Statistical Theories and Machine Learning Using Geometric Methods

Osaka Central Advanced Mathematical Institute (OCAMI)
Osaka Metropolitan University
MEXT Joint Usage/Research Center on Mathematics and Theoretical Physics

OCAMI Reports Vol. 2 (2024)

Statistical Theories and Machine Learning Using Geometric Methods

Organized by

Koichi Tojo

Hideto Nakashima

Yoshihiko Konno

Hideyuki Ishi

Kenji Fukumizu

December 14–15, 2023

Abstract

This workshop was held on December 14–15, 2023 in order to connect researchers in several fields, in particular Statistics, Machine Learning and Mathematics, and to share problems and researches in these fields interdisciplinary.

2020 *Mathematics Subject Classification.*
20G05, 22F30, 53B12, 60E05, 62E10, 62H12, 62J07

Key words and Phrases.
Lie group, information geometry, exponential families,
machine learning, geometric analysis, kernel methods

OCAMI. Statistical Theories and Machine Learning Using Geometric Methods. OCAMI Reports. Vol. 2, Osaka Central Advanced Mathematical Institute, Osaka Metropolitan University. 2024, 174 pp.

Preface

This is a proceedings of the international workshop "Statistical Theories and Machine Learning Using Geometric Methods" held from December 14th to December 15th in 2023. This workshop aimed to connect researchers in several fields, in particular Statistics, Machine Learning and Mathematics, and to share problems and researches in these fields in an interdisciplinary manner.

This workshop was supported by Osaka Central Advanced Mathematical Institute (MEXT Promotion of Distinctive Joint Research Center Program JPMXP0723833165), Osaka Metropolitan University.

This workshop was held in a hybrid format. Domestic speakers were gathered in Academic Extension Center (Osaka Metropolitan University), and overseas speakers participated by Zoom. We had 11 talks, 9 of which were from Japan and the others were from abroad, and 51 people participated in this workshop.

Organizers

Koichi Tojo

RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

Email address: koichi.tojo@riken.jp

Hideto Nakashima

Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Email address: hideto@ism.ac.jp

Yoshihiko Konno

Department of Mathematics, Osaka Metropolitan University, 1-1, Gakunen-cho, Naka-ku, Sakai-shi, 599-8531, Japan

Email address: konno@omu.ac.jp

Hideyuki Ishi

Department of Mathematics, Osaka Metropolitan University, 3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585, Japan

Email address: hideyuki-ishi@omu.ac.jp

Kenji Fukumizu

Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Email address: fukumizu@ism.ac.jp

Contents

[1]Slides are not included in this report

Mean-variance joint statistic valued in a real Siegel domain

Hiroto Inoue
Nishinippon Institute of Technology

Abstract

Let $\overline{x}$ and s_x^2 be the mean and the variance of given data $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. The main idea of this talk is to characterize the pair of statistics $x \mapsto (\overline{x}, s_x^2)$ as a quotient map $\mathbb{R}^n \to \mathbb{R}^n/H$ with the $n-1$ dimensional orthogonal group $H = \mathrm{O}(n-1)$. This is realized by a combination of a quadratic form and the Cholesky decomposition, or the QR decomposition for matrices. Since the quotient space $\mathbb{R}^n/H$ can be identified with the affine group $G = \mathbb{R}_+ \ltimes \mathbb{R}$ or a real Siegel domain Ω, the joint statistics $(\overline{x}, s_x^2)$ is considered to be valued in G or Ω.

From this characterization, we find some fundamental properties and advantages of the joint statistic $(\overline{x}, s_x^2)$ as follows.

- $(\overline{x}, s_x^2)$ is invariant under the actions of H, especially under the permutations between $x_1, \ldots, x_n$. More precisely, it is the maximal invariant statistic in the context of statistical inference.
- The quotient map is equivariant under the affine transformation, which is consequently followed by the transformation rule $ax + b \mapsto (a\overline{x} + b, a^2 s_x^2)$. Moreover, under the assumption of the normal distributions for the data $x \in \mathbb{R}^n$, the parameter estimators and test statistics are written in the product and inverse in the affine group G.
- The probability density functions (PDFs) of each statistic which is a function of the joint statistic $(\overline{x}, s_x^2)$ can be calculated by using the integral formula on the space $\mathbb{R}^n/H$ in terms of the Haar measure on G. The familiar distributions such as the χ^2, t and F distributions are derived as the marginal ones. It enables us to describe the simultaneous hypothesis testing for the pair of parameters (μ, σ^2) of the normal distribution.

We notice that the PDF of the joint statistic $(\overline{x}, s_x^2)$ is obtained from the Bartlett decomposition of the classical Wishart distribution restricted onto Ω. Then it might be meaningful to consider the same kind of statistic valued in other homogeneous cones or Siegel domains as our joint statistic.

Mean-variance joint statistic valued in a real Siegel domain

Hiroto Inoue

Nishinippon Institute of Technology

OCAMI Workshop "Statistical Theories and Machine Learning Using Geometric Methods" Dec. 14, 2023

Basic idea

For a set of real data $x_1, x_2, \ldots, x_n \in \mathbb{R}$, we consider a $2 \times n$ matrix

$$x = \begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{2\times n}.$$

Consider the QR decompositon of x in the following form

$$x = RU \; ; \quad R = \begin{pmatrix} s & t \\ 0 & 1 \end{pmatrix}, \; s > 0, \quad UU^T = nI_2.$$

Then, it can be checked that

$$s = s_x := \left\{ \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})^2 \right\}^{\frac{1}{2}}, \quad t = \overline{x} := \frac{1}{n} \sum_{k=1}^{n} x_k.$$

By the definition, the factor $R = R(x)$ has the following properties.

- Equivariance

$$R(gx) = gR(x), \quad \forall g = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}, \, a > 0.$$

This is equivalent to $(s_{ax+b}, \overline{ax+b}) = (as_x, a\overline{x}+b)$.

- Invariance

$$R(xh) = R(x), \quad \forall h \in \mathrm{O}(n) \text{ s.t. } (1,\dots,1)h = (1,\dots,1).$$

Especially, $R(x)$ is symmetric in $x_1, \dots, x_n$.

▷ We will obtain the test statistics and their PDFs in a group theoretical way.

Setting

- Data space : a set $E_n = E_{n,r}$ $(n \geq 2, r \geq 1)$ of matrices s.t.

$$E_n = \left\{ x = \begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(r+1)\times n} \;\middle|\; \operatorname{rank} x = r+1 \right\}^{1}$$

- Range of statistic : a region $\Omega_n \subset \mathrm{Sym}^+_{r+1}(\mathbb{R})$,

$$\Omega_n = \left\{ \boldsymbol{w} = (w_{ij}) \in \mathrm{Sym}^+_{r+1}(\mathbb{R}) \;\middle|\; w_{r+1,r+1} = n \right\}$$

- Quadratic statistic : a map $\mathcal{Q} : E_n \to \Omega_n$,

$$\mathcal{Q}(x) := xx^T \quad (x \in E_n)$$

[1] The condition of rank can be omitted if we ignore the zero-measure set in the following argument.

- Transformation group : a group $G_0 = B \ltimes \mathbb{R}^r$ by

$$(s,t)\cdot(a,b) := (sa, sb+t) \quad ((s,t),(a,b) \in B \times \mathbb{R}^r),$$

where

$$B = \left\{ s \in \mathrm{GL}_r(\mathbb{R}) \mid \begin{array}{ll} s_{ij} = 0 & (i < j), \\ s_{ii} > 0 & (i = 1, \dots, r) \end{array} \right\},$$

the group of the upper triangular matrices with positive diagonals.

- Actions of G_0 on E_n, Ω_n by

$$G_0 \curvearrowright E_n : \quad (s,t)x = \begin{pmatrix} s & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \quad (x \in E_n)$$

$$G_0 \curvearrowright \Omega_n : \quad (s,t)\boldsymbol{w} = \begin{pmatrix} s & t \\ 0 & 1 \end{pmatrix} \boldsymbol{w} \begin{pmatrix} s & t \\ 0 & 1 \end{pmatrix}^T \quad (\boldsymbol{w} \in \Omega_n)$$

Notice that $G_0 \curvearrowright \Omega_n$ is simply transitive due to the Cholesky decomposition.

Proposition 1.1 (QR decomposition)

Any $x \in E_n$ is uniquely decomposed as follows.

$$x = gu, \quad g \in G_0, \quad u \in \mathcal{Q}^{-1}(n\boldsymbol{e}). \tag{1}$$

Here $\boldsymbol{e} \in \Omega_n$ is the identity matrix.

So the element g is uniquely determined from x and we denote it by $g = R(x)$.

Proposition 1.2

For $x \in E_n$, it holds that $R(x) = (s_x, \overline{x})$, where

$$s_x s_x^T = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(x_k - \overline{x})^T, \quad \overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{2}$$

By the definition of R, we see again

- G_0-Equivariance

$$R(gx) = gR(x), \quad \forall g \in G_0.$$

- $\mathrm{O}(n-1)$-Invariance

$$R(xh) = R(x), \quad \forall h \in \mathrm{O}(E_n),$$

where

$$\mathrm{O}(E_n) = \{ h \in \mathrm{O}(n) \mid E_n h = E_n \} \cong \mathrm{O}(n-1).$$

$$\begin{tikzcd}
E_n \arrow[r, dash, "\cong"] \arrow[d, "\mathcal{Q}"] \arrow[dr, dashed, "R"] & G_0 \times \mathcal{Q}^{-1}(n\boldsymbol{e}) \arrow[d, "\mathrm{pr}"] \\
\Omega_n \arrow[r, dash, "\cong"] & G_0
\end{tikzcd}$$

Notations for volume elements For $\mathbf{s} = (s_1, \dots, s_r) \in \mathbb{R}^r$ and $\boldsymbol{w} \in \Omega_n$, we put[2]

$$\Delta_{\mathbf{s}}(\boldsymbol{w}) := \Delta_1(\boldsymbol{w})^{s_1 - s_2} \Delta_2(\boldsymbol{w})^{s_2 - s_3} \cdots \Delta_r(\boldsymbol{w})^{s_r},$$

$$\Delta_k(\boldsymbol{w}) = \det(\boldsymbol{w}_{ij})_{r-k+1 \le i,j \le r+1}.$$

We notice that

$$\Delta_{\mathbf{s}}(g\boldsymbol{e}) = \prod_{i=1}^{r} (s_{ii}^2)^{s_{r-i+1}} \quad (g = (s,t) \in G_0).$$

For convenience of the index, we let

$$[r] = (1, \dots, r), \quad \mathbf{s} + \lambda = \mathbf{s} + (\lambda, \dots, \lambda) \quad (\mathbf{s} \in \mathbb{R}^r,\ \lambda \in \mathbb{R}).$$

[2] Notations by Faraut-Korányi [1]

We take a left Haar measure $d\mu(g)$ on G_0 as

$$d\mu(g) = \Delta_{-\frac{1}{2}[r]}(g\boldsymbol{e}) \prod_{i=1}^{r} s_{ii}^{-1} ds_{ii} \prod_{i<j}^{r} ds_{ij} \prod_{i=1}^{r} dt_i, \quad g = (s,t).$$

It can be checked the following transformation rule.

$$\begin{cases} d\mu(h_1 g h_2) = \Delta_G(h_2) d\mu(g) & (h_1, h_2 \in G_0), \\ d\mu(g^{-1}) = \Delta_G(g)^{-1} d\mu(g), \end{cases}$$

where $\Delta_G(g) = \Delta_{\frac{1}{2}r-[r]}(g\boldsymbol{e})$.

Mean-variance joint statistic valued in a real Siegel domain
└ Mean-Variance Joint Statistic
└ Decomposition of volume element

Next we denote by dx the Lebesgue measure on $E_n \subset \mathbb{R}^{rn}$;

$$dx = \prod_{k=1}^{n} dx_k, \quad x = \begin{pmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix}$$

We notice the transformation rule

$$d(hx) = |h|^n dx \quad (h \in G_0), \quad \text{where } |h| = \Delta_{+\frac{1}{2}}(h\boldsymbol{e}).$$

Proposition 1.3

Under the relation $x = gu$ in (1), there exists a volume element du on $\mathcal{Q}^{-1}(ne)$ such that

$$dx = |g|^n d\mu(g)\, du. \tag{3}$$

The integral of du is evaluated as follows.

$$\int_{\mathcal{Q}^{-1}(ne)} du = \frac{\sqrt{2\pi}^{rn}}{c_n}, \tag{4}$$

$$c_n = \frac{1}{2^r}\left(\frac{n}{2}\right)^{-\frac{1}{2}rn} \sqrt{\pi}^{\frac{1}{2}r(r+1)} \prod_{i=1}^{r} \Gamma\left(\frac{1}{2}(n-i)\right).$$

Here $\Gamma(x)$ be the Gamma function.

Let $\Phi_{\boldsymbol{G}}(g)$ denote the PDF of random variables $\boldsymbol{G}$ on G_0 ;

$$P(\boldsymbol{G} \in \mathcal{V}) = \int_{\mathcal{V} \subset G_0} \Phi_{\boldsymbol{G}}(g) d\mu(g).$$

We define an expectation $E[\boldsymbol{G}] \in G_0$ by

$$E[\boldsymbol{G}]\boldsymbol{e} = \int_{G_0} (h\boldsymbol{e}) \Phi_{\boldsymbol{G}}(h)\, d\mu(h).$$

Proposition 2.1

For a random variable $\boldsymbol{G}$ on G_0 and $h \in G_0$, it holds that

$$E[h\boldsymbol{G}] = hE[\boldsymbol{G}].$$

Distributions on G_0 derived from normal distributions

Suppose $\boldsymbol{X}$ is a random variable on E_n and

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n \\ 1 & \cdots & 1 \end{pmatrix}, \quad \boldsymbol{X}_i \sim N_r(t, ss^T) \quad \text{i.i.d..}$$

We denote it by $\boldsymbol{X} \sim N_r(h)^n$, where $h = (s,t) \in G_0$.

- Parameter change

$$g \in G_0,\ \boldsymbol{X} \sim N_r(h)^n \ \Rightarrow\ g\boldsymbol{X} \sim N_r(gh)^n.$$

- Unbiased estimator

$$\boldsymbol{X} \sim N_r(h)^n \ \Rightarrow\ E[R(\boldsymbol{X})] = h.$$

The PDF of $\boldsymbol{X} \sim N_r(g)^n$ is written as

$$f_{\boldsymbol{X}}(x) = \frac{1}{\sqrt{2\pi}^{rn}} e^{-\frac{1}{2}\mathrm{tr}\mathcal{Q}(g^{-1}x)+\frac{1}{2}n}|g|^{-n}.$$

The PDF of $R(\boldsymbol{X})$ is obtained as a marginal distribution.

Proposition 2.2

If $\boldsymbol{X} \sim N_r(e)^n$, then $\Phi_{R(\boldsymbol{X})}(h)$ is given by

$$\Phi_n(h) := \frac{1}{c_n} e^{-\frac{1}{2}n\,\mathrm{tr}(h\boldsymbol{e})+\frac{1}{2}n}|h|^n. \tag{5}$$

Mean-variance joint statistic valued in a real Siegel domain
└Probability distributions on G_0

Proposition 2.3

If $\boldsymbol{X} \sim N_r(e)^n$, $\boldsymbol{Y} \sim N_r(e)^m$ independently, then $\Phi_{R(\boldsymbol{X})^{-1}R(\boldsymbol{Y})}(h)$ is given as

$$\Phi_{n,m}(h) := \frac{c_{n+m}}{c_n c_m}|h|^m \Delta_{[r]-\frac{1}{2}(n+m+r)}\left(\frac{n}{n+m}\boldsymbol{e} + \frac{m}{n+m}h\boldsymbol{e}\right). \tag{6}$$

Sketch of proof. It is obtained by calculating the convolution of the PDFs

$$\Phi_{R(\boldsymbol{X})^{-1}R(\boldsymbol{Y})}(h) = \int_{G_0} \Phi_n(g)\Phi_m(gh)d\mu(g).$$

□

Marginal distributions on $\mathbb{R}_+, \mathbb{R}$ when $r=1$

When $r=1$, for $h=(s,t)\in G_0$ we have

$$\Phi_{R(\boldsymbol{X})}(h)=\frac{1}{c_n}e^{-\frac{1}{2}n(s^2+t^2)}s^n, \quad c_n=\frac{\sqrt{\pi}}{2}\left(\frac{n}{2}\right)^{-\frac{n}{2}}\Gamma\left(\frac{n-1}{2}\right).$$

Corollary 2.1

For $R(\boldsymbol{X})=(S_{\boldsymbol{X}},\overline{\boldsymbol{X}})$, it holds that

$$\begin{cases} \Phi_{S_{\boldsymbol{X}}}(s)=\dfrac{\sqrt{2\pi/n}}{c_n}e^{-\frac{1}{2}ns^2}s^{n-2} & : \chi^2 \text{ dist.} \\ \Phi_{\overline{\boldsymbol{X}}}(t)=\dfrac{1}{\sqrt{2\pi/n}}e^{-\frac{1}{2}nt^2} & : \text{normal dist.} \end{cases}$$

Mean-variance joint statistic valued in a real Siegel domain
└Probability distributions on G_0

We have

$$\Phi_{R(\boldsymbol{X})^{-1}}(h)=\Phi_n(h^{-1})\Delta_G(h)^{-1}=\frac{1}{c_n}e^{-\frac{1}{2s^2}n(1+t^2)}s^{-n+1}$$

Corollary 2.2

For $R(\boldsymbol{X})^{-1}=(S_{\boldsymbol{X}}^{-1},-S_{\boldsymbol{X}}^{-1}\overline{\boldsymbol{X}})$, it holds that

$$\begin{cases} \Phi_{S_{\boldsymbol{X}}^{-1}}(s)=\dfrac{\sqrt{\pi/n}}{c_n}e^{-\frac{n}{2s^2}}s^{-n-1} & : \text{inv. } \chi^2 \text{ dist.} \\ \Phi_{S_{\boldsymbol{X}}^{-1}\overline{\boldsymbol{X}}}(t)=\dfrac{1}{B\left(\frac{n-1}{2},\frac{1}{2}\right)}\left(1+t^2\right)^{-\frac{n}{2}} & : t \text{ dist.} \end{cases}$$

We have

$$\Phi_{R(\boldsymbol{X})^{-1}R(\boldsymbol{Y})}(h) = \frac{c_{n+m}}{c_n c_m} s^m \left(\frac{n}{n+m} + \frac{m}{n+m}s^2 + \frac{nm}{(n+m)^2}t^2\right)^{-\frac{n+m-1}{2}}.$$

Let $B(x,y)$ be the beta function.

Corollary 2.3

For $R(\boldsymbol{X})^{-1}R(\boldsymbol{Y}) = (S_{\boldsymbol{X}}^{-1}S_{\boldsymbol{Y}}, S_{\boldsymbol{X}}^{-1}(\overline{\boldsymbol{Y}} - \overline{\boldsymbol{X}}))$, it holds that

$$\begin{cases} \Phi_{S_{\boldsymbol{X}}^{-1}S_{\boldsymbol{Y}}}(s) = \dfrac{2n^{\frac{n-1}{2}}m^{\frac{m-1}{2}}}{B\left(\frac{n-1}{2},\frac{m-1}{2}\right)} s^{m-2}\left(\dfrac{1}{n+ms^2}\right)^{\frac{n+m}{2}-1} & : F \text{ dist.} \\ \Phi_{S_{\boldsymbol{X}}^{-1}(\overline{\boldsymbol{Y}}-\overline{\boldsymbol{X}})}(t) = \dfrac{1}{B\left(\frac{n-1}{2},\frac{1}{2}\right)}\left(\dfrac{m}{n+m}\right)^{\frac{1}{2}}\left(1+\dfrac{m}{n+m}t^2\right)^{-\frac{n}{2}} & : t \text{ dist.} \end{cases}$$

On the orthogonal factor $U(x)$, when $r = 1$

For $x \in E_n$, we write its decomposition as

$$x = R(x)U(x), \quad U(x) \in \mathcal{Q}^{-1}(n\boldsymbol{e}).$$

For $x, y \in E_n$, the determinant of transposed product

$$r_{xy} := \det\left(U(x)U(y)^T\right)$$

is the correlation coefficient of x and y.

Assume that $\boldsymbol{X} \sim N_r(e)^n$. Then,

- $R(\boldsymbol{X})$ and $U(\boldsymbol{X})$ are independent of each other.
- $U(\boldsymbol{X}) = ((x_i - \overline{x})/\sigma_x)$ follows the uniform distribution on a sphere S^{n-1}.[3]

  $\Rightarrow$ For $y \in E_n$, $\boldsymbol{R} = \det\Big(U(\boldsymbol{X})U(y)^T\Big)$ has a following PDF

$$f(r) = c(1-r^2)^{\frac{n}{2}-1} \quad (-1 \leq r \leq 1),\ c > 0.$$

  $\Rightarrow$ $\boldsymbol{T} := \dfrac{\sqrt{n-2}\boldsymbol{R}}{\sqrt{1-\boldsymbol{R}^2}}$ follows the t distribution (used in correlation analysis).

[3] Each $(x_i - \overline{x})/\sigma_x$ is called an ancillary statistic.

One-sample testing

We consider one-sample test for a normal distribution $N_1(h)$,

$$H_0 : h = h_0$$

with a set of sample data $\boldsymbol{X} \sim N_1(h)^n$.

- We define a G_0-valued test statistic $\boldsymbol{U}$ as

$$\boldsymbol{U} = h_0^{-1} R(\boldsymbol{X}).$$

Now H_0 implies $\Phi_{\boldsymbol{U}}(g) = \Phi_n(g)$.

- We take an acceptance region $\mathcal{V}$: a neighbor of the identity element $e \in G_0$ such that

$$\int_{\mathcal{V}} \Phi_n(g) d\mu(g) = 1 - \alpha \quad (\alpha \text{ is a significance level})$$

Mean-variance joint statistic valued in a real Siegel domain
└Hypothesis test of normal distributions
└One-sample testing

It realizes the following testings depending on the choice of $\mathcal{V}$.

- $\mathcal{V} = \mathcal{I} \times \mathbb{R}$:

$$\boldsymbol{U} \in \mathcal{V} \Leftrightarrow \frac{S_{\boldsymbol{X}}}{\sigma_0} \in \mathcal{I} \quad (\ \chi^2\text{-test for } \sigma^2)$$

- $\mathcal{V} = \mathbb{R}_+ \times \mathcal{J}$:

$$\boldsymbol{U} \in \mathcal{V} \Leftrightarrow \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma_0} \in \mathcal{J} \quad (Z\text{-test for } \mu \text{ with known } \sigma^2)$$

- $\mathcal{V} = (\mathbb{R}_+ \times \mathcal{J})^{-1} := \{h \in G_0;\ h^{-1} \in \mathbb{R}_+ \times \mathcal{J}\}$

$$\boldsymbol{U} \in \mathcal{V} \Leftrightarrow -\frac{\overline{\boldsymbol{X}} - \mu_0}{S_{\boldsymbol{X}}} \in \mathcal{J} \quad (\ t\text{-test for } \mu \text{ with unknown } \sigma^2)$$

Here $R(\boldsymbol{X}) = (S_{\boldsymbol{X}}, \overline{\boldsymbol{X}})$, $h_0 = (\sigma_0, \mu_0)$.

Two-sample testing

We consider a two-sample test for normal distributions $N_1(h_x)$, $N_1(h_y)$,

$$H_0 : h_x = h_y$$

with two sets of data $\boldsymbol{X} \sim N_1(h_x)^n$, $\boldsymbol{Y} \sim N_1(h_y)^m$.

- We define a G_0-valued test statistic $\boldsymbol{U}$ as

$$\boldsymbol{U} = R(\boldsymbol{X})^{-1} R(\boldsymbol{Y}).$$

  Now H_0 implies $\Phi_{\boldsymbol{U}}(g) = \Phi_{n,m}(g)$.
- We take an acceptance region $\mathcal{V}$: a neighbor of the identity element $e \in G_0$ such that

$$\int_{\mathcal{V}} \Phi_{n,m}(g) d\mu(g) = 1 - \alpha \quad (\alpha \text{ is a significance level})$$

Mean-variance joint statistic valued in a real Siegel domain
└Hypothesis test of normal distributions
└Two-sample testing

Depending on the choice of $\mathcal{V}$, it realizes the following testings.

- $\mathcal{V} = \mathcal{I} \times \mathbb{R}$:

$$\boldsymbol{U} \in \mathcal{V} \Leftrightarrow \frac{S_{\boldsymbol{Y}}}{S_{\boldsymbol{X}}} \in \mathcal{I} \quad (\text{ } F\text{-test for the ratio } \sigma_x^2/\sigma_y^2)$$

- $\mathcal{V} = \mathbb{R}_+ \times \mathcal{J}$:

$$\boldsymbol{U} \in \mathcal{V} \Leftrightarrow \frac{\overline{\boldsymbol{Y}} - \overline{\boldsymbol{X}}}{S_{\boldsymbol{X}}} \in \mathcal{J} \quad (\text{ } t\text{-test for the difference } \mu_x - \mu_y)^4$$

Here $R(\boldsymbol{X}) = (S_{\boldsymbol{X}}, \overline{\boldsymbol{X}})$, $R(\boldsymbol{Y}) = (S_{\boldsymbol{Y}}, \overline{\boldsymbol{Y}})$.

[4] This statistic is however not common for testing the difference $\mu_x - \mu_y$.

Remarks

- We notice the similarity of $\mathcal{Q}(\boldsymbol{X}), R(\boldsymbol{X})$ to the Wishart matrix and its Bartlett decomposition.
- Ω_n, the range of $\mathcal{Q}(\boldsymbol{X})$, is isomorphic to the following convex domain D,

$$\Omega_n \cong D = \left\{ (\Sigma, \mu) \in \mathrm{Sym}_n^+(\mathbb{R}) \times \mathbb{R}^n \;\middle|\; \Sigma - \mu\mu^T \in \mathrm{Sym}_n^+(\mathbb{R}) \right\}$$

It is one of so-called real Siegel domains.

▷ Future work Take other data space E to define a similar statistic $\tilde{\mathcal{Q}} : E^{\oplus n} \to \Omega$ valued in a convex domain Ω.

References

J. Faraut and A. Korányi.
Analysis on symmetric cones.
Oxford mathematical monographs. Clarendon Press,Oxford University Press, 1994.

Bayesian Learning with Lie Groups

Eren Mehmet Kıral

Abstract

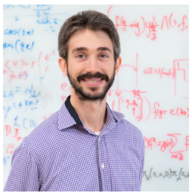
There is uncertainity in the data, and there can be more than one explanation for the same data. Two models can distinguish between a cat picture and a dog picture but one may focus on the fur texture, while the other may make its decisions mostly based on the pointedness of the ears. These can be equivalently good explanations of the data. In this talk I will talk about how we can learn multiple models "at the same time", by introducing some uncertainity to our model parameters. I will present a way of updating probability distributions on the model parameters using Lie Groups, and how desired structures can be preserved by the action of a Lie group.

This is joint work with Thomas Möllenhoff and M. Emtiyaz Khan. It was presented in the AISTATS 2023 conference.

Lie Group Bayesian Learning Rule

E. Mehmet Kıral, joint w. Thomas Möllenhoff and M. Emtiyaz Khan.

Dec 14, 2023 OCAMI Workshop
Statistical Theories and Machine Learning Using Geometric Methods

The work is supported mainly by the Bayes-duality project, JST CREST Grant Number JPMJCR2112.
Presented at **AISTATS 2023**, arxiv # 2303.04397

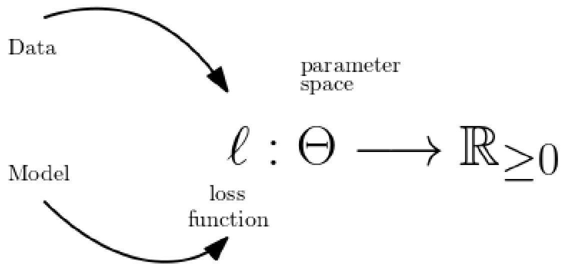
The classical and Bayesian learning setups

Data

Model

$$\underbrace{\ell}_{\text{loss function}} : \overset{\text{parameter space}}{\Theta} \longrightarrow \mathbb{R}_{\geq 0}$$

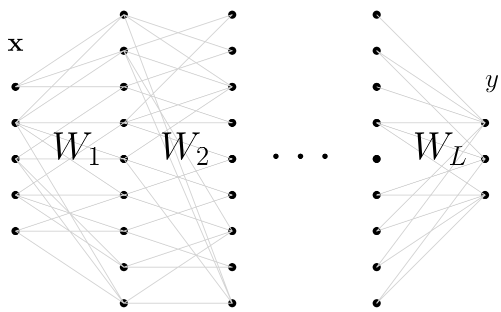
The classical and Bayesian learning setups



Classically: find $\theta^* \in \Theta$ minimizing ℓ.

Bayesian : find a distribution $q \in \mathcal{P}(\Theta)$ ....

Example of loss function



$$f(\mathbf{x}, \theta) = W_L(\sigma(W_{L-1}(\cdots \sigma(W_1 \mathbf{x})))))$$

e.g. $\sigma : x_i \mapsto \max\{0, x_i\}$ componentwise. The tunable parameters are

$$\theta = (W_1, W_2, \ldots, W_L) \in \mathbb{R}^P = \Theta.$$

For data $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N}$ we search for a parameter $\theta \in \Theta$ s.t. $f(\mathbf{x}_i, \theta) \approx y_i$ for all i.

Let $c(\overline{y}, y) = \|y - \overline{y}\|^2$, say. Then using gradient, minimize,

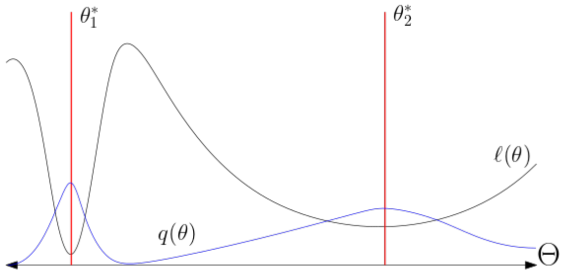
$$\ell(\theta) = \sum_{(\mathbf{x}_i, y_i) \in \mathscr{D}} \underbrace{c(f(\mathbf{x}_i, \theta), y_i)}_{\ell_i(\theta)} + \underbrace{\tfrac{1}{2}\lambda\|\theta\|^2}_{R(\theta)}.$$

Classical vs. Bayesian learning

The loss function is highly nonconvex. Usually

$$\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta) + R(\theta)$$

where $\ell_i(\theta)$ is the loss contribution from the i^{th} data point and $R(\theta)$ regularizer.



θ_1^* and θ_2^* are both equally valid explanations of the same data.
A distribution over the data considers both explanations "*at the same time*".

Betting it all on one outcome

Say two dice are thrown and I tell you that the sum is greater than 7. ⚄,⚂ satisfies this.

We could say the result was definitely ⚄,⚂.

But there are a total of 15 possibilities

⚅,⚁	⚄,⚂	⚃,⚃	⚂,⚄	⚁,⚅
⚅,⚂	⚄,⚃	⚃,⚄	⚂,⚅	
⚅,⚃	⚄,⚄	⚃,⚅		
⚅,⚄	⚄,⚅			
⚅,⚅				

It is much more sensible to say it is one of these 15 outcomes, with equal probability.

(principle of indifference, principle of maximum entropy)

The Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ. We solve

$$q_* \in \underset{q \in \mathcal{Q}}{\arg\min}\ \mathbb{E}_q[\ell] - \tau \mathcal{H}(q)$$

for some family of distributions $\mathcal{Q} \subseteq \mathcal{P}_\nu(\Theta) = \{q(\theta)\mathrm{d}\nu(\theta)\}$ on the parameters.

- The expectation $\mathbb{E}_q[\ell] = \int_\Theta \ell(\theta)q(\theta)\mathrm{d}\nu(\theta)$ prefers regions with low loss.
- The entropy $\mathcal{H}_\nu(q) = -\int_\Theta q(\theta)\log q(\theta)\mathrm{d}\nu(\theta)$ prefers a higher spread of q.
- The temperature $\tau > 0$ is a balancing term.

Constrained maximization: Statistical mechanics interpretation

Assume θ to be a kind of "microstate" with energy level $\ell(\theta)$. So Θ is some "state space".

Statistical mechanics: Assume a distribution of the microstates (across "particles") maximizing entropy, constrained to have expected energy $\leq E_0$.

Lagrange multiplier $\beta \geq 0$:

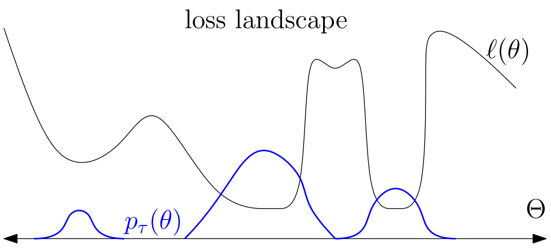
$$\underset{q \in \mathcal{P}_\nu(\Theta)}{\arg\min} -\mathcal{H}_\nu(q) + \beta(\mathbb{E}_{q\mathrm{d}\nu}[\ell] - E_0) = \underset{q \in \mathcal{P}_\nu(\Theta)}{\arg\min}\, \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \frac{1}{\beta}\mathcal{H}_\nu(q)$$

$\tau = \frac{1}{\beta}$ corresponds to the thermodynamical notion of temperature.

