Estimating a probability mass function with unknown labels

Richard Gill
Mathematical Institute, University Leiden

Research initiated by Erik van Zwet with Allard Veldman, leading to

http://arxiv.org/abs/1312.1200
by Dragi Anevski, Richard Gill, and Stefan Zohren;

continuing with Maikel Bargpeter and Giulia Cereda

Estimating a probability mass function with ~~unknown~~ labels

deliberately discarded

Richard Gill
Mathematical Institute, University Leiden

http://arxiv.org/abs/1312.1200

Dragi Anevski, Richard Gill, and Stefan Zohren

The problem

- Notation: $\boldsymbol{X} = (X_1, X_2, \ldots)$, $\boldsymbol{p} = (p_1, p_2, \ldots)$
- Model: $\boldsymbol{X} \sim \text{Multinomial}(N, \boldsymbol{p})$, where:
  - very many p_k are very small
  - no further structure assumed:
    - $k = 1, 2, \ldots$ are mere labels

The problem

- Problem: estimate functionals of $\boldsymbol{p}$ such as
    - $\sum_k p_k \log p_k$
    - $\sum_k p_k{}^2$
    - $\log \left(\sum_k (1-p_k)^N p_k \;/\; \sum_k (1-p_k)^N p_k^2 \right)$, …

Note: invariant under permutations of labels!

The problem

- Problem: estimate functionals of $\boldsymbol{p}$ such as …
- Standard solution (“naieve estimator”):
  - Estimate $\boldsymbol{p}$ with MLE = empirical mass function $\boldsymbol{p}_N$
  - Plug-in to functional

Applications

- Biodiversity (ecology)
- Computer science (coding an unknown language in an unknown alphabet)
- Forensic science (Good-type estimators for problem of quantifying the evidential value of a rare Y-STR haplotype, rare mitochondrial DNA haplotype, …)
- Literature (how many words did Shakespeare know?

Hi-profile estimator

- Notation: (1), (2), … are the (backwards) ranks
- ((1), (2), …) is a ranking (a bijection $\mathbb{N} \to \mathbb{N}$)
- Reduce data to $\dot{\boldsymbol{X}} = (X_{(1)}, X_{(2)}, \ldots)$
- Reduce parameter to $\dot{\boldsymbol{p}} = (p_{(1)}, p_{(2)}, \ldots)$
- $\dot{\boldsymbol{X}}$ is $\boldsymbol{X}$ ordered by decreasing size, …
- Now estimate $\dot{\boldsymbol{p}}$ from $\dot{\boldsymbol{X}}$ by MLE, and plug-in…

Hi-profile = MLE for reduced problem

- If (wlog) $\boldsymbol{p} = \dot{\boldsymbol{p}}$, likelihood = $\sum_{\text{rankings}}$ (N choose $\boldsymbol{X}$) $\prod_k p_k^{X_k}$

- Hi-profile estimator proposed by computer scientist Alon Orlitsky and explored in many very short papers with many collaborators

