Maximization of the information divergence from multinomial distributions

Jozef Juríček
Charles University in Prague
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics

Supervisor: **Ing. František Matúš, CSc.**
Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation
Department of Decision-Making Theory

SUMMARY Explicit solution of the problem of maximization of information divergence from the family of multinomial distributions is presented. General problem of maximization of information divergence from an exponential family has emerged in probabilistic models for evolution and learning in neural networks, based on infomax principles. The maximizers admit interpretation as stochastic systems with high complexity w.r.t. exponential family.

1 Introduction

PROBLEM *"Find all empirical distributions of data, which lies farthest from the model, when modelling by the multinomial family; in the sense of information divergence and the method of maximum likelihood."*

- *Exponential family*

$$\mathcal{E}_{\mu,f} = \left\{ Q_{\mu,f,\vartheta} \sim \left(e^{\langle \vartheta, f(z) \rangle} \mu(z) \right)_{z \in Z} : \vartheta \in \mathbb{R}^d \right\}$$

  - ⋆ μ nonzero *reference measure* on a finite set Z
  - ⋆ $f : Z \to \mathbb{R}^d$ the *directional statistics*

- *Divergence* of a pm P (on Z) from ν (on Z)

$$D(P\|\nu) = \begin{cases} \sum_{z \in \mathsf{s}(P)} P(z) \ln \dfrac{P(z)}{\nu(z)}, & \mathsf{s}(P) \subseteq \mathsf{s}(\nu), \\ +\infty, & \text{otherwise,} \end{cases}$$

  - ⋆ $\mathsf{s}(\cdot)$ the *support*, from now on, let $\mathsf{s}(\mu) = Z$

- *Divergence* of a pm P *from exponential family* $\mathcal{E} = \mathcal{E}_{\mu,f}$

$$D(P\|\mathcal{E}) = \inf_{Q \in \mathcal{E}} D(P\|Q) = \min_{Q \in \overline{\mathcal{E}}} D(P\|Q) \qquad \text{(last "=" } \Leftarrow \text{ Thm 1)}$$

THEOREM 1 *There exist **unique*** rI-projection (generalized MLE) $P^{\mathcal{E}} = \arg\min_{Q \in \overline{\mathcal{E}}} D(P\|Q)$.
For P empirical distribution, s.t. $P^{\mathcal{E}} \in \mathcal{E}$, $P^{\mathcal{E}}$ is the MLE for data with emprical distribution P. Details in [2].

- *Multinomial family* (N indep. trials, each with n outcomes)

$$\overline{\mathcal{M}} = \left\{ \left(Q(z) = \binom{N}{z} \prod_{j=1}^{n} p_j^{z_j} \right)_{z \in Z} : p \in \overline{\mathcal{P}}([1:n]) \right\}$$

  - ⋆ $\overline{\mathcal{P}}([1:n])$ all pm's on $[1:n] = [n] = \{1,\dots,n\}$
  - ⋆ $Z = \{z \in [0:N]^n : \sum_{j=1}^n z_j = N\}$
  - ⋆ $\overline{\mathcal{M}} = \overline{\mathcal{E}}_{\mu,f}$ with $f(z) = z$, $\mu(z) = \binom{N}{z}$

PROBLEM *Calculate* $\sup_{P \in \overline{\mathcal{P}}(Z)} D(P\|\overline{\mathcal{M}})$ *and find all maximizers* $\arg\sup_{P \in \overline{\mathcal{P}}(Z)} D(P\|\overline{\mathcal{M}})$. Generalization of [3].

2 Preliminaries

- $\pi : [1:N] \overset{1-1}{\leftrightarrow} [1:N]$, set of all permutations $[1:N]!$
  $x \in X = [1:n]^N$, $x^\pi = (x_{\pi(1)}, \dots, x_{\pi(N)})^\top$, $Q^\pi(x) = Q(x^\pi)$
- *Exchangable* distributions' family
  $\overline{\mathcal{E}} := \{P \in \overline{\mathcal{P}}(X) : P(x) = P(x^\pi), x \in X; \forall \pi \in [1:N]!\}$
- *1-factorizable* distributions' family
  $\overline{\mathcal{F}} := \{Q \in \overline{\mathcal{P}}(X) : Q(x) = \prod_{i=1}^N Q_i(x_i), x \in X\}$, Q_i marginal
- $X^z := \{x \in X : \forall j \in [1:n] : |\{i \in [1:N] : x_i = j\}| = z_j\}$

LEMMA 2 $P \in \overline{\mathcal{P}}(Z)$: $h(P) = P'$, $P'(x) = \dfrac{P(z)}{\binom{N}{z}}$, $\begin{matrix} z \in Z, \\ x \in X^z \end{matrix}$.