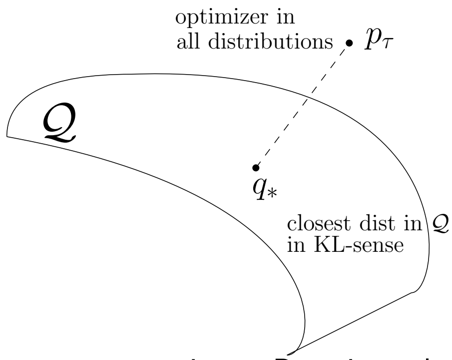
The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\theta) \propto e^{-\frac{1}{\tau}\ell(\theta)}$:

$$\arg\min_{q\in\mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) = \arg\min_{q\in\mathcal{Q}} \mathbb{D}_\nu(q\|p_\tau).$$



Minimize the objective $\mathcal{E}(q) := \mathbb{D}(q\|p_\tau)$ for $q \in \mathcal{Q}$...



...an approximate Bayesian solution.

Previously... BLR: The Bayesian Learning Rule, [KR21][1]

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$. $T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

$$\lambda \longleftarrow \lambda - \alpha F(\lambda)^{-1}\nabla_\lambda \mathcal{E}(q_\lambda)$$

$\alpha > 0$ step size
$F(\lambda)$ the Fisher matrix

BLR is Natural Gradient Descent on λ parameters.

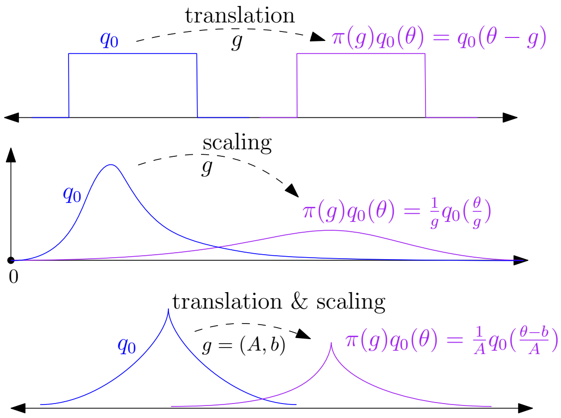
- Issue 1 The candidates $\mathcal{Q}$ is required to be an exponential family,
- Issue 2 Not every λ is allowed as a natural parameter, and the linear update rule could overshoot the constraints.
- Issue 3 Computing $\nabla_\lambda \mathcal{E}(q_\lambda)$ is not efficient in general but for special exponential families.

[1][KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Parametrizing $\mathcal{Q}$ by groups.

Group G acting on the parameter set Θ, also acts on distributions on Θ.
$\mathcal{Q}$ is formed as the orbit of such an action for any base distribution q_0:

$$\mathcal{Q} = \{\pi(g)q_0 : g \in G\}. \qquad \text{Call } q_g = \pi(g)q_0.$$



- $G = (\mathbb{R}, +)$, $\Theta = \mathbb{R}$,
- $G = (\mathbb{R}_{>0}, \times)$, $\Theta = \mathbb{R}_{>0}$,
- $G = \mathrm{Aff}(\mathbb{R}) = \mathbb{R}_{>0} \ltimes \mathbb{R}$, $\Theta = \mathbb{R}$

Optimization on the group

We now solve

$$\arg\min_{g \in G} \mathcal{E}(q_g) = \arg\min_{g \in G} \int_\Theta q_g \log\left(\frac{q_g}{e^{-\frac{1}{\tau}\ell}}\right)$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of X is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q_{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\int_\Theta q_{ge^{tX}}(\theta)\tfrac{1}{\tau}\ell(\theta)\mathrm{d}\nu(\theta)}_{\text{data contribution}} + \underbrace{\int_\Theta q_{ge^{tX}}(\theta)\log q_g(\theta)\mathrm{d}\nu(\theta)}_{\text{entropy contribution}}\Big|_{t=0}$$

The data contribution can be rewritten as

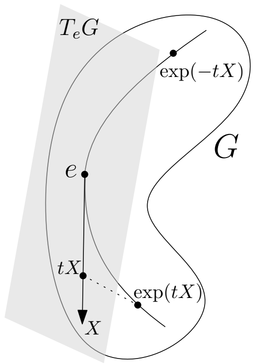
$$\int_\Theta q_g(\theta)(\nabla_\theta \ell(\theta))^\top(\mathrm{Ad}_g(X)\cdot\theta)\mathrm{d}\nu(\theta) \approx \frac{1}{K}\sum_{\substack{i=1 \\ \theta_i \sim q_g}}^{K} \nabla\ell(\theta_i)^\top(\mathrm{Ad}_g(X)\cdot\theta_i)$$

Important aspects of a Lie-Group.

A Lie group is also a **smooth manifold** and **directions** makes sense.
Can take derivatives of paths on the group, giving tangent vectors.

There is a special map $\exp : T_eG \to G$, which works as a global **retraction**. Multiplying the whole path by a $g \in G$ carries it to a neighborhood of g. So, directions at g can also be parametrized by the tangents at identity.

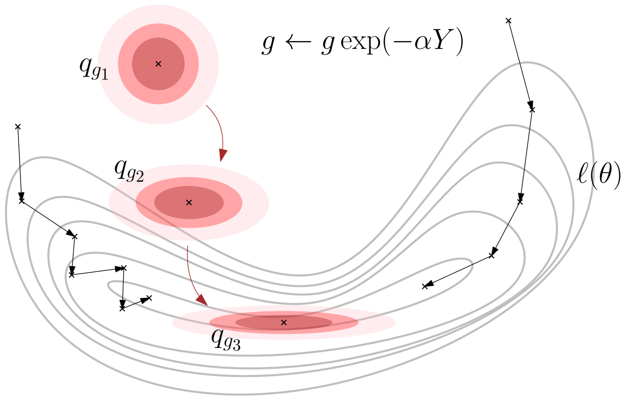
But also other retractions can be created using the exponential map.



A rich mine of examples of Lie Groups are **matrix groups** (e.g. invertible matrices $\mathrm{GL}(n)$, orthogonal matrices $\mathrm{O}(n)$, matrices with determinant one $\mathrm{SL}(n)$) with matrix multiplication.

Classical Learning vs. Learning via Group

The *point based* gradient descent updates parameters: $\theta \leftarrow \theta - \alpha\nabla\ell(\theta)$
Bayesian Learning Rule(s) update the distribution over the parameters θ.



$Y \in T_eG$ is the direction of fastest ascent of $\mathcal{E}(q_g)$ w.r.t. the Fisher metric.

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization
$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q_g}[\ell] = \int_\Theta q_0(\theta)(\nabla_\theta \ell(g \cdot \theta))^\top \frac{\mathrm{d}g \cdot \theta}{\mathrm{d}g}\mathrm{d}\theta$
Also called *pathwise gradient estimators*[2]

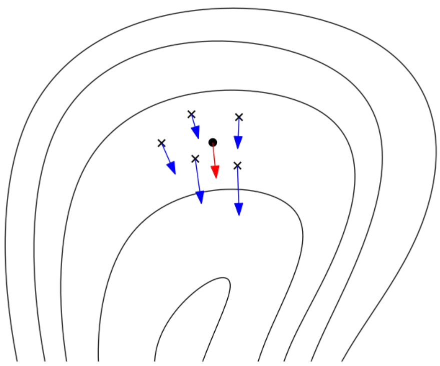
Bonus 1 The Fisher metric is invariant under translations by G.

Bonus 2 The tangent directions Y at each step lie in the same vector space T_eG, so they can be accumulated from previous steps.

[2]Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Specific Update Formulas: The Additive Group

$$g \in \mathbb{R}^P \text{ additive} \quad \Longrightarrow \quad g \longleftarrow g - \alpha \mathbb{E}_{q_g}\left[\nabla_\theta \ell\right]$$



Instead of going in the direction of the derivative at g, the direction is chosen by consensus with at points sampled from q_g.

Multiplicative and Affine Update Formulas

$g \in \mathbb{R}_{>0}$ multiplicative $\Longrightarrow$

$$g \longleftarrow g \exp\Big(-\alpha\big(\mathbb{E}_{q_g}[\theta\partial_\theta \ell] - \tau\big)\Big)$$

$(A, b) \in \mathrm{Aff}(\mathbb{R})$ affine group $\Longrightarrow$

$$b \longleftarrow b + \tfrac{c_X}{c_y} A \tfrac{\exp(-\alpha U)-1}{U} V$$

$$A \longleftarrow A \exp(-\alpha U)$$

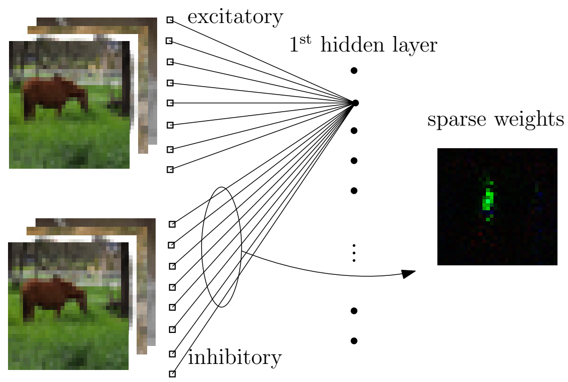
where $\quad U = \mathbb{E}_{q_g}[(\theta - b)\partial_\theta \ell] - \tau$

$\qquad\quad V = A\mathbb{E}_{q_g}[\partial_\theta \ell]$

Filters of the multiplicative group

Label nodes in a neural network "excitatory" or "inhibitory" like biology.

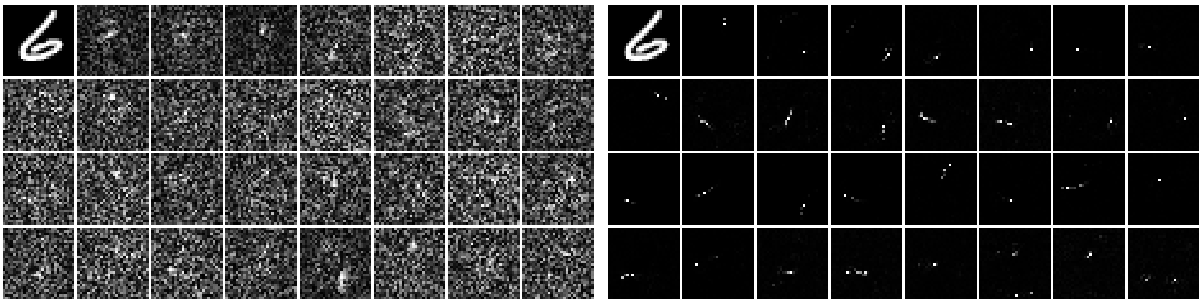
Magnitudes of the weights (in $\mathbb{R}_{>0}$) are the parameters (signs are fixed).



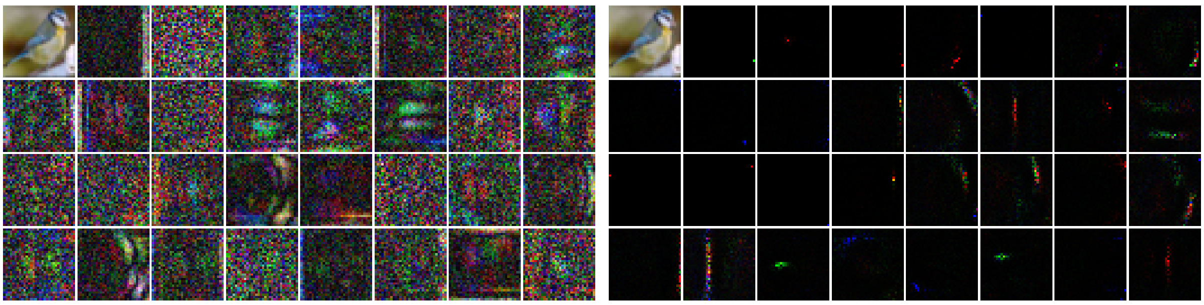
Multiplicative vs Additive filters

Model & Dataset	Method	Accuracy ↑ (higher is better)	NLL ↓ (lower is better)	ECE ↓ (lower is better)
MNIST MLP	add.	$98.38_{\pm 0.02}$	$0.083_{\pm 0.001}$	$0.012_{\pm 0.000}$
	mult.	$98.59_{\pm 0.02}$	$0.058_{\pm 0.001}$	$0.006_{\pm 0.000}$
CIFAR–10 MLP	add.	$58.85_{\pm 0.08}$	$1.236_{\pm 0.002}$	$0.085_{\pm 0.001}$
	mult.	$59.19_{\pm 0.07}$	$1.160_{\pm 0.001}$	$0.026_{\pm 0.001}$

Additive rule is similar to SGD with momentum, multiplicative is different. They both learn.



The additive vs multiplicative filters for RGB images



Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.
- Yet there are some issues with BLR such as: closure in the statistical manifold under updates, and calculation of derivatives w.r.t. distribution parameters.
- The group framework solves the closure problem by design and is able to very generally employ the **reparametrization trick**.
- Each new group would deserve an empirical study to investigate their learning behaviours (like multiplicative vs. additive)
- There may be implementation problems with arbitrary Lie groups, e.g. the exponential map may not always be feasible to compute, so approximations may be necessary.

Teşekkürler

Stiefel Manifold Update

Assume parameters are given as a matrix and want to preserve orthogonality of columns.

$$\Theta = \mathrm{St}(n,m) = \{\theta \in \mathrm{Mat}(n,m) : \theta^\top \theta = I_{m\times m}\}$$

The group $S = \mathrm{SO}(n)$ preserves this manifold. And given a loss function $\ell : \Theta \to \mathbb{R}_{\geq 0}$

$Y \in \mathfrak{so}(n)$ the update direction $\quad Y = \mathrm{Skew} Y_0 = \frac{Y_0 - Y_0^\top}{2} \quad\quad Y_0 = \mathbb{E}_{q_\Lambda}[\nabla \ell \theta^\top]$

Here the distributions are parametrized by $\Lambda \in \mathrm{Mat}(n,m)$

$$q_\Lambda(\theta) \propto e^{-\operatorname{Tr}(\Lambda^\top \theta)}$$

and the update is given by

$$\Lambda \leftarrow e^{-\alpha Y}\Lambda \qquad \text{(actually an efficient variation is used)}$$

Koichi Tojo, Taro Yoshino's: "*Harmonic Exponential Families*".

G a Lie group $H \leq G$. Let ν be a relatively invariant measure on G $\pi : G \to \mathrm{GL}(V)$ a representation of G. Let α be a 1-cocycle of π such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g). \qquad \text{So } \alpha : \underbrace{G/H}_{:=\Theta} \to V$$

Let ν be a relatively invariant measure on Θ, meaning $\nu(gE) = \chi(g)\nu(E)$ for some homomorphism χ. Let $\lambda \in V^\vee$ s.t. $A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta)\rangle} \mathrm{d}\nu(\theta) < \infty$. For such λ

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta)\rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

forms an exponential family satisfying

$$\frac{1}{\chi(g)} q_\lambda(g^{-1}\theta) := q_{\pi^\vee(g)\lambda}(\theta) \qquad \text{where } \langle \pi^\vee(g)\lambda, v\rangle = \langle \lambda, \pi(g)v\rangle.$$

Why call it Bayesian?

Let $\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta) + R(\theta)$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution ℓ_{new}.

How to update p_τ? Take $\tau = 1$

Bayes' rule is about conditional probabilities, and updating priors:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Interpret $e^{-\ell_i(\theta)}$ as the likelihood of observing label y_i given the model parameter θ and $\mathbf{x}_i$.
Interpret $\pi(\theta) \propto e^{-R(\theta)}$ as the prior on the parameters.

After one round of learning the posterior $p \propto e^{-\sum_i \ell_i}\pi$ is our prior belief about θ distribution.
According to Bayes rule updated belief should be *after* a new data point.

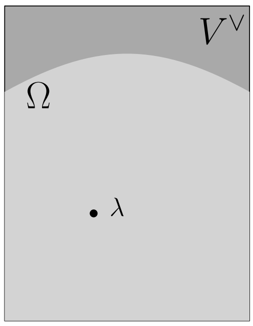
$$p_{\text{updated}}(\theta) \propto e^{-\ell_{\text{new}}(\theta)} p(\theta).$$

This is also the optimizer if we had initially considered the loss function $\ell_{\text{updated}} = \ell + \ell_{\text{new}}$.

Exponential Families

Let $T : \Theta \to V$, called the sufficient statistic. Call

$$\Omega = \left\{ \lambda \in V^\vee : Z(\lambda) := \int_\Theta e^{-\langle \lambda, T(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty \right\}.$$



Then $q_\lambda(\theta) = \frac{1}{Z(\lambda)} e^{-\langle \lambda, T(\theta) \rangle}$ form an exponential family of distributions.

$$-\frac{\partial \ln Z}{\partial \lambda_i} = \int_\Theta T_i(\theta) \frac{1}{Z(\lambda)} e^{-\langle \lambda, T(\theta) \rangle} \mathrm{d}\nu(\theta) = \mathbb{E}_{q_\lambda \mathrm{d}\nu}[T_i] =: \mu_i$$

$$\frac{\partial^2 \ln Z}{\partial \lambda_i \partial \lambda_j} = \int_\Theta (T_i(\theta) - \mu_i)(T_j(\theta) - \mu_j) q_\lambda(\theta) \mathrm{d}\nu(\theta)$$

$$= \mathbb{E}_{q_\lambda}\left[\left(\tfrac{\partial}{\partial \lambda_i} \log q_\lambda \right) \left(\tfrac{\partial}{\partial \lambda_j} \log q_\lambda \right) \right] =: F_{i,j}(\lambda) \qquad \text{Fisher Matrix}$$

Example: If $T(\theta) = \left[\begin{smallmatrix} \theta \\ \theta^2 \end{smallmatrix} \right]$ then we get 1-D Gaussians $q_\lambda(\theta) \propto e^{-\lambda_1 \theta - \lambda_2 \theta^2}$ for $\lambda_2 > 0$.

The generalized Pythagorean theorem on the compactifications of certain dually flat spaces via toric geometry

Hajime Fujita (Japan Women's University)

The *dually flat space* is a fundamental object in information geometry, which appear as a geometric structure of a family of probability measures with "nice properties" such as exponential family. The dually flat space is a data consisting of a Riemannian manifold (X, g) and two flat connections (∇, ∇^*) which are dual to each other with respect to the metric g. The Riemannian metric g of the dually flat space is a Hesse metric, and hence, it admits a (local) potential function φ for a ∇-affine coordinate and a dual potential ψ for a ∇^*-affine coordinate. The potential functions determine the *Bregman divergence* $D(\cdot|\cdot) : X \times X \to \mathbb{R}$. One important feature of the Bregman divergence is the *generalized Pythagorean theorem*, which is a fundamental tool in statistical inference.

Toric geometry is an area which have a position in the intersection of algebraic geometry, complex geometry and symplectic geometry. We focus on its complex and symplectic aspects, i.e., the Kähler structure. Any toric manifold has a structure of a singular Lagrangian torus fibration by its *moment map* and it associates a convex polytope as its image which is of spacial class called *Delzant polytopes*. There is a dictionary between toric manifolds and Delzant polytopes, and several geometric quantities of given toric manifold can be converted into them on the Delzant polytope. The description of the Riemannian metric on the toric manifold as in [1] tells us that the interior of the corresponding Delzant polytope has a natural structure of a dually flat space. This observation can be understood as a variant of Dombrowski's construction ([2]) which asserts a correspondence between Kähler manifold and dually flat space, and it leads to the notion of the *torification* introduced in [4].

In this talk we explain an extension of the dually flat structure of the Delzant polytope to its boundary following [3]. In particular we give a continuous extension of the Bregman divergence to the boundary so that the generalized Pythagorean theorem holds including the boundary points. A typical example is the complex projective space $\mathbb{C}P^n$ and the probability simplex Δ^n as its Delzant polytope. In the language of information geometry this setting is the family of probability measures on the finite set $[n+1] = \{1, \cdots, n+1\}$ together with the expectation parameter. Our result offers an framework to handle the generalized Pythagorean theorem including the zero probabilities.

References

[1] M. Abreu, *Kähler geometry of toric manifolds in symplectic coordinates*, Symplectic and contact topology: interactions and perspectives (Toronto, ON/Montreal, QC, 2001), 2003, pp. 1-24.

[2] P. Dombrowski, *On the geometry of the tangent bundle*, J. Reine Angew. Math. 210 (1962), 73-88.

[3] H. Fujita, *The generalized Pythagorean theorem on the compactifications of certain dually flat spaces via toric geometry*, arXiv:2305.08422v3, to appear in Information Geometry (DOI : 10.1007/s41884-023-00123-y).

[4] M. Molitor, *Kähler toric manifolds from dually flat spaces.* arXiv:2109.04839v1.

The generalized Pythagorean theorem on the compactifications of certain dually flat spaces via toric geometry

Hajime Fujita
(Japan Women's university)

based on : arXiv:2305.08422 [math.SG], Information Geometry (published online).

Self-introduction & personal background of today's talk

My research area : symplectic geometry
(Geometric Quantization of Hamiltonian torus manifold using Dirac operators, metric geometry of toric simplectic manifold)

[Submitted on 4 Mar 2020]

Distance functions on convex bodies and symplectic toric manifolds

Hajime Fujita, Yu Kitabeppu, Ayato Mitsuishi

In this paper we discuss three distance functions on the set of con
in symplectic toric geometry. By using these observations, we deri
distance.

Comments: 19 pages
Subjects: **Metric Geometry (math.MG)**

[Submitted on 7 Jan 2020 (v1), last revised 1 Mar 2021 (this version, v3)]

Deformation of Dirac operators along orbits and quantization of non-compact Hamiltonian torus manifolds

Hajime Fujita

We give a formulation of a deformation of Dirac operator along orbits of a group action on a possibly non-compact manifold to get an equivariant index and a K-homology cycle representing the index. We apply this framework to non-compact Hamiltonian torus manifolds to define geometric quantization from the view point of index theory. We give two applications. The first one is a proof of a [Q,R]=0 type theorem, which can be regarded as a proof of the Vergne conjecture for Abelian case. The other is a Danilov-type formula for toric case in the non-compact setting, which shows that this geometric quantization is independent of the choice of polarization. The proofs are based on the localization of index to lattice points.

Comments: 27pages. Due to referee's comments several expositions are rewritten, and typos are corrected. Especially descriptions for non-abelian case are withdrawn. References uploaded. To appear in Canadian Journal of Mathematics
Subjects: **Differential Geometry (math.DG)**; K-Theory and Homology (math.KT); Symplectic Geometry (math.SG)

I am an armature of statistical theories and machine learning.

Two years ago I started studying information geometry just from curiosity.

I leaned a connection between information geometry and Kăhler geometry.

Last year I met Tojo-san (he was a part-time lecturere at my university) and I attended his lecture on information geometry. At that time we had a conversation, and Tojo-san presented me a paper :

Self-introduction & personal background of today's talk

[Submitted on 10 Sep 2021]

Kahler toric manifolds from dually flat spaces

Mathieu Molitor

We present a correspondence between real analytic Kähler toric manifolds and dually flat spaces, similar to Delzant correspondence in symplectic geometry. This correspondence gives rise to a lifting procedure: if $f : M \to M'$ is an affine isometric map between dually flat spaces and if N and N' are Kähler toric manifolds associated to M and M', respectively, then there is an equivariant Kähler immersion $N \to N'$. For example, we show that the Veronese and Segre embeddings are lifts of inclusion maps between appropriate statistical manifolds. We also discuss applications to Quantum Mechanics.

Subjects: **Differential Geometry (math.DG)**; Information Theory (cs.IT); Mathematical Physics (math-ph); Symplectic Geometry (math.SG)

In this paper a correspondence between a specific dually flat space and a toric Kähler manifold is discussed.
Molitor treated the "open dense part" of toric manifold. I tried to capture the behavior on the "compactification of it".

As a result I got an application for the divergence on the boundary, the extension of "the extented Pythagorean theorem".
Unofortunetely I do not have any statistical applications so far.
Please let me inform and have a diusscusion if you have any idea or insight!!

Contents

§1. Dually flat space and its example

§2. Dually flat structure on a convex polytope

§3. Delzant polytope and toric symplectic manifold

§4. The "generalized" generalized Pythagorean theorem

§5. Further discussions

§1. Dually flat space and its examples

geometric structure which appears in statistical model including exponential family.

§1. Dually flat space and its example

Def. A dually flat space is a data (X, h, ∇, ∇^*) consisting of

- (X, h) : Riemannian manifold
- ∇ and ∇^* are dual to each other w.r.t. h.

$$\Updownarrow$$

$$V_1(h(V_2, V_3)) = h(\nabla_{V_1} V_2, V_3) + h(V_2, \nabla^*_{V_1} V_3)$$
$$\forall V_1, V_2, V_3 \in \Gamma(TX)$$

- ∇ and ∇^* are flat connections of TX.

  ↳ torsion $\equiv 0$, curvature $\equiv 0$

(h, ∇) is called a dually flat structure on X

§1. Dually flat space and its example

Prop. (X, h, ∇, ∇^*) : dually flat

$\Rightarrow \begin{cases} {}^\exists x = (x_1, \cdots, x_n) : \text{(local) coordinate of } X \\ {}^\exists \varphi = \varphi(x) : C^\infty\text{- function on } X. \end{cases}$

s.t. $\begin{cases} (1)\ x \text{ is } \nabla\text{-affine} ; \nabla_{\partial x_i} \partial x_j = 0 \quad ({}^\forall i, j) \\ (2)\ h = \mathrm{Hess}_x \varphi ; h(\partial x_i, \partial x_j) = \dfrac{\partial^2 \varphi}{\partial x_i \partial x_j} \end{cases}$

Def. φ is called a potential of (h, ∇).

§1. Dually flat space and its example

Prop. (X, h, ∇, ∇^*) : dually flat

$x = (x_1, \cdots, x_n)$, $\varphi = \varphi(x)$ as in (1) & (2).

$\Rightarrow \begin{cases} {}^\exists y = (y_1, \cdots, y_n) : \text{(local) coordinate of } X \\ {}^\exists \psi = \psi(y) : C^\infty\text{- function on } X. \end{cases}$

s.t. $\begin{cases} (1)\ y \text{ is } \nabla^*\text{-affine} ; \nabla^*_{\partial y_i} \partial y_j = 0 \ ({}^\forall i, j) \\ (2)\ y = \mathrm{grad}_x \varphi \\ (3)\ \psi \text{ is the Legendre dual of } \varphi ; \\ \qquad \varphi(x) + \psi(y) - x \cdot y = 0 \end{cases}$

Def. y is called the dual affine coordinate.
ψ is called the dual potential of φ.

§1. Dually flat space and its example

Def. For a dually flat space (X, h, ∇, ∇^*) the two-variable function

$$\begin{array}{rccc} D(\cdot,\cdot): & X \times X & \longrightarrow & \mathbb{R} \\ & \rotatebox{90}{$\in$} & & \rotatebox{90}{$\in$} \\ & (\xi, \xi') & \longmapsto & D(\xi, \xi') \end{array}$$

$$:= \varphi(x(\xi)) + \psi(y(\xi')) - x(\xi)\cdot y(\xi')$$

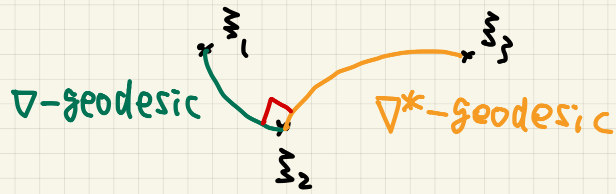
is called the (Bregman) divergence.

The divergence can be thought as a variant of the "square of the distance function" on (X, h).

§1. Dually flat space and its example

Prop. $D(\xi, \xi') \geqq 0$ for $\forall (\xi, \xi') \in X \times X$. and $D(\xi, \xi') = 0 \iff \xi = \xi'$.

Thm. (The generalized Pythagorean theorem)
For $\xi_1, \xi_2, \xi_3 \in X$, if:



then $D(\xi_1, \xi_3) = D(\xi_1, \xi_2) + D(\xi_2, \xi_3)$.

§1. Dually flat space and its example

The following gives the example of our dually flat space.

Prop U : open subset in an affine space with the trivial flat connection ∇_0

$\varphi : U \longrightarrow \mathbb{R}$: smooth convex function

$\mathrm{Hess}\,\varphi = \nabla_0 (d\varphi) > 0$

$\Rightarrow$ $(\mathrm{Hess}\,\varphi, \nabla_0)$ is a dually flat str. on U with dual affine coord. $y = \mathrm{grad}\,\varphi$.

Ex $U \subset \mathbb{R}^n$ with std. coordinate $(x_1, \cdots, x_n)$

$$\Rightarrow \quad \nabla_0 = \sum_i \frac{\partial}{\partial x_i} dx_i \, , \quad \mathrm{Hess}\,\varphi = \left(\frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right)_{i,j}$$

§1. Dually flat space and its example

Summary

Any dually flat space $(X, \underset{\mathrm{Hess}\,\varphi}{\underset{\shortparallel}{h}}, \nabla)$

determines the divergence

$$D : X \times X \longrightarrow \mathbb{R}.$$

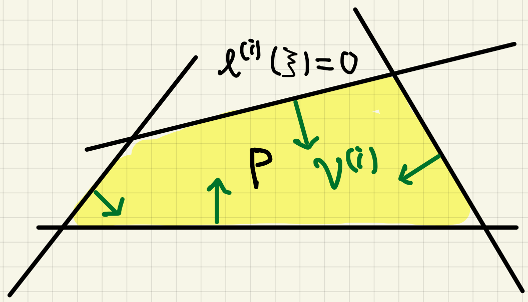
which satisfies the generalized Pythagorean thm.

§2. Dually flat structure on a convex polytope

§2. Dually flat structure on a convex polytope

P ; convex polytope in $\mathbb{R}^n$ defined by
N- inequalities ;

$$\boxed{\ell^{(i)}(\xi) = \underbrace{\nu^{(i)}}_{\text{inward normal vector } \in \mathbb{R}^n} \cdot \xi + \overbrace{\lambda^{(i)}}^{\in \mathbb{R}} \geqq 0 \quad (i=1,2,\cdots,N) \quad (\xi \in \mathbb{R}^n)}$$



§2. Dually flat structure on a convex polytope

Def. Define $\varphi_P : P \longrightarrow \mathbb{R}$ by

$$\boxed{\varphi_P(\xi) := \sum_{i=1}^{N} \ell^{(i)}(\xi) \log \ell^{(i)}(\xi)}$$

called the Guillemin potential of P

Prop φ_P is continuos on P and smooth convex on $\mathring{P} = \mathrm{int}(P)$.

$\leftarrow \ell^{(i)}(\xi) > 0. \ (\forall i)$

$\hookrightarrow$ $\boxed{\text{dually flat sp. } (\mathring{P}, \mathrm{Hess}\,\varphi_P)}$

§2. Dually flat structure on a convex polytope

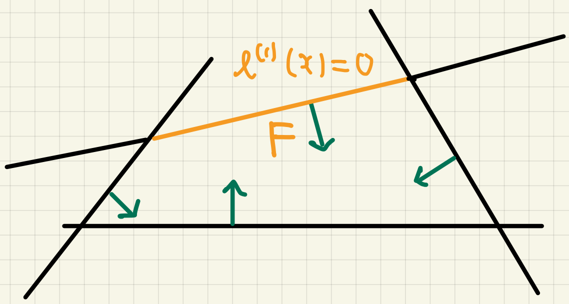
Take a face F of P determined by

$$\ell^{(1)}(\xi) = \cdots = \ell^{(n-k)}(\xi) = 0 \quad (k = \dim F).$$

$$\ell^{(n-k+1)}(\xi), \cdots, \ell^{(N)}(\xi) \geqq 0$$

Define $\varphi_F : F \longrightarrow \mathbb{R}$ by

$$\boxed{\varphi_F(\eta) = \sum_{i=n-k+1}^{N} \ell^{(i)}(\eta) \log \ell^{(i)}(\eta)}$$



§2. Dually flat structure on a convex polytope

Prop $\varphi_F : F \longrightarrow \mathbb{R}$ is continuos on F and smooth convex on $\mathring{F} = \mathrm{int}(F)$.

