GSI 2021

Statistical Bundle of the Transport Model

Giovanni Pistone

DE CASTRO STATISTICS

DISMA Dipartimento di Scienze Matematiche G. L. Lagrange
ECCELLENZA 2018 · 2022

Paris, July 23 2021

web-page: www.giannidiorestino.it
e-mail: giovanni.pistone@carloalberto.org

Plan

- Affine space, exponential statistical bundle
- ANOVA: simple effect, interaction
- Transportation model
- Natural gradient flow in Optimal Transport

1921 H. Weyl. *Space- time- matter / by Hermann Weyl*. Dover, New York, 1952. translation of the 1921 RAUM ZEIT MATERIE

1942 L. Kantorovitch. On the translocation of masses. *Management Sci.*, 5:1–4, 1958

1980 S.E. Fienberg. *The Analysis of Cross-classified Categorical Data* 2nd Edition. M.I.T. Press, Cambridge, MA. 1980

1985 S. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985

1994 K. Nomizu and T. Sasaki. *Affine differential geometry: geometry of affine immersions*. Number 111 in Cambridge Tracts in Mathematics. Cambridge University Press, 1994

2019 G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. arXiv:1803.00567v2

Affine space

Given a set M and a real finite dimensional vector space V, Hermann Weyl considers a displacement or difference mapping

$$M \times M \ni (p, q) \mapsto \overrightarrow{pq} \in V \ ,$$

such that

- for each p the mapping $s_p \colon M \ni q \mapsto \overrightarrow{pq}$ is 1-to-1 and unto, and
- the parallelogram law, $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$, holds.

In particular, it follows $\overrightarrow{pq} + \overrightarrow{qp} = \overrightarrow{pp} = 0$.

Weyl's affine space

The structure $(M, V, \overrightarrow{\cdot\cdot})$ is an affine space. The atlas of charts $s_p \colon M \to V$, $p \in M$, defines an affine manifold. All transition mappings of the atlas are vector translations.

$$s_q \circ s_p^{-1} \colon v \mapsto s_q(p) + v$$

(p_1, q_1) is parallel to (p_2, q_2) if $s_{p_1}(q_1) = s_{p_2}(q_2)$.

Affine space with a parallel transport

In IG, it is more intrinsic and more natural to allow general parallel transports and rephrase Weyls's definition. That is, we expand the trivial bundle $M \times V$ into a nontrivial bundle.

Affine space

Let M be a set and B_μ, $\mu \in M$, a family of topological vector spaces (toplinear spaces). Let $(\mathbb{U}_\nu^\mu)$, $\nu, \mu \in M$ be a family of toplinear isomorphism $\mathbb{U}_\nu^\mu \colon B_\nu \to B_\mu$, $\mathbb{U}_\mu^\nu \mathbb{U}_\nu^\mu = I$. Define the difference mapping

$$\mathbb{S} \colon M \times M \ni (\nu, \mu) \mapsto s_\nu(\mu) \in B_\nu$$

so that:

1. For each fixed ν the mapping $\mu \mapsto s_\nu(\mu) = \mathbb{S}(\nu, \mu)$ is injective
2. Parallelogram law: $\mathbb{S}(\mu_1, \mu_2) + \mathbb{U}_{\mu_2}^{\mu_1} \mathbb{S}(\mu_2, \mu_3) = \mathbb{S}(\mu_1, \mu_3)$

Parallelogram law with $\mu_1 = \mu_3 = \nu$ and $\mu_2 = \mu$ becomes

$$\mathbb{S}(\nu, \mu) + \mathbb{U}_\mu^\nu \mathbb{S}(\mu, \nu) = 0 \ .$$

Affine manifold

Let us compute, where defined, the change-of-origin map $s_\mu \circ s_\nu^{-1}$ in an affine space. At $\rho = s_\nu^{-1}(w)$, $w \in B_\nu$, it holds

$$s_\mu \circ s_\nu^{-1}(w) = s_\mu(\rho) = s_\mu(\nu) + \mathbb{U}_\nu^\mu s_\nu(\rho) = s_\mu(\nu) + \mathbb{U}_\nu^\mu w \ .$$

The change-of-origin map extends to an affine map.