(i) $h : \overline{\mathcal{P}}(Z) \overset{1-1}{\leftrightarrow} \overline{\mathcal{E}}$, $\quad h\restriction_{\overline{\mathcal{M}}} : \overline{\mathcal{M}} \overset{1-1}{\leftrightarrow} \overline{\mathcal{E}} \cap \overline{\mathcal{F}} = \overline{\mathcal{E} \cap \mathcal{F}}$
(ii) $Q \in \overline{\mathcal{M}}$: $D(P\|Q) = D(h(P)\|h(Q))$
(iii) $P \in \overline{\mathcal{E}}$, $Q \in \overline{\mathcal{F}} \setminus \overline{\mathcal{E} \cap \mathcal{F}}$: $\exists \pi \in [1:N]!$,
$Q^\pi \not\equiv Q$ & $D(P\|Q) = D(P\|Q^\pi)$
(iv) $P \in \overline{\mathcal{E}}$: $P^{\mathcal{F}} = P^{\mathcal{E} \cap \mathcal{F}}$

- $P \in \mathcal{P}(X)$: $D(P\|\mathcal{F}) = M(P)$, the multi-information

THEOREM 3 (Maximization of multi-information)
$\arg\sup_{P \in \mathcal{P}(X)} D(P\|\mathcal{F}) = \{P_\Pi = \frac{1}{n} \sum_{j=1}^n \delta_{(j,\pi_2(j),\dots,\pi_N(j))^\top} :$
$\Pi = (\pi_2, \dots, \pi_N) \in [1:n]!^{(N-1)}\}$

$$D(P_\Pi\|\mathcal{F}) = (N-1)\ln n,\ P_\Pi^{\mathcal{F}} = U^X = \sum_{x \in X} \frac{\delta_x}{n^N}, \forall \Pi \in [1:n]!^{(N-1)}$$

3 Result

- $e^j = e^{j,j} = (0,\dots,0,1_j,0,\dots,0_n)^\top$, $\epsilon^{j,j} = \delta_{2e^{j,j}}$
  $e^{k,l} = (0,\dots,0,1_k,0,\dots,0,1_l,0,\dots,0_n)^\top$, $\epsilon^{k,l} = 2\delta_{e^{k,l}}, k < l$

CORROLARY 4 (Maximization of $D(\cdot\|\overline{\mathcal{M}})$)

$$\arg\max_{P \in \mathcal{P}(Z)} D(P\|\overline{\mathcal{M}}) = h^{-1}\left(\overline{\mathcal{E}} \cap \arg\sup_{P \in \mathcal{P}(X)} D(P\|\mathcal{F}) \right)$$

If $N = 2$

$$= \left\{ P_\pi = \frac{1}{n} \left(\sum_{j \in [1:n]: j \le \pi(j)} \epsilon^{j,\pi(j)} \right),\ \begin{matrix} \pi \in [n]! : \forall j,k \in [n] \\ [\pi(j) = k] \Rightarrow [\pi(k) = j] \end{matrix} \right\},$$

if $N > 2$
$= \{P_{\mathrm{Id}} = \frac{1}{n} \sum_{j=1}^n \delta_{Ne^j}\}$. $\forall N : \max_{P \in \mathcal{P}(Z)} D(P\|\overline{\mathcal{M}}) = (N-1)\ln n$.

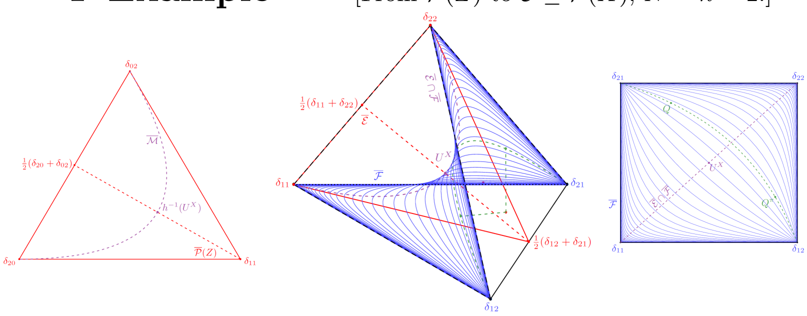
For every maximizer P_π : $P_\pi^{\mathcal{M}}(z) = \dfrac{\binom{N}{z}}{n^N}, z \in Z$.

(!) More general situation and esentially simpler proof than in [3]
$\Leftarrow$ Application of Theorem 1, Lemma 2 & Theorem 3

(?) $D(\cdot\|\overline{\mathcal{M}_k})$, where $\overline{\mathcal{M}_k} = h^{-1}(\overline{\mathcal{E}} \cap \overline{\mathcal{F}_k})$ and $\overline{\mathcal{F}_k}$, the *k-factorizable*

4 Example

[From $\overline{\mathcal{P}}(Z)$ to $\overline{\mathcal{E}} \subseteq \overline{\mathcal{P}}(X)$; $N = n = 2$.]



References

[1] Ay, N., Knauf, A. (2006). Maximizing multi-information. *Kybernetika* **45** 517-538.

[2] Csiszár, I., Matúš, F. (2003). Information projections revisited. *IEEE Transactions Information Theory* **49** 1474-1490.

[3] Matúš, F. (2004). Maximization of information divergences from binary i.i.d. sequences. *Proceedings IPMU* (2004) **2** 1303-1306. Perugia, Italy.