$\rightsquigarrow \boxed{\text{dually flat sp. } (\mathring{F}, \mathrm{Hess}\,\varphi_F)}$

$P \supset F$

$\varphi_P \rightsquigarrow \varphi_F = \varphi_P|_F$

$(\mathring{P}, \mathrm{Hess}\,\varphi_P)$

$\rightsquigarrow D_P : \mathring{P} \times \mathring{P} \longrightarrow \mathbb{R}$

$(\mathrm{Hess}\,\varphi_P)|_F$

$\ne$

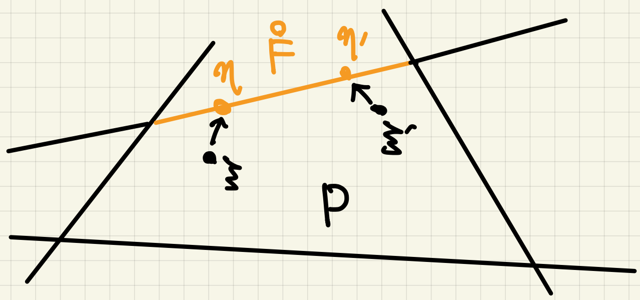
$(\mathring{F}, \mathrm{Hess}\,\varphi_F)$

$\rightsquigarrow D_F : \mathring{F} \times \mathring{F} \longrightarrow \mathbb{R}$

§2. Dually flat structure on a convex polytope

Prop We have the following "continuity".

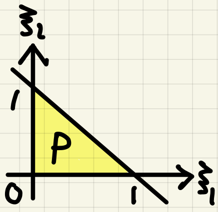
$$\boxed{\lim_{\xi' \to \eta'} \lim_{\xi \to \eta} D_P(\xi, \xi') = D_F(\eta, \eta')}$$



Rem. $\lim_{\xi \to \eta} \lim_{\xi' \to \eta'}$ is not valid.

§2. Dually flat structure on a convex polytope

Example $P:\ \xi_1 \geqq 0,\ \xi_2 \geqq 0,\ 1-(\xi_1+\xi_2) \geqq 0$



Guillemin potential

- $\varphi_P(\xi) = \xi_1 \log \xi_1 + \xi_2 \log \xi_2 + (1-\xi_1-\xi_2)\log(1-\xi_1-\xi_2)$

metric

- $\operatorname{Hess} \varphi_P = \begin{pmatrix} \frac{1}{\xi_1} + \frac{1}{1-\xi_1-\xi_2} & \frac{1}{1-\xi_1-\xi_2} \\ \frac{1}{1-\xi_1-\xi_2} & \frac{1}{\xi_2} + \frac{1}{1-\xi_1-\xi_2} \end{pmatrix}$

dual coord

- $y(\xi) = \operatorname{grad} \varphi_P = \left(\log \frac{\xi_1}{1-\xi_1-\xi_2},\ \log \frac{\xi_2}{1-\xi_1-\xi_2} \right) = (y_1, y_2)$

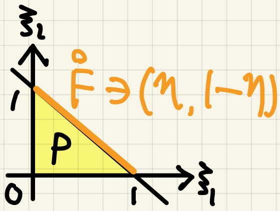
divergence

- $D_P(\xi, \xi') = \xi_1 \log \frac{\xi_1}{\xi_1'} + \xi_2 \log \frac{\xi_2}{\xi_2'} + (1-\xi_1-\xi_2)\log \frac{1-\xi_1-\xi_2}{1-\xi_1'-\xi_2'}$

$\xi = (\xi_1, \xi_2),\ \xi' = (\xi_1', \xi_2') \in P$

§2. Dually flat structure on a convex polytope

Example (continued) Take a face F as $\xi_1 + \xi_2 = 0$.



- $\varphi_F(\eta) = \eta \log \eta + (1-\eta)\log(1-\eta)$
- $\operatorname{Hess} \varphi_F = \frac{1}{\eta} + \frac{1}{1-\eta}$
- $D_F(\eta, \eta') = \eta \log \frac{\eta}{\eta'} + (1-\eta)\log \frac{1-\eta}{1-\eta'}$

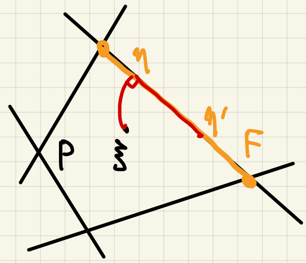
$$\boxed{\text{You can see: } \lim_{\substack{\xi_1' \to \eta' \\ \xi_2' \to 1-\eta'}} \lim_{\substack{\xi_1 \to \eta \\ \xi_2 \to 1-\eta}} D_P(\xi, \xi') = D_F(\eta, \eta') \quad \text{for } \eta, \eta' > 0.}$$

Question

Can we genelarize
the generalized Pythagorean thm.
among interior and boundary points?

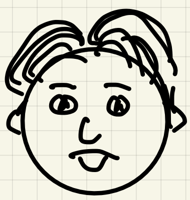
My answer

I did it when P: Delzant
by using "toric geometry".

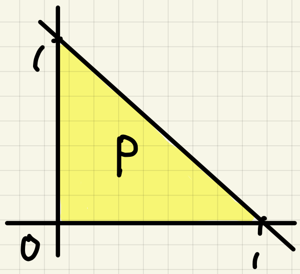


By the way:

Statistician



This P is the set of all probabilities on the three points $[3] = \{1, 2, 3\}$.
D_P is the KL-divergence.



Symplectic geometer

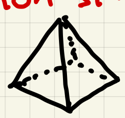
This P is a Delzant polytope
$=$ moment map image of $\mathbb{C}P^2$.

§3. Delzant polytope and toric Kähler manifold

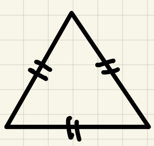
Def. $P \subset \mathbb{R}^n$: convex polytope is **Delzant** if:

- P is **simple**, i.e., each vertex has exactly n-edges.
- P is **rational**, i.e., each edge has directionals vectors $\in \mathbb{Z}^n$.
- P is **smooth**, i.e., directional vectors at each vertex form a $\mathbb{Z}$-basis of $\mathbb{Z}^n$.

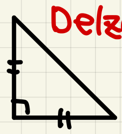
non-simple

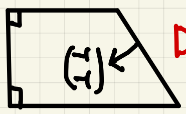


non-Delzant



Delzant





Delzant

§3. Delzant polytope and toric Kahler manifold

FACT (Delzant construction)

P: Delzant polytope in $\mathbb{R}^n$

$\rightsquigarrow \exists M_P$: toric Kähler mfd

$\overset{\shortparallel}{M}$ $\quad \dim_{\mathbb{R}} M_P = 2n$

What is a toric Kähler mfd?

$\updownarrow$

Symplectic mfd with

"nice" $\begin{cases} \bullet \text{ torus action} \\ \bullet \text{ Riemannian metric} \\ \bullet \text{ complex structure} \\ \bullet \text{ map } \mu : M_P \twoheadrightarrow P \end{cases}$

EXCUSE 😣

I do not give the precise definition of toric symplectic/kähler manifold because of the constraint of time.

§3. Delzant polytope and toric Kahler manifold

FACT (Delzant construction)

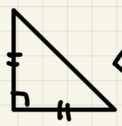
P: Delzant polytope

$\rightsquigarrow \exists M_P$: toric Kähler mfd

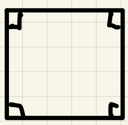
$\overset{\shortparallel}{M}$ $\quad \dim_{\mathbb{R}} M_P = 2n$

In fact, the converse direction holds. $(\mu(M) = P)$

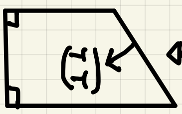
Ex.



$\leftrightsquigarrow M_P = \mathbb{C}P^2$



$\leftrightsquigarrow M_P = \mathbb{C}P^1 \times \mathbb{C}P^1$



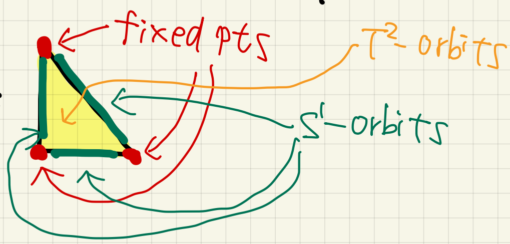
$\leftrightsquigarrow M_P =$ Hirzebruch surface

§3. Delzant polytope and toric Kahler manifold

Fundamental relation between $M=M_P$ & P.

(1) The map $\mu: M \twoheadrightarrow P$ is called the moment map which is in fact the quotient map.

$$T^2 \curvearrowright \mathbb{CP}^2 \longrightarrow P =$$

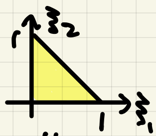


(2) $\mu: M \longrightarrow P$ is a singular Lagrangian torus fibration.

(3) $\mathring{P}$ parametrizes the principal orbits.

§3. Delzant polytope and toric Kahler manifold

Ex.



$$(U(1))^2 = T^2 \curvearrowright M = \mathbb{CP}^2 = (\mathbb{C}^3 \setminus 0)/\mathbb{C}^* \xrightarrow[\mu]{(1)} P \subset \mathbb{R}^2$$

$$(t_1, t_2) \in T^2, \qquad [z_1, z_2, z_3] \longmapsto \frac{(|z_1|^2, |z_2|^2)}{|z_1|^2+|z_2|^2+|z_3|^2}$$

$$[z_1, z_2, z_3] \longmapsto [t_1 z_1, t_2 z_2, z_3]$$

$$F: \xi_1 + \xi_2 = 1$$

$$\rightsquigarrow M_F = \mu^{-1}(F) = \{[z_1, z_2, z_3] \mid z_3 = 0\} \cong \mathbb{CP}^1 \quad (4)$$

$$\downarrow$$

$$F \longleftarrow P_F = [0,1] \subset \mathbb{R}$$

$$(\eta, 1-\eta) \longleftarrow\!\shortmid\ \eta$$

$$M \xleftarrow[(6)]{s} P$$

$$[\sqrt{\xi_1}, \sqrt{\xi_2}, \sqrt{1-\xi_1-\xi_2}] \longleftarrow\!\shortmid\ (\xi_1, \xi_2)$$

§3. Delzant polytope and toric Kahler manifold

Summary

$$\begin{tikzcd} & M \supset M_F \arrow[dl, "\mu"'] \arrow[dr, "\mu_F"] & \\ P \supset F \arrow[rr, leftrightarrow] & & P_F \end{tikzcd}$$

potential

$\varphi_P : P \longrightarrow \mathbb{R}$

divergence

$D_P : \mathring{P} \times \mathring{P} \longrightarrow \mathbb{R}$

potential

$\varphi_F : P_F \longrightarrow \mathbb{R}$

divergence

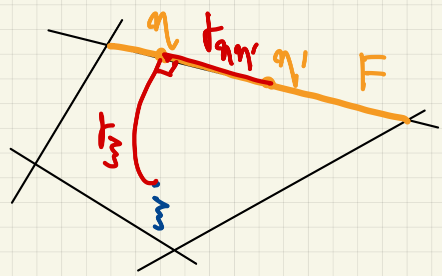
$D_F : \mathring{P}_F \times \mathring{P}_F \longrightarrow \mathbb{R}$

§4. The "generalized" generalized Pythagorean theorem

§4. The “generalized”generalized Pythagorean theorem

Thm. P : Delzant polytope $\rightsquigarrow$ M : toric Kähler
$\cup$
F : face. Take $\xi \in \mathring{P}$, $\eta, \eta' \in \mathring{F}$.

Let γ_ξ and $\gamma_{\eta\eta'}$ be ∇^*, ∇-geodesics s.t.



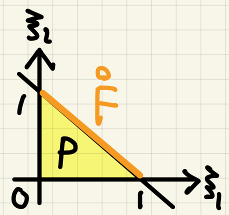
- $\lim_{t\to\infty} s(\gamma_\xi(t)) = s(\eta) \in M$
- $\lim_{t\to\infty} \frac{d}{dt} s(\gamma_\xi(t)) \perp \frac{d}{dt} s(\gamma_{\eta\eta'}(t))$ at $s(\eta) \in M_F \subset M$

Then: $D'_F(\eta', \xi) = D_F(\eta', \eta) + D'_F(\eta, \xi)$

$$\underset{\left(\lim_{P \ni \xi' \to \eta'} D(\xi', \xi)\right)}{\underset{\|}{D'_F(\eta', \xi)}}$$

§4. The “generalized” generalized Pythagorean theorem

Ex $\xi = (a, b) \in \mathring{P}$



- $y = \operatorname{grad}\varphi = \left(\log \frac{\xi_1}{1-\xi_1-\xi_2}, \log \frac{\xi_2}{1-\xi_1-\xi_2}\right)$
- $x = x(y) = \left(\frac{e^{y_1}}{1+e^{y_1}+e^{y_2}}, \frac{e^{y_1}}{1+e^{y_1}+e^{y_2}}\right)$

∇^*-geodesic γ_ξ
$\|$
affine map w.r.t. y

$$t \mapsto \left(\log \frac{a}{1-a-b} + t v_1, \log \frac{b}{1-a-b} + t v_2\right) =: \Gamma_\xi(t)$$

$$\gamma_\xi(t) = x(\Gamma_\xi(t)) \quad (\leftarrow \nabla^*\text{-geod. in } x\text{-coord.})$$

$$= \frac{1}{\underbrace{1-a-b+ae^{tv_1}+be^{tv_2}}_{A}} (e^{tv_1}, e^{tv_2})$$

$$\xrightarrow[t\to\infty]{} \left(\frac{a}{a+b}, \frac{b}{a+b}\right) \in \mathring{F}$$

if $v_1 = v_2 > 0$

§4. The "generalized" generalized Pythagorean theorem

Ex (continued)

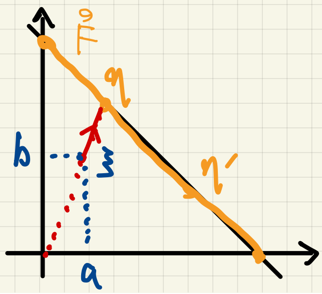
$$S: P \longrightarrow \mathbb{C}P^2$$
$$(\xi_1, \xi_2) \longmapsto [\sqrt{\xi_1}, \sqrt{\xi_2}, \sqrt{1-\xi_1-\xi_2}]$$

$$S(t_\xi(t)) = \left[\sqrt{\frac{a e^{t v_1}}{A}}, \sqrt{\frac{b e^{t v_1}}{A}}, \sqrt{\frac{1-a-b}{A}}\right]$$
$$= [\sqrt{a}, \sqrt{b}, e^{-\frac{t}{2} v_1}\sqrt{1-a-b}] \in \mathbb{C}P^2$$

perpendicular to $\hookrightarrow M_F \cong \mathbb{C}P^1$

$\xi = (a, b) \in \mathring{P}$,
$\eta = \left(\frac{a}{a+b}, \frac{b}{a+b}\right) \in \mathring{F}$,
$\eta' = (\eta', 1-\eta') \in \mathring{F}$.

satisfy the assumptions in the theorem



§4. The "generalized" generalized Pythagorean theorem

Ex (continued) By the direct computations we have;

- $D'_F(\eta', \xi) = \cdots = \eta' \log \frac{\eta'}{a} + (1-\eta') \log \frac{1-\eta'}{b}$
- $D_F(\eta', \eta) = \cdots = \eta' \log \frac{\eta'}{a} + (1-\eta') \log \frac{1-\eta'}{b} - \log(a+b)$
- $D'_F(\eta, \xi) = \cdots = \log(a+b)$

$D'_F(\eta', \xi) = D_F(\eta', \eta) + D'_F(\eta, \xi)$ holds!!

§5. Further discussions

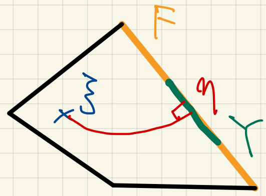
§5. Further discussions

Toward the "projection thm".

For $\xi \in \mathring{\mathcal{P}}$ and $Y \subset F$: submfd.

$\eta \in \mathrm{Crit}(f_\xi) \overset{??}{\Longleftrightarrow}$ orthogonality of the ∇^*-geod. and Y

$$\left(\begin{array}{cccc} f_\xi : & Y & \longrightarrow & \mathbb{R} \\ & \rotatebox{90}{$\in$} & & \rotatebox{90}{$\in$} \\ & \eta & \longmapsto & D_F'(\eta, \xi) \end{array} \right)$$

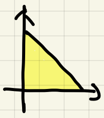


<u>Rem</u> $\mathbb{C}P^2$ ($\mathbb{C}P^n$) case is confirmed.

§5. Further discussions

Applications to statistical inference?

$\mathbb{C}P^n$ $\downarrow$ Some P has an interpretation as probability densities on a finite set.

$\Delta^2 =$ 

- $\Delta^n = \{\xi = (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n \mid {}^\forall \xi_i \geq 0,\ \sum_{i=1}^{n} \xi_i \leq 1\}$

  $\hookrightarrow$ prob. densities on $[n+1] = \{1, \cdots, n+1\}$ with the expectation parameter ξ.

- $D_{\Delta^n} = D_{KL}$ ⇜ The Pytagorean thm including 0-probabilities.

- $\Delta^{n_1} \times \Delta^{n_2} \longleftrightarrow$ a subfamily on $[n_1 n_2]$ (mixture)

§5. Further discussions

Relation with Nakajima-Ohmoto's theory?

The dually flat structure for singular models

Naomichi Nakajima, Toru Ohmoto

The dually flat structure introduced by Amari-Nagaoka is highlighted in information geometry and related fields. In practical applications, however, the underlying pseudo-Riemannian metric may often be degenerate, and such an excellent geometric structure is rarely defined on the entire space. To fix this trouble, in the present paper, we propose a novel generalization of the dually flat structure for a certain class of singular models from the viewpoint of Lagrange and Legendre singularity theory - we introduce a quasi-Hessian manifold endowed with a possibly degenerate metric and a particular symmetric cubic tensor, which exceeds the concept of statistical manifolds and is adapted to the theory of (weak) contrast functions. In particular, we establish Amari-Nagaoka's extended Pythagorean theorem and projection theorem in this general setup, and consequently, most of applications of these theorems are suitably justified even for such singular cases. This work is motivated by various interests with different backgrounds from Frobenius structure in mathematical physics to Deep Learning in data science.

Comments:	29pages, 5figures
Subjects:	**Differential Geometry (math.DG)**; Mathematical Physics (math-ph); Statistics Theory (math.ST)
MSC classes:	53B12(Primary)

They gave a reformulation of dually flat spaces so that they can handle singular models.

§5. Further discussions

Relation with Nakajima-Ohmoto's theory?

The basis of their theory are contact manifold and Legendre submfd.

Legendre contact

$$\begin{tikzcd} L \arrow[d] & \subset & T^*\mathbb{R}^n \times \mathbb{R} \arrow[d] \\ X & \subset & \mathbb{R}^n \end{tikzcd}$$

dually flat

$\overset{??}{\longleftrightarrow}$

toric

$$\begin{tikzcd} T^*\mathring{P} \arrow[r, dashed] \arrow[d] & M \arrow[d, shift left, "\mu"] & \\ \mathring{P} & \subset \; P \arrow[u, shift left, "s"] & \subset \mathbb{R}^n \end{tikzcd}$$

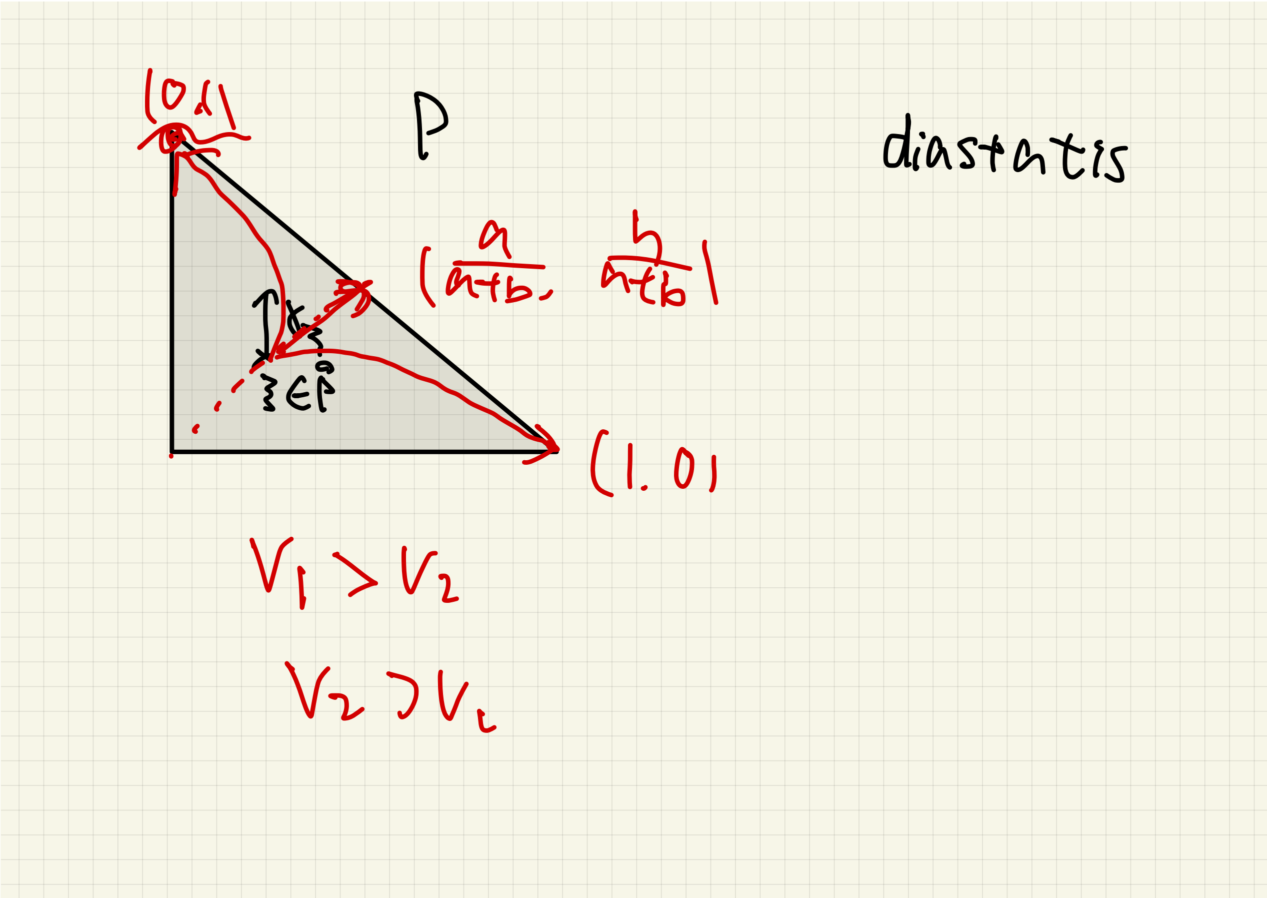
(compactification)

dually flat

Thank you for your attension!!

Summary

- Each convex polytope P has a natural dually flat str. by the Guillemin potential φ_P
- We can extend D_P to any face F.
- If P is Delzant, then we can generalize the Pythagorean thm. to F.


(0,1)
P
diastatis
(a/(a+b), b/(a+b))
ξ ∈ P̊
(1, 0)
V1 > V2
V2 ⊃ V1

Doubly autoparallel structure and curvature integrals: An application to iteration complexity analysis of convex optimization

Atsumi Ohara

University of Fukui, Japan

ohara@fuee.u-fukui.ac.jp

Abstract

On a statistical manifold, we can define autoparallel submanifolds and path integrals of the second fundamental forms (curvature integrals) for its primal and dual affine connections, respectively. A submanifold is called doubly autoparallel if it is simultaneously autoparallel with respect to the both connections.

In this presentation we first discuss common properties of such submanifolds. In particular we next give an algebraic characterization of them in Jordan algebras and show their applications. Further, we exhibit that both curvature integrals induced from dually flat structure are interestingly related to an unexpected quantity, i.e., iteration-complexity of the interior-point algorithms for convex optimization defined on a submanifold that is *not* doubly autoparallel.

This is a joint work with Hideyuki Ishi and Takashi Tsuchiya.

References

[1] A. Ohara, H. Ishi and T. Tsuchiya, Doubly autoparallel structure and curvature integrals: An application to iteration complexity for solving convex programs, *Information Geometry*, 2023. https://doi.org/10.1007/s41884-023-00116-x (Open Access)

Doubly autoparallel structure and curvature integrals

- An application to iteration-complexity analysis of convex optimization -

Atsumi Ohara University of Fukui

Joint work with
H. Ishi (Osaka Metropolitan Univ.),
T. Tsuchiya (GRIPS)

Statistical Theories and Machine Learning Using Geometric Methods
December 14-15, 2023 @Osaka Metropolitan Univ.

1

Introduction

- Doubly autoparallel (DA) submanifold
  - natural notion in information geometry with symmetry
  - sometimes appears but much attention has not been paid
  - important applications in statistics and optimization
    - convex optimization on symmetric cones (e.g. SemiDef. Prog.)
    - MLE of structured covariance matrices
    - means on symmetric cones
    - and so on.

2

Introduction

- A goal of the presentation is to demonstrate
  - Characterization of DA submfds in a symmetric cone
  - If the feasible region of SDP is DA

    ➔ an explicit formula for the optimal solutions
  - If the feasible region of SDP is not DA

    ➔ curvature integral evaluates an iteration-complexity to obtain the optimal solutions with IP algorithm

3

Outline

- Doubly autoparallel submanifolds
- Preliminaries: Dually flat structure on a symmetric cone
- Characterization of DA submfds in a symmetric cone
  - Applications to

    - SemiDef. Prog. (SDP) - MLE for structured cov. mtx.

    ➔ explicit formulas for an optimal solution for them
- Non DA case: curvature integrals
  - Iteration-complexity analysis for conic linear program (an extension of SDP) via an Interior-Point algorithm
  - Application to primal-dual path following methods
- Concluding remark

4

Information geometry [Amari & Nagaoka 00]

Def. Statistical manifold: $(\mathcal{S}, g, \nabla, \nabla^*)$

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$$

X, Y and Z : arbitrary vector fields on $\mathcal{S}$

★ g : Riemannian metric

★ (∇, ∇^*) : torsion-free affine connections

$R^{\nabla} = 0,\ R^{\nabla^*} = 0$ ➡ dually flat

★ $\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*$: α-connections

5

Doubly autoparallel submanifolds

[Uohashi&O 04] [OIT23]

- Def. Let (S, g, ∇, ∇^*) be a statistical manifold and M be its submanifold. We call M a doubly autoparallel submanifold in S when the followings hold:

  - $\forall X, Y \in \mathcal{X}(M),\ \nabla_X Y \in \mathcal{X}(M)$

    i.e. $H_M(X,Y) = 0$

  - $\forall X, Y \in \mathcal{X}(M),\ \nabla^*_X Y \in \mathcal{X}(M)$

    i.e. $H^*_M(X,Y) = 0$

6

Important Properties

Proposition 1 The following statements are equivalent:

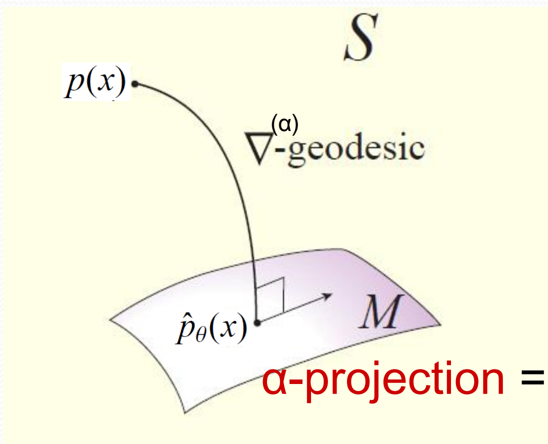
- 1) A submanifold M is doubly autoparallel (DA)
- 2) M is autoparallel w.r.t. the α-connections

$$\nabla^{(\alpha)} = \{(1+\alpha)\nabla + (1-\alpha)\nabla^*\}/2$$

  for two different α's.
- 3) M is autoparallel w.r.t. all the α-connections.
- 4) all the α-geodesics connecting two points on M lay in M (if it is simply connected).
- 5) M is affinely constrained in both ∇- and ∇^*-affine coordinates if S is dually flat.

Furthermore, for a parametric statistical model S

- If M is DA in S, then α-projections from p to M are unique for all α.



M is simultaneously an exponential and mixture family

α-projection =arg min $D^{(\alpha)}(p, M)$

Related topics and applications

Symmetric cones

- MLE for structured covariance matrices is tractable (cast to convex program: inversely linear structure) [Anderson 70, Malley 94]
- Explicitly solvable Semi-Definite Programs [O 99]
- Structure of α-power means on symmetric cones [O 04]

9

Related topics and applications

Probability simplex

- Statistical models Markov-isomorphic to the probability simplex [Nagaoka 17]
- Characterization and classification of DA submfds in the probability simplex via Hadamard algebra [O&Ishi 18]
- Learning theory [Mutus&Ay 03]

Miscellaneous

- The self-similar (*Barenblatt–Pattle*) solution for the porous medium equation [O&Wada 10]

General statistical manifolds

- Purely geometric study [Satoh *et al*. 21]

10

Preliminaries

[Faraut&Korani 94]

- Euclidian space E with an inner product $(.|.)$
- Symmetric cone Ω : open and convex in E
  - homogeneous

    $G(\Omega) = \{\tau \in GL(E) \mid \tau(\Omega) = \Omega\}$ acts transitively
  - self-dual w.r.t. an inner product of E

    $\Omega = \Omega^*, \quad \Omega^* = \{y \in E \mid (x|y) > 0, \forall x \in \overline{\Omega} \backslash \{0\}\}$
- Euclidean Jordan algebra $(V, *)$
  - commutative
  - $x^2 * (x * y) = x * (x^2 * y), \quad$ where $x^2 = x * x$
  - $\exists$ associative inner-product $(x * y|z) = (y|x * z)$ on V

Prop. $\Omega = \mathrm{int}\{x^2 \mid x \in V\}$ is a symmetric cone in $V(=E)$.

11

- $L(x) : V \to V \quad L(x)y = x * y$
- $P(x, y) := L(x)L(y) + L(y)L(x) - L(x * y)$
- Mutation: $x \perp_a y := P(x, y)a$

  isomorphic to $*$, the unit element: a^{-1}

Ex. the set of real symmetric pos. def. matrices PD(n,$\mathbf{R}$)

$$V = \mathrm{Sym}(n; \mathbf{R}), \quad X * Y = (XY + YX)/2$$

$$\tau_G(X) = GXG^T, \quad G \in GL(n, \mathbf{R})$$

$(X|Y) = \mathrm{tr}(XY)$, the unit: I, the inverse: X^{-1}

$$X \perp_A Y = (XAY + YAX)/2$$

12

Dually flat structure on Ω

- Logarithmic characteristic function on Ω

$$\psi(x) := \log \int_{\Omega^*} e^{-\langle s,x\rangle} ds,$$

  - positive definite Hessian on Ω
  - $x^{-1} = -\mathrm{grad}\,\psi(x)$, $\quad (\mathrm{grad} f(x)|u) = D_u f(x)$

- a coordinate system (x^i) : $x = \sum_{i=1}^{n} x^i e_i$, $\{e_i\}_{i=1}^n$: a basis of E

- a dual coordinate system (s_i) :
  $x^{-1} = \sum_{i=1}^n s_i e^i$, $\{e^i\}_{i=1}^n$: a basis of E with $(e^i|e_j) = -\delta^i_j$

13

- D : the canonical flat affine connection on E
- $\{x^1, \cdots, x^n\}$: affine coordinate system, i.e., $D_{\frac{\partial}{\partial x^i}} \dfrac{\partial}{\partial x^j} = 0$
- g : Riemannian metric on Ω

$$g = Dd\psi = \sum_{i,j} \frac{\partial^2 \psi}{\partial x^i \partial x^j} dx^i dx^j.$$

- D': the dual affine connection on Ω

$$Xg(Y,Z) = g(D_X Y, Z) + g(Y, D'_X Z)$$

(g, D, D') : dually flat structure on Ω

14

Expression via Jordan algebra

Dually flat structure on Ω [Uohashi&O 04]

- Potential= log. char. func. $\psi(x) = -\log\det x$,
- Riemanian metric: $g_x\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = (P(x)^{-1}e_i | e_j)$, $P(x) := P(x, x)$
- α-connections: $\left(\nabla^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)_x = (\alpha - 1)(e_i \perp_{x^{-1}} e_j)$

Ex. On PD$(n,\mathbf{R})$

- $\psi(x)$ = -log det X, $(X = \sum_{i=1}^N x^i E_i)$
- $g_X\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \mathrm{tr}(X^{-1}E_i X^{-1} E_j)$
- $\left(\nabla^{(\alpha)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)_X = \frac{\alpha - 1}{2}(E_i X^{-1} E_j + E_j X^{-1} E_i)$

15

Characterization of DA submfds in Ω

Let W be a linear subspace in Jordan algebra $(V, *)$ and $p = q * q$ in a symmetric cone Ω.

Thm. [OIT23] The following 1)-3) are equivalent:

1) A submanifold $M = (W + p) \cap \Omega$ is DA, where
$$W + p = \{w + p \mid w \in W\}$$

2) For all x in M, $u \perp_{x^{-1}} v \in W$, $(u, v \in W)$

3) The subspace $P(q)^{-1}W$ is a Jordan subalgebra.