The affine space provides a family of charts $s_\nu \colon M \to B_\nu$, $\nu \in M$, that we want to use as an atlas.

Affine manifold

Assume that the vector fibers of the affine space $(M, (B_\mu), \mathbb{S})$ are Banach spaces and assume that for each ν, $s_\nu M$ is a neighborhood of 0 in B_μ. Define $U_\nu = s_\nu^{-1}\left(s_\nu(M)^\circ\right)$. Then $(S\nu, U, \nu)$ is a chart on M. The charts are compatible and the resulting manifold is the affine manifold of the affine space.

Affine bundle

The specific form of the atlas defining the affine manifold allows the extension of the same atlas to define an affine bundle.

Affine bundle

Given the affine manifold $\mathcal{M}$, consider the set

$$\{(\mu, v) | \mu \in M, v \in B_\mu\} \tag{1}$$

and, for each ν define the chart

$$s_\nu(\mu, v) = (s_\nu(\mu), \mathbb{U}_\mu^\nu v) \in B_\nu \times B_\nu \tag{2}$$

to define the manifold $S\,\mathcal{M}$. Equivalently, we can say that $S\,\mathcal{M}$ is a linear bundle with trivialization

$$s_\nu \colon (\mu, v) \mapsto (s_\nu(\mu), \mathbb{U}_\mu^\nu v) \ . \tag{3}$$

Kinematics

Velocity

In an affine manifold, the velocity of a smooth curve $t \mapsto \gamma(t)$ is the section $t \mapsto (\gamma(t), \overset{\star}{\gamma}(t)) \in S\,\mathcal{M}$ defined by

$$\overset{\star}{\gamma}(t) = \lim_{h \to 0} h^{-1}(s_{\gamma(t)}(\gamma(t+h)) - s_{\gamma(t)}(\gamma(t)))\ . \tag{4}$$

Consider a duality pairing on the fibers of the affine bundle and define the dual bundle in the standard way.

Gradient

If $\phi \colon M \to \mathbb{R}$, the natural gradient $\operatorname{grad} \phi$ is defined on M with values in the dual fibers and such that for each smooth curve γ

$$\frac{d}{dt}\phi(\gamma(t)) = \langle \operatorname{grad} \phi(\gamma(t)), \overset{\star}{\gamma}(t) \rangle_{\gamma(t)}$$

Exponential affine space

Now M is the convex set of positive probability functions of a finite sample space. The difference mapping is

$$\mathbb{S} \colon (p, q) \mapsto \log \frac{q}{p} - \mathbb{E}_p \left[\log \frac{q}{p} \right] \in B_p = L_0^2(p)$$

The transports and the dual transports are

$${}^{\mathrm{e}}\mathbb{U}_p^q \colon u \mapsto u - \mathbb{E}_q [u] \ , \quad {}^{\mathrm{m}}\mathbb{U}_q^p \colon v \mapsto \frac{q}{p} v$$

The parallelogram law holds,

$$\log \frac{q}{p} - \mathbb{E}_p \left[\log \frac{q}{p} \right] + \mathbb{U}_q^p \left(\log \frac{r}{q} - \mathbb{E}_q \left[\log \frac{r}{q} \right] \right) = \log \frac{r}{p} - \mathbb{E}_p \left[\log \frac{r}{p} \right]$$

The inverse charts are

$$s_p^{-1} \colon u \mapsto \mathrm{e}^{u - K_p(u)} \ , \quad K_p(u) = \mathbb{E}_p \left[\mathrm{e}^u \right] \ .$$

For each smooth curve $t \mapsto \gamma(t) \in \Delta^\circ(\Omega)$ the velocity (or score) is $\overset{\star}{\gamma}(t) = \frac{d}{dt} \log \gamma(t)$.

Evolution equation

In the duality we have

$$\frac{d}{dt}\mathbb{E}_{\gamma(t)}[u] = \left\langle u - \mathbb{E}_{\gamma(t)}[u], \overset{\star}{\gamma}(t)\right\rangle_{\gamma(t)} \quad . \tag{5}$$

The mapping $\gamma \mapsto u - \mathbb{E}_\gamma[u]$ is the gradient mapping of $\gamma \mapsto \mathbb{E}_\gamma[u]$. It is a section of the statistical bundle.

