Property Estimation ++

with Yi Hao, UCSD

Including past work with
J. Acharya, H. Das, A.T. Suresh, K. Viswanathan

Probability and Information Theory Workshop, Hong Kong

21 August 2019

Outline

Property estimation

Prior work

Plug-in estimators

Maximum likelihood

Profile maximum likelihood

Simple unified approach

Proof elements

Results

Discrete support set $\mathcal{X}$

$\{\text{heads, tails}\} = \{\text{h, t}\} \qquad \mathbb{Z}$

Distribution p over $\mathcal{X}$, probability p_x for $x \in \mathcal{X}$

$p_x \geq 0 \qquad \sum_{x \in \mathcal{X}} p_x = 1$

$p = (p_{\text{h}}, p_{\text{t}}) \qquad p_h = .6, \ p_t = .4$

$\mathcal{P}$ collection of distributions

$\mathcal{P}_{\mathcal{X}}$ all distributions over $\mathcal{X}$

$\mathcal{P}_{\{\text{h, t}\}} = \{(p_{\text{h}}, p_{\text{t}})\} = \{(.6, .4), (.4, .6), (.5, .5), (0, 1), \ldots\}$

$f : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$

Maps distribution to real value, also called *functional*

Shannon entropy	$H(p)$	$\sum_x p_x \log \frac{1}{p_x}$
Support size	$S(p)$	$\sum_x \mathbb{1}_{p_x>0}$
Support coverage	$S_m(p)$	$\sum_x (1-(1-p_x)^m)$
Expected # distinct symbols in m samples		
Distance to uniformity	$L_{\text{uni}}(p)$	$\sum_x \left\lvert p_x - \frac{1}{\lvert\mathcal{X}\rvert}\right\rvert$
Rényi entropy	$H_\alpha(p)$	$\frac{1}{1-\alpha} \log \left(\sum_x p_x^\alpha\right)$
Highest probability	$\max(p)$	$\max\{p_x : x \in \mathcal{X}\}$

Many applications

f invariant under label permutations

$$H(p) \qquad H_\alpha(p) \qquad S(p) \qquad S_m(p) \qquad L_{\mathsf{uni}}(p) \qquad \mathsf{max}(p)$$

Non-symmetric: f depends on labels

$$p_{\mathsf{h}} \qquad \frac{p_{\mathsf{h}}}{p_{\mathsf{t}}} \qquad p_{\mathsf{h}} \cdot p_{\mathsf{t}}, \text{ if } |\mathcal{X}| > 2$$

$f(p) = \sum_x f(p_x)$

$S(p) := \sum_x \mathbb{1}_{p_x>0}$

$H(p) = \sum_x p_x \log \frac{1}{p_x}$

$S_r(p)$

$L_{\mathsf{uni}}(p)$

Non-additive

$H_\alpha(p) := \frac{1}{1-\alpha} \log \left(\sum_x p_x^\alpha\right)$

$\mathsf{max}(p) := \max\{p_x : x \in \mathcal{X}\}$

Most results apply to additive symmetric properties

Given: support set $\mathcal{X}$, property f

Unknown: $p \in \mathcal{P}_{\mathcal{X}}$

Estimate: $f(p)$

Entropy of English words

Given: $\mathcal{X} = \{\text{English words}\}$, $f = H$, unknown: p, estimate: $H(p)$

species in habitat

Given: $\mathcal{X} = \{\text{bird species}\}$, $f = S$, unknown: p, estimate: $S(p)$

Learn from examples

Observe n independent samples $X^n = X_1, \ldots, X_n \sim p$

Estimate $f(p)$

Estimator: $f^{\text{est}} : \mathcal{X}^n \to \mathbb{R}$

Estimate: $f^{\text{est}}(X^n)$

Simple two-step estimators

Use X^n to derive estimate $p^{\text{est}}(X^n)$ of p

Plug-in $f(p^{\text{est}}(X^n))$ to estimate $f(p)$

If as $n \to \infty$, $p^{\text{est}}(X^n) \to p$, then $f(p^{\text{est}}(X^n)) \to f(p)$

What is the simplest p^{est}?

n samples

N_x # times x appears

$p_x^{\text{emp}} := \frac{N_x}{n}$

Entropy estimation

$\mathcal{X} = \{a, b, c\} \qquad p = (p_a, p_b, p_c) = (.5, .3, .2)$

Estimate $H(p)$ from $n = 10$ samples

$X^{10} = c, a, b, a, b, a, b, a, b, c$

$p^{\text{emp}} = (.4, .4, .2)$

$H^{\text{emp}}(X^{10}) = H(.4, .4, .2)$

Best-known, most widely-used distribution estimator

Relatively easy to analyze

Min-max formulation

Given: Property f, collection $\mathcal{P}$ of distributions over $\mathcal{X}$

n i.i.d. samples X^n from unknown $p \in \mathcal{P}$

Property value $f(p)$ – unknown

Estimator's value $f^{\text{est}}(X^n)$

Estimator's absolute loss $|f^{\text{est}}(X^n) - f(p)|$

Expected loss $L_f(f^{\text{est}}, p, n) := \mathbb{E}_{X^n \sim p}|f^{\text{est}}(X^n) - f(p)|$

