Maximum likelihood estimators for discrete exponential families and random graphs

Tomasz Skalski

Wrocław University of Science and Technology

Faculty of Pure and Applied Mathematics

(joint work with K. Bogdan, M. Bosy and T. Stroiński)

8$^{\text{TH}}$ PCC, 14 – 18 September 2020

Discrete exponential families – Notation

- $\mathcal{X} = \{x_1, \ldots, x_K\}$ – finite state space
- $\mathcal{B} \subset \mathbb{R}^{\mathcal{X}}$ – linear space of functions ($\phi \equiv \mathbb{1} \in \mathcal{B}$)
- $\mathcal{B}_+ = \{\phi \in \mathcal{B} : \phi \geq 0\}$ – subclass (cone) of non-negative functions
- $e(\phi) = exp\{\phi\}/Z(\phi)$ – exponential density
- $Z(\phi)$ – normalising constant (partition function)
- $e(\mathcal{B}) = \{p = e(\phi) : \phi \in \mathcal{B}\}$ – exponential family
- $L_p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$, $x_i \in \mathcal{X}$ – likelihood function

MLE

Definition

Let $x_1, \ldots, x_n$ 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:

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

To facilitate the calculations the log-likelihood function is being often used: $\ell_p(x_1, \ldots, x_n) = \log L_p(x_1, \ldots, x_n)$.

Definition

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

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

Existence of MLE

History

- *O. Barndorff-Nielsen (1978) – criterion of existence of MLE of a parameter (from the exponential family) for a given sample* $x_1, \ldots, x_n$.
  - *Beautiful, but cumbersome to apply*
- *S. J. Haberman (1974) – criterion of MLE existence in hierarchical log-linear models*
- *N. Eriksson, S. E. Fienberg, A. Rinaldo, S. Sullivant (2006) – interpreting the above criterion in terms of polyhedral geometry*
- *K. Bogdan, M. Bogdan (2000) – criterion of MLE existence for exponential families of continuous functions on* $[0,1]$.
- *A. Rinaldo, S. E. Fienberg, Y. Zhou (2009) – application of MLE existence in exponential random graph models (ERGM).*
- *K. Bogdan, M. Bosy, TS (2020+) – criterion of MLE existence in discrete exponential families.*

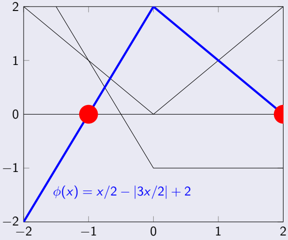
Sets of uniqueness

Definition

We say that $U \subset \mathcal{X}$ is a set 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) on $\{-2, -1, 0\}$ and on $\{0, 1, 2\}$.



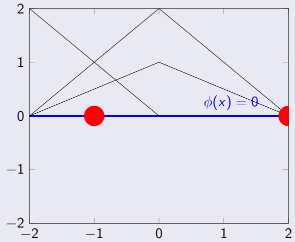
Then the set $\{-2, 2\}$ is of uniqueness for $\mathcal{B}$, but the set $\{-1, 2\}$ is not.

Remark

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

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 $\mathcal{B}_+$.

MLE existence – main criterion

Theorem (K. Bogdan, M. Bosy, TS (2020+))

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* $\mathcal{B}_+$*.*

Applications

Maximization of likelihood is fundamental in estimation, model selection and testing. In many procedures it is important to know if MLE actually exists for given data $x_1 \ldots, x_n$ and the linear space of exponents.
There are two types of results obtained with sets of uniqueness we propose below:

- Conditions for existence of MLE
- Probability bounds for MLE for an i.i.d. sample

For the i.i.d. random variables valued in $\mathcal{X}$ it will be useful to define the random (stopping) time:

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

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

Let $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$. As $\mathcal{X}$ is the only one set of uniqueness in this setting, one may observe that

Lemma

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

Thus, the problem of obtaining the existence of MLE for $\{x_1, \ldots, x_n\}$ resembles the Coupon Collector Problem. Therefore

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 $c \in \mathbb{R}$,

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

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

Applications – Rademacher functions

For $k \in \mathbb{N}$ consider the hypercube $\mathcal{X} = Q_k = \{-1, 1\}^k$. Let $K = |\mathcal{X}| = 2^k$.
For $j = 1, \ldots, k$ and $\chi = (\chi_1, \ldots, \chi_k) \in Q_k$ we define Rademacher functions:

$$r_j(\chi) = \chi_j,$$

and we denote $r_0(\chi) = 1$. In other words, the Rademacher functions may be seen as affine transforms of the indicators of half-cubes.

Observation

Moreover, the linear space of products of q Rademacher functions corresponds to the linear space of indicators of sub-cubes of Q_k created by fixing q of k coordinates.

Applications – Rademacher functions

Theorem (K. Bogdan, M. Bosy, TS (2020+))

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

In other words, the condition above is satisfied if and only if $\{x_1, \ldots, x_n\}$ intersects with every half-cube of Q_k.

Theorem (K. Bogdan, M. Bosy, TS (2020+))

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

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

and $\log_2 k = \log_2 \log_2 K$ 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.