- Much numerical work, many conjectures

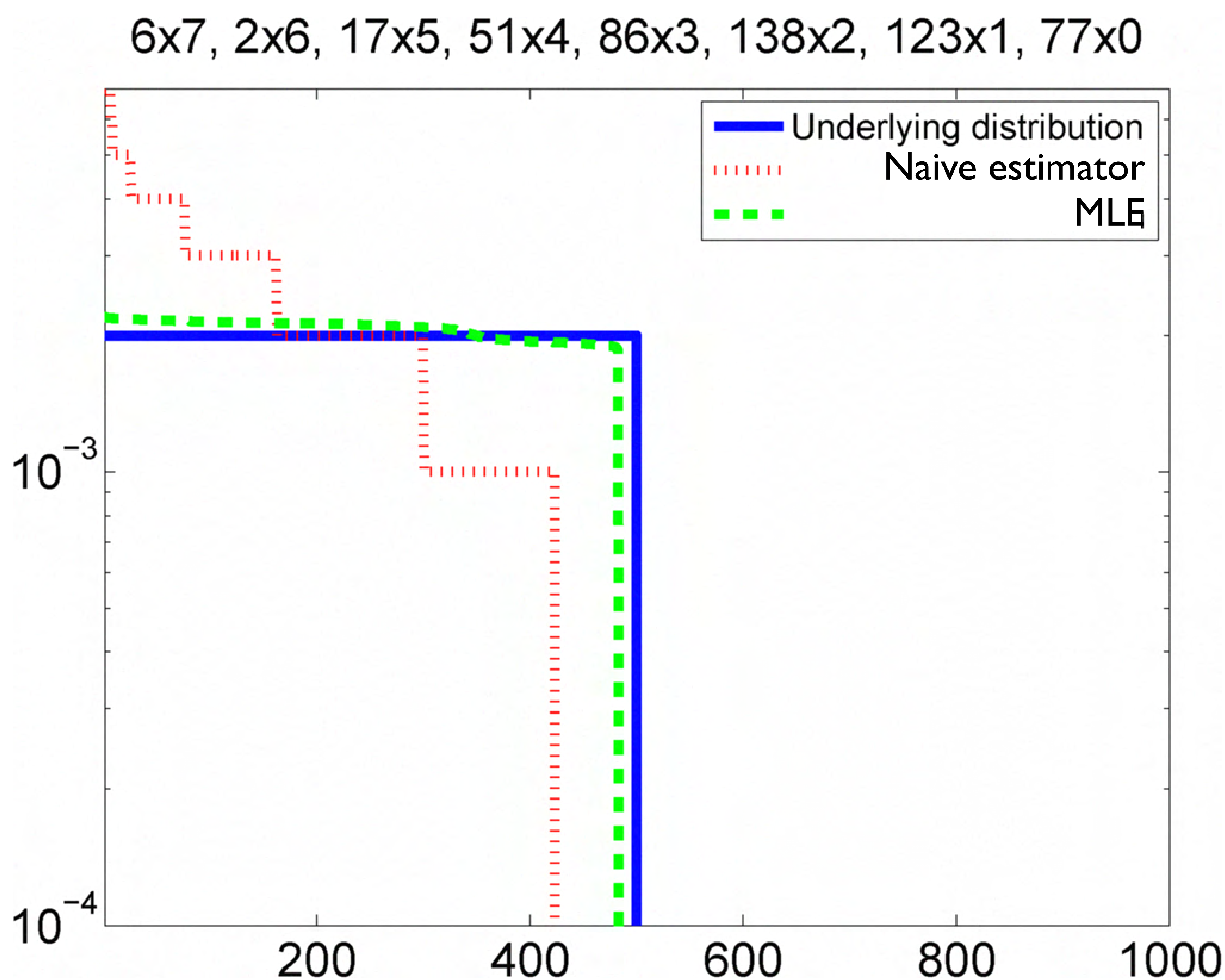
- Incomprehensible outline proof of L_1 consistency … (obviously totally wrong, but containing brilliant ideas!)

The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

Jayadev Acharya
ECE Department, UCSD
Email: jayadev@ucsd.edu

Alon Orlitsky
ECE & CSE Departments, UCSD
Email: alon@ucsd.edu

Shengjun Pan
CSE Department, UCSD
Email: s1pan@ucsd.edu



N=1000

The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

Jayadev Acharya
ECE Department, UCSD
Email: jayadev@ucsd.edu

Alon Orlitsky
ECE & CSE Departments, UCSD
Email: alon@ucsd.edu

Shengjun Pan
CSE Department, UCSD
Email: s1pan@ucsd.edu

Canonical $\overline{\psi}$	$\widehat{P}_{\overline{\psi}}$	Reference
1	any distribution	Trivial
11, 111, 111, ...	(1)	Trivial
12, 123, 1234, ...	$()$	Trivial
112, 1122, 1112, 11122, 111122	$(1/2, 1/2)$	[12]
11223, 112233, 1112233	$(1/3, 1/3, 1/3)$	[13]
111223, 1112223,	$(1/3, 1/3, 1/3)$	Corollary 5
1123, 1122334	$(1/5, 1/5, \ldots, 1/5)$	[12]
11234	$(1/8, 1/8, \ldots, 1/8)$	[13]
11123	(3/5)	[15]
11112	$(0.7887.., 0.2113..)$	[12]
111112	$(0.8322.., 0.1678..)$	[12]
111123	(2/3)	[15]
111234	(1/2)	[15]
112234	$(1/6, 1/6, \ldots, 1/6)$	[13]
112345	$(1/13, \ldots, 1/13)$	[13]
1111112	$(0.857.., 0.143..)$	[12]
1111122	$(2/3, 1/3)$	[12]
1112345	(3/7)	[15]
1111234	(4/7)	[15]
1111123	(5/7)	[15]
1111223	$\left(\frac{1}{\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}\right)$	Corollary 7
1123456	$(1/19, \ldots, 1/19)$	[13]
1112234	$(1/5, 1/5, \ldots, 1/5)?$	Conjectured