For each section F we define the evolution equation $\overset{\star}{\gamma} = F(\gamma)$. By writing $\overset{\star}{\gamma} = \dot{\gamma}/\gamma$, we see that the evolution equation in our sense is equivalent to the Ordinary Differential Equation (ODE) $\dot{\gamma} = \gamma F(\gamma)$.

Given a sub-manifold of $\Delta^\circ(\Omega)$, each fiber S_γ of the statistical bundle splits to define the proper sub-statistical bundle.

Transport model

Given positive probability functions $\mu_1 \in \Delta^\circ(\Omega_1)$ and $\mu_2 \in \Delta^\circ(\Omega_2)$, the transport model with margins μ_1 and μ_2 is the statistical model

$$\Gamma(\mu_1, \mu_2) = \{\gamma \in \Delta(\Omega) | \gamma(\cdot, +) = \mu_1, \gamma(+, \cdot) = \mu_2\} \ .$$

Our sub-manifold is the open transport model

$$\Gamma^\circ(\mu_1, \mu_2) = \{\gamma \in \Delta^\circ(\Omega) | \gamma(\cdot, +) = \mu_1, \gamma(+, \cdot) = \mu_2\} \ .$$

Tangent vectors

If $t \mapsto \gamma(t)$ is a smooth curve in the open transport model, then

$$0 = \frac{d}{dt}\mathbb{E}_{\mu_1}[f] = \frac{d}{dt}\mathbb{E}_{\gamma(t)}[f(X)] = \left\langle f(X), \overset{\star}{\gamma}(t) \right\rangle_{\gamma(t)} = \left\langle f, \overset{\star}{\gamma}(t)_1 \right\rangle_{\mu_1} \ ,$$

with $\overset{\star}{\gamma}(t)_1(X) = \mathbb{E}_{\gamma(t)}[\overset{\star}{\gamma}(t)|X]$. Similarly on the other projection. It follows that

$$\mathbb{E}_{\gamma(t)}[\overset{\star}{\gamma}(t)|X] = 0 \quad \text{and} \quad \mathbb{E}_{\gamma(t)}[\overset{\star}{\gamma}(t)|Y] = 0$$

ANOVA: definition

Effects

The linear sub-spaces of $L^2(\gamma)$ which , respectively, express the γ-grand-mean, the two γ-simple effects, and the γ-interactions, are

$$
\begin{aligned}
B_0(\gamma) &\sim \mathbb{R}, \\
B_1(\gamma) &= \left\{ f \circ X \middle| f \in L^2_0(\gamma_1) \right\}, \\
B_2(\gamma) &= \left\{ f \circ Y \middle| f \in L^2_0(\gamma_2) \right\}, \\
B_{12}(\gamma) &= \left(B_0(\gamma) + B_1(\gamma) + B_2(\gamma) \right)^{\perp},
\end{aligned}
\tag{6}
$$

where the orthogonality is computed in the γ weight, that is in the inner product of $L^2(\gamma)$, $\langle u, v \rangle_\gamma = \mathbb{E}_\gamma [uv]$.

Each element of the space $B_0(\gamma) + B_1(\gamma) + B_2(\gamma)$ has the form $u = u_0 + f_1(X) + f_2(Y)$, where $u_0 = \mathbb{E}_\gamma [u]$ and f_1, f_2 are uniquely defined.

ANOVA: conditions

Conditions

For each $\gamma \in \Delta(\Omega)$ there exist a unique orthogonal splitting

$$L^2(\gamma) = \mathbb{R} \oplus (B_1(\gamma) + B_2(\gamma)) \oplus B_{12}(\gamma) \ .$$

Namely, each $u \in L^2(\gamma)$ can be written uniquely as

$$u = u_0 + (u_1 + u_2) + u_{12} \ ,$$

where $u_0 = \mathbb{E}_\gamma [u]$ and $(u_1 + u_2)$ is the γ-orthogonal projection of $u - u_0$ unto $(B_1(\gamma) + B_2(\gamma))$. That is

$$\mathbb{E}_{\gamma(t)} [u_{12}|X] = 0 \quad \text{and} \quad \mathbb{E}_{\gamma(t)} [u_{12}|Y] = 0$$

