant under probabilistic morphisms (Le2020).

• Diffeological Fisher metric (Fisher metric on diffeological statistical models) are invariant under sufficient probabilistic morphisms (Le2020).

- J. Jost, H. V. Lê, , D. H. Luu and T. D. Tran, Probabilistic mappings and Bayesian nonparametrics, arXiv:1905.11448.

- H. V. Lê, Diffeological statistical models, the Fisher metric and probabilistic mappings, arXiv:1912.02090, Mathematics 2020, 8(2), 167.

Thank you for your attention!