Worst-case loss $L_f(f^{\text{est}}, \mathcal{P}, n) := \max_{p \in \mathcal{P}_{\mathcal{X}}} L_f(f^{\text{est}}, p, n)$

Minimum worst-case loss $L_f(\mathcal{P}, n) := \min_{f^{\text{est}}} L_f(f^{\text{est}}, \mathcal{P}, n)$

Symmetric properties

$\mathcal{P}_{\mathcal{X}}$ all distributions over $\mathcal{X}$

Dependence on $\mathcal{X}$ only through $k = |\mathcal{X}|$

H over {cat, dog} same as over {ma, shu}

$L_f(\mathcal{P}_{\mathcal{X}}, n) \;\to L_f(k, n)$

Prior work: Min-max Error up to Constant Factors

References: P03, VV11a/b, WY14/19, JVHW14, AOST14, OSW16, ADOS17, JVW18

Property	Base function	$L(f^{\text{emp}}, k, n)$	$L(k, n)$
Entropy [1]	$p(x)\log\frac{1}{p(x)}$	$\frac{k}{n}+\frac{\log n}{\sqrt{n}}$	$\frac{k}{n\log n}+\frac{\log n}{\sqrt{n}}$
Supp. coverage[2]	$(1-(1-p(x))^r)$	$r\exp\left(-\Theta\left(\frac{n}{r}\right)\right)$	$r\exp\left(-\Theta\left(\frac{n\log n}{r}\right)\right)$
Power sum [3] [4]	$p(x)^\alpha,\ \alpha\in(0,\frac{1}{2}]$	$\frac{k}{n^\alpha}$	$\frac{k}{(n\log n)^\alpha}$
	$p(x)^\alpha,\ \alpha\in(\frac{1}{2},1)$	$\frac{k}{n^\alpha}+\frac{k^{1-\alpha}}{\sqrt{n}}$	$\frac{k}{(n\log n)^\alpha}+\frac{k^{1-\alpha}}{\sqrt{n}}$
Dist. to uniform[5]	$\lvert p(x)-\frac{1}{k}\rvert$	$\sqrt{\frac{k}{n}}$	$\sqrt{\frac{k}{n\log n}}$
Support size[6]	$\mathbb{1}_{p(x)>0}$	$k\exp\left(-\Theta\left(\frac{n}{k}\right)\right)$	$k\exp\left(-\Theta\left(\sqrt{\frac{n\log n}{k}}\right)\right)$

n to $n\log n$ when comparing the worst-case performances

[1] $n \gtrsim k$ for empirical; $n \gtrsim k/\log k$ for minimax

[2] $n \gtrsim r$ for empirical; $n \gtrsim r/\log r$ for minimax

[3] $\alpha\in(0,\frac{1}{2}]$: $n \gtrsim k^{1/\alpha}$ for empirical; $n \gtrsim \frac{k^{1/\alpha}}{\log k}$ and $\log k \gtrsim \log n$ for minimax

[4] $\alpha\in(\frac{1}{2},1)$: $n \gtrsim k^{1/\alpha}$ for empirical; $n \gtrsim \frac{k^{1/\alpha}}{\log k}$ for minimax

[5] $n \gtrsim k$ for empirical; $n \gtrsim k/\log k$ and $\log k \gtrsim \log n$ for minimax

[6] consider $\mathcal{P}_{\geq 1/k}$ instead of $\mathcal{P}_{\mathcal{X}}$; $k\log k \gtrsim n \gtrsim k/\log k$ for minimax

Why is Empirical Suboptimal?

Intuitive, simple

Why does it work at all?

For i.i.d. $p \in \mathcal{P}_{\mathcal{X}}$, the probability of observing $x^n \in \mathcal{X}^n$

$$p(x^n) := \Pr_{X^n \sim p}(X^n = x^n) = \prod_{i=1}^{n} p(x_i)$$

Maximum likelihood estimator: $x^n \to$ dist. p maximizing $p(x^n)$

$$p^{\mathsf{ml}}(x^n) = \arg\max_p p(x^n)$$

$$p^{\mathsf{ml}}(h,t,h) = \arg\max_{p_h} p_h^2 \cdot (1 - p_h) \quad \to \quad p_h = 2/3,\ p_t = 1/3$$

Identical to empirical estimator – always

Good: distribution that best explains observation

Sub-optimal for all properties in table

ML / EF work well for small alphabets large sample

Overfit data when alphabet is large relative to sample size iid: Do not care about order

Symmetric properties: Do not care about specific values

(h,h,t), (t,t,h), (h,t,h), (t,h,t), (t,h,h), (h,t,t) same entropy

Care only: # of elements appearing any given number of times

Three samples: 1 element appeared once, 1 element appeared twice

Profile: $\varphi = \{1, 2\}$

Profile $\varphi(x^n)$ of x^n is the multiset of its symbol frequencies

$$x^n = a\,b\,a\,c\,c\,d\,e \implies a\,c \text{ appears twice, } b\,d\,e \text{ appear once}$$
$$\implies \varphi(x^n) = \{2, 2, 1, 1, 1\}$$