Rem. (a) 3) is able to be checked at the single point p

(b) $M = \{(W' + p^{-1}) \cap \Omega\}^{-1}$ with $W' = P(p)^{-1}W$

The proof is based on the property 5) in Prop. 1 (p.6)

16

(c) Implication: Classification of DA submflds in Ω reduces to that of Jordan subalgs of $(V, *)$.
(For V=Sym(n,**R**) cf.[Jacobson 87], [Malley 87])

- Ex. - Jordan subalgebras in Sym(n,**R**)
  1) fixed eigen vectors, 2) doubly symmetric, etc.
  - Two bases $\{E_i\}_{i=1}^m$ and $\{F^i\}_{i=1}^m$ of Sym(n,**R**)

$$\mathcal{M} = \{P \mid P = E_0 + \sum_{i=1}^m x^i E_i,\ \exists x = (x^i) \in \mathbf{R}^m\} \cap \mathrm{PD}(n)$$

$$\mathcal{M} = \{P \mid P^{-1} = F^0 + \sum_{i=1}^m s_i F^i,\ \exists s = (s_i) \in \mathbf{R}^m\} \cap \mathrm{PD}(n)$$

17

Application(1) Means on Positive Operators

[Kubo & Ando 80]

- Def. (Axioms of means)
  σ is a mean on self-conjugate positive operators
  - i) $A \le C,\ B \le D \Rightarrow A\sigma B \le C\sigma D$
  - ii) $C(A\sigma B)C = (CAC)\sigma(CBC)$
  - iii) $A_n \downarrow A,\ B_n \downarrow B \Rightarrow A_n\sigma B_n \downarrow A\sigma B$

  where $A_n \downarrow A \overset{\text{def}}{\Leftrightarrow} (A_i \ge A_{i+1}, \forall i)\ \&\ (A_n \to A)$
  - iv) $I\sigma I = I$

18

α-geodesics on PD(n)

- α-geodesic $P(s)$ boundary conds. : $P(0)=A$, $P(1)=B$

$$P^{(\alpha)}(s) = A^{1/2}\left\{[(A^{-1/2}BA^{-1/2})^{\alpha} - I]s + I\right\}^{1/\alpha} A^{1/2}$$

$$\begin{array}{ll} \alpha = 1 & P(s) = A + s(B - A) \\ \alpha = 0 & \hat{P}(s) = A^{1/2}\exp(s \log A^{-1/2}BA^{-1/2})A^{1/2} \\ \alpha = -1 & P^*(s) = \{A^{-1} + s(B^{-1} - A^{-1})\}^{-1} \end{array}$$

$$\begin{array}{lcl} A\mathrm{a}B & := & P(1/2) \\ A\mathrm{g}B & := & \hat{P}(1/2) \\ A\mathrm{h}B & := & P^*(1/2) \end{array}$$

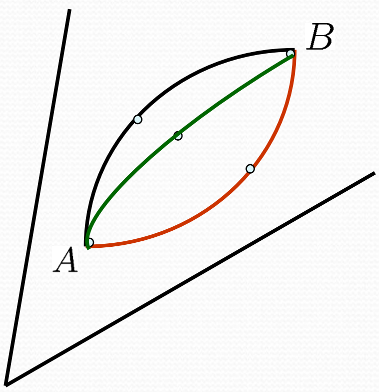
$P^{(\alpha)}(1/2)$: a power mean

19

Means and α-geodesics on PD(n) [O 04]

Thm. Points on α-geodesics for s in [0,1] and α in [-1,1] are 2-param. family of means, i.e.,

$$A\sigma_s^{(\alpha)}B = P^{(\alpha)}(s)$$



In particular, for fixed s in [0, 1]

$$P^{(\alpha)}(s) > P^{(\beta)}(s),$$
$$1 \geq \alpha > \beta \geq -1$$ AGH ineq. (s=1/2)

Cor. A and B are in a DA submanifold M

$$\Rightarrow \quad A\sigma_s^{(\alpha)}B \in M, s \in [0,1], \alpha \in [-1,1]$$

20

App.(2) MLE for structured covariance matrices

- a sample covariance S in PD$(n,\mathbf{R})$
- a zero-mean Gaussian p.d.f. with a covariance mtx. Σ

$$p(x) = (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp\{-\tfrac{1}{2}x^T\Sigma^{-1}x\}$$

- structured covariance mtx. (with linear constraints)

$$\Sigma \in \mathcal{M} = (E_0 + \mathcal{W}) \cap \mathrm{PD}(n,\mathbf{R})$$

- Ex.
  - Toeplitz matrices: $\{T = (t_{ij}) | t_{ij} = t_{ji} = y_{|i-j|}\}$
  - zero-patterns : $\{\Sigma = (\sigma_{ij}) | \sigma_{ij} = \sigma_{ji} = 0,\ (i,j) \in \mathcal{E}\}$
  - etc...

21

MLE for structured covariance matrices

- Negative logarithmic likelihood func (up to const.):

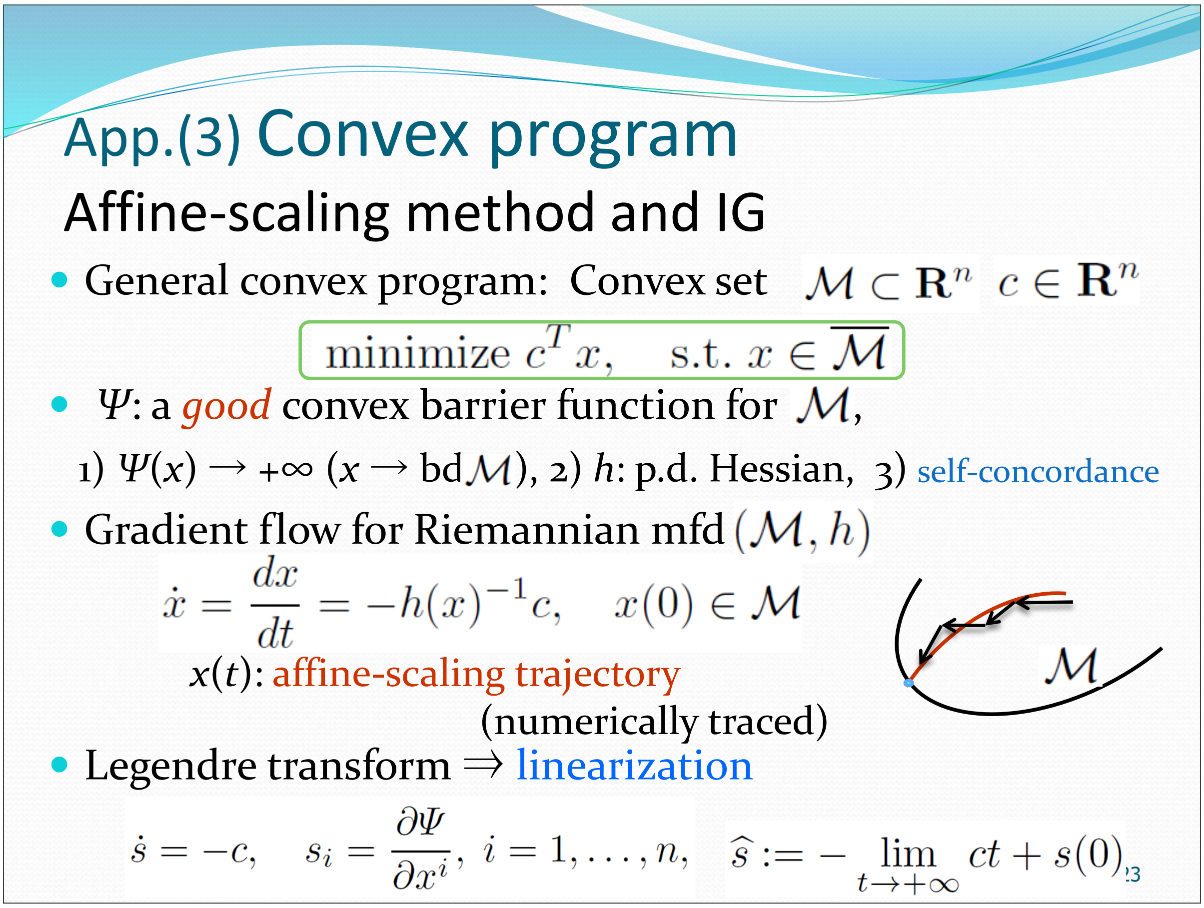
$$\ell(\Sigma) := -\log\det\Sigma^{-1} + \mathrm{tr}(\Sigma^{-1}S) \to \min$$

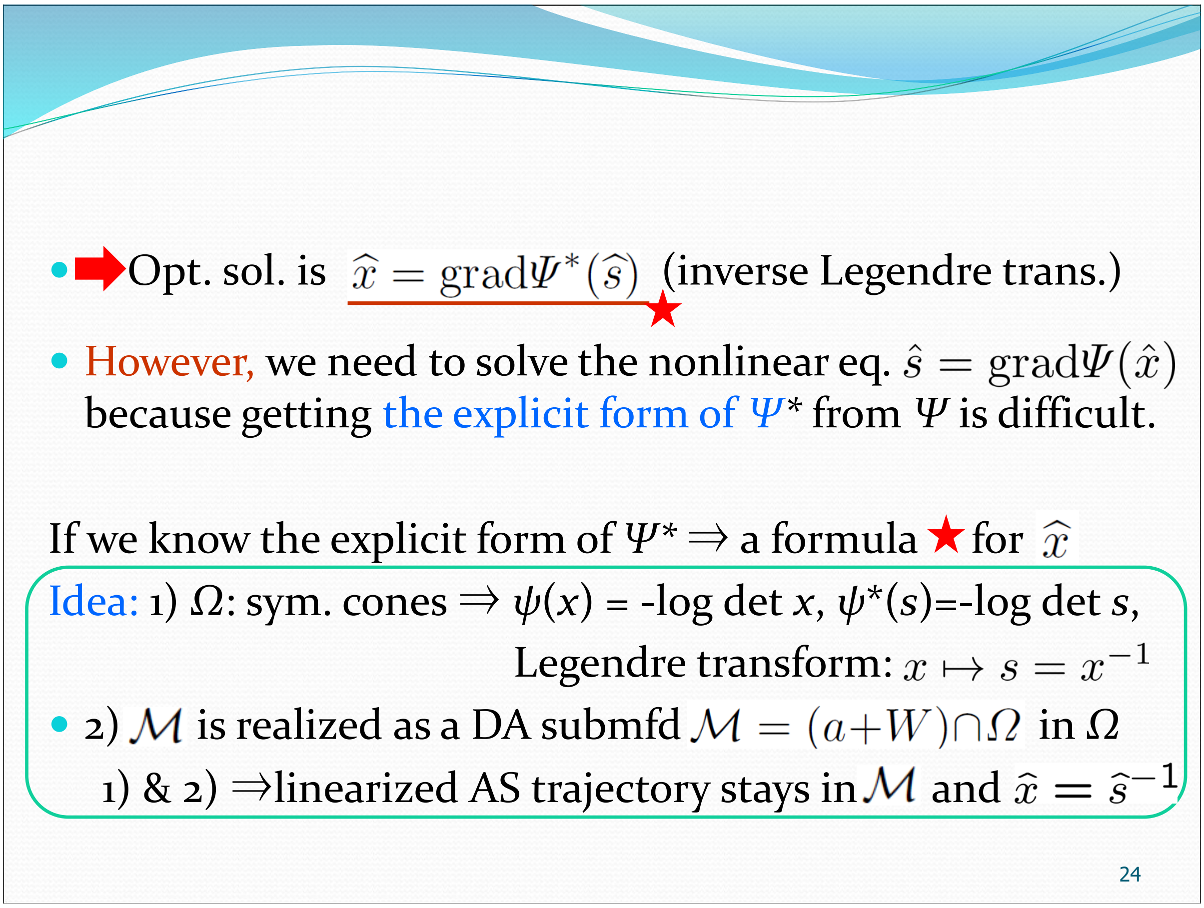
  - Rem Note that $-\log\det$ is a convex function.

- If $\mathcal{M}$ is DA (inversely linear structure), then the minimization problem of $\ell(\Sigma)$ (MLE) s.t. $\Sigma \in \mathcal{M}$ is a strictly convex program.

$\Rightarrow$ Unique solution if it exists,
Numerically tractable (optimality eq. is linear)

22


App.(3) Convex program
Affine-scaling method and IG
General convex program: Convex set $\mathcal{M} \subset \mathbf{R}^n$ $c \in \mathbf{R}^n$
minimize $c^T x$, s.t. $x \in \overline{\mathcal{M}}$
Ψ: a good convex barrier function for $\mathcal{M}$,
1) $\Psi(x) \to +\infty$ ($x \to$ bd$\mathcal{M}$), 2) h: p.d. Hessian, 3) self-concordance
Gradient flow for Riemannian mfd $(\mathcal{M}, h)$
$\dot{x} = \frac{dx}{dt} = -h(x)^{-1}c, \quad x(0) \in \mathcal{M}$
$\mathcal{M}$
$x(t)$: affine-scaling trajectory
(numerically traced)
Legendre transform $\Rightarrow$ linearization
$\dot{s} = -c, \quad s_i = \frac{\partial \Psi}{\partial x^i}, \; i = 1, \ldots, n, \quad \widehat{s} := -\lim_{t \to +\infty} ct + s(0)$
23


Opt. sol. is $\widehat{x} = \mathrm{grad}\Psi^*(\widehat{s})$ (inverse Legendre trans.)
However, we need to solve the nonlinear eq. $\hat{s} = \mathrm{grad}\Psi(\hat{x})$ because getting the explicit form of Ψ^* from Ψ is difficult.
If we know the explicit form of $\Psi^* \Rightarrow$ a formula ★ for $\widehat{x}$
Idea: 1) Ω: sym. cones $\Rightarrow$ $\psi(x)$ = -log det x, $\psi^*(s)$=-log det s,
Legendre transform: $x \mapsto s = x^{-1}$
2) $\mathcal{M}$ is realized as a DA submfd $\mathcal{M} = (a+W) \cap \Omega$ in Ω
1) & 2) $\Rightarrow$linearized AS trajectory stays in $\mathcal{M}$ and $\widehat{x} = \widehat{s}^{-1}$
24

- Typical Example: SemiDefinite Program (SDP)

$$\underset{P}{\text{minimize}}\ (C|P),\ \text{s.t. } P = E_0 + \sum_{i=1}^{m} x^i E_i \in \overline{\mathcal{M}} = \overline{(E_0 + \mathcal{W}) \cap \mathrm{PD}(n)}$$

- If $\mathcal{M}$ is DA in PD(n) and $P \in \mathcal{M}$
  - 1. Set $F^0 = P^{-1}, F^i = -P^{-1}E_iF^{-1}$, then

$$\mathcal{M} = \{P \,|\, P^{-1} = F^0 + \sum_{i=1}^{m} s_i F^i,\ \exists s = (s_i) \in \mathbf{R}^m\} \cap \mathrm{PD}(n)$$

  - 2. Solve $\widetilde{C} \in \mathrm{span}\{F^i\}_{i=1}^m$ meeting

$$\forall P \in \mathcal{M}, \quad (C|P) = (\widetilde{C}|P) + \text{const.}$$

25

  - 3. Spectral decomposition

$$\widetilde{C} = (V_1\ V_2)\begin{pmatrix} \Sigma_1 & O \\ O & O \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} = V_1 \Sigma_1 V_1^T$$

  - 4. For $\forall P_0 \in \mathcal{M}$ with $S_0 = P_0^{-1}$, the opt. sol. is

$$\widehat{P} = \lim_{t\to\infty} S(t)^{-1} = \lim_{t\to\infty}(-\widetilde{C}t + S_0)^{-1} = P_0 - P_0V_1(V_1^TP_0V_1)^{-1}V_1^TP_0$$

Rem. Independent of the objective function $(C|P)$ and an initial value P_0

26

Interior point method (IP) for Conic linear program

Notation :

- Vector space E of dimension n
- the dual vector space E^*
- $\langle s, x \rangle$: Paring $\quad E^* \ni s \quad E \ni x$
- Ω : proper open convex cone in E
- Ω^* : the dual cone of Ω
  $$\Omega^* := \{s \in E^* | \langle s, x \rangle > 0, \forall x \in \overline{\Omega} \backslash \{0\}\}$$
- T^*: (Orthogonal) dual subspace of $T \subset E$
  $$T^* = \{s \in E^* | \langle s, x \rangle = 0, \ \forall x \in T\}$$

27

Conic Linear Program

Given

$$c \in E^*, \ f \in E \text{ and } T \subset E$$

- Primal problem

$$\text{(P)} \quad \begin{aligned} &\text{minimize } \langle c, x \rangle, \text{ s.t. } x \in \overline{\mathcal{P}}, \\ &\qquad\qquad \text{where } \mathcal{P} := (f + T) \cap \Omega, \end{aligned}$$

- Dual problem

$$\text{(D)} \quad \begin{aligned} &\text{maximize } \langle s, f \rangle, \text{ s.t. } s \in \overline{\mathcal{D}}, \\ &\qquad\qquad \text{where } \mathcal{D} := (c + T^*) \cap \Omega^*. \end{aligned}$$

28

Typical Examples

- Linear program (LP):

$$E = E^* = \mathbf{R}^n,\ \Omega = \Omega^* = \mathbf{R}^n_{++}$$

- Semidefinite program (SDP):

  $E=E^*$: the set of real symmetric matrices

  $\Omega = \Omega^*$: the set of positive definite matrices

- Second order cone (Lorentz cone) program (SOCP)
- Mixture of the aboves

29

ϑ-normal barrier on an open convex cone Ω

[Nesterov & Nemirovski 94]

- Def. θ-normal barrier ψ on Ω
  - A (smooth) convex function ψ satisfying, at each x in Ω,

    1) $\psi(tx) = \psi(x) - \vartheta \log t,$

    2) $|(D^2 d\psi)_x(X,X,X)| \leq 2((Dd\psi)_x(X,X))^{3/2}$ self-concordance

    for $\vartheta \geq 1,\ \forall t > 0$ and $\forall X \in T_x\Omega \cong E$

    3) $\psi(x) \to +\infty$ ($x \to$ bd Ω),

Rem. (a) Existence for all Ω, but not with explicit forms

(symmetric cones ➔ Yes, log. char. func.)

(b) the Hessian is positive definite

(c) Self-concordance ⇒the Newton method is efficient

- Ex. $\psi(x) = -\sum_{i=1}^{n} \log x^i$ (LP), $\quad \psi(x) = -\log\det X$ (SDP)

Dually flat structure on Ω (revisited)

- D : the canonical flat affine connection on E
- $\{x^1, \cdots, x^n\}$: affine coordinate system, i.e., $D_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = 0$
- g : Riemannian metric on Ω

$$g = Dd\psi = \sum_{i,j} \frac{\partial^2 \psi}{\partial x^i \partial x^j} dx^i dx^j .$$

- D': the dual affine connection on Ω

$$Xg(Y, Z) = g(D_X Y, Z) + g(Y, D'_X Z)$$

(g, D, D') : dually flat structure on Ω

31

Remark

- $\{s_1, \cdots, s_n\}$: dual coordinate system on E^* , s.t.

$$\langle s, x\rangle = \sum_i s_i(s) x^i(x)$$

- Gradient map $\iota : \Omega \to \Omega^*$ defined by

$$s_i \circ \iota = -\frac{\partial \psi}{\partial x^i}$$

Legendre transform

induces dually flat structure on Ω^* from (g, D, D')

(1) D^* : the canonical flat affine connection on E^*

$$D^*_{\iota_*(X)} \iota_*(Y) = \iota_*(D'_X Y) \qquad (\iota^* D^* = D')$$

D^* -autoparallel in Ω^* $\Longleftrightarrow$ D'-autoparallel in Ω

32

Remark

(2) Riemannian metric $g^* := D^* d\psi^*$ on Ω^*

$$g = \iota^* g^*$$

(3) $\langle \iota_*(X), Y \rangle = -g_x(X, Y)$

Hessian norm : We denote the length of X in $T_x\Omega \cong E$ by

$$\|X\|_x := \|Z\|_s := \sqrt{g_x(X,X)} = \sqrt{g_s^*(Z,Z)},$$

where $s = \iota(x)$ and $Z = \iota_*(X)$.

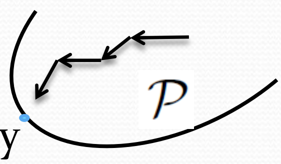
33

Curvature integral and iteration-complexity of IP

One of important computational performance indices for optimization algorithms is the iteration-complexity.

- Ω: sym. cone and $\mathcal{P} := (f+T) \cap \Omega$ is DA

  $\Rightarrow$ iteration-complexity=0 for (P)



- General case? Iter.-comp. is characterized by
  - Curvature integrals along the central trajectory $\gamma_{\mathcal{P}}$

$$\int_{t_1}^{t_2} \|H_{\mathcal{P}}^*(\dot\gamma_{\mathcal{P}}(t), \dot\gamma_{\mathcal{P}}(t))\|_{\gamma_{\mathcal{P}}(t)}^{1/2} dt$$

  - Similarly, for (D) curvature integrals along the dual c. t. $\gamma_{\mathcal{D}}$

$$\int_{t_1}^{t_2} \|H_{\mathcal{D}}(\dot\gamma_{\mathcal{D}}(t), \dot\gamma_{\mathcal{D}}(t))\|_{\gamma_{\mathcal{D}}(t)}^{1/2} dt$$

34

Central trajectory (= the special A-S trajectory)

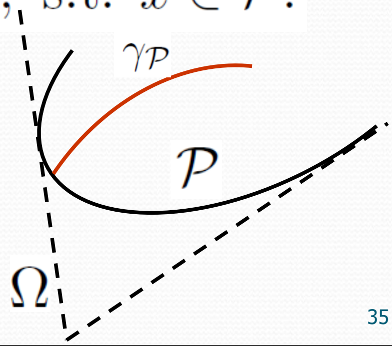
- Primal problem: minimize $\langle c, x\rangle$, s.t. $x \in \overline{\mathcal{P}}$,
  where $\mathcal{P} := (f + T) \cap \Omega$,

- $x(t) := \gamma_{\mathcal{P}}(t)$: the unique minimizer of

$$\text{minimize } \underline{\underline{t}}\langle c, x\rangle + \psi(x), \text{ s.t. } x \in \overline{\mathcal{P}}.$$

  for each $t > 0$

- $\gamma_{\mathcal{P}} := \{\gamma_{\mathcal{P}}(t) | t > 0\}$:
  (Primal) central trajectory



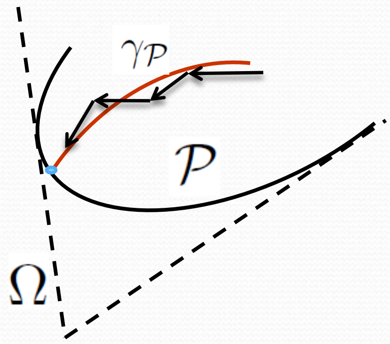
35

Central trajectory $\gamma_{\mathcal{P}}$

- Homotopy path to the opt. sol. of the primal problem, i.e., $x(t)$ converges when $t \to \infty$.
- Numerically tracing $\gamma_{\mathcal{P}}$ is the standard and efficient way to solve the primal problem.

  Path-following method

Idea: consider the problem in the dual cone Ω^* in order to relate the complexity with the curvature



36

(1)Representation of feasible region

- a linear surj. operator $A : E \to \mathbf{R}^m$ s.t. Ker A =T

$$\mathcal{P} = \{x \in \Omega | Ax = b\},$$
$$\mathcal{D} = \{s \in \Omega^* | s = c - A^* y,\ y \in \mathbf{R}^m\}$$

where $A^* : \mathbf{R}^m \to E^*$ satisfying $y^T(Ax) = \langle A^* y, x\rangle$,
$b := Af \in \mathbf{R}^m$

- dim $\mathcal{P}$ =*n*-*m*, dim $\mathcal{D}$ =*m*

$\mathcal{P}$ is *D*-autoparallel and $\mathcal{D}$ is *D**-autoparallel

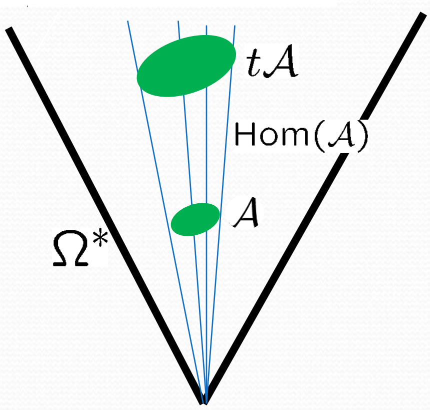
37

(2)Homogenization (conic hull)

- homogenization of $\mathcal{D}$ in Ω^*

$$\mathrm{Hom}(\mathcal{D}) := \bigcup_{t>0} t\mathcal{D},\ t\mathcal{D} := \{s \in \Omega^* | s = t\tilde{s},\ \tilde{s} \in \mathcal{D}\}$$

- *D**-autoparallel (because $\mathcal{D}$ is.)
- dim $\mathrm{Hom}(\mathcal{D})$=*m*+1



Homogenization

38

Lemma

The following relations hold in Ω^*:

$$\iota(\gamma_{\mathcal{P}}) = \iota(\mathcal{P}) \cap \mathrm{Hom}(\mathcal{D})$$

$$s(t) := \iota(x(t)) = \iota(\mathcal{P}) \cap t\mathcal{D}$$

proved using the Lagrange function

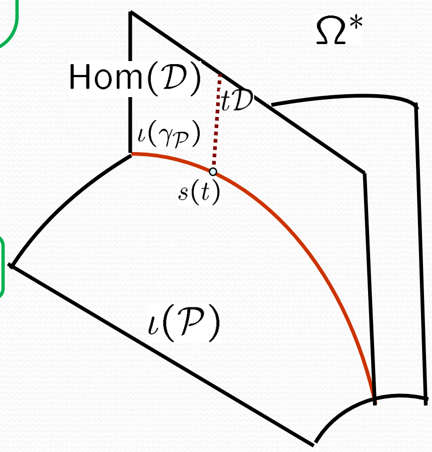
$$L(x,y) := t\langle c, x\rangle + \psi(x) + y^T(b - Ax)$$

$$\partial L/\partial x = 0 \ \rightarrow s \in t\mathcal{D}$$

$\mathcal{P}$ is DA ➔ $\iota(\gamma_{\mathcal{P}})$ is D^*-autoparallel

Remark

$\iota(\mathcal{P})$ and $t\mathcal{D}$ are orthogonal w.r.t. g^* at $s(t)$ by definition.



39

3. Geometric predictor-corrector algorithm (tracing $\gamma_{\mathcal{P}}$ in $\mathrm{Hom}(\mathcal{D})$)

Ideal case

- Predictor

  From $s(t) \in \iota(\gamma_{\mathcal{P}})$

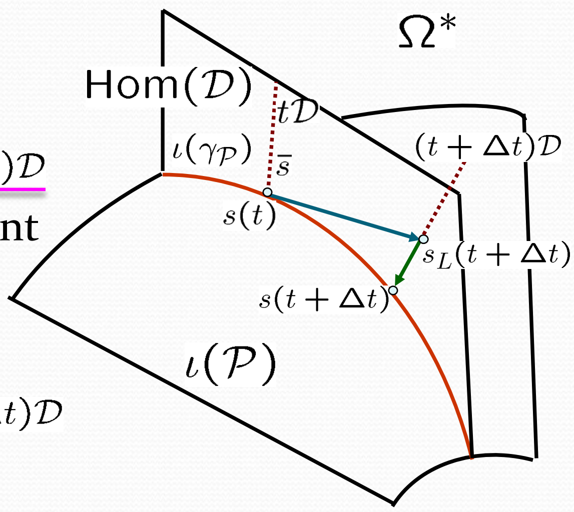
  to $s_L(t+\Delta t) \in (t+\Delta t)\mathcal{D}$

  with the direction tangent to $\iota(\gamma_{\mathcal{P}})$

- Corrector

  From $s_L(t+\Delta t) \in (t+\Delta t)\mathcal{D}$

  to $s(t+\Delta t) \in \iota(\gamma_{\mathcal{P}})$



40

Intuitive observation

- $H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}(t), \dot{\gamma}_{\mathcal{P}}(t))$: the Euler-Schouten embedding curvature (second fundamental form) of $\iota(\gamma_{\mathcal{P}})$with respect to D^*

- If $H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}(t), \dot{\gamma}_{\mathcal{P}}(t))$ is small at t, so is expected the iteration number !?

Actually,

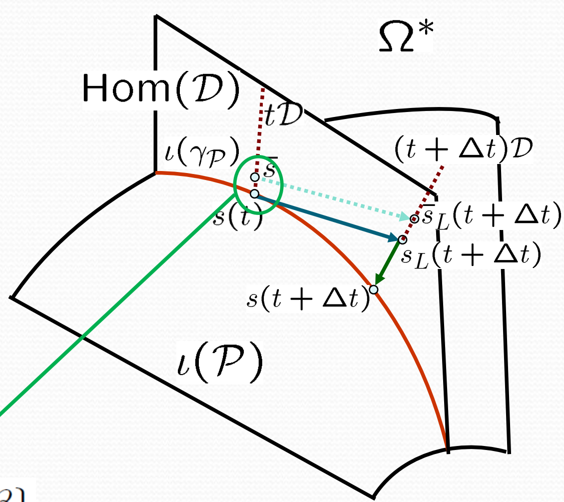
$$\boxed{\ddot{s} = D^*_{\dot{s}}\dot{s} = \iota_*(H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}, \dot{\gamma}_{\mathcal{P}}))}$$

41

Remark: practical case

- Cannot expect that the corrector returns precisely on $\iota(\gamma_{\mathcal{P}})$

- Consider the point $\bar{s}$ in the neighborhood of $s(t) \in \iota(\gamma_{\mathcal{P}})$ in the sense of Riemannian metric



$$\mathcal{N}_t(\beta) := \{s \in t\mathcal{D} | \delta(s) \le \beta\}$$

42

Predictor

- The differential equation expressing $\iota(\gamma_{\mathcal{P}})$:

$$\dot{s} = (\mathrm{id} - \Pi_s^{\perp})c = \frac{1}{t}(\mathrm{id} - \Pi_s^{\perp})s$$

  where $\Pi_s^{\perp}$ is the orthogonal projection w.r.t. g^* from E^* to $T^* = \mathrm{Range} A^*$ at s.

- Hence, the predictor is defined by

$$\bar{s}_L(t + \Delta t) := \bar{s} + \Delta t(I - \Pi_{\bar{s}}^{\perp})c \in (t + \Delta t)\mathcal{D}$$

43

Corrector

- Reduces to the following convex optimization on $t\overline{\mathcal{D}}$:

$$\text{minimize } F(s) := \langle s, f \rangle + \psi^*(s), \text{ s.t. } s \in t\overline{\mathcal{D}}$$

- Newton direction N for this opt. problem:

$$D^* dF(X, N) = -dF(X), \ \forall X \in \mathcal{X}(t\mathcal{D})$$

  - Newton decrement: measure of approximation of s to the optimal sol. s^*

$$\delta(s) := \|N\|_s$$

- We define the corrector with a single Newton step by:

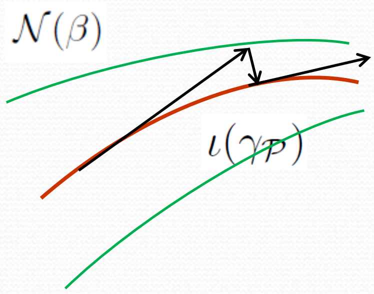
$$\bar{s}_L^+(t + \Delta t) := \bar{s}_L(t + \Delta t) + N_{\bar{s}_L(t+\Delta t)}$$

44

Tubular neighborhood

- The standard analysis technique in IP ensures the polynomiality of the complexity for this path-following strategy if all the generated points are near to $\iota(\gamma_{\mathcal{P}})$.

- Introduce the tubular neighborhood $\mathcal{N}(\beta)$ of $\iota(\gamma_{\mathcal{P}})$



$$\mathcal{N}(\beta) := \bigcup_{t \in (0,\infty)} \mathcal{N}_t(\beta),$$

where $\mathcal{N}_t(\beta) := \{ s \in t\mathcal{D} | \delta(s) \leq \beta \}$.