ANOVA splitting of the S-bundle

Let us write the ANOVA decomposition of the statistical bundle as

$$S_\gamma \Delta^\circ(\Omega) = (B_1(\gamma) + B_2(\gamma)) \oplus B_{12}(\gamma) \ .$$

Check

1. Let $t \mapsto \gamma(t) \in \Gamma^\circ(\mu_1, \mu_2)$ be a smooth curve with $\gamma(0) = \gamma$. Then the velocity at γ belongs to the interactions, $\overset{\star}{\gamma}(0) \in B_{12}(\gamma)$.
2. Given any interaction $v \in B_{12}(\gamma)$, the curve $t \mapsto \gamma(t) = (1 + tv)\gamma$ stays in $\Gamma^\circ(\mu_1, \mu_2)$ for t in a neighborhood of 0 and $v = \overset{\star}{\gamma}(0)$.

Statistical Bundle of the Transport Model

Transport Model Bundle

- The TMB with margins μ_1 and μ_2 is the sub-statistical bundle

$$S\Gamma^\circ(\mu_1, \mu_2) = \{(\gamma, v) | \gamma \in \Gamma^\circ(\mu_1, \mu_2), \mathbb{E}_\gamma[v|X] = \mathbb{E}_\gamma[v|Y] = 0\} \ .$$

- The transport ${}^{\mathrm{m}}\mathbb{U}_\gamma^{\bar{\gamma}}$ maps the fiber at γ to the fiber at $\bar{\gamma}$.

The sub-manifold of the transport model is flat in the mixture geometry and there is no simple expression of the exponential coordinate.

The splitting of the statistical bundle suggests a mixed parameterization of $\Delta^\circ(\Omega)$. This is a classical topic in the statistics of contingency tables.

Gradient flow of the OT problem

Let us discuss the Optimal Transport OT problem in the framework of the transport model bundle. Let $c \colon \Omega_1 \times \Omega_2 = \Omega \to \mathbb{R}$ be a cost function and define the expected cost function

$$C \colon \Delta(\Omega) \ni \gamma \mapsto \mathbb{E}_\gamma [c]$$

Gradient

The function $\gamma \mapsto C(\gamma)$ restricted to the open transport model $\Gamma^\circ(\mu_1, \mu_2)$ has statistical gradient in $S\Gamma^\circ(\mu_1, \mu_2)$ given by

$$\operatorname{grad} C \colon \gamma \mapsto c_{12,\gamma} = c - c_{0,\gamma} - (c_{1,\gamma} + c_{2,\gamma}) \in s_\gamma \Gamma^\circ(\mu_1, \mu_2)$$

Gradient flow

The equation of the gradient flow of C is

$$\overset{\star}{\gamma} = -(c - c_{0,\gamma} - (c_{1,\gamma} + c_{2,\gamma})) = -c_{12,\lambda}$$

Kantorovich potential

The gradient mapping grad $C(\gamma)$ is defined to be the orthogonal projection of the cost c onto the space of γ-interactions $B_{12}(\gamma)$.

Assume $\gamma \mapsto c_{12,\gamma}$ extends to all $\hat{\gamma} \in \Gamma(\mu_1, \mu_2)$.

Stationary point

If $\hat{\gamma}$ is a zero of the extended gradient map, grad $C(\hat{\gamma}) = 0$, then it holds

$$c(x,y) = c_{0,\gamma} + c_{1,\gamma}(x) + c_{2,\gamma}(y) \ , \quad (x,y) \in \operatorname{Supp} \hat{\gamma} \ .$$

We expect any solution $t \mapsto \gamma(t)$ of the gradient flow to converge to a coupling $\bar{\gamma} = \lim_{t \to \infty} \gamma(t) \in \Delta(\Omega)$ such that $\mathbb{E}_{\bar{\gamma}}[c]$ is the value of the Kantorovich optimal transport problem.

Kantorovich Theorem

$\hat{\gamma}$ is optimal for the cost c in $\Gamma(\mu_1, \mu_2)$ if, and only if, there exists potentials $u_i \colon \Omega_i \to \mathbb{R}$ such that

$$u_1(x) + u_2(y) \leq c(x,y) \quad \text{and} \quad c(x,y) = u_1(x) + u_2(y)$$

for all $(x,y) \in \operatorname{Supp} \hat{\gamma}$.