Applications – ERGM

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

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

Also, by $G \subset H$ we mean that $E_1 \subset E_2$.
We define $\chi_{r,s} : \mathcal{G}_N \to \{-1, 1\}$ by $\chi_{r,s}(G) = 1 - 2\mathbb{1}_G(r, s)$ and consider the following linear space

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

We also consider coefficients $c \in \mathbb{R}^{\binom{V}{2}}$ corresponding to all edges of a complete graph K_N and the following exponential family:

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

where

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

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

Observation

Fix $c \in \mathbb{R}^{\binom{V}{2}}$. Then in the random graph $\mathbb{G}$ sampled from $\mathcal{G}_{N,c}$ 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 (2020+))

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 (2020+))

Let $\{\mathbb{G}_1, \ldots, \mathbb{G}_n\}$ *be independently distributed random graphs from* $\mathcal{G}_{N,c}$*. 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, $\log N$ *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 $k \in \mathbb{N}, 1 \leq q \leq k$, and $\mathcal{B}_q^k = \mathrm{Lin}\{w_S : S \subset \{1,\dots,k\} \text{ and } |S| \leq q\}$, where $w_S(x) = \prod_{i \in S} r_i(x), x \in Q_k, S \subset \{1,\dots,k\}$, are the Walsh functions.

Recall the previously mentioned observation:

Observation

The linear space of products of q Rademacher functions corresponds to the linear space of indicators of sub-cubes of Q_k created by fixing q of k coordinates.

- $q = 1$: Rademacher functions (already considered)
- $q = 2$: connections with the Ising model

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

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

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 (2020+))

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

Dynkin system

Definition

Let $\mathcal{X}$ be a non-empty set, and let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$. $\mathcal{D}$ is called a Dynkin system if the following conditions hold:

- $\mathcal{X} \in \mathcal{D}$;
- if $A, B \in \mathcal{D}$ and $A \subset B$, then $B \setminus A \in \mathcal{D}$;
- if $A_1, A_2, A_3, \dots \in \mathcal{D}$ and $A_j \subset A_{j+1}$ for $j \in \mathbb{N}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{D}$.

Let $\mathcal{Y}$ be a collection of subsets of $\mathcal{X}$. The smallest Dynkin system containing $\mathcal{Y}$ is called the Dynkin system generated by $\mathcal{Y}$, denoted $\mathcal{D}(\mathcal{Y})$.

Let $1 \leq q \leq k$. Define S_q^k as set of all $(k-q)$-dimensional subcubes of Q_k.

Lemma

If U is a set of uniqueness for $\mathcal{B}_{q+}^k$, then U intersects with every non-empty element of $\mathcal{D}(S_q^k)$.

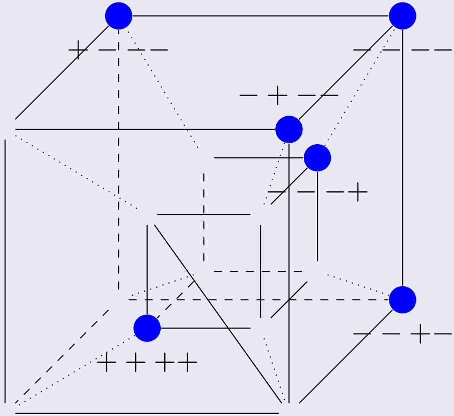
Lemma

If U is a set of uniqueness for $\mathcal{B}^k_{q_+}$, then U intersects with every non-empty element of $\mathcal{D}(S^k_q)$.

For $q = 1$ and $q = k - 1$ the above implication can be replaced with an equivalence. For general q the converse implication is false:

Counterexample ($k = 4, q = 2$)

Let $k = 4$ and $q = 2$. The following set (which consists of all points with exactly $0, 1$ or 4 positive coordinates) intersects with every non-empty element of $\mathcal{D}(S_2^4)$, but its not of uniqueness for $\left(\mathcal{B}_2^4\right)_+$.



Dynkin multi-system

Definition

Let $\mathcal{X}$ be a non-empty set, and let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$. $\mathcal{D}$ is called a multi-Dynkin system if the following conditions hold:

- $\mathcal{X} \in \mathcal{D}$;
- if $A, B \in \mathcal{D}$ and $A \subset B$, then $B - A \in \mathcal{D}$;
- if $A_1, A_2, A_3, \cdots \in \mathcal{D}$, then $\sum_{j=1}^{\infty} A_j \in \mathcal{D}$.

Analogously, let $\mathcal{Y}$ be a collection of multi-sets in $\mathcal{X}$. The smallest multi-Dynkin system containing $\mathcal{Y}$ is called the multi-Dynkin system generated by $\mathcal{Y}$, denoted $\mathcal{MD}(\mathcal{Y})$.

Observation

For each family $\mathcal{F}$ of sets we have

$$\mathcal{D}(\mathcal{F}) \subset \mathcal{MD}(\mathcal{F}).$$

Theorem (K. Bogdan, TS, T. Stroiński (in preparation))

Let $\mathcal{X}$ be a finite set and let $\mathcal{S} \subset 2^{\mathcal{X}}$. Then the set $U \subset \mathcal{X}$ is of uniqueness for $(\mathcal{B}_{\mathcal{S}})_+$ if and only if U has a non-empty intersection with each non-empty member of a multi-Dynkin system $\mathcal{MD}(\mathcal{S})$ generated by $\mathcal{S}$.

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. (2+++)

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

Bogdan, K., Bosy, M., Skalski, T. (2020+)

"'Maximum likelihood estimation for discrete exponential families and random graphs"', https://arxiv.org/abs/1911.13143

Erdős, P., Rényi, A. (1961)

"'On a classical problem of probability theory"', Magyar Tud. Akad. Mat. Kutató Int. Közl., 6:215-220.

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.

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