45

4. Curvature integral and asymptotic iteration-complexity (Main result)

- Assumption: $\iota(\gamma_{\mathcal{P}})$ is not D^*-autoparallel(i.e. $\mathcal{P}$ is not DA)

  In this case, $\beta \to 0$ implies that $\Delta t \to 0$

- Theorem

  For $0 < t_1 < t_2$ and $s_1 \in \mathcal{N}(\beta) \cap t_1\mathcal{D}$, let $\sharp(s_1, t_2, \beta)$ be the iteration number to find $s_2 \in \mathcal{N}(\beta) \cap t_2\mathcal{D}$. Then,

$$\lim_{\beta \to 0} \frac{\sqrt{\beta} \times \sharp(s_1, t_2, \beta)}{I_{\mathcal{P}}(t_1, t_2)} = 1,$$

where

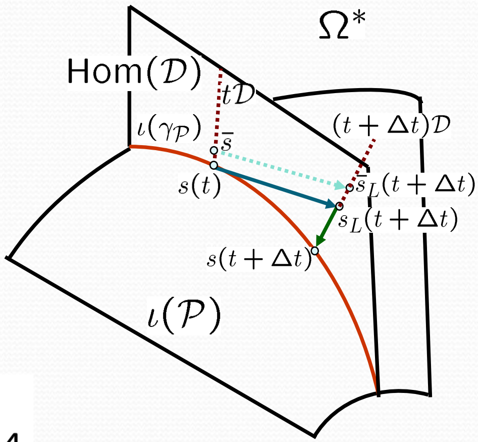
$$I_{\mathcal{P}}(t_1, t_2) := \frac{1}{\sqrt{2}} \int_{t_1}^{t_2} \| H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}(t), \dot{\gamma}_{\mathcal{P}}(t)) \|^{1/2}_{\gamma_{\mathcal{P}}(t)} dt.$$

46

Outline of the proof

- Evaluate the Newton dec. of the predictor $\bar{s}_L(t+\Delta t)$ by $\|\ddot{s}(t)\|_{s(t)}$ (For each iteration)

$$\delta(\bar{s}_L(t+\Delta t))$$
$$= \|s(t+\Delta t) - \bar{s}_L(t+\Delta t)\|_{\bar{s}_L(t+\Delta t)} + r_4$$
$$= \frac{(\Delta t)^2}{2}\underline{\|\ddot{s}(t)\|}_{s(t)} + \delta(\bar{s}) + r_1 + r_2 + r_3,$$



Rem. By the assumption

When $\beta \to 0$,

$\delta(\bar{s}) \to 0, r_i \to 0$ for $i = 1, \ldots, 4$,

47

Outline of the proof

- Self-concordance implies the following two inequalities:

Intermediate two relations for sufficiently small Δt and β.

(For each iteration)

- $$\sqrt{(1-\eta)\beta}(1 - O(\sqrt{\beta})) \leq \sqrt{w} - \sqrt{M_3}\delta(\bar{s}) \leq \frac{\Delta t}{\sqrt{2}}\|\ddot{s}(t)\|^{1/2}_{s(t)} + \sqrt{|r_1|} + \sqrt{M_3}(\Delta t)^2,$$
- $$\frac{\Delta t}{\sqrt{2}}\|\ddot{s}(t)\|^{1/2}_{s(t)} - \sqrt{|r_1|} - \sqrt{M_3}(\Delta t)^2 \leq \sqrt{w} + \sqrt{M_3}\delta(\bar{s}) \leq \sqrt{\beta}(1 + O(\sqrt{\beta}))$$

48

Outline of the proof

- Take summations of iterations
  - $$\sqrt{(1-\eta)\beta} \sum_{k=1}^{\sharp(s_1,t_2,\beta)} (1 - O(\sqrt{\beta})) \leq \frac{1}{\sqrt{2}} \int_{t_1}^{t_2} \|\ddot{s}(t)\|_{s(t)}^{1/2}\, dt + M'\sqrt{\Delta t_{\max}},$$
  - $$\frac{1}{\sqrt{2}} \int_{t_1}^{t_2} \|\ddot{s}(t)\|_{s(t)}^{1/2}\, dt - M'\sqrt{\Delta t_{\max}} \leq \sqrt{\beta} \sum_{k=1}^{\sharp(s_1,t_2,\beta)} (1 + O(\sqrt{\beta}))$$
- Recall $\boxed{\ddot{s} = D^*_{\dot{s}}\dot{s} = \iota_*(H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}, \dot{\gamma}_{\mathcal{P}}))}$

49

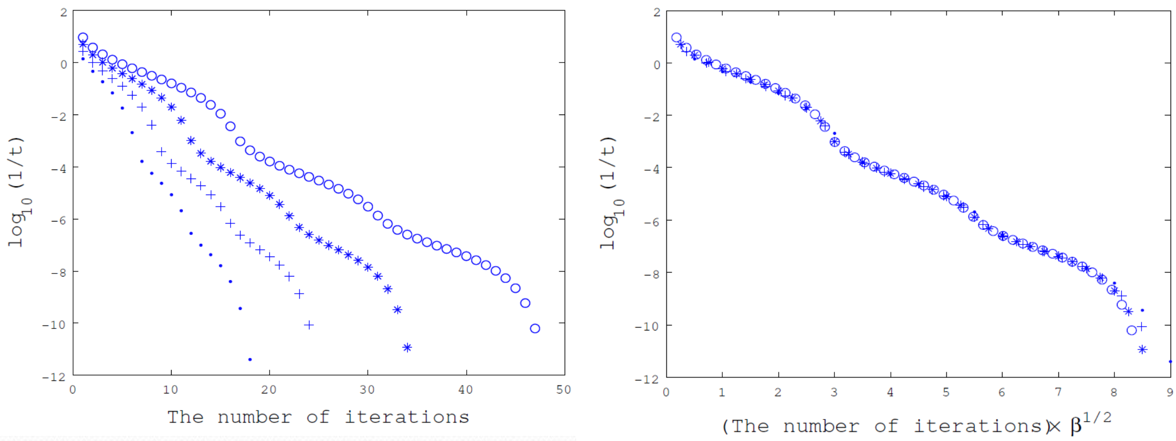
Remark

- An asymptotic result for $\beta \to 0$ (and hence, $\Delta t \to 0$)
  - $\mathcal{P}$ is DA $\Rightarrow \iota(\gamma_{\mathcal{P}})$ is DA (D^*-autoparallel) $\Rightarrow$ $\Delta t \to \infty$ $\Rightarrow$ explicit sol.
- The same argument holds for the dual problem.
- The results are valid for general convex cones

50

Numerical experiment

- Curvature structure of CT for a certain LP



$$(\cdot : \beta = 1/4,\ + : \beta = 1/8,\ * : \beta = 1/16,\ \circ : \beta = 1/32)$$

51

Curved part is Straight and Straight part is Curved?(1)

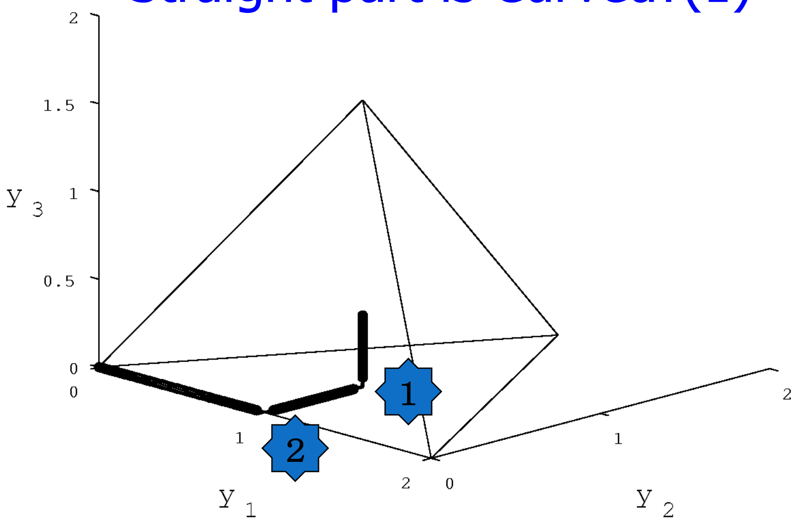


Figure 7: Exaple Figure

52

Proposition

It holds that $\|H^*_{\mathcal{P}}(\dot{\gamma}_{\mathcal{P}}, \dot{\gamma}_{\mathcal{P}})\|^{1/2}_{\gamma_{\mathcal{P}}(t)} \leq \dfrac{\sqrt{2\vartheta}}{t}$

ϑ : a constant determined by $\psi(x)$

- Remark

  The above proposition gives the upper bound:

$$I_{\mathcal{P}}(t_1, t_2) \leq \sqrt{\vartheta}\log(t_2/t_1)$$

53

Further study for LP case

- Primal and Dual Linear Program:

$$\min\ c^T x$$

$$\text{s.t.}\quad Ax = b,\quad x \geq 0,\ A \in R^{m\times n}, b \in R^n$$

$$\max\ b^T y$$

$$\text{s.t.}\quad s = c - A^T y,\quad s \geq 0,$$

54

Application to Primal-dual path-following (PDPF) method

- current main-stream IP (cheap in each iteration)
- The following quantity has been known to play an important and similar role in complexity analysis of PDPF method:

$$I_{PD}(t_1, t_2) = \int_{t_1}^{t_2} h_{PD}(t)^{1/2} dt$$

where $h_{PD}(t)$ is given by

$$h_{PD}(t) := \frac{1}{t^2}((I_n - Q(t))e) * (Q(t)e).$$

e: the unit element of Jordan product $*$

$Q(t)$: a certain projection matrix

55

Proposition

It holds that

$$h_{PD}(t)^2 = \left(\frac{1}{2}\|H^*_{\mathcal{P}}(\dot\gamma_{\mathcal{P}}(t), \dot\gamma_{\mathcal{P}}(t))\|_{\gamma_{\mathcal{P}}(t)}\right)^2 + \left(\frac{1}{2}\|H_{\mathcal{D}}(\dot\gamma_{\mathcal{D}}(t), \dot\gamma_{\mathcal{D}}(t))\|_{\gamma_{\mathcal{D}}(t)}\right)^2$$

Remark :

- geometric implication of the quantity of $I_{PD}(t_1, t_2)$
- inequalities

$$\max\{I_{\mathcal{P}}(t_1, t_2), I_{\mathcal{D}}(t_1, t_2)\} \le I_{PD}(t_1, t_2) \le I_{\mathcal{P}}(t_1, t_2) + I_{\mathcal{D}}(t_1, t_2).$$

56

Concluding Remark

- Tractable characterization of DA submfds in symmetric cones Ω
- Application to conic linear programs
  - Explicit sol. when the feasible region M is DA in Ω.
    - M is DA $\Rightarrow$ AS (CT) traj. is DA (D^*-autoparallel) $\Rightarrow$ $\Delta t \to \infty$ $\Rightarrow$ explicit sol.
- Extension: # of iterations and curvature integral of CT
  - Asymptotic analysis ($\beta \to 0$)
    - Complemented by numerical experiment for finite β
  - Geometric structure of CT has an influence on complexity of the IP algorithm

57

- Relation among iteration-complexities of primal-, dual- and primal-dual algorithms.
- DA submanifolds in the set of invertible elements of Jordan algebras [OIT23]
- Future work: Geometrical study for general stat. mfd.
  - Various geometrical concepts for mutually dual connections and their characterizations (Furuhata *et al.*)
  - Classifications
  - Families of continuous probability densities
  - Applications (Ex. Study of ODE's on manifolds?)

58

Thank you for your attention

References:

[OIT23] A. Ohara, H. Ishi and T. Tsuchiya,
Doubly autoparallel structure and curvature integrals: Applications to iteration complexity for solving convex programs,
Information Geometry (2023),
https://doi.org/10.1007/s41884-023-00116-x (Open Access).

59

Uncovering Data Symmetries: Estimating Covariance Matrix in High-Dimensional Setting With 'gips' R Package

Adam Przemysław Chojecki[1]
Hideyuki Ishi[2]

[1]Warsaw University of Technology (Poland)
[2]Osaka Metropolitan University

In high-dimensional settings, where the number of variables exceeds the number of observations, accurately estimating the covariance matrix poses a significant challenge. This talk presents a novel approach that leverages the identification of symmetries within the data to improve covariance matrix estimation. In the 'gips' R package [2], we implement the Bayesian model selection procedure within Gaussian vectors (invariant under the permutation group) introduced in [1].

Our method aims to capture the underlying low-dimensional structure by exploring the permutation symmetries within the data. Identifying symmetries enables us to interpret relationships in data in a new and natural way. The 'gips' package provides a comprehensive set of functions that facilitate identifying and utilizing symmetries, making it a valuable resource for researchers working with high-dimensional data.

We demonstrate the effectiveness of our approach through simulations and real-world data examples. Our results show that incorporating data symmetries leads to more reliable covariance matrix estimates, enabling better inference and decision-making. More results can be found in [3].

The presented novel approach contributes to the growing field of statistical methods for $p > n$, offering promising avenues for future research and practical applications.

References

[1] P. Graczyk, H. Ishi, B. Kołodziejek and H. Massam., *Model selection in the space of Gaussian models invariant by symmetry*, The Annals of Statistics, 2022

[2] A. Chojecki, P. Morgen, B. Kołodziejek, *gips: Gaussian Model Invariant by Permutation Symmetry*, CRAN, 2022, `https://CRAN.R-project.org/package=gips`, `https://przechoj.github.io/gips/`

[3] A. Chojecki, P. Morgen and B. Kołodziejek, *Learning permutation symmetries with gips in R*, `https://arxiv.org/abs/2307.00790`

Uncovering Data Symmetries: Estimating Covariance Matrix in High-Dimensional Setting With 'gips' R Package

Adam Przemysław Chojecki

MiNI PW

14.12.2023

What is gips?

gips

lifecycle stable | CRAN 1.2.1 | R-CMD-check passing | codecov 96%

gips - Gaussian model Invariant by Permutation Symmetry

`gips` is an R package that looks for permutation symmetries in the multivariate Gaussian sample. Such symmetries reduce the free parameters in the unknown covariance matrix. This is especially useful when the number of variables is substantially larger than the number of observations.

`gips` will help you with two things:

1. Finding hidden symmetries between the variables. `gips` can be used as an exploratory tool for searching the space of permutation symmetries of the Gaussian vector. Useful in the Exploratory Data Analysis (EDA).
2. Covariance estimation. The Maximum Likelihood Estimator (MLE) for the covariance matrix is known to exist if and only if the number of variables is less or equal to the number of observations. Additional knowledge of symmetries significantly weakens this requirement. Moreover, the reduction of model dimension brings the advantage in terms of precision of covariance estimation.

Installation

From CRAN:

```
# Install the released version from CRAN:
install.packages("gips")
```

Figure: Documentation of our R package, gips, https://przechoj.github.io/gips/

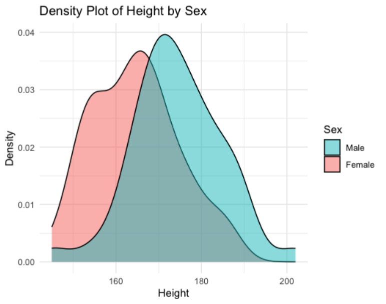
Inspiring citation

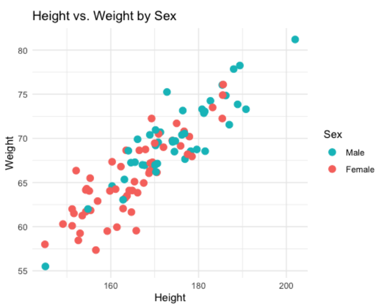
Life is the art of drawing sufficient conclusions from insufficient premises.

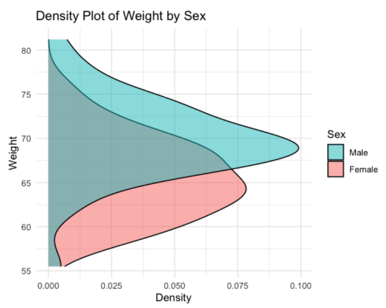
— Samuel Butler

Presentation plan

Example of conditional Independence





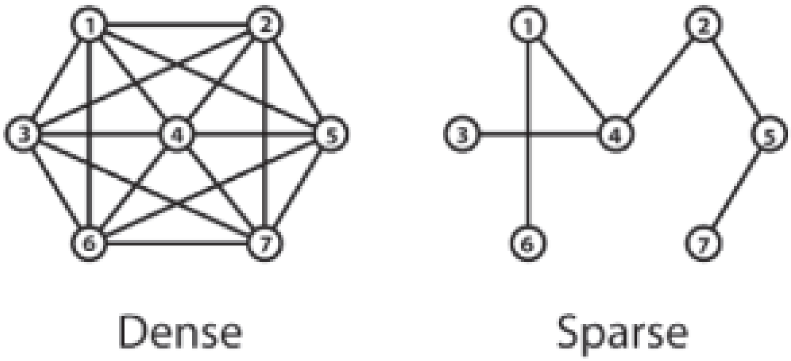


Example assumption

$(W \perp\!\!\!\perp S)|H$

$$p(W, H, S) = \frac{p(W, H)p(H, S)}{p(H)}$$

Sparse Graph Models



Covariance matrix estimation

$$\mathrm{X}_1, \mathrm{X}_2, \dots \mathrm{X}_\mathrm{n} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$$

When $n > p$:

- $\hat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{X}_i)(\mathrm{X}_i)^T$; $\hat{\boldsymbol{K}}_{MLE} = \hat{\boldsymbol{\Sigma}}_{MLE}^{-1}$
- If we know the graph G & $n >$ maximum clique, then $\hat{\boldsymbol{K}}_{G,MLE} = \dots \hat{\boldsymbol{K}}_{MLE}$ with zeros putted in places where G has no edge,
  $\hat{\boldsymbol{\Sigma}}_{G,MLE} = \dots \hat{\boldsymbol{K}}_{G,MLE}^{-1}$

More can be found in article [5] or in the book [4].
Authors of LASSO [6] also made the paper [1] where they introduce a GLASSO method for graph estimation.

Empirical data example

Table 1. Empirical concentrations $\times 1000$ (on or above the diagonal) and partial correlations (below the diagonal) for the examination marks in five mathematical subjects.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.24	−2.44	−2.74	0.01	−0.14
Vectors	0.33	10.43	−4.71	−0.79	−0.17
Algebra	0.23	0.28	26.95	−7.05	−4.70
Analysis	−0.00	0.08	0.43	9.88	−2.02
Statistics	0.02	0.02	0.36	0.25	6.45

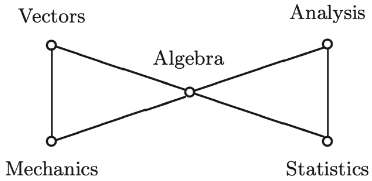


Fig. 1. Conditional independence structure of examination marks for 88 students.

Figure: Table and Fig., from [2]

Empirical data example

Table 1. Empirical concentrations $\times 1000$ (on or above the diagonal) and partial correlations (below the diagonal) for the examination marks in five mathematical subjects.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.24	−2.44	−2.74	0.01	−0.14
Vectors	0.33	10.43	−4.71	−0.79	−0.17
Algebra	0.23	0.28	26.95	−7.05	−4.70
Analysis	−0.00	0.08	0.43	9.88	−2.02
Statistics	0.02	0.02	0.36	0.25	6.45

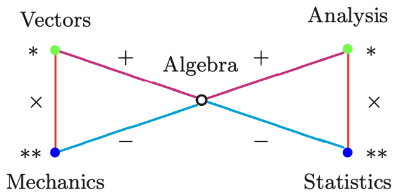


Fig. 8. Coloured graph of an RCOP symmetry model for the examination marks of 88 students. The distribution of the marks is unchanged if we simultaneously replace Vectors with Analysis and Mechanics with Statistics.

Figure: Table and Fig., from [2]

Permutational symmetry example

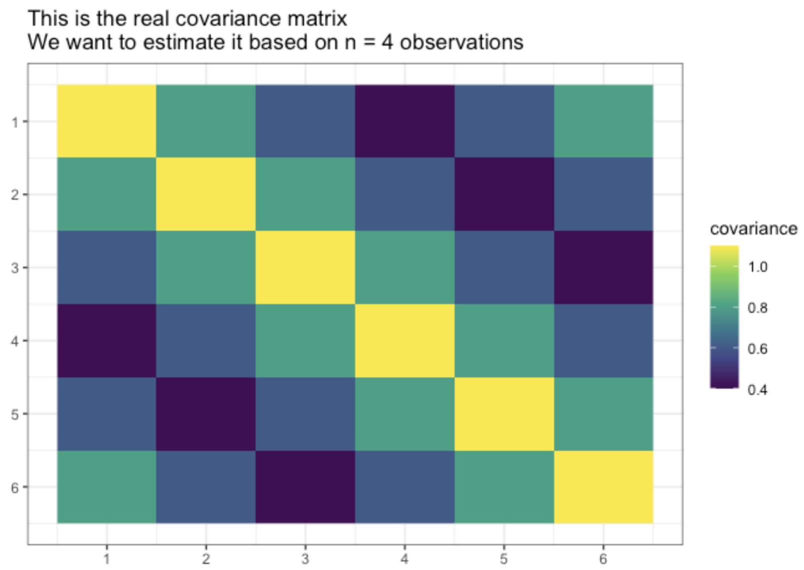


Figure: Fig., from [3]; Symmetry $\Gamma = < (1, 2, 3, 4, 5, 6) >$

Permutational symmetry example

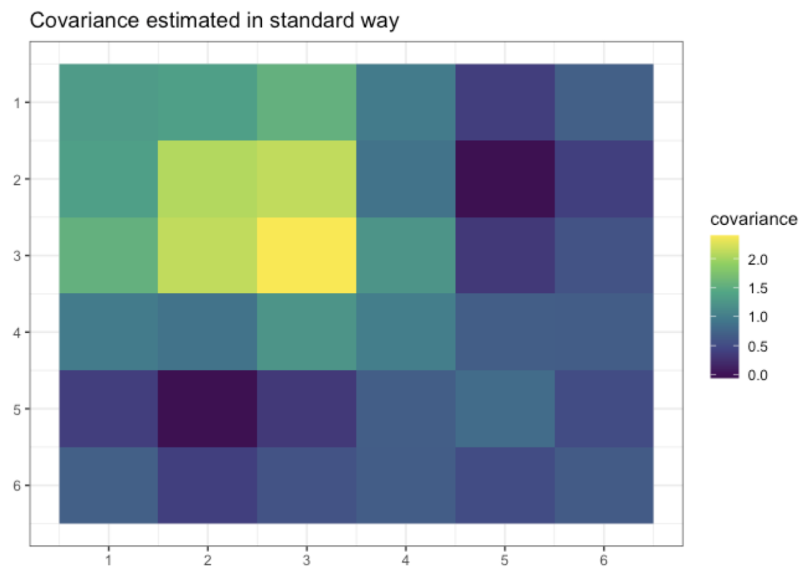


Figure: Fig., from [3]; no symmetry assumed

Permutational symmetry example

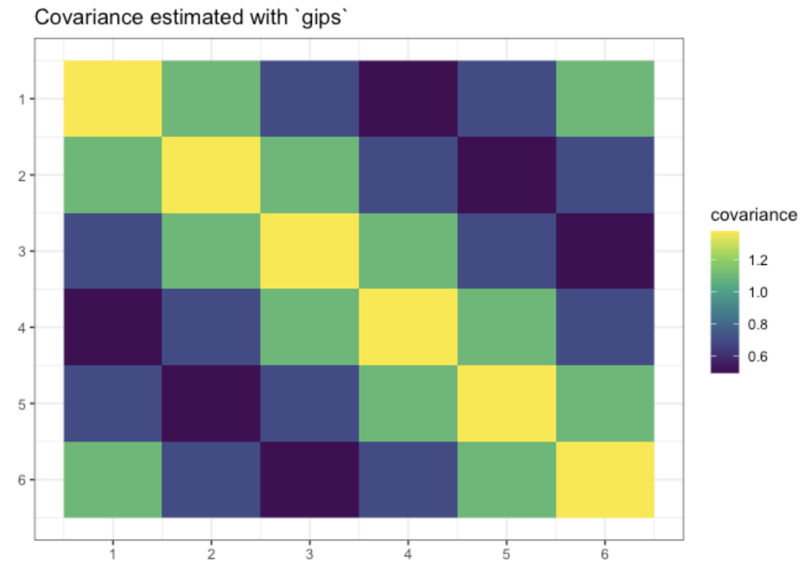


Figure: Fig., from [3]; gips automatically found symmetry $\Gamma = <(1, 2, 3, 4, 5, 6)>$

Recognition on big matrix

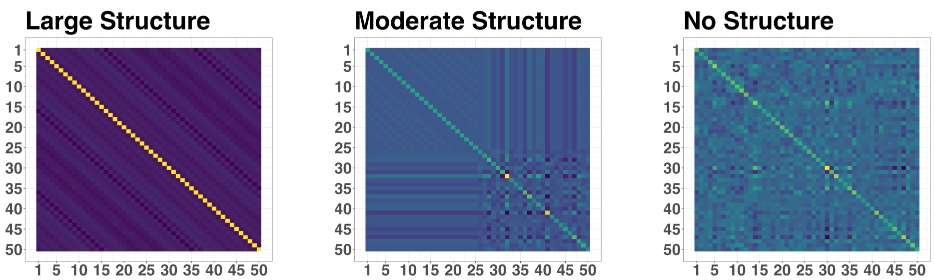


Figure: True covariance matrices corresponding to the three scenarios:
left panel - large structure $\pi_{\langle(1,2,\ldots,50)\rangle}(S)$,
middle panel - moderate structure $\pi_{\langle(1,2,3,\ldots,25)\rangle}(S)$,
right panel - no structure S.

Performance of matrix restoration

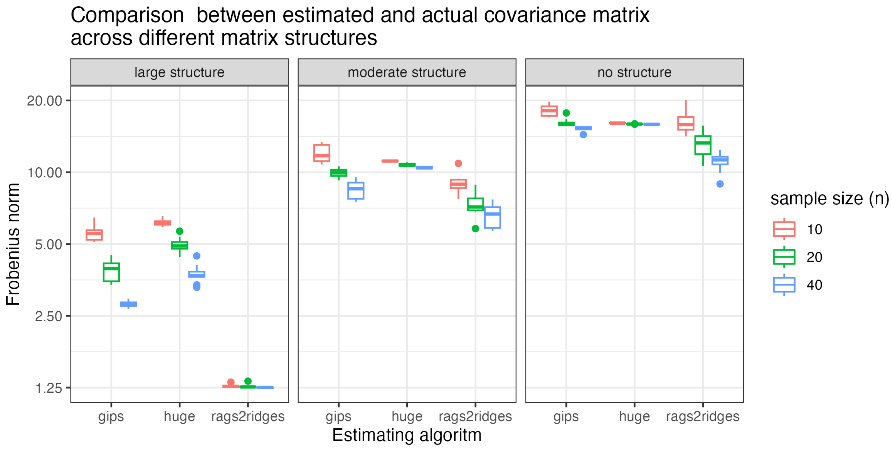


Figure: Frobenius norm (on a logarithmic scale) of the difference of the estimate and the true covariance matrix. 10 runs for each configuration.

Inspiring citation

> *If you torture the data long enough, it will confess*
>
> — Ronald Coase

Conclusions

1. Graphs are a convenient language for recording relationships in data.
2. Expert knowledge is an integral part of statistical analysis.
3. Expert knowledge can be captured with a graph and used in modeling.
4. There are methods for learning graphs from data, i.e., automatic learning of expert knowledge.
5. There are many ways to put constraints on parameters. The appropriate choice of these depends on the nature of the collected data.
6. It is possible, when one tries, to draw false conclusions from the data.

Bibliography I

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (Dec. 2007), pp. 432–441.

[2] Søren Højsgaard and Steffen L. Lauritzen. "Graphical Gaussian Models with Edge and Vertex Symmetries". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70.5 (2008), pp. 1005–1027. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/20203867 (visited on 10/30/2023).

[3] *R package: gips - Gaussian model Invariant by Permutation Symmetry*. przechoj.github.io/gips. [Online; 14-12-2023].

Bibliography II

[4] Alberto Roverato. *Graphical Models for Categorical Data*. SemStat Elements. Cambridge University Press, 2017. DOI: 10.1017/9781108277495.

[5] KAYVAN SADEGHI and STEFFEN LAURITZEN. "Markov properties for mixed graphs". In: *Bernoulli* 20.2 (2014), pp. 676–696. ISSN: 13507265. URL: http://www.jstor.org/stable/42919409.

[6] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

Undirected graphs - independence
Colored graphs - equality; gips
Conclusions
References

Thanks for attention
Happy modelling!

Adam Przemysław Chojecki Uncovering Data Symmetries with 'gips'

Mathematical complement

$\sigma \in \mathfrak{S}_p \qquad R(\sigma) := \sum_{i=1}^{p} E_{\sigma(i)\,i} \in GL(p, \mathbb{R}) \qquad \left(\begin{array}{l} \text{e.g. } \sigma = (1\ 2) \in \mathfrak{S}_3 \\ \Rightarrow R(\sigma) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \right)$

permutation matrix

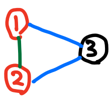
$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N(0, \Sigma) \qquad (\Sigma \in \mathrm{Sym}^+(p, \mathbb{R}))$

$R(\sigma) X \overset{d}{=} X \iff {}^t R(\sigma) \Sigma R(\sigma) = \Sigma$

identically distributed

$\Gamma \subset \mathfrak{S}_p$ (subgr.) $\qquad \mathcal{P}_\Gamma := \{ \Sigma \in \mathrm{Sym}^+(p, \mathbb{R}) \,;\, \forall_{\sigma \in \Gamma}\ {}^t R(\sigma) \Sigma R(\sigma) = \Sigma \}$

Ex. $p = 3$, $\Gamma = \langle (1\ 2) \rangle$



$\mathcal{P}_\Gamma = \left\{ \Sigma = \begin{pmatrix} a & b & c \\ b & a & c \\ c & c & d \end{pmatrix} \in \mathrm{Sym}^+(3, \mathbb{R}) \,;\, a, b, c, d \in \mathbb{R} \right\}$

$\mathcal{M}_\Gamma := \{ N(0, \Sigma) \,;\, \Sigma \in \mathcal{P}_\Gamma \}$: statistical model of Gaussian distr. with Γ-invariance

Problem For given samples $x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{R}^p$,

find the most reasonable subgroup $\Gamma \subset \mathfrak{S}_p$.

(model selection)

Bayesian approach:

$\Gamma \subset \mathfrak{S}_p$ and $K := \Sigma^{-1}$ are RANDOM VARIABLES

$$p(\Gamma) := \frac{1}{\#(\text{subgroups of } \mathfrak{S}_p)} \quad \text{uniform distribution}$$

Fix $\delta > 0$ and $D \in \mathrm{Sym}^+(p, \mathbb{R})$ (hyperparameters)

$$p(K|\Gamma) := \frac{1}{I_\Gamma(\delta, D)} e^{-\mathrm{tr}\, KD/2} (\det K)^{\frac{\delta-2}{2}} 1_{\mathcal{P}_\Gamma}(K)$$

Diaconis-Ylvisaka conjugate prior

where

$$I_\Gamma(\delta, D) := \int_{\mathcal{P}_\Gamma} e^{-\mathrm{tr}\, KD/2} (\det K)^{\frac{\delta-2}{2}} dK$$

$$p(x^{(1)}, \cdots, x^{(n)} \mid \underset{\Sigma^{-1}}{\underset{\parallel}{K}}) = (2\pi)^{-np/2} (\det K)^{\frac{n}{2}} e^{-\mathrm{tr}(KY)/2}$$

$$\left(Y := \sum_{k=1}^{n} x^{(k)}\, {}^t x^{(k)} \in \mathrm{Sym}(p, \mathbb{R}) \right)$$

Then

$$p(x^{(1)}, \cdots, x^{(n)}, \Sigma, \Gamma) = p(x^{(1)}, \cdots, x^{(n)} | \Sigma) p(\Sigma | \Gamma) p(\Gamma)$$

We want to know $p(\Gamma | x^{(1)}, \cdots, x^{(n)}) = \dfrac{p(x^{(1)}, \cdots, x^{(n)} | \Gamma)\, p(\Gamma)}{p(x^{(1)}, \cdots, x^{(n)})}$

$$\propto p(x^{(1)}, \cdots, x^{(n)} | \Gamma)$$

$$p(x^{(1)}, \cdots, x^{(n)} | \Gamma) = \int_{\mathcal{P}_\Gamma} p(x^{(1)}, \cdots, x^{(n)} | K) p(K|\Gamma)\, dK$$

$$= (2\pi)^{-np/2} \frac{I_\Gamma(\delta + n, D + Y)}{I_\Gamma(\delta, D)}$$

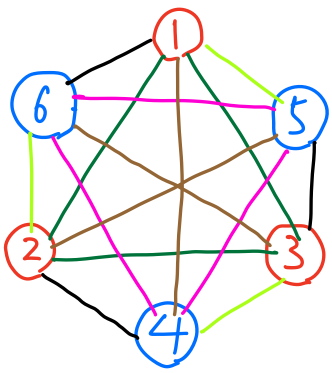
GIKM 2022 : EXACT FORMULA for

$$I_\Gamma(\delta, D) = \int_{\mathcal{P}_\Gamma} e^{-\mathrm{tr}(KD)} (\det K)^{\frac{\delta-2}{2}} \, dK$$

$\rightsquigarrow$ implemented in GIPS.