TABLE I
PML DISTRIBUTIONS OF ALL PATTERNS OF LENGTH ≤ 7

Computation

- We propose SA-MH-EM (Orlitsky et al: MH within EM)
- SA = Stochastic approximation (solve score equations)
- MH = Metropolis-Hastings (sample from conditional law of complete data given incomplete)
- EM = Expectation Maximization (missing data problem)
- First we reduced data and parameter; now we put both back again!
- In our new complete data problem we <u>pretend</u> $\boldsymbol{p} = \dot{\boldsymbol{p}}$

Computation

- SA-MH-EM
- To guarantee <u>existence</u> of MLE we need to extend the model
    - Extension: allow <u>blob</u> of infinitely many zero probability categories, together having positive probability
- To make computation feasible, we have to <u>sieve</u> extended parameter space
    - Reduction: finite dimensional, assume positive lower bounds, but keeping blob

Our main theorem

- (Almost) root-N L_1-consistency of (sieved extended) Hi-profile estimator of $\boldsymbol{\dot{p}}$
- Ingredients: Dvoretsky-Kiefer-Wolfowitz inequality: exponential probability bound for $||\boldsymbol{p}_N - \boldsymbol{p}||_\infty$
- Hardy's asymptotic formula for # partitions of N
- Hardy's lemma: monotone re-ordering is an L_∞ contraction
- A new Lemma about MLE, reminiscent of Neyman-Pearson

Lemma

- Suppose P and Q are two probability measures, both members of a statistical model $\mathcal{P}$ for observed data $\dot{\boldsymbol{X}}$, mass functions p and q, (corresponding to parameters $\boldsymbol{p}$ and $\boldsymbol{q}$)

- Suppose A is some event in the sample space of the observed data

- Suppose $P(A) \geq 1 - \delta$ and $Q(A) \leq \varepsilon$

- Then P (The MLE is Q) $\leq \delta + \varepsilon$

Proof of Lemma

- P (The MLE is Q) $\leq P(p \leq q)$
- $P(A^c) \leq \delta$
- $Q(A) \leq \varepsilon$ hence $P(A \cap \{p \leq q\}) \leq \varepsilon$
- $P(p \leq q) \leq P(A^c) + P(A \cap \{p \leq q\}) \leq \delta + \varepsilon$

Putting the pieces together

- Dvoretsky-Kiefer-Wolfowitz $\Rightarrow P(B^c)$ exponentially small, $B = \{\| \boldsymbol{p}_N - \boldsymbol{p} \|_\infty \leq c \}$
- Hardy (monotone ordering) $\Rightarrow P(A^c)$ exponentially small, $A = \{\| \dot{\boldsymbol{p}}_N - \dot{\boldsymbol{p}} \|_\infty \leq c \} \supseteq B$
- Repeat (with care!) for Q, $C = \{\| \boldsymbol{q}_N - \boldsymbol{q} \|_\infty \leq c \} \subseteq A^c$, where $\boldsymbol{q}$ is at least a certain L_1 distance from $\boldsymbol{p}$
- Lemma $\Rightarrow P$ (The MLE is Q) is exponentially small

Putting the pieces together

- Sample space is finite ⇒ set of possible MLE's is finite

  Hardy (# partitions of N) ⇒ # possible MLE's is of

  smaller order than $\exp(+b\sqrt{N})$)

- Sum over all $\boldsymbol{q}$ outside of an L_1 ball around $\boldsymbol{p}$

- $\exp(-\,a\,N)$ wins from $\exp(+b\sqrt{N})$

- P (MLE is outside L_1 ball around $\boldsymbol{p}$) is exponentially small

Is that result any good?

- It's far too weak: MLE of $\boldsymbol{p} = \dot{\boldsymbol{p}}$ based on $\dot{\boldsymbol{X}}$ does not have better rate than naive estimator: $\dot{\boldsymbol{p}}_N$!

- We conjecture it truly is (or can be) a whole lot better

- <u>Challenge 1</u>: refine this proof, or build a second stage on top of it

- So far we used <u>*almost nothing*</u> about the model!

- <u>Challenge 2</u>: better computational algorithm