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

by Dragi Anevski, Richard Gill, and Stefan Zohren;
continuing with Maikel Bargpeter and Giulia Cereda
Estimating a probability mass function with unknown labels

Richard Gill
Mathematical Institute, University Leiden

Dragi Anevski, Richard Gill, and Stefan Zohren
The problem

- **Notation**: \( \mathbf{X} = (X_1, X_2, \ldots), \mathbf{p} = (p_1, p_2, \ldots) \)

- **Model**: \( \mathbf{X} \sim \text{Multinomial}(N, \mathbf{p}) \), 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 $p$ such as

\[ \sum_k p_k \log p_k \]

\[ \sum_k p_k^2 \]

\[ \log \left( \frac{\sum_k (1-p_k)^N p_k}{\sum_k (1-p_k)^N p_k^2} \right), \ldots \]

**Note**: invariant under permutations of labels!
The problem

- **Problem**: estimate functionals of $p$ such as ...

- **Standard solution** ("naive estimator"): 
  - Estimate $p$ with MLE = empirical mass function $p_N$
  - 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 \( \mathbb{N} \rightarrow \mathbb{N} \))

- Reduce data to \( \dot{X} = (X_{(1)}, X_{(2)}, \ldots ) \)

- Reduce parameter to \( \dot{p} = (p_{(1)}, p_{(2)}, \ldots ) \)

- \( \dot{X} \) is \( X \) ordered by decreasing size, …

- Now estimate \( \dot{p} \) from \( \dot{X} \) by MLE, and plug-in…
Hi-profile = MLE for reduced problem

• If (wlog) $p = \hat{p}$, likelihood = $\sum_{\text{rankings}} \binom{N}{X} \prod_k p_k^{x_k}$

• 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: sjpan@ucsd.edu

6x7, 2x6, 17x5, 51x4, 86x3, 138x2, 123x1, 77x0

N=1000
The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

<table>
<thead>
<tr>
<th>Canonical $\overline{\psi}$</th>
<th>$\hat{P}_{\overline{\psi}}$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>any distribution</td>
<td>Trivial</td>
</tr>
<tr>
<td>11, 111, 111, ...</td>
<td>(1)</td>
<td>Trivial</td>
</tr>
<tr>
<td>12, 123, 1234, ...</td>
<td>()</td>
<td>Trivial</td>
</tr>
<tr>
<td>112, 1122, 1112, 11122, 111222</td>
<td>(1/2, 1/2)</td>
<td>[12]</td>
</tr>
<tr>
<td>11223, 112233, 11122233</td>
<td>(1/3, 1/3, 1/3)</td>
<td>[13]</td>
</tr>
<tr>
<td>111223, 1112223</td>
<td>(1/3, 1/3, 1/3)</td>
<td>Corollary 5</td>
</tr>
<tr>
<td>1123, 1122334</td>
<td>(1/5, 1/5, ... , 1/5)</td>
<td>[12]</td>
</tr>
<tr>
<td>11234</td>
<td>(1/8, 1/8, ... , 1/8)</td>
<td>[13]</td>
</tr>
<tr>
<td>11123</td>
<td>(3/5)</td>
<td>[15]</td>
</tr>
<tr>
<td>1111212</td>
<td>(0.7887.., 0.2113..)</td>
<td>[12]</td>
</tr>
<tr>
<td>11111212</td>
<td>(0.8322.., 0.1678..)</td>
<td>[12]</td>
</tr>
<tr>
<td>111123</td>
<td>(2/3)</td>
<td>[15]</td>
</tr>
<tr>
<td>111234</td>
<td>(1/2)</td>
<td>[15]</td>
</tr>
<tr>
<td>112234</td>
<td>(1/6, 1/6, ... , 1/6)</td>
<td>[13]</td>
</tr>
<tr>
<td>112345</td>
<td>(1/13, ... , 1/13)</td>
<td>[13]</td>
</tr>
<tr>
<td>11111112</td>
<td>(0.857.., 0.143..)</td>
<td>[12]</td>
</tr>
<tr>
<td>11111122</td>
<td>(2/3, 1/3)</td>
<td>[12]</td>
</tr>
<tr>
<td>1112345</td>
<td>(3/7)</td>
<td>[15]</td>
</tr>
<tr>
<td>1111234</td>
<td>(4/7)</td>
<td>[15]</td>
</tr>
<tr>
<td>11111123</td>
<td>(5/7)</td>
<td>[15]</td>
</tr>
<tr>
<td>1111223</td>
<td>$(\frac{1}{\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}})$</td>
<td>Corollary 7</td>
</tr>
<tr>
<td>1123456</td>
<td>(1/19, ... , 1/19)</td>
<td>[13]</td>
</tr>
<tr>
<td>1112234</td>
<td>(1/5, 1/5, ... , 1/5)?</td>
<td>Conjectured</td>
</tr>
</tbody>
</table>

TABLE I
PML distributions of all patterns of length $\leq 7$
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 pretend $p = \hat{p}$
Computation

• SA-MH-EM

• To guarantee existence of MLE we need to extend the model

  • Extension: allow blob of infinitely many zero probability categories, together having positive probability

• To make computation feasible, we have to sieve 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 $\hat{p}$

• Ingredients: Dvoretzky-Kiefer-Wolfowitz inequality: exponential probability bound for $\|p_N - p\|_\infty$

• Hardy’s asymptotic formula for $\#$ partitions of $N$

• Hardy’s lemma: monotone re-ordering is an $L_\infty$ 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 \( \hat{X} \), mass functions \( p \) and \( q \), (corresponding to parameters \( p \) and \( q \))

• 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 \( P (\text{The MLE is } Q) \leq \delta + \varepsilon \)
Proof of Lemma

• $P(\text{The MLE is } Q) \leq P(p \leq q)$

• $P(A^c) \leq \delta$

• $Q(A) \leq \varepsilon$ 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** ⇒ $P(B^c)$ exponentially small, $B = \{||p_N - p||_\infty \leq c\}$

- **Hardy (monotone ordering)** ⇒ $P(A^c)$ exponentially small, $A = \{||p'_N - p'||_\infty \leq c\} \supseteq B$

- Repeat (with care!) for $Q$, $C = \{||q_N - q||_\infty \leq c\} \subseteq A^c$, where $q$ is at least a certain $L_1$ distance from $p$

- Lemma ⇒ $P$ (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√N)

• Sum over all q outside of an L₁ ball around p

• exp(– a N) wins from exp(+b√N)

• P (MLE is outside L₁ ball around p) is exponentially small
Is that result any good?

- It’s far too weak: MLE of $p = \dot{p}$ based on $\dot{X}$ does not have better rate than naive estimator: $\dot{p}_N$!
- We conjecture it truly is (or can be) a whole lot better
- **Challenge 1**: refine this proof, or build a second stage on top of it
- So far we used *almost nothing* about the model!
- **Challenge 2**: better computational algorithm