Point : Group representation of Γ

Ex. The case $\Gamma = \langle \sigma \rangle \subset \mathfrak{S}_6$ with $\sigma = (1\ 2\ 3)(4\ 5\ 6)$



$$K = \begin{pmatrix} a & b & b & e & f & g \\ b & a & b & g & e & f \\ b & b & a & f & g & e \\ e & g & f & c & d & d \\ f & e & g & d & c & d \\ g & f & e & d & d & c \end{pmatrix} \in \mathcal{P}_\Gamma$$

Put

$$U_\Gamma := \begin{pmatrix} 1/\sqrt{3} & 0 & 2/\sqrt{6} & 0 & 0 & 0 \\ 1/\sqrt{3} & 0 & -1/\sqrt{6} & 1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{3} & 0 & -1/\sqrt{6} & -1/\sqrt{2} & 0 & 0 \\ 0 & 1/\sqrt{3} & 0 & 0 & 2 & 0 \\ 0 & 1/\sqrt{3} & 0 & 0 & -1/\sqrt{6} & 1/\sqrt{2} \\ 0 & /\sqrt{3} & 0 & 0 & -1/\sqrt{6} & 1/\sqrt{2} \end{pmatrix}$$

Then

$${}^tU_\Gamma K U_\Gamma = \left(\begin{array}{cc:cccc} A & B & 0 & 0 & 0 & 0 \\ B & C & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & D & 0 & F & G \\ 0 & 0 & 0 & D & -G & F \\ 0 & 0 & F & -G & E & 0 \\ 0 & 0 & G & F & 0 & E \end{array}\right) \sim \begin{pmatrix} A & B \\ B & C \end{pmatrix} \oplus \begin{pmatrix} D & F-iG \\ F+iG & E \end{pmatrix}$$

U_Γ comes from an irreducible decomposition of $\mathbb{R}^6$ as Γ-m

$$\mathbb{R}^6 = \mathbb{R}\begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \\ 0 \\ 0 \\ 0 \end{pmatrix} \oplus \mathbb{R}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} \oplus \left\langle \begin{pmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ -1/\sqrt{6} \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \\ 0 \\ 0 \end{pmatrix} \right\rangle_{\mathbb{R}} \oplus \left\langle \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2/\sqrt{6} \\ -1/\sqrt{6} \\ -1/\sqrt{6} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \right\rangle_{\mathbb{R}}$$

$$\begin{array}{ccccc} \mathcal{P}_\Gamma & \simeq & \mathrm{Sym}^+(2,\mathbb{R}) & \oplus & \mathrm{Herm}^+(2,\mathbb{C}) \\ \rotatebox{90}{$\in$} & & & & \rotatebox{90}{$\in$} \\ K & \longmapsto & \begin{pmatrix} A & B \\ B & C \end{pmatrix} & \oplus & \begin{pmatrix} D & F-iG \\ F+iG & E \end{pmatrix} \end{array}$$

$$K = \underbrace{\pi_\Gamma}_{\text{ortho. proj. w.r.t. the trace inner product}}(X) \in \mathcal{P}_\Gamma \quad (X \in \mathrm{Sym}^+(6,\mathbb{R}))$$

$$\Rightarrow I_\Gamma(\delta, X) = (2\pi)^{3/2}\, 2^{-\delta}\, \Gamma\!\left(\tfrac{\delta+1}{2}\right)\Gamma\!\left(\tfrac{\delta}{2}\right)\Gamma(\delta)\Gamma(\delta-1) \times (AC-B^2)^{-(\delta+1)/2}(DE-F^2-G^2)^{-\delta}$$

find U_Γ (only for cyclic group in GIKM)

$\Gamma \subset \mathfrak{S}_p$

$$\curvearrowright \mathbb{R}^p = \bigoplus_{i=1}^{L} V_i^{\oplus r_i} \quad \Rightarrow \quad \mathcal{P}_\Gamma \simeq \bigoplus_{i=1}^{L} \mathrm{Herm}^+(r_i, \mathbb{K}_i)$$

$\Rightarrow$ formula for $I_\Gamma(\delta, D)$

Maximum likelihood estimation for discrete exponential families, its geometry and combinatorics

Tomasz Skalski

Wrocław University of Science and Technology, Poland

We discuss the existence of the maximum likelihood estimator for discrete exponential family on finite set. Using the newly introduced notion of sets of uniqueness, we present new criterion for the existence of MLE. We show how this criterion can be applied in various discrete settings, in which it can be easily solved using the tools from random graph theory and discrete geometry. Most notably, we discuss the MLE existence for exponential models of random graphs and for linear spaces spanned by Rademacher functions. Additionally, we give a few remarks concerning the existence of MLE for spaces spanned by products of Rademacher functions. As an application, we discuss the asymptotics of the size of independent identically distributed samples for which the maximum likelihood estimator exists with high probability.

References

[1] K. Bogdan, M. Bosy, T. Skalski, *Maximum likelihood estimation for discrete exponential families and random graphs*, ALEA Lat. Am. J. Probab. Math. Stat. 19 (2022), no. 1, 1045–1070.

Maximum likelihood estimation for discrete exponential families, its geometry and combinatorics

Tomasz Skalski

Wrocław University of Science and Technology

(joint research with Krzysztof Bogdan and Michał Bosy)

Osaka & on-line
2023/12/14

Estimation

Estimation

Estimation

- Waiting time τ – random variable
- Number of waiting passengers N – observed data
- $\hat{\tau}(N)$ – estimator of τ

When the estimator is "good"?
Popular choice – to maximise the likelihood function
(Maximum Likelihood Estimation, MLE)

In other words, we are looking for the parameter, under which the given situation is most likely to be present.

Notation

- $\mathcal{X} = \{x_1, \dots, x_K\}$ – finite state space, $K = |\mathcal{X}|$
- $\mu : \mathcal{X} \to (0, \infty)$ – weight function
- $\mathcal{B} \subset \mathbb{R}^{\mathcal{X}}$ – linear space of functions ($\phi = \mathbb{1} \in \mathcal{B}$)
- $\mathcal{B}_+ = \{\phi \in \mathcal{B} : \phi \geq 0\}$ – subclass (cone) of non-negative functions
- $N(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x)$ – normalising constant (partition function)
- $p = e(\phi) = \frac{e^{\phi}}{N(\phi)}$ – exponential density
- $e(\mathcal{B}) = \{p = e(\phi) : \phi \in \mathcal{B}\}$ – exponential family

MLE

Definition

Let $x_1, \dots, x_n$ be a sample from the finite set $\mathcal{X}$ and let $\phi \in \mathcal{B}$. The likelihood function of $p = e(\phi)$ is defined as:

$$L_p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i).$$

Joint density: function of $(x_1, \dots, x_n)$
Likelihood: function of p

Definition

The $\hat{p} \in e(\mathcal{B})$ is called the maximum likelihood estimator (MLE), if

$$L_{\hat{p}}(x_1, \dots, x_n) = \sup_{p \in e(\mathcal{B})} L_p(x_1, \dots, x_n).$$

Existence of MLE

History

- *S. J. Haberman (1974) – criterion of existence of MLE in hierarchical log-linear models*
- *B. R. Crain (1976) – sufficient criterion of a.s. existence of MLE*
- *O. Barndorff-Nielsen (1978) – criterion of existence of MLE for the exponential families in terms of convex geometry*
- *M. Jacobsen (1989) – alternative condition for discrete exponential families*
- *K. Bogdan, M. Bogdan (2000) – criterion of existence of for exponential families of continuous functions on* $[0, 1]$ *in terms of the sets of uniqueness.*
- *N. Eriksson, S. E. Fienberg, A. Rinaldo, S. Sullivant (2006) – interpretation of the criterion in terms of polyhedral geometry*
- *A. Rinaldo, S. E. Fienberg, Y. Zhou (2009) – application to exponential models of random graphs (ERGM).*
- *K. Bogdan, M. Bosy, TS (2022) – criterion of existence of MLE in discrete exponential families in terms of sets of uniqueness.*

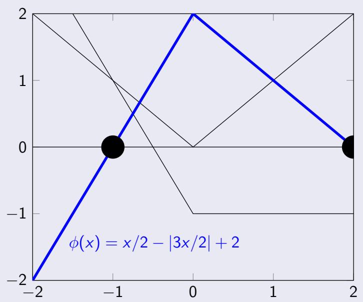
Sets of uniqueness

Definition

The set $U \subset \mathcal{X}$ *is of uniqueness for* $\mathcal{B}$, *if* $\phi \equiv 0$ *is the only function in* $\mathcal{B}$ *such that* $\phi(U) = 0$.

Example

Let $\mathcal{X} = \{-2, -1, 0, 1, 2\}$. *Let* $\mathcal{B}$ *denote the class of all the real functions on* $\mathcal{X}$ *that are linear (affine) both on* $\{-2, -1, 0\}$ *and on* $\{0, 1, 2\}$.



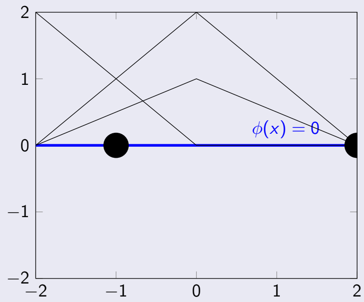
Then the set $\{-1, 1, 2\}$ *is of uniqueness for* $\mathcal{B}$, *but the set* $\{-1, 2\}$ *is not.*

Definition

U is a set of uniqueness for $\mathcal{B}_+$, if $[\phi \in \mathcal{B}_+,\ \phi\,(U) = 0\,] \Rightarrow [\,\phi \equiv 0\,]$.

Example

Again, let $\mathcal{X} = \{-2, -1, 0, 1, 2\}$ and let $\mathcal{B}$ be the class of all the real functions on $\mathcal{X}$ that are linear (affine) on $\{-2, -1, 0\}$ and on $\{0, 1, 2\}$.



Then the set $\{-1, 2\}$ is of uniqueness for $\mathcal{B}_+$.

Existence of MLE – main criterion

Theorem (K. Bogdan, M. Bosy, TS (2022))

The maximum likelihood estimator for $e(\mathcal{B})$ and $x_1, \dots, x_n \in \mathcal{X}$ exists if and only if $\{x_1, \dots, x_n\}$ is a set of uniqueness for $\mathcal{B}_+$.

Proof.

($\Rightarrow$) If $\{x_1, \dots, x_n\}$ is not of uniqueness for $\mathcal{B}_+$, we may subtract from every candidate for MLE ϕ a non-negative function ψ vanishing on $\{x_1, \dots, x_n\}$, so $\psi - \phi = \psi$ on $\{x_1, \dots, x_n\}$. Thus $N(\psi - \phi) < N(\psi)$ and the resulting likelihood is increased. $\square$

Existence of MLE – main criterion

Theorem (K. Bogdan, M. Bosy, TS (2022))

The maximum likelihood estimator for $e(\mathcal{B})$ and $x_1, \ldots, x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ is a set of uniqueness for $\mathcal{B}_+$.

Proof.

($\Leftarrow$) We introduce a seminorm

$$\lambda_U(\phi) = \max_{\mathcal{X}}(\phi) - \min_U(\phi)$$

related to given set $U \subset \mathcal{X}$ and compare it with an oscillation seminorm $\lambda_{\mathcal{X}}(\phi)$. Both seminorms are comparable since U is a set of uniqueness for $\mathcal{B}_+$. □

Applications

Two types of proposed applications:

- Conditions for the existence of MLE for specific exponential families
- Probability bounds for MLE for i.i.d. samples

For the i.i.d. random variables $X_1, X_2, \ldots$ valued in $\mathcal{X}$ we define the following (random) time:

$$\nu_{uniq} = \inf\{n \geq 1 : \{X_1, \ldots, X_n\} \text{ is a set of uniqueness for } \mathcal{B}_+\}.$$

Probabilistic tool — Threshold functions

Definition (Threshold)

A function $n^* = n^*(K)$ is a threshold of the size of the sample $\mathbb{X} = (X_1, \ldots, X_n)$ for a given (monotone) property $\mathscr{P}$ if

$$\lim_{K\to\infty} \mathbb{P}(\mathbb{X} \in \mathscr{P}) = \begin{cases} 0 & \text{if } n(K)/n^*(K) \to 0, \qquad K \to \infty, \\ 1 & \text{if } n(K)/n^*(K) \to \infty, \qquad K \to \infty. \end{cases}$$

Definition (Sharp threshold)

A function $n^* = n^*(K)$ is a sharp threshold of the size of the sample $\mathbb{X} = (X_1, \ldots, X_n)$ for a given (monotone) property $\mathscr{P}$ if for every $\varepsilon > 0$

$$\lim_{K\to\infty} \mathbb{P}(\mathbb{X} \in \mathscr{P}) = \begin{cases} 0 & \text{if } n(K)/n^*(K) < 1-\varepsilon, \\ 1 & \text{if } n(K)/n^*(K) > 1+\varepsilon. \end{cases}$$

Applications – $\mathbb{R}^{\mathcal{X}}$

Let $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$. As $\mathcal{X}$ is the only set of uniqueness for $\mathcal{B}_+$, we observe that

Lemma

MLE for $e(\mathbb{R}^{\mathcal{X}})$ and $x_1, \ldots, x_n$ exists if and only if $\{x_1, \ldots, x_n\} = \mathcal{X}$,

i.e. each element of $\mathcal{X}$ has to be reached by the sample (Coupon Collector Problem).

Corollary

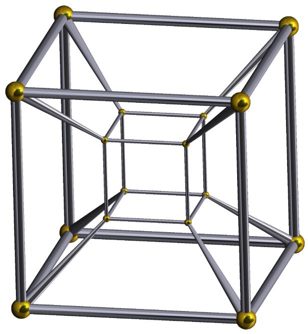
Let $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$ and $K = |\mathcal{X}|$. Let $X_1, X_2, \ldots$ be independent random variables, each with uniform distribution on $\mathcal{X}$. Then, for every $c \in \mathbb{R}$,

$$\lim_{K\to\infty} \mathbb{P}\left(\nu_{uniq} < K\log K + Kc\right) = e^{-e^{-c}}.$$

In particular, $n^*(K) = K\log K$ is a sharp threshold of the sample size for the existence of MLE for $e(\mathcal{X})$.

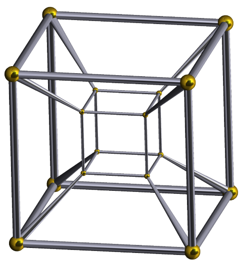
Examples – Rademacher functions

Vertices of a discrete hypercube $\{-1,+1\}^k \sim$ sequences of (-1) and $(+1)$ with length k



Examples – Discrete hypercube

Rademacher function $\sim$ indicator of a half-cube in $\{-1,+1\}^k$



Product of Rademacher functions $\sim$ indicator of a subcube of fixed size

Applications – Rademacher functions

For $k \in \mathbb{N}$ consider the discrete hypercube $\mathcal{X} = Q_k = \{-1, 1\}^k$.
For $j = 1, \dots, k$ we define Rademacher functions:

$$r_j(\chi) = \chi_j, \quad \chi = (\chi_1, \dots, \chi_k) \in Q_k$$

and denote $r_0(\chi) = 1$.
Here $K = |\mathcal{X}| = 2^k$.

Theorem (K. Bogdan, M. Bosy, TS (2022))

Let $\mathcal{B}^k = Lin\{r_0, r_1, \dots, r_k\}$. MLE for $e(\mathcal{B}^k)$ and $x_1, \dots, x_n \in Q_k$ exists if and only if for all $j = 1, \dots, k$ we have $\{r_j(x_1), \dots, r_j(x_n)\} = \{-1, 1\}$.

Satisfied if and only if $\{x_1, \dots, x_n\}$ intersects with every half-cube of Q_k, e.g. $\{x_1, -x_1\}$.

Applications – Rademacher functions

Theorem (K. Bogdan, M. Bosy, TS (2022))

Let $k \in \mathbb{N}$, $n(k) = \log_2 k + b + o(1)$. Let $X_1, \dots, X_{n(k)}$ be independent random variables, each with uniform distribution on Q_k. Then

$$\lim_{k \to \infty} \mathbb{P}(\{X_1, \dots, X_{n(k)}\} \text{ is of uniqueness for } \mathcal{B}_+) = \exp\{-2^{1-b}\}.$$

and $n^*(K) = \log_2 k = \log_2 \log_2 K$ is a sharp threshold of the sample size for the existence of MLE for $e(\mathcal{B}^k)$ and i.i.d. uniform samples on Q_k.

Proof idea: asymptotics of the maximum of i.i.d. geometric variables.

Applications – ERGM

We consider simple undirected graphs containing no loops or multiple edges. Let N and m denote the number of vertices and edges of the graph. By $\mathcal{G}_N$ we denote the family of all graphs with N vertices.
For graphs $G = (V, E_1)$, $H = (V, E_2)$ we let, as usual,

$$G \cup H := (V, E_1 \cup E_2), \qquad G \cap H := (V, E_1 \cap E_2),$$

$$G \subset H \quad \equiv \quad E_1 \subset E_2.$$

We define $\chi_{r,s}(G) = 1 - 2\mathbb{1}_G(r, s)$ and consider the following linear space

$$\mathcal{B}^{\mathcal{G}_N} = \mathrm{Lin}\left\{ 1, \chi_{r,s}(G) : 1 \le r < s \le N \right\}.$$

Consider coefficients $c \in \mathbb{R}^{\binom{V}{2}}$, indexed by the edges of the complete graph K_N, and the following exponential family:

$$e(\mathcal{B}^{\mathcal{G}_N}) = \left\{ p_c := e^{\phi_c - \psi(\phi_c)} : c \in \mathbb{R}^{\binom{V}{2}} \right\},$$

where

$$\phi_c(G) = \sum_{(r,s) \in \binom{V}{2}} c_{r,s}\chi_{r,s}(G), \qquad \psi(\phi_c) = \log \sum_{G \in \mathcal{G}_N} e^{\phi_c(G)},$$

and $G \in \mathcal{G}_N$.

Observation

Fix $c \in \mathbb{R}^{\binom{V}{2}}$. In the random graph $\mathbb{G}$ sampled from $p_c \in e(\mathcal{B}^{\mathcal{G}_N})$, each edge (r, s) appears independently with probability

$$p_{r,s} = \frac{e^{c_{r,s}}}{1 + e^{c_{r,s}}}.$$

Applications – ERGM

Theorem (K. Bogdan, M. Bosy, TS (2022))

MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and $G_1, \ldots, G_n \in \mathcal{G}_N$ exists if and only if

$$\bigcup_{i=1}^{n} G_i = K_N \qquad \text{and} \qquad \bigcap_{i=1}^{n} G_i = \overline{K_N}.$$

Lemma (K. Bogdan, M. Bosy, TS (2022))

Let $\{\mathbb{G}_1, \ldots, \mathbb{G}_n\}$ be independent random graphs from $p_c \in e(\mathcal{B}^{\mathcal{G}_N})$. Then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ equals

$$\prod_{1 \le r < s \le N} \left(1 - p_{r,s}^n - (1 - p_{r,s})^n\right).$$

In particular, $n^*(N) = \log N$ is a threshold of the sample size n for the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$.

Applications – Products of Rademacher functions

Let $k \in \mathbb{N}, 1 \le q \le k$, and $\mathcal{B}_q^k = \mathrm{Lin}\{w_S : S \subset \{1, \ldots, k\} \text{ and } |S| \le q\}$, where $w_S(x) = \prod_{i \in S} r_i(x), x \in Q_k, S \subset \{1, \ldots, k\}$, are the Walsh functions.

Observation

$\mathcal{B}_q^k$ is the linear space spanned by indicator functions of the sub-cubes of Q_k, obtained by fixing q out of k coordinates.

- $q = 1$: Rademacher functions (already discussed)
- $q = 2$: Ising model (open problem)
- $\ldots$: open problems
- $q = k - 1$: see the next slide

Applications – Products of $(k-1)$ Rademacher functions

$\mathcal{B}^k_{k-1}$ corresponds to indicators of edges of Q_k. Consider the following partition: $Q_k = \mathcal{E} \cup \mathcal{O}$:

Definition

- $\mathcal{E} := \{\chi \in Q_k : \chi \textit{ has even number of positive coordinates}\}$
- $\mathcal{O} := \{\chi \in Q_k : \chi \textit{ has odd number of positive coordinates}\}$

Theorem (K. Bogdan, M. Bosy, TS (2022))

MLE exists for $e(\mathcal{B}^k_{k-1})$ and $x_1, \dots, x_n \in Q_k$ if and only if $\mathcal{E} \subset \{x_1, \dots, x_n\}$ or $\mathcal{O} \subset \{x_1, \dots, x_n\}$.

Thank you for your attention!

References

Barndorff-Nielsen, O. (1978)
"'*Information and exponential families in statistical theory*"', John Wiley & Sons Ltd., Chichester. Wiley Series in Probability and Mathematical Statistics.

Bogdan, K., Bogdan, M. (2000)
"'*On existence of maximum likelihood estimators in exponential families*"', Statistics, 34(2):137–149.

Bogdan, K., Bosy, M., Skalski, T. (2022)
"'*Maximum likelihood estimation for discrete exponential families and random graphs*"', ALEA Lat. Am. J. Probab. Math. Stat. 19 (2022), no. 1, 1045–1070.

Crain, B. R. (1976)
"'*Exponential models, maximum likelihood estimation, and the Haar condition*"', J. Amer. Statist. Assoc., 71 (355), 737–740.

Eriksson, N., Fienberg, S. E., Rinaldo, A., Sullivant, S. (2006)
"'*Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models*"', J. Symbolic Comput., 41(2):222–233.

Haberman, S. J. (1974)
"'*The analysis of frequency data*"', The University of Chicago Press, Chicago, Ill.-London. Statistical Research Monographs, Vol. IV.

Jacobsen, M. (1989)
"'*Existence and unicity of MLEs in discrete exponential family distributions*"', Scand. J. Statist., 16 (4), 335–349.

Rinaldo, A., Fienberg, S. E., Zhou, Y. (2009)
"'*On the geometry of discrete exponential families with application to exponential random graph models*"', Electron. J. Stat., 3:446–484

Some open problems on minimum information dependence models

Tomonari Sei (The University of Tokyo)*

This is joint work with Keisuke Yano (The Institute of Statistical Mathematics).

1. Minimum information dependence model

Consider σ-finite measure spaces $(\mathcal{X}_i, \mathcal{F}_i, \mathrm{d}x_i)$ for $i \in [d] = \{1, \ldots, d\}$. Denote their product space by $(\mathcal{X}, \mathcal{F}, \mathrm{d}x)$. Let r_i be a probability density function on $\mathcal{X}_i$ for each i, which may have a nuisance parameter.

Suppose that we have a measurable map $h : \mathcal{X} \to \mathbb{R}^K$, which describes dependence among the d variables $(x_1, \ldots, x_d) \in \mathcal{X}$. A minimum information dependence model [2] is defined by a set of probability density functions

$$p(x; \theta) = \exp\left(\theta^\top h(x) - \sum_{i=1}^{d} a_i(x_i; \theta) - \psi(\theta)\right) \prod_{i=1}^{d} r_i(x_i), \quad \theta \in \mathbb{R}^K,$$

with respect to $\mathrm{d}x$, where $a_i(x_i; \theta)$ and $\psi(\theta)$ are determined by conditions

$$\int_{\mathcal{X}_{-i}} p(x; \theta)\mathrm{d}x_{-i} = r_i(x_i), \quad i \in [d], \quad x_i \in \mathcal{X}_i,$$

and $\int_{\mathcal{X}} \sum_{i=1}^{d} a_i(x_i; \theta) p(x; \theta)\mathrm{d}x = 0$. Here, $-i$ denotes the removal of the i-th coordinate. We call θ the canonical parameter, $h(x)$ the canonical statistic, $a_i(x_i; \theta)$ the adjusting function and $\psi(\theta)$ the potential function. An existence and uniqueness theorem under mild conditions is established by [2]. If the marginal distributions are uniform on the unit interval $[0, 1] \subset \mathbb{R}$, the model is called the minimum information copula model [1].

The potential function ψ is convex because it is characterized by

$$\psi(\theta) = \sup_{p \in \mathcal{M}} \left\{\theta^\top \int h(x)p(x)\mathrm{d}x - \int p(x) \log \frac{p(x)}{\prod_i r_i(x_i)} \mathrm{d}x\right\},$$

where $\mathcal{M}$ is the set of probability density functions with fixed marginals $r_i(x_i)$. The pair of the parameters θ and $\eta = \int h(x)p(x; \theta)\mathrm{d}x$ induces a dually flat structure.

2. Open problems

There are a couple of open questions about the model. See [2] for details.

1. Is $\psi(\theta)$ analytic?
2. Find examples of $p(x; \theta)$ with closed expressions. Known examples are the Gaussian, multinomial and a class of circular distributions.
3. Construct an asymptotically efficient estimator of θ (without knowing r_i's).

References

[1] T. Bedford and K. J. Wilson (2014). On the construction of minimum information bivariate copula families, *Annals of the Institute of Statistical Mathematics*, **66**, 703–723.

[2] T. Sei and K. Yano (2023). Minimum information dependence modeling, *Bernoulli*, accepted. (arXiv:2206.06792)

*e-mail: sei@mist.i.u-tokyo.ac.jp

Some open problems on minimum information dependence models

Tomonari Sei   Keisuke Yano

The University of Tokyo   The Institute of Statistical Mathematics

Dec 14–15, 2022 OCAMI Workshop
"Statistical Theories and Machine Learning Using Geometric Methods"

Introduction

- In this talk, we introduce a novel statistical model called the minimum information dependence model.
- The model can describe dependence among variables separately from marginal distributions. (cf. copula theory)
- We discuss some open problems.

Reference: T. Sei and K. Yano (2023). Minimum information dependence modeling, *Bernoulli*, accepted. (arXiv:2206.06792)

Table of contents

Setup

- Consider σ-finite measure spaces $(\mathcal{X}_i, \mathcal{F}_i, \mathrm{d}x_i)$ for $i \in [d] = \{1, \ldots, d\}$.
- Denote their product as $(\mathcal{X}, \mathcal{F}, \mathrm{d}x) = (\prod_i \mathcal{X}_i, \prod_i \mathcal{F}_i, \prod_i \mathrm{d}x_i)$.
- How to construct a statistical model on the product space $\mathcal{X}$?

Example

Fisher's iris data consists of 4 continuous and 1 categorical variables:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮
5.9	3.0	5.1	1.8	virginica

A natural choice is

$$\mathcal{X}_1 = \cdots = \mathcal{X}_4 = \mathbb{R}, \quad \mathcal{X}_5 = \{\text{setosa}, \text{versicolor}, \text{virginica}\}.$$

equipped with the Lebesgue and counting measures, respectively.

Gaussian model

- Let us begin with the Euclidean case $\mathcal{X}_1 = \cdots = \mathcal{X}_d = \mathbb{R}$.
- The d-dimensional Gaussian model on $\mathcal{X} = \mathbb{R}^d$ is

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

  where μ and Σ are parameters.
- Properties:
  1. The density function is written as

$$p(x;\mu,\Sigma) = \underbrace{\exp\left(-\sum_{i<j}\sigma^{ij}x_i x_j\right)}_{\text{dependence}} \underbrace{\prod_{i=1}^d A_i(x_i;\mu,\Sigma)}_{\text{independence}}$$

  2. The marginal distributions are Gaussian.
- These properties characterize the Gaussian model. We use this idea for constructing more general statistical models.

5 / 29

The minimum information dependence model

- Now consider the general space $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$.
- Suppose that we have
  1. a measurable map $h : \mathcal{X} \to \mathbb{R}^K$ and
  2. marginal density functions $r_i : \mathcal{X}_i \to \mathbb{R}_{>0}$ for $i \in [d]$.

Definition

A minimum information dependence model is defined by

$$p(x;\theta) = \exp\left(\theta^\top h(x) - \sum_{i=1}^d a_i(x_i;\theta) - \psi(\theta)\right) \prod_{i=1}^d r_i(x_i), \quad \theta \in \mathbb{R}^K,$$

where a_i and ψ are determined by conditions $\int_{\mathcal{X}_{-i}} p(x;\theta)\mathrm{d}x_{-i} = r_i(x_i)$ and $\int_{\mathcal{X}} \sum_{i=1}^d a_i(x_i;\theta)p(x;\theta)\mathrm{d}x = 0$.

- We call θ the canonical parameter, h the canonical statistic, a_i the adjusting functions and ψ the potential function.
- $\theta = 0$ corresponds to the independence model $\prod_{i=1}^d r_i(x_i)$.

6 / 29

Comparison

exp family	min info dep model
$p(x;\theta) = e^{\theta^\top h(x) - \psi(\theta)} p_0(x)$	$p(x;\theta) = e^{\theta^\top h(x) - \sum_i a_i(x_i;\theta) - \psi(\theta)} \prod_i r_i(x_i)$
θ	θ
h	h
ψ	ψ
—	a_i
$\int p dx = 1$	$\int p dx_{-i} = r_i$

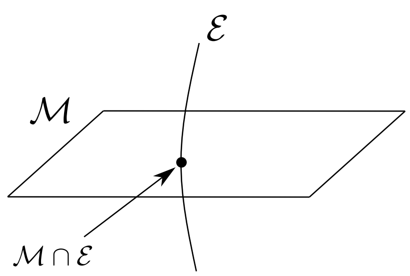
Picture

- Let $\mathcal{P}$ be the set of probability density functions on $\mathcal{X}$.
- The density $p(x;\theta)$ is the unique intersection point of two manifolds

$$\mathcal{M} = \mathcal{M}(r_1, \ldots, r_d) = \{p \in \mathcal{P} \mid \int p(x) dx_{-i} = r_i(x_i)\}$$

and

$$\mathcal{E} = \mathcal{E}(\theta) = \{e^{\theta^\top h(x) - \sum_i b_i(x_i)} \in \mathcal{P} \mid b_i : \mathcal{X}_i \to \mathbb{R}\}.$$



- Uniqueness follows from the Pythagorean theorem (discussed later).

Notation

Before giving examples, we define additional notations.

- The marginal $r_i(x_i)$ often has parameters. We denote it as $r_i(x_i;\nu)$, where ν is the (nuisance) parameter.
- The model is written as

$$p(x;\theta,\nu)=\exp\left(\theta^\top h(x)-\sum_{i=1}^d a_i(x_i;\theta,\nu)-\psi(\theta,\nu)\right)\prod_{i=1}^d r_i(x_i,\nu).$$

- We also use an abbreviation

$$p(x;\theta,\nu)=e^{\theta^\top h(x)}\prod_{i=1}^d A_i(x_i;\theta,\nu).$$

Example (1/3)

Multivariate Gaussian

- The d-dimensional normal distribution $N(\mu,\Sigma)$ is written as

$$p(x;\theta,\nu)=e^{\sum_{i<j}\theta_{ij}x_ix_j}\prod_{i=1}^d A_i(x_i;\theta,\nu),$$
$$\int p(x;\theta,\nu)dx_{-i}=\phi(x_i;\mu_i,\sigma_i^2),\quad i\in[d],$$

where the nuisance parameter is $\nu=(\mu_1,\ldots,\mu_d,\sigma_1^2,\ldots,\sigma_d^2)$.
- We obtain a bijection

$$(\mu,\Sigma)\mapsto(\theta,\nu)$$
$$\mathbb{R}^d\times\mathrm{Sym}^+(d,\mathbb{R})\to\mathbb{R}^{d(d-1)/2}\times\mathbb{R}^d\times\mathbb{R}^d_{>0}.$$

Lemma (Dempster 1972)

If A and B are positive definite matrices, then there exists a unique positive definite matrix C such that $C_{ii}=A_{ii}$ and $(C^{-1})_{ij}=(B^{-1})_{ij}$ $(i\neq j)$.

Example (2/3)

Contingency table model (log-linear model)

- Consider finite spaces $\mathcal{X}_1 = [I]$ and $\mathcal{X}_2 = [J]$.
- A statistical model on $I \times J$ contingency tables is written as

$$
\begin{aligned}
& p_{ij} = e^{\theta_{ij}} A_{1i}(\theta, \nu) A_{2j}(\theta, \nu), \quad (i,j) \in [I] \times [J], \\
& \textstyle\sum_j p_{ij} = \nu_{1i}, \quad \sum_i p_{ij} = \nu_{2j}.
\end{aligned}
$$

- We obtain a bijection

$$
\begin{aligned}
(p_{ij}) &\mapsto (\theta, \nu) \\
\Delta_{IJ-1} &\to \mathbb{R}^{(I-1)(J-1)} \times \Delta_{(I-1)} \times \Delta_{(J-1)}
\end{aligned}
$$

- Relevant studies:
  - Sinkhorn and Knopp (1967): an algorithm for finding A_{1i}, A_{2j}
  - Amari (2001): orthogonal foliation structure
  - Piantadosi et al. (2012): maximum entropy checkerboard copulas
  - Geenens (2020): similar construction of discrete families

Example (3/3)

Directional statistics

Circula

- Let $\mathcal{X}_1 = \mathcal{X}_2 = S^1$ (circle), identified with $[0, 2\pi)$.
- A distribution satisfying conditions

$$
\int_0^{2\pi} p(x_1, x_2) dx_2 = \int_0^{2\pi} p(x_1, x_2) dx_1 = \frac{1}{2\pi}
$$

  is called a circula (Jones et al. 2015).
- The following density is circula for any function $h_0 : S^1 \to \mathbb{R}$.

$$
e^{\theta h_0(x_1 - x_2) - \psi(\theta)}, \quad \psi(\theta) = \log \textstyle\int_0^{2\pi} \exp(\theta h_0(z)) dz,
$$

  This is a minimum information dependence model with a shift-invariant canonical statistic $h(x_1, x_2) = h_0(x_1 - x_2)$.
- In this case, the adjusting functions are zero.

Existence and uniqueness theorem

- The examples so far have explicit density functions. But this is not the case in general. We need an existence and uniqueness theorem.
- Denote $H(x) = \theta^\top h(x)$. Suppose that $H \in L_1(\prod_i r_i \, \mathrm{d}x)$.

Theorem 1 (SY 2023)

If there are integrable functions $b_i \in L_1(r_i \mathrm{d}x_i)$ satisfying

$$\int e^{H(x)-\sum_i b_i(x_i)} \prod_i r_i(x_i)\mathrm{d}x < \infty,$$

then there exist measurable $a_i : \mathcal{X}_i \to \mathbb{R}$ and $\psi \in \mathbb{R}$ such that

$$p(x) = e^{H(x)-\sum_i a_i(x_i)} \prod_i r_i(x_i) \in \mathcal{M}(r_1, \ldots, r_d)$$

and $\int p \sum_i a_i \mathrm{d}x = 0$. The function p is unique.

- The proof relies on Csiszár (1975) and Borwein et al. (1994).
- Remark: each a_i may not be integrable. This point was missed in Csiszár's paper.

Application of Theorem 1

3-factor interaction on $\mathbb{R}^3$

- Let $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathbb{R}$ equipped with the Lebesgue measure.
- Suppose that $\int |x_i|^3 r_i(x_i)\mathrm{d}x_i < \infty$ for each i.
- The minimum information dependence model

  $$p(x_1, x_2, x_3) = e^{\theta x_1 x_2 x_3 - a_1(x_1) - a_2(x_2) - a_3(x_3) - \psi} r_1(x_1) r_2(x_2) r_3(x_3)$$

  is well defined for arbitrary $\theta \in \mathbb{R}$.
- Indeed, the inequality $|x_1 x_2 x_3| \leq \frac{1}{3}(|x_1|^3 + |x_2|^3 + |x_3|^3)$ implies

  $$\iiint e^{\theta x_1 x_2 x_3 - |\theta|(|x_1|^3+|x_2|^3+|x_3|^3)/3} r_1(x_1) r_2(x_2) r_3(x_3) dx_1 dx_2 dx_3 \leq 1 < \infty.$$

  So $b_i(x_i) = |\theta||x_i|^3/3$ satisfy the sufficient condition of the theorem.

In a similar manner, we can take any polynomial $h(x)$ as canonical statistics whenever the marginal distributions have all finite moments.

Pros and cons

Advantages of our model:

- Various type of dependence (such as conditional independence) can be incorporated.
- The domain and marginal distributions are almost arbitrary.
- Multivariate Gaussian and multinomial models are contained as particular instances.

Drawbacks:

- Computation of the adjusting and potential functions are difficult.
  $\rightarrow$ We can avoid them in estimation of θ as described later.

From practical side:

- How to choose h? $\rightarrow$ polynomial, eigenfunctions etc.
- How to interpret θ? $\rightarrow$ regression (not discussed today)

The first open problem

Problem

Find examples that have explicit density functions.

Observation

- In the circular case, we used the shift-invariant canonical statistic $h_0(x_1 - x_2)$. This may be generalized to other compact groups.

Table of contents

Pythagorean relationship

- Recall that $\mathcal{M} = \{p \in \mathcal{P} \mid \int p(x) \mathrm{d}x_{-i} = r_i(x_i)\}$
- Let $\mathcal{E} = \mathcal{E}(\theta) = \{e^{\theta^\top h(x) - \sum_i b_i(x_i)} \prod_i r_i \in \mathcal{P} \mid b_i \in L_1(r_i \mathrm{d}x_i)\}$
- Let $D(p|q) = \int p \log(p/q) \mathrm{d}x$ (Kullback–Leibler divergence).

Generalized Pythagorean theorem (Csiszár 1975)

If $p \in \mathcal{M}, q \in \mathcal{M} \cap \mathcal{E}, s \in \mathcal{E}$, then

$$D(p|s) = D(p|q) + D(q|s)$$

Proof:

$$\begin{aligned} D(p|s) - D(p|q) - D(q|s) &= \int (p-q) \log(q/s) \mathrm{d}x \\ &= \int (p-q)\Big(-\sum_i b_i^q + \sum_i b_i^s\Big) \mathrm{d}x \\ &= \sum_i \int (r_i - r_i)\Big(-\sum_i b_i^q + \sum_i b_i^s\Big) \mathrm{d}x_i = 0. \end{aligned}$$

Consequence of Pythagorean

- For $p \in \mathcal{M}, q \in \mathcal{M} \cap \mathcal{E}, s \in \mathcal{E}$,
$$D(p|s) = D(p|q) + D(q|s).$$
In particular,
$$D(p|s) \geq D(q|s).$$
The equality is attained iff $p = q$.
- By substituting $q = e^{\theta^\top h - \sum_i a_i - \psi} \prod_i r_i$ and $s = e^{\theta^\top h - \sum_i b_i} \prod_i r_i$, we obtain
$$\int p\Big(\log p - \theta^\top h + \sum_i b_i - \sum_i \log r_i\Big) \geq \int q\Big(-\sum_i a_i - \psi + \sum_i b_i\Big).$$
- Since $\int pb_i = \int qb_i$ and $\int q \sum_i a_i = 0$, we have
$$\psi \geq \theta^\top \int ph - \int p \log \frac{p}{\prod_i r_i}$$

Potential function

- Now we obtained the following characterization of ψ:
$$\psi(\theta) = \sup_{p \in \mathcal{M}} \left\{ \theta^\top \int hp - \int p \log \frac{p}{\prod_i r_i} \right\},$$
which implies ψ is convex.

Open problem 2

Are $a_i(x_i; \theta)$ and $\psi(\theta)$ analytic with respect to θ?

- Observation: For exponential families $e^{\theta^\top h(x) - \psi(\theta)} p_0(x)$, the potential function ψ is analytic on int(dom(ψ)). This is directly proved by Taylor expansion
$$\int e^{(\theta+\delta)^\top h(x)} p_0(x) \mathrm{d}x = \sum_k \frac{\delta^k}{k!} \int h(x)^k e^{\theta^\top h(x)} p_0(x) \mathrm{d}x.$$

Derivatives of the potential

- The model
$$\log p(x|\theta) = \theta^\top h(x) - \sum_i a_i(x_i;\theta) - \psi(\theta)$$
- Suppose that a_i and ψ are smooth, and integrals and derivatives are exchangeable.
- The derivative of ψ is the expectation of h:
$$\partial_\alpha \psi(\theta) = \int h_\alpha(x)p(x)\mathrm{d}x$$
- The Hessian matrix is equal to the Fisher information
$$\partial_\alpha\partial_\beta\psi(\theta) = \int \{h_\alpha(x) - \textstyle\sum_i \partial_\alpha a_i(x_i;\theta) - \partial_\alpha\psi(\theta)\} \cdot \{h_\beta(x) - \textstyle\sum_i \partial_\beta a_i(x_i;\theta) - \partial_\beta\psi(\theta)\}p(x)\mathrm{d}x,$$
which is not equal to the covariance of h.
- The quantity $\sum_i \partial_\alpha a_i + \partial_\alpha\psi$ is the orthogonal projection of $h_\alpha(x)$ to the space of additive functions $\{\sum_i b_i(x_i)\}$.

Comparison

exp. family	"mindemo"
θ	θ
h	h
ψ	ψ
—	a_i
$\nabla\psi = \mathrm{E}(h)$	$\nabla\psi = \mathrm{E}(h)$
$\nabla\nabla^\top\psi = \mathrm{Cov}(h)$	$\nabla\nabla^\top\psi \neq \mathrm{Cov}(h)$

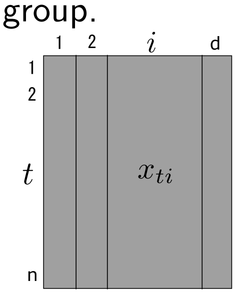
Table of contents

Conditional inference

- Consider a random sample $x(t) = (x_i(t))_{i=1}^d$ for $1 \leq t \leq n$.
- Decompose the data into "marginal" and rank statistics:

$$M = (\{x_{1i}, \ldots, x_{ni}\})_{1 \leq i \leq d}, \quad \pi = (\pi_1, \ldots, \pi_d) \in S_n^d$$

where S_n is the permutation group.



- Decomposition of the likelihood

$$\prod_{t=1}^{n} p(x(t); \theta, \nu) = f(\pi | M; \theta) \times g(M; \theta, \nu)$$

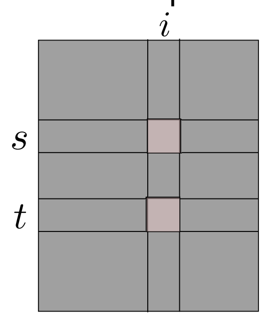
- We use the conditional likelihood $f(\pi|M;\theta)$ for inference of θ.
- M must have almost no information about θ! (ancillary statistic)

Parameter estimation

- The conditional likelihood is an exponential family on S_n^d:

$$f(\pi|M;\theta) = \frac{e^{\sum_{t=1}^n \theta^\top h((M\circ\pi)(t))}}{\sum_{\bar{\pi}\in S_n^d} e^{\sum_{t=1}^n \theta^\top h((M\circ\bar{\pi})(t))}}, \quad \pi \in S_n^d.$$

- The maximizer of $f(\pi|M;\theta)$ is called the conditional maximum likelihood estimator.
- We can sample π by MCMC. The implementation is not difficult.



- MCMC + conditional MLE (Geyer and Thompson 1992).
- Other option: Besag's pseudo likelihood estimator (Mukherjee 2016).

Consistency of conditional MLE

- The conditional MLE has consistency ($\hat{\theta} \to \theta$ as $n \to \infty$).
- Assumptions
  1. The domain $\mathcal{X}$ satisfies some metric entropy condition. The function h is Lipscitz. The parameter space Θ is bounded.
  2. a_i is smooth in θ.

Theorem 1 (SY 2023)

Under Assumption 1, the likelihood ratio is close to the conditional likelihood ratio. (a rough statement)

Corollary (SY 2023)

Under Assumptions 1 and 2, we have $\hat{\theta} \to \theta$ in probability.

Open problem 3

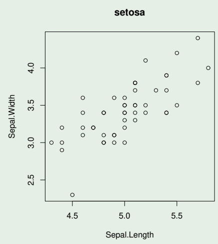
Is the conditional MLE asymptotically normal and efficient?

- For finite-set cases, Haberman (1977) proved this is correct.

Conditional inference works

Example (Fisher's iris data)

- Consider sepal length and sepal width of Setosa. $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$.



- We compared the MLE of Gaussian model and the conditional MLE of the proposed model with $h(x_1, x_2) = x_1 x_2$. Only the difference is whether the marginal model is specified or not.
- Result: $\hat{\theta}_{\mathrm{MLE}} = 1.66$ and $\hat{\theta}_{\mathrm{CLE}} = 1.72$ (standard errors are 0.4).
- Almost the same!

Other applications: penguin data and earthquake data. See SY (2023).

Conclustion

- We proposed the minimum information dependence model specified by canonical statistic h and the marginal densities r_i.
- Estimation is performed via conditional likelihood and permutation.
- The model is applicable to any domain of data.

Open problems:

- Further examples
- Properties of adjusting and potential functions
- Proof of asymptotic efficiency of the conditional MLE (or other estimators)

Thank you for your attention.

References I

- Amari, S. (2001). Information geometry on hierarchy of probability distributions, *IEEE Trans. Info. Theo.*, **47** (5), 1701–1711.
- Arnold, R. and Jupp, P. (2013). Statistics of orthogonal axial frames, *Biometrika*, **100** (3), 571–586.
- Bedford, T. (2002). Interactive expert assignment of minimally-informative copulae, In: Lindqvist B, editor. Proceedings of mathematical methods in reliability.
- Bedford, T. and Wilson, K. J. (2014). On the construction of minimum information bivariate copula families, *Ann. Inst. Stat. Math.*, **66**, 703–723.
- Bedford, T., Daneshkhah, A. and Wilson, K. J. (2016). Approximate uncertainty modeling in risk analysis with vine copulas, *Risk Analysis*, **36** (4), 792–815.
- Borwein, J. M., Lewis, A. S. and Nussbaum, R. D. (1994). Entropy minimization, DAD problems, and doubly stochastic kernels, *J. Funct. Anal.*, **123**, 264–307.
- Chen, Y. and Sei, T. (2022). A proper scoring rule for minimum information copulas, 2022, preprint. arxiv:2204.03118.
- Choi, L., Blume, J. D. and Dupont, W. D. (2015). Elucidating the foundations of statistical inference with 2 x 2 tables, *PLoS ONE*, **10** (4), e0121263, 1–22.
- Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Ann. Probab.*, **3** (1), 146–158.
- Dempster, A. P. (1972). Covariance selection, *Biometrics*, **28**, 157–175.

27 / 29

References II

- Eckstein, S. and Nutz, M. (2021). Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm, arXiv:2110.06798.
- Geenens, G. (2020). Copula modeling for discrete random vectors, *Dependence Modeling*, **8** (1), 417–440.
- Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, *J. Roy. Statist. Soc. Ser. B*, **54** (3), 657–683.
- Haberman, S. J. (1977), *The Analysis of Frequency Data*, Univ. of Chicago Press.
- Hayakawa, J. and Takemura, A. (2016). Estimation of exponential-polynomial distribution by holonomic gradient descent, *Comm. Statist. Theo. Meth*, **45** (23), 6860–6882.
- Jansen, M. J. W. (1997). Maximum entropy distributions with prescribed marginals and normal score correlations, In: Beneš, V. and Štěpán, J. (ed.), *Distributions with given Marginals and Moment Problems*, 87–92, Springer.
- Jaynes, E. T. (1957). Information theory and statistical mechanics, *Phys. Rev.*, **106** (4), 620–630.
- Jones, M.C., Pewsey, A. and Kato S. (2015). On a class of circulas: copulas for circular distributions, *Ann. Inst. Stat. Math*, **67**, 843–862.
- Meeuwissen, A. M. H. and Bedford, T. (1997). Minimally informative distributions with given rank correlation for use in uncertainty analysis, *J. Statist. Comput. Simul.*, **57**, 143–174.

28 / 29

References III

- Mukherjee, S. (2016). Estimation in exponential families on permutations, *Ann. Statist.*, **44** (2), 853–875.
- Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science, *Foundations and Trends in Machine Learning*, **11** (5-6), 355–607. (Preprint arxiv:1803.00567)
- Piantadosi, J., Howlett, P. and Borwein, J. (2012). Copulas with maximum entropy, *Optim. Lett.*, **6**, 99–125.
- Sei, T. and Yano, K. (2022). Minimum information dependence modeling, preprint, arxiv:2206.06792.
- Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices, *Pacific J. Math.*, **21** (2), 343–348.
- Takayama, N., Kuriki, S., Takemura, A. (2018) A-hypergeometric distributions and Newton polytopes, *Advances in Applied Mathematics*, **99**, 109–133.
- Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, **16** (1), 3813–3847.
- Zhang, M. and Bedford, T. (2018). Vine copula approximation: a generic method for coping with conditional dependence, *Stat. Comput.*, **28**, 219–237.
- Zhu, Y. and Reid, N. (1994). Information, ancillary, and sufficiency in the presence of nuisance parameters, *Canad. J. Stat.*, **22**, 111–123.

Stein identity, Poincaré inequality and exponential integrability on a metric measure space

Tomonari Sei (The University of Tokyo)*1
Ushio Tanaka (Osaka Metropolitan University)*2

Stein identity due to Stein [3] characterises a Gaussian random variable. The identity further implies Poincaré inequality in terms of the Gaussian random variable:

Theorem (Chernoff [1])**.** Let g be absolutely continuous function and X a random variable following standard normal distribution such that $g(X)$ has finite variance. Then it follows that

$$\mathrm{Var}(g(X)) \leq \mathbb{E}[g'(X)^2]$$

with equality if and only if $g(X)$ is linear.

We showed a discrete Stein identity based on the idea of Sei [2] and that it implies discrete Poincare inequality. We study them from the point of view of geometric analysis; in particular, we restrict our attention to implications with regard to discrete Poincaré inequality to show exponential integrability of a function on a discrete metric measure space.

References

[1] H. Chernoff, A note on an inequality involving the normal distribution, *Ann. Probab.* **9** (1981), 533–535.

[2] T. Sei, Coordinate-wise transformation of probability distributions to achieve a Stein-type identity, *Inf. Geom.* **5** (2022), 325–354.

[3] C.M. Stein, Estimation of the mean of a multivariate normal distribution, *Ann. Statist.* **9** (1981), 1135–1151.

*1 e-mail: `tomonari-sei@g.ecc.u-tokyo.ac.jp`
*2 e-mail: `utanaka@omu.ac.jp`

Infinite Dimentional Parameterized Measure Models

Hikaru Watanabe

The University of Tokyo

Shunichi Amari pointed out that generalization of finite dimentional Information Geometry to infinite dimentional Information Geometry is an important problem in his book “Methods of Information Geometry” (1993) [1]. After that, in 2015, Nihat Ay, Jürgen Jost, Hông Vân Lê and Lorenz Schwachhöfer proposed the notion of (infinite dimensional) parameterized measure models and statistical models, which is a natural generalization of finite dimensional statistical manifolds ([2]). Moreover, they proved that Exponential statistical manifold, which is built by Giovanni Pistone and Carlo Sempi ([4]), can be regarded as an example of infinite dimensional statistical models.

In this talk, we prove that Exponential manifold by reproducing kernel Hilbert spaces, which is constructed by Kenji Fukumizu ([3]), is also an example of infinite dimensional statistical models. Furthermore, we replace reproducing kernel Hilbert spaces with reproducing kernel Banach spaces (see [5] for reproducing kernel Banach spaces). The aim of this replacement is to show that existing finite dimensional open exponential family is obtained when a certain reproducing kernel Banach space is set.

Moreover, if a parameterized measure model or a statistical model satisfies the condition called n-integrability, we can obtain covariant n-tensor. Especially, covariant 2-tensor is Fisher metric if it is positive definite, and covariant 3-tensor is Amari-Chentsov tensor. However we must be carefull about tensor fields on infinite dimentional manifolds because there are two kinds of defnition of tensor fields on infinite dimentional manifolds, unlike tensor fields on finite dimensional manifolds. Nihat Ay, Jürgen Jost, Hông Vân Lê Lorenz Schwachhöfer introduced covariant n-tensor as C^0 “weak tensor fields”. However in this talk we show that one can introduce covariant n-tensors as C^∞ “strong tensor fields” when parameterized measure model or statistical model is C^∞.

References

[1] Amari, S and Nagaoka, H.: Methods of Information Geometry . Iwanami (1993)

[2] Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry, Ergebnisse der Mathematikund ihrer Grenzgebiete. Springer, Berlin (2017)

[3] Fukumizu, K.: Exponential Manifold by Reproducing Kernel Hilbert Spaces, pp. 291–306. Cambridge University Press, Cambridge (2010). https://doi.org/10.1017/CBO9780511642401.019

[4] Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. 23(5), 1543–1561 (1995)

[5] Lin, R., Zhang, H., Zhang, J.: On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions, (2019). arXiv:1901.01002.

Infinite dimensional parameterized measure model

Hikaru Watanabe

The University of Tokyo

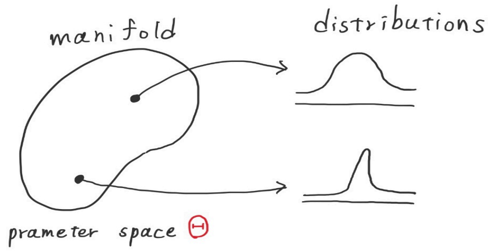
December 15, 2023

table of contents

Rao(1945)

Rao: Regarding statistical model as manifold

Fisher metric: Riemannian metric



exponential family and Fisher metric

Example 1 (exponential family)

Ω: m'ble set
$\mu \in P(\Omega)$
$\Theta \subset \mathbb{R}^n$: parameter space
$f_1, \cdots, f_n$: m'ble function on Ω
Some assumptions
Exponential family is a family of distributions

$$P := \{ \frac{\exp(\sum_{k=1}^{n} \theta_k f_k)}{\int_\Omega \exp(\sum_{k=1}^{n} \theta_k f_k) d\mu} \mu \mid \theta \in \Theta \}$$

Example 2

examples of exponential family:
Normal distributions, Bernoulli distributions, etc.

Proposition 3 (Fisher metric)

$$g_\theta(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}) = E_\theta[(f_i - E_\theta[f_i])(f_j - E_\theta[f_j])]$$

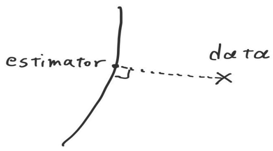
$\hookrightarrow$variance form

previous works
questions
result

previous works about finite dimensional information geometry
previous works about infinite dimensional information geometry

Efron(1975)

Efron : connecting geometry and statistics
statistical curvature: geometrical
statistical estimation by statistical curvature



Hikaru Watanabe Infinite dimensional parameterized measure model

previous works
questions
result

previous works about finite dimensional information geometry
previous works about infinite dimensional information geometry

Amari(1980)

Amari : formalizing connecting geometry and statistics
α, e, m-connection
statistical estimation by e, m-connection
dually flat structure
AC-tensor

Hikaru Watanabe Infinite dimensional parameterized measure model

AC-tensor

Example 4 (exponential family)

Ω: m'ble set
$\mu \in P(\Omega)$
$\Theta \subset \mathbb{R}^n$: parameter space
$f_1, \cdots, f_n$: m'ble function on Ω
Some assumptions
Exponential family is a family of distributions

$$P := \{\frac{\exp(\sum_{k=1}^n \theta_k f_k)}{\int_\Omega \exp(\sum_{k=1}^n \theta_k f_k) d\mu} \mu \mid \theta \in \Theta\}$$

Proposition 5 (AC-tensor)

$$\Gamma_f(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}, \frac{\partial}{\partial \theta_k}) = E_\theta[(f_i - E_\theta[f_i])(f_j - E_\theta[f_j])(f_k - E_\theta[f_k])]$$

$\hookrightarrow$3-rd moment around the average value

Amari(1980)

Amari: pointing out the importance of a generalization to infinite dimension

This is because

1. Finite dimensional statistical models are small compared to P(Ω).
2. We want to solve estimation problems in infinite dimensional spaces.
3. If mathematical rigorousness is ignored, infinite dimensional information geometry can be built.

problems based on the pointout

problems based on the pointout

1. How to formalize infinite dimensional manifold?
2. How to regard infinite dimensional distributions as infinite dimensional manifold?
3. How to define geometrical notions such as Fisher metric and AC-tensor on infinite dimensional information geometry?

Lang(1962)

(1)How to formalize infinite dimensional manifold?
$\hookrightarrow$ Lang: Banach manifold
Banach manifold: manifold modeled on Banach space

three studies

(2)How to regard infinite dimensional distributions as infinite dimensional manifolds?

↪ There are three studies.

(2-1) Pistone and Sempi(1995):
exponential statistical manifold by Orlicz space

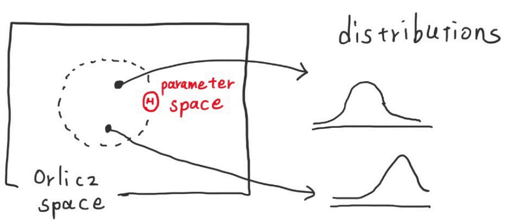
(2-2) Fukumizu(2010):
exponential manifold by reproducing kernel Hilbert space

(2-3) Ay, Jost, Lê and Schwachhöfer(2017):
parameterized measure model and statistical model

Pistone and Sempi(1995)

(2)How to regard infinite dimensional distributions as infinite dimensional manifolds?

↪ Pistone and Sempi: exponential statistical manifold by Orlicz space



Pistone and Sempi(1995)

Definition 6 (Pistone and Sempi(1995))

$\mu \in P(\Omega)$
Θ: a subset of a certain Orlicz sp.
exponential statistical manifold is a distributions

$$P := \{\frac{e^f}{\int_\Omega e^f \, d\mu}\mu \mid f \in \Theta\}$$

is a generalization of

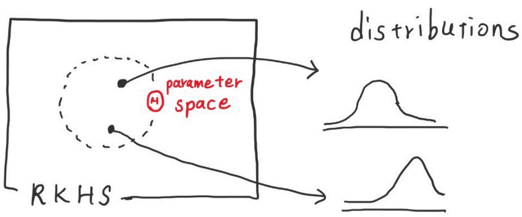
Example 7 ((existing finite dimensional) exponential family)

$$P := \{\frac{\exp(\sum_{k=1}^n \theta_k f_k)}{\int_\Omega \exp(\sum_{k=1}^n \theta_k f_k) d\mu}\mu \mid \theta \in \Theta\}$$

Fukumizu(2010)

(2)How to regard infinite dimensional distributions as infinite dimensional manifolds?

$\hookrightarrow$ Fukumizu: exponential manifold by reproducing kernel Hilbert space. reproducing kernel Hilbert space$\rightarrow$ RKHS



RKHS

Definition 8 (RKHS)

Ω: set
$(H, \langle \cdot, \cdot \rangle)$: Hilbert space of functions on Ω
$^{\exists}K : \Omega \times \Omega \to \mathbb{R}$: positive definite (kernel function)

1. $K(x, \cdot) \in H \quad (x \in \Omega)$
2. $f(x) = \langle f, K(x, \cdot) \rangle \quad (x \in \Omega, f \in H)$

exponential manifold

Definition 9 (Fukumizu(2010))

Ω: top sp (m'ble sp by Borel sets)
$\mu \in P(\Omega)$
H: an RKHS on Ω
Θ: a subset of H
Some assumptions
exponential manifold is a family of distributions

$$P := \{ \frac{e^f}{\int_\Omega e^f d\mu} \mu \mid f \in \Theta \}$$

is a generalization of

Example 10 ((existing finite dimensional) exponential family)

$$P := \{ \frac{\exp(\sum_{k=1}^n \theta_k f_k)}{\int_\Omega \exp(\sum_{k=1}^n \theta_k f_k) d\mu} \mu \mid \theta \in \Theta \}$$

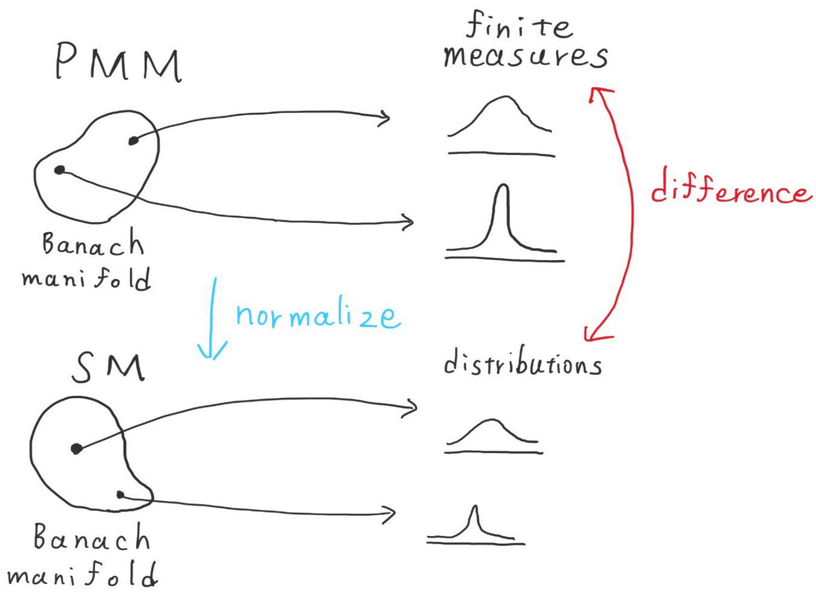
Ay, Jost, Lê and Schwachhöfer(2017)

(2)How to regard infinite dimensional distributions as infinite dimensional manifolds?

$\hookrightarrow$ Ay, Jost, Lê and Schwachhöfer(2017) : parameterized measure model and statistical model

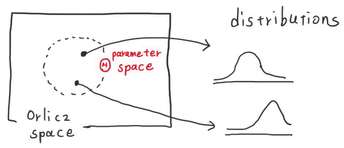
parameterized measure model $\rightarrow$ PMM
statistical model $\rightarrow$ SM
PMM and SM : a generalization of (existing finite dimensional) statistical models

PMM and SM

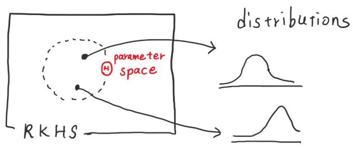


examples of PMM and SM

Exponential statistical model by Pistone and Sempi



is a SM.
Exponential manifold by Fukumizu



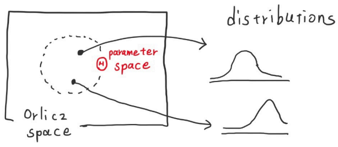
is a SM.

Ay, Jost, Lê and Schwachhöfer(2017)

(3)How to define geometrical notions such as Fisher metric and AC-tensor on infinite dimensional information geometry?
$\hookrightarrow$ Ay, Jost, Lê and Schwachhöfer: covariant n-tensor
covariant n-tensor: a generalization of Fisher metric and AC-tensor.
covariant 2-tensor: Fisher metric in PMM and SM
covariant 3-tensor: AC-tensor in PMM ans SM

Ay, Jost, Lê and Schwachhöfer(2017)

In the case of the PMM before normalized to exponential statistical model



Proposition 11 (Ay, Jost, Lê and Schwachhöfer(2017))

τ^n : covariant n-tensor

$$\tau_f^n(v_1, \cdots, v_n) = E_f[v_1 \cdots v_n]$$

where $f \in \Theta, v_1, \cdots, v_n \in$ (this Orlicz space).
Especially covariant $2, 3$-tensor is as follows.

$$\tau_f^2(v_1, v_2) = E_f[v_1 v_2]$$

$$\tau_f^3(v_1, v_2, v_3) = E_f[v_1 v_2 v_3]$$

questions

We want to understand

I Are there any other good examples of parameterized measure models and statistical models?

II Give explicit formula of covariant n-tensor on normalized statistical models.

other good examples

I Are there any other good examples of parameterized measure models and statistical models?
(I think) we want more examples.

covariant $2, 3$-tensor

In finite dimensional exponential family

$$g_\theta(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}) = E_\theta[(f_i - E_\theta[f_i])(f_j - E_\theta[f_j])]$$

$$\Gamma_f(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}, \frac{\partial}{\partial \theta_k}) = E_\theta[(f_i - E_\theta[f_i])(f_j - E_\theta[f_j])(f_k - E_\theta[f_k])]$$

In the PMM before normalized to exponential statistical manifold

$$\tau_f^2(v_1, v_2) = E_f[v_1 v_2]$$

$$\tau_f^3(v_1, v_2, v_3) = E_f[v_1 v_2 v_3]$$

other good examples

- Are there any other good examples of parameterized measure models and statistical models?

 $\hookrightarrow$ We can construct exponential manifolds by reproducing kernel Banach space.

reproducing kernel Banach space(RKBS): Banach spaces of functions, a generalization of RKHS to Banach space

RKBS

Definition 12

Ω: set
B: Banach space of functions on Ω.
$B \to \mathbb{R}, f \mapsto f(x)$ is continuous.

Is RKBS really generalization of RKHS?

Definition 13

Ω: set
$(H, \langle \cdot, \cdot \rangle)$: Hilbert space of functions on Ω
$^{\exists}K : \Omega \times \Omega \to \mathbb{R}$: positive definite (kernel function)

1. $K(x, \cdot) \in H \quad (x \in \Omega)$
2. $f(x) = \langle f, K(x, \cdot) \rangle \quad (x \in \Omega, f \in H)$

$\Leftrightarrow H \to \mathbb{R}, f \mapsto f(x)$ is continuous.

difficulty

What is the difficulty of RKHS $\to$ RKBS?
$\hookrightarrow$ non existing of kernel function K

Proposition 14 (2023)

$\sqrt{K(x,x)} = \|e_x\|_{B^*}$

Theorem 15 (2023)

We can obtain exponential manifold by RKBS when we replace RKHS with RKBS and $\sqrt{K(x,x)}$ with $\|e_x\|_{B^*}$ in the discussion of Fukumizu.

merit

What is a benefit of RKBS against RKHS?
$\hookrightarrow$ examples of RKBS not RKHS Ω: top sp

Example 16

$$C_b(\Omega) := \{f : \Omega \to \mathbb{R} \mid f \text{ is continuous and bounded.}\}$$

$$\|f\|_{C_b(\Omega)} := \sup|f(x)|$$

is one of examples of RKBS.

Example 17

$$C_0(\Omega) := \{f : \Omega \to \mathbb{R} \mid f \text{ is continuous and vanishing at } \infty\}$$

$$\|f\|_{C_0(\Omega)} := \sup|f(x)|$$

is one of examples of RKBS.

covariant n-tensor

II Give explicit formula of covariant n-tensor on normalized statistical models.

Theorem 18 (2023)

X: Banach manifold
$p\colon X \to P(\Omega)$
(X, p): a parameterized measure model
(Y, q): the normalization of (X, p) (statistical model)
Under some assumption

$$\tau_y^n(v_1, \cdots, v_n) = E_\mu[(f_1 - E_\mu[f_1]) \cdots (f_n - E_\mu[f_n])$$

where $y \in Y, \mu := q(y), f_i\mu := dp_y(v_i)$, τ^n:covariant n-tensor on (Y, q).

In short, using this formula from the calculation on PMM we can obtain covariant n-tensor of normalized SM.

previous works
questions
result
other good examples
covariant n-tensor

covariant n-tensor

Theorem 19 (2023)

τ^n: covariant n-tensor on exponential statistical manifold

$$\tau_f^n[v_1, \cdots, v_n] = E_f[(v_1 - E_f[v_1]) \cdots (v_n - E_f[v_n])]$$

Especially covariant $2, 3$-tensor are

$$\tau_f^2[v_1, v_2] = E_f[(v_1 - E_f[v_1])(v_2 - E_f[v_2])]$$

$$\tau_f^3[v_1, v_2, v_3] = E_f[(v_1 - E_f[v_1])(v_2 - E_f[v_2])(v_3 - E_f[v_3])]$$

Roughly covariant 2-tensor is a form of variance and covariant 3-tensor is a form of 3-rd moment around the average value.

reference

- Amari, S and Nagaoka, H.: Methods of Information Geometry . Iwanami (1993)
- Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry, Ergebnisse der Mathematikund ihrer Grenzgebiete. Springer, Berlin (2017)
- Fukumizu, K.: Exponential Manifold by Reproducing Kernel Hilbert Spaces, pp. 291–306. Cambridge University Press, Cambridge (2010). https://doi.org/10.1017/CBO9780511642401.019
- Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. 23(5), 1543–1561 (1995)
- Lin, R., Zhang, H., Zhang, J.: On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions, (2019). arXiv:1901.01002.

Neural-Kernel Conditional Mean Embeddings

Eiki Shimizu (SOKENDAI)

With a positive definite kernel, conditional distributions can be embedded into an associated reproducing kernel Hilbert space (RKHS). Such approaches are referred to as Kernel Conditional Mean Embeddings (CMEs), and have been applied to various applications such as causal inference and kernelized Bayes rule.

Although several appealing theoretical properties have been shown, CMEs are yet to be suitable for modern ML tasks: the empirical estimates involve inversion of a Gram matrix, and the hyperparameter selection is not straightforward. To address these issues, we propose a method that combines neural networks (NNs) with CMEs.

Our approach is simple, and can be interpreted as a NN trained with a RKHS loss. This allows us to replace the matrix inversion with a NN model, while taking advantage of NN's ability to learn useful features. We also provide a strategy to efficiently optimize the hyperparameter of the kernel, without relying on median heuristics or cross validation. We demonstrate the effectiveness of our method with ML related tasks, where the estimation of conditional distribution plays an important role.

Neural-Kernel Conditional Mean Embeddings

Eiki Shimizu, SOKENDAI

Intro

- PhD student working with Prof. Kenji Fukumizu
- Interested in : Kernel methods, Bayesian Inference
  and their applications to Deep Learning models
- This presentation is about
  Kernel method + DeepNN = useful conditional density estimator

Outline

1. Brief review of kernel methods and kernel mean embeddings

2. Propose a new conditional density estimator,
   Share experimental results

3. Applications to more complicated ML tasks: Reinforcement Learning

Section 1

RKHS and Kernel Mean Embeddings

RKHS and Notations

RKHS

A symmetric function $k_{\mathcal{X}}: \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}_{\mathcal{X}}$ if and only if

- $\forall x \in \mathcal{X},\ k_{\mathcal{X}}(x,\cdot) \in \mathcal{H}_{\mathcal{X}}$
- $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}_{\mathcal{X}}, f(x) = \langle f, k_{\mathcal{X}}(x,\ \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}$

A space $\mathcal{H}_{\mathcal{X}}$ is called reproducing kernel Hilbert space (RKHS)

Notations

Let (X, Y) be a random variable on $\mathcal{X}\times\mathcal{Y}$ with distribution P and density function $p(x, y)$, and $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive definite kernel corresponding to $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ respectively.

We denote feature maps as $\psi(x) = k_{\mathcal{X}}(x,\ \cdot)$ and $\phi(y) =\ k_{\mathcal{Y}}(y,\ \cdot)$

Kernel Mean Embeddings

$m_{P(X)} = \mathrm{E}_P[\psi(X)] \in \mathcal{H}_{\mathcal{X}},\ \langle f, m_{P(X)}\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathrm{E}_P\,[f(X)]$

The embedding uniquely defines the probability distribution (the mapping is injective) if the kernel k_x is *characteristic*. Popular kernels like Gaussian kernel has this property.

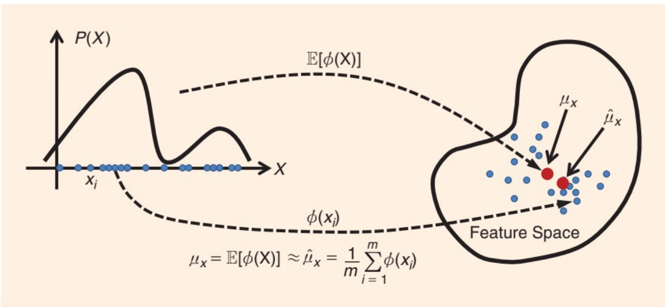


Figure taken from [1]

Kernel Covariance Operators

$C_{XX} = \mathrm{E}[\psi(X) \otimes \psi(X)]$, $C_{XY} = \mathrm{E}[\psi(X) \otimes \phi(Y)]$

Generalizes finite-dimensional covariance matrices to the case of infinite feature spaces. Always exist for bounded kernels.

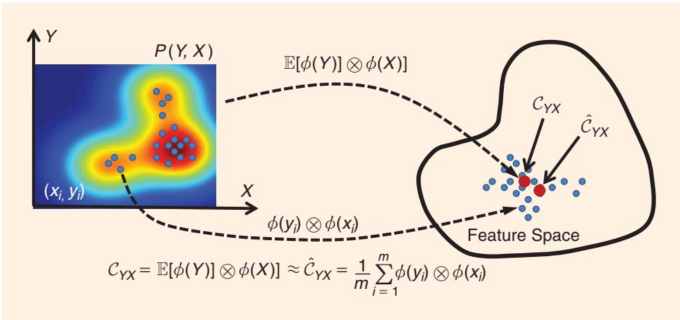


Figure taken from [1]

Kernel Conditional Mean Embeddings (CME)

$m_{P(Y|X)}(x) = \mathrm{E}_P[\phi(Y)|X = x] \in \mathcal{H}_\mathcal{Y}$

Under condition $\mathrm{E}_P[g(Y)|X = x] \in \mathcal{H}_\mathcal{X}$ for all $g \in \mathcal{H}_\mathcal{Y}$, there exists an operator $C_{Y|X}$ such that $m_{P(Y|X)}(x) = C_{Y|X}\psi(x)$.

This is a minimizer of the RKHS loss l

$$l(C_{Y|X}) = \mathrm{E}_P\left[\left\|\phi(Y) - C_{Y|X}\psi(X)\right\|_{\mathcal{H}_\mathcal{Y}}^2\right]$$

The closed form solution is

$$C_{Y|X} = C_{YX}(C_{XX})^{-1}$$

Empirical Estimate

$$\hat{l}(C_{Y|X}) = \frac{1}{n}\sum_{i=1}^{n}\left\|\phi(y_i) - C_{Y|X}\psi(x_i)\right\|_{\mathcal{H}_y}^2 + \lambda\left\|C_{Y|X}\right\|_{HS}$$

The solution is:

$$\hat{C}_{Y|X} = \boldsymbol{\phi}(\mathbf{K}_{\mathcal{X}} + \lambda I)^{-1}\boldsymbol{\psi}^{\top}$$

where, $\boldsymbol{\phi} = (\phi(y_1), \ldots, \phi(y_n))$, $\boldsymbol{\psi} = (\psi(x_1), \ldots, \psi(x_n))$, $\mathbf{K}_{\mathcal{X}} = \boldsymbol{\psi}^{\top}\boldsymbol{\psi}$

Thus,

$$\widehat{m}_{P(Y|X)}(x) = \sum_{i=1}^{n}\beta_i(x)\phi(y_i) = \boldsymbol{\phi}\boldsymbol{\beta}(x)$$

where, $\boldsymbol{\beta}(x) = (\mathbf{K}_{\mathcal{X}} + \lambda I)^{-1}\mathbf{k}_{\mathcal{X}}$

Interpretations

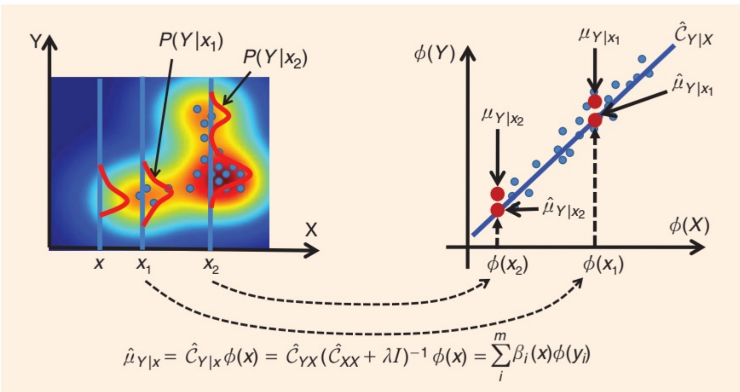


Figure taken from [1]

- Regression in the function space
- If $k_{\mathcal{Y}}$ is a linear kernel, this is just a Kernel Ridge Regression
- Weighted particle view

$$\widehat{m}_{P(Y|X)}(x) = \sum_{i=1}^{n}\beta_i(x)\phi(y_i)$$

$\beta_i(x)$ weights particles $\phi(y_i)$
Note that the weights does not necessarily have to be positive nor sum up to one

Kernel herding

Can we "sample" from the KMEs?

$$\widehat{m}_P = \sum_{i=1}^{N} w_i k(x_i,\cdot) \to \hat{x}_{1,\ldots,}\ \hat{x}_m \text{ (samples)}$$

Kernel herding is a deterministic sampling approach used to obtain super samples $\hat{x}_m$. Selects $\tilde{x}$ greedily and iterates the following:

$$\hat{x}_{l+1} = \mathrm{argmin}_{\tilde{x}} \left\| \frac{1}{l+1} \left\{ \sum_{j=1}^{l} k(\hat{x}_j,\cdot) + k(\tilde{x},\cdot) \right\} - \widehat{m}_P \right\|_{\mathcal{H}_k}^2$$

Practical limitations

- The computational cost of $(\mathbf{K}_{\mathcal{X}} + \lambda I)^{-1}$ is $\mathrm{O}(n^3)$, and does not scale to large dataset
- RKHS features are pre-specified feature maps. This may lead to poor performance when input variables are high-dimensional, or possess highly non-linear structure
- Hyperparameter selection for $k_{\mathcal{X}}$, $k_{\mathcal{Y}}$ and λ is not straightforward. While the choice significantly affects the performance, particularly for $k_{\mathcal{Y}}$, standard procedures like cross-validation can not be applied

Section 2

Proposal: Integrate CME with DNNs

Proposal: Big idea

CME Recap:

$$\widehat{m}_{P(Y|X)}(x) = \sum_{i=1}^{n} \beta_i(x)\phi(y_i) = \boldsymbol{\Phi}\boldsymbol{\beta}(x)$$

where, $\boldsymbol{\beta}(x) = (\mathbf{K}_{\mathcal{X}} + \lambda I)^{-1}\mathbf{k}_{\mathcal{X}}$

$\boxed{\text{Why don't we just replace } \boldsymbol{\beta}(x) \text{ with a DNN model?}}$

$$\widehat{m}_{P(Y|X)}(x) = \sum_{a=1}^{M} \phi(\eta_a) f_a(x;\theta)$$

where, $f(x;\theta)\colon \mathcal{X} \to \mathbb{R}^M$ is DNN, and $\eta \in \mathcal{Y}$ corresponds to M atoms/particles

Objective function (for CDE)

$$\min_\theta \frac{1}{n}\sum_{i=1}^{n}\left\|\phi(y_i)-\sum_{a=1}^{M}\phi(\eta_a)f_a(x_i;\theta)\right\|_{\mathcal{H}_{\mathcal{Y}}}^2$$

$$\Leftrightarrow \boxed{\min_\theta \frac{1}{n}\sum_i\left\{k_{\mathcal{Y}}(y_i,y_i)-2\sum_a k_{\mathcal{Y}}(y_i,\eta_a)f_a(x_i;\theta)+\sum_{a,b}k_{\mathcal{Y}}(\eta_a,\eta_b)f_a(x_i,\theta)f_b(x_i,\theta)\right\}}$$

Simply use this loss function, and everything else (e.g. implementation, training procedures) is the same as the standard DNN!

"What's there to be happy about? Job's not finished." Kobe Bryant

Positives (solves first two issues):
- No matrix inversion!
- Can use mini-batch optimization for efficient training
- DNN models learn features: $k_{\mathcal{X}}$ implicitly tuned!

Negatives:
- We still have $k_{\mathcal{Y}}$ left to be tuned, and this is not easy
- The objective function is defined in terms of the RKHS norm of $\mathcal{H}_{\mathcal{Y}}$. If we change the kernel parameter, the definition of the objective function also changes

What should we do with k_y?

Empirically, we find the following strategy to work quite well

1. Use a positive definite kernel that also has density interpretation

We use the Gaussian-Density Kernel

$$k_\sigma(y, y') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|y - y'\|^2}{2\sigma^2}\right)$$

2. Use the following objective function (will just call it the RKHS loss):

$$\min_{\theta,\sigma} \frac{1}{n} \sum_i \left\{-2 \sum_a k_\sigma(y_i, \eta_a) f_a(x_i; \theta) + \sum_{a,b} k_\sigma(\eta_a, \eta_b) f_a(x_i, \theta) f_b(x_i, \theta)\right\}$$

Experiments (Only to be shared during the presentation)

- Note that this work is still in progress
- Experiments on 1-dimensional conditional density estimation
  (output variables are 1-dimensional, but input variables can be multi-dimensional)
- Show preliminary results on:

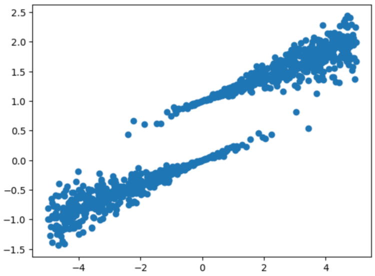
1. Toy data simulations

2. More realistic setting with the UCI dataset

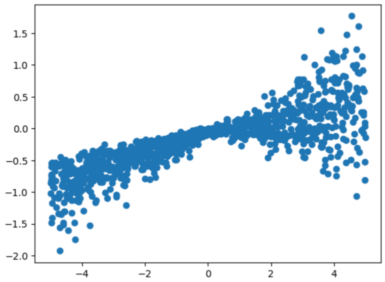
Toy data settings

- 3 settings

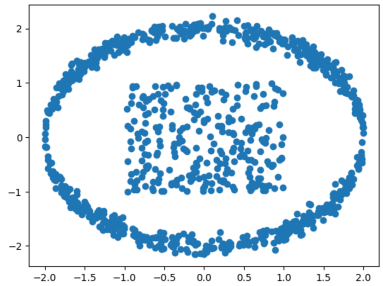
Bimodal (MoG with hetero noise)



Skewed normal



Ring & Box



Evaluation

1. Train each model with 5000 training data
2. Generate enough samples from $\hat{p}(y|X = x)$
3. Calculate the Wasserstein distance between samples and the true conditional distribution for each x, and average them in the end

Wasserstein distance:

$$l_1(u, v) = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| d\pi(x, y)$$

where, $\Gamma(u, v)$is the set of distribution whose marginals are u and v

Competitors Part1

Deep Feature approach (DF)

Learn adaptive feature $\psi_\theta(x)$ represented by DNN:$\mathcal{X} \to \mathbb{R}^d$

$\boldsymbol{\beta}(x) = \boldsymbol{\Psi}_{\boldsymbol{\theta}}^\top \left(\boldsymbol{\Psi}_{\boldsymbol{\theta}} \boldsymbol{\Psi}_{\boldsymbol{\theta}}^\top + \lambda I\right)^{-1} \psi_\theta(x)$

Several successful applications such as in causal inference (IV regression) [3]

- Positives

Explicitly learn feature, The matrix inverse can be done with $O(d^3)$, Compatible with mini-batch optimization

- Negatives

Still need to tune λ and the kernel hyperparameter of k_y

For k_y, typically rely on the median heuristic

$$\sigma = \text{median}\{\|X_i - X_j\| \mid i, j = 1, \dots, n\}$$

Competitors Part2

Mixture Density Networks (MDN)

DNN learns the weight, the mean and the variance of a GMM

$$\hat{p}(y|x) = \sum_{k=1}^{K} w_k(x;\theta)\mathcal{N}(y|\mu_k(x,\theta), \sigma_k^2(x;\theta))$$

Several successful applications such as in likelihood-free inference [4]

- Positives

Easy to implement, sample and evaluate the log-likelihood

- Negatives

Optimizing 3 parameters of the GMM jointly may lead to numerical instability and over-fitting

UCI datasets

- "Real world" dataset

Dataset	Boston	Concrete	Energy	Kin8nm	Naval	Power	Protein
(N, P)	$(506, 13)$	$(1030, 8)$	$(768, 8)$	$(8192, 8)$	$(11, 934, 16)$	$(9568, 4)$	$(45, 730, 9)$

- Both input and output variables are normalized before training

Evaluation

Quantile Interval Coverage Error (QICE) [5]:

1. Generate enough samples from $\hat{p}(y|X = x)$
2. Divide them into M=10 bins and get 10 quantile intervals with the boundary $\hat{y}_n^{\mathrm{low}_m}$ and $\hat{y}_n^{\mathrm{high}_m}$
3. Calculate

QICE $= \frac{1}{M}\sum_{m=1}^{M} \left| r_m - \frac{1}{M} \right|$, where $r_m = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{y_n \geq \hat{y}_n^{\mathrm{low}_m}} \cdot \mathbf{1}_{y_n \geq \hat{y}_n^{high_m}}$

In the optimal scenario, about 10% of true data shall fall into each of the 10 quartile intervals, and QICE reaches 0

New competitor: The Big Boss

Diffusion model

Used in "Generative AI", super strong performance on image generation tasks

[5] proposed conditional version of this model, enabling flexible conditional density modelling

Though this model demonstrates SOTA level performance on several tasks, it is computationally costly (requires two NNs/optimizations), and the sampling may be slow

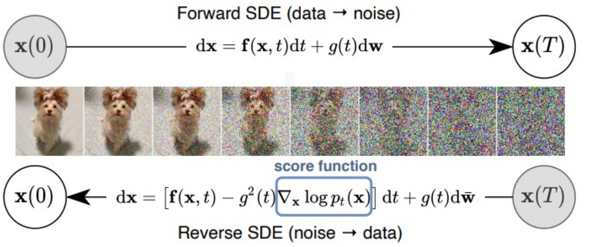


Figure 1: **Solving a reverse-time SDE yields a score-based generative model.** Transforming data to a simple noise distribution can be accomplished with a continuous-time SDE. This SDE can be reversed if we know the score of the distribution at each intermediate time step, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.

Figure taken from [6]

Section 3

Application to Distributional RL

Reinforcement Learning framework

1. An agent on S_t interacts with the environment (takes an action A_t)
2. Moves to the next state S_{t+1}, and gets a reward R_{t+1}
3. Repeat until the agent learns a good policy

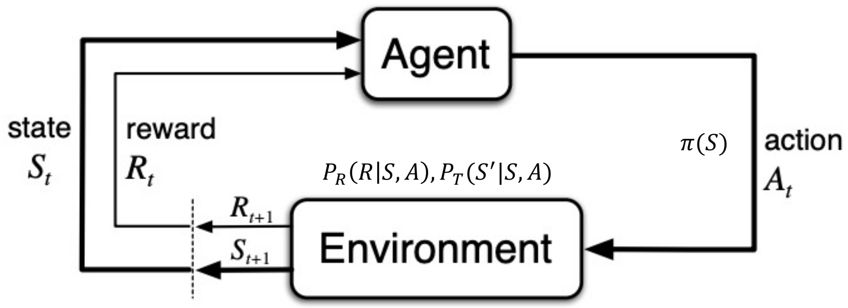


Figure taken from [7]

Q-learning: basic idea

Learn state-action value function Q^{π} of a policy π, which corresponds to expected discounted sum of rewards

$$Q^{\pi}(s,a) = \mathrm{E}_{s,r}\left[\sum_{t}^{\infty} \gamma^t r_t\right], a_t = \pi(s_t)$$

where, $\gamma \in [0,1)$ is the discount factor.

This satisfies the Bellman equation/backup:

$$Q^{\pi}(s,a) \leftarrow \mathrm{E}_{s',r}[r + \gamma Q^{\pi}(s', \pi(s'))]$$

Similarly, define Bellman operator $\mathcal{T}^{\pi}$:

$$\mathcal{T}^{\pi} Q^{\pi}(s,a) = \mathrm{E}_{s',r}[r + \gamma Q^{\pi}(s', \pi(s'))]$$

Q-learning: algorithm

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Figure taken from [7]

Notes
- $Q(s, a)$ can be approximated well by DNN
- In that case, we stop the gradient for $\max_a Q(s, a)$

Distributional RL (DRL)

Model distribution over state-action value instead of just expectation

$$Z^\pi(s, a) \overset{D}{=} \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$$

Distributional Bellman equation [8]:

$$Z^\pi(s, a) \leftarrow R(s, a) + \gamma P^\pi Z^\pi(s, a)$$

Distributional Bellman operator $\mathcal{T}^\pi$:

$$\mathcal{T}^\pi Z^\pi(s, a) \overset{D}{=} R(s, a) + \gamma P^\pi Z^\pi(s, a)$$

Categorical DQN (CDQN)

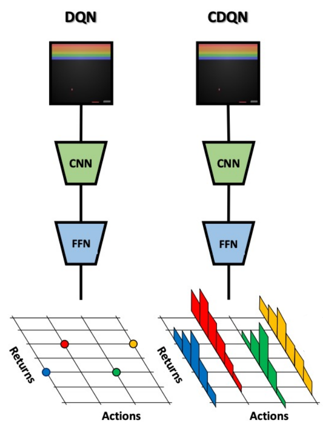


Figure taken from [9]

CDQN [8] basically models the distribution with histogram, and use cross-entropy loss for the loss function

A famous approach named "Rainbow" uses this CDQN structure and achieved SOTA performance on Atari games few years ago

Technically, there needs to be a heuristic step to enable cross-entropy loss to be used

Can we
- Represent the distribution better, and
- Compare two distributions in a more principled way?

Proposal: MCMD DRL

Maximum Conditional Mean Discrepancy (MCMD)[10]:

$$\widehat{\mathrm{MCMD}}^2\left(P_{Y|X}, P_{Y'|X'}\right) = \left\| \widehat{m}_{P(Y|X)}(x) - \widehat{m}_{P(Y'|X')}(x) \right\|^2_{\mathcal{H}_y}$$

Empirical estimate with our proposed model:

$$\sum_{a,b} k_y(\eta_a, \eta_b) f_a(x_i,\theta) f_b(x_i,\theta) - 2\sum_{a,b} k_y(\eta_a, \eta'_b) f_a(x_i,\theta) f'_b(x_i,\theta) + \sum_{a,b} k_y(\eta'_a, \eta'_b) f'_a(x_i,\theta) f'_b(x_i,\theta)$$

To make this Distributional Bellman equation, simply calculate

$$\widehat{\mathrm{MCMD}}\left(P_{Y|X}, \mathcal{T}^{\pi} P_{Y|X}\right)$$

Why DRL is the perfect application

In Q-learning framework,

- Evaluate the learned state-action value every step
- This is done for millions and billions of steps
- The accumulated “data”/experience can be as large as steps taken

In this case, we never want to see matrix inversion. Our approach can be applied to DQN-style learning in a straight manner, and offers principled way to represent and compare distributions.

Experimental setups/results only shared during the presentation

- Test on “Classic Control” provided by Gymnasium environment

- Though “classic”, SOTA DeepRL approaches can easily fail in these environments, with bad modelling or hyperparameters

- Run for 5,000,000 steps, and evaluate agents every 100 steps

Summary

- New conditional density estimator by combining KME and DNN
- Overcomes some of the KME limitations:
  matrix inverse, hyperparameter tunings
- Good performances on density estimation tasks
- Promising performances on RL tasks, address bandwidth selection issues at the same time

Future work: make it Bayesian

- Large scale settings / multi-dimensional density estimation
- Application to kernelized Bayes Rule
- Combine with Gaussian Process, Bayesian Deep Learning?
- Can we do the Bayesian model selection for hyperparameter tuning?
- Could some "Geometric methods" incorporated into our work/KMEs

References

[1] Song et.al. Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Methods. *IEEE Signal Proceeding Magazine*, 2013

[2] Hsu and Ramos. Bayesian Learning of Conditional Kernel Mean Embeddings for Automatic Likelihood-Free Inference. *Proceeding of the 22nd International Conference on Artificial Intelligence and Statistics*. 2019

[3] Xu et.al. Learning Deep Features in Instrumental Variable Regression. *International Conference on Learning Representations*. 2020

References

[4] Papamakorios and Murray. Fast ϵ-free inference of Simulation Models with Bayesian Conditional Density Estimation. *Advances in Neural Information Processing Systems*. 2016

[5] Han et.al. CARD: Classification and Regression Diffusion Models. *Advances in Neural Information Processing Systems*. 2022

[6] Song et.al. Score-Based Generative Modelling through Stochastic Differential Equations. *International Conference on Learning Representations*. 2021

[7] Sutton and Barto. Reinforcement Learning: An Introduction *MIT Press*. 2018

References

[8] Bellemare et.al. A Distributional Perspective on Reinforcement Learning. *Proceedings of 34th International Conference on Machine Learning*. 2017

[9] Theate et.al. Distributional Reinforcement Learning with Unconstrained Monotonic Neural Networks. *Neurocomputing*. 2023

[10] Park and Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. *Advances in Neural Information Processing Systems*. 2020

[11] Biggs et.al. MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting. *Advances in Neural Information Processing Systems*. 2023

Bonferroni method and tube method for heavy-tailed distributions

Satoshi Kuriki (Inst. Statist. Math.), Evgeny Spodarev (Ulm Univ.)

The Bonferroni method is the simplest method for approximating the maximum distribution of the statistics $T_1,\ldots,T_n$:

$$\mathbb{P}\Big(\max_i T_i > c\Big) \lessapprox \sum_i \mathbb{P}\big(T_i > c\big).$$

When the joint distribution of $(T_i)_{1\le i\le n}$ is Gaussian, the relative approximation error

$$\Delta(c) = \frac{\sum_i \mathbb{P}(T_i > c) - \mathbb{P}(\max_i T_i > c)}{\sum_i \mathbb{P}(T_i > c)}$$

is exponentially small as $c\to\infty$. This is shown in the tube method that includes the Bonferroni method as a special case. However, when the statistics are studentized by a common standard deviation estimator $\widehat{\sigma}$, the distribution of $(T_i)_{1\le i\le n}$ becomes a heavy-tailed distribution such as the multivariate t-distribution. In this talk, we evaluate the relative error $\Delta(c)$ in such cases.

We first set the class of random vectors $(T_i)_{1\le i\le n}$ with a correlation structure (ρ_{ij}) and a specified tail behavior. Let $x_1,\ldots,x_n\in\mathbb{R}^n$ be unit vectors such that $\langle x_i,x_j\rangle=\rho_{ij}$. By using standard Gaussian random vector $\xi\sim\mathcal{N}_n(0,I_n)$ and $\widehat{\sigma}$ independent of ξ, we let $T_i=\langle x_i,\xi\rangle/\widehat{\sigma}$. The marginal distribution function of T_i^2 is denoted by F (identical for all i), and the tail distribution of F is parameterized with parameters (β,ℓ) ($\beta\le 1$, ℓ is a slowly varying function):

$$1-F(x)\sim C\exp\left(-\int_{x_0}^{x} q(t)\mathrm{d}t\right),\quad q(t)=\frac{\ell(t)}{t^\beta},$$

where we assume that the limit $\lim_{x\to\infty}\ell(x)=\gamma\in(0,\infty)\cup\{\infty\}$ exists. This distribution family includes the light-tailed distribution (exponential, super-exponential), the long-tailed distribution $\mathrm{RV}_{-\gamma}$ (regularly varying distribution with index $-\gamma$), and intermediate cases (subexponential distribution) $\mathcal{S}$.

β	$\beta=0$	$\beta=0$	$\beta\in(0,1)$	$\beta=1$	$\beta=1$
γ	$\gamma=\infty$	$\gamma<\infty$	$\gamma\le\infty$	$\gamma=\infty$	$\gamma<\infty$
	super-exponential	exponential	$\mathcal{S}$	$\mathcal{S}$	$\mathrm{RV}_{-\gamma}$

Theorem 1. Suppose that $\beta<1$ or $\gamma=\infty$. Then,

$$\log\Delta(c)\sim -c^{2(1-\beta)}\ell(c^2)g_\beta(\cos^2\theta_{\mathrm{cri}}),\quad c\to\infty,$$

where $g_\beta(y)=\frac{y^{\beta-1}-1}{1-\beta}$ ($\beta<1$), $-\log y$ ($\beta=1$), and $\theta_{\mathrm{cri}}=\frac{1}{2}\min_{i<j}\cos^{-1}\rho_{ij}$ is the critical radius (reach). Moreover, $\Delta(c)\to 0$ ($c\to\infty$).

Example 1. If F is the chi-square distribution ($\beta=0$, $\gamma=1/2$), or the log-normal distribution ($\beta=1$, $\ell(x)=\log x$, $\gamma=\infty$), then $\log\Delta(c)\sim-(1/2)c^2\tan^2\theta_{\mathrm{cri}}$, or $-\log(c^2)(-\log\cos^2\theta_{\mathrm{cri}})$, respectively.

Theorem 2 (Regulary varying distributions)**.** Suppose that $\beta=1$ and $\gamma<\infty$. By using independent random variables $\widetilde{B}\sim B_{\gamma+\frac{1}{2},\frac{n-1}{2}}$ and $V_i\sim\mathrm{Unif}\big(\{x\in\mathbb{R}^n\mid \|x\|=1,\,\langle x,x_i\rangle=0\}\big)$, we have

$$\lim_{c\to\infty}\Delta(c)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\Big(\widetilde{B}<\cos^2\theta_i(V_i)\Big),\quad \theta_i(v)=\tan^{-1}\min_{j\ne i}\frac{1-\rho_{ij}}{\langle x_j,v\rangle}.$$

Moreover, $0\le\lim_{c\to\infty}\Delta(c)\le\bar{\Delta}:=\mathbb{P}\big(\widetilde{B}<\cos^2\theta_{\mathrm{cri}}\big)$.

These evaluations for $\Delta(c)$ can be generalized to the tube formula.

Statistical Theories and Machine Learning Using Geometric Methods

Date : December 14-15, 2023 (Japan Standard Time)

Venue : Academic Extension Center (Osaka Metropolitan University)

Contents : Workshop (Hybrid: physical/virtual)

- This workshop is supported by Osaka Central Advanced Mathematical Institute (MEXT Promotion of Distinctive Joint Research Center Program JPMXP0723833165), Osaka Metropolitan University.

Organizers: Koichi Tojo (RIKEN AIP: koichi.tojo@riken.jp), Hideto Nakashima (ISM), Yoshihiko Konno (OMU), Hideyuki Ishi (OMU), Kenji Fukumizu (ISM)

Program

- December 14 (Thursday):

13:00–13:50 **Hiroto Inoue** (Nishinippon Institute of Technology)
Mean-variance joint statistic valued in a real Siegel domain

14:00–14:50 **Eren Mehmet Kiral** (Keio University)
Bayesian Learning with Lie Groups

15:00–15:50 **Hajime Fujita** (Japan Women's University)
The generalized Pythagorean theorem on the compactifications of certain dually flat spaces via toric geometry

16:10–17:00 **Atsumi Ohara** (University of Fukui)
Doubly autoparallel structure and curvature integrals: An application to iteration complexity analysis of convex optimization

17:10–18:00 **Adam Chojecki** (Warsaw University of Technology), **Hideyuki Ishi** (Osaka Metropolitan University)
Uncovering Data Symmetries: Estimating Covariance Matrix in High-Dimensional Setting With 'gips' R Package

18:10–19:00 **Tomasz Skalski** (Wroclaw University of Science and Technology)
Maximum likelihood estimation for discrete exponential families, its geometry and combinatorics

- December 15 (Friday):

10:00-10:50 **Tomonari Sei** (The University of Tokyo)
Some open problems on minimum information dependence models

11:00–11:50 **Tomonari Sei** (The University of Tokyo), **Ushio Tanaka** (Osaka Metropolitan University)
Stein identity, Poincaré inequality and exponential integrability on a metric measure space

13:50–14:40 **Hikaru Watanabe** (The University of Tokyo)
Infinite dimensional parameterized measure models

14:50–15:40 **Eiki Shimizu** (SOKENDAI)
Neural-Kernel Conditional Mean Embeddings

15:50–16:40 **Satoshi Kuriki** (The Institute of Statistical Mathematics)
Bonferroni method and tube method for heavy-tailed distributions