Probability of observing a profile φ when sampling from p is

$$p(\varphi) := \sum_{y^n:\varphi(y^n)=\varphi} p(y^n) = \sum_{y^n:\varphi(y^n)=\varphi} \prod_{i=1}^{n} p(y_i)$$

[OSVZ04] Profile maximum likelihood maps x^n to

$$p^{\mathsf{ml}}_{\varphi(x^n)} := \operatorname*{argmax}_{p \in \mathcal{P}_{\mathcal{X}}} \; p(\varphi(x^n))$$

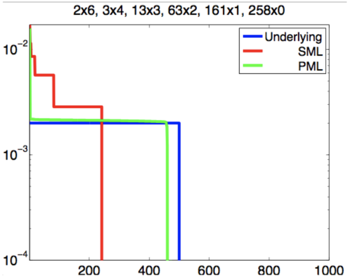
Uniform

500 symbols

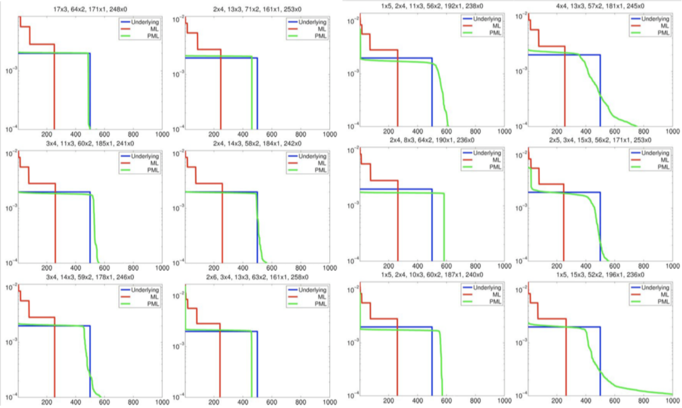
350 samples

2x6, 3x4, 13x3, 63x2, 161x1

242 appeared, 258 did not



U[500], 350x, 12 experiments



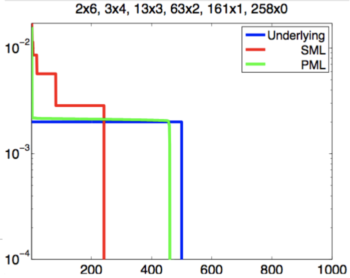
Uniform

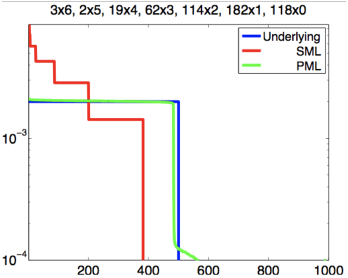
500 symbols

350 samples

2x6, 3x4, 13x3, 63x2, 161x1

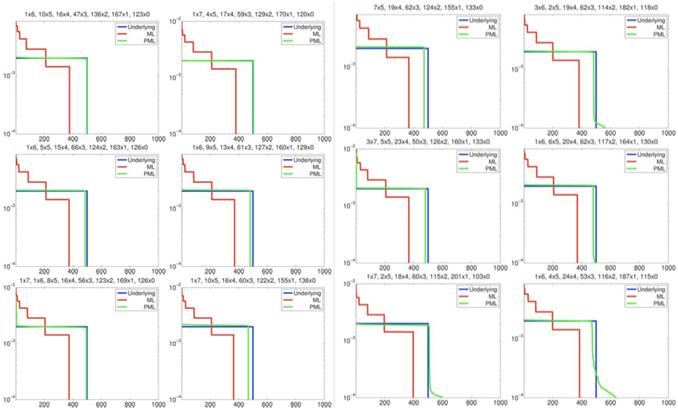
248 appeared, 258 did not





700 samples

U[500], 700x, 12 experiments



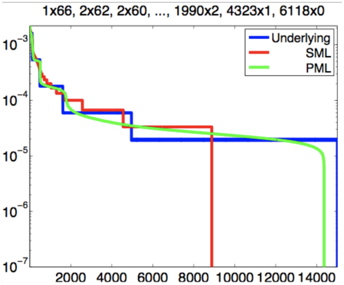
PML: Experimental performance

15K elements, 5 steps, ~3x

30K samples

Observe 8,882 elts

6,118 missing



PML: Experimental performance

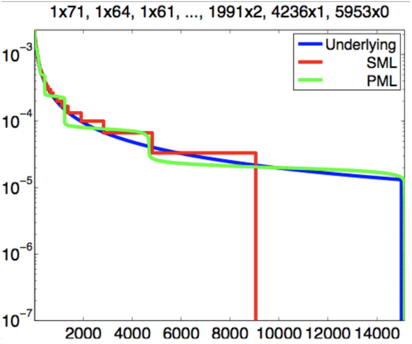
Underlies many natural phenomena

$p_i = C/i$, i=100...15,000

30,000 samples

Observe 9,047 elts

5,953 missing



1990 Census - Last names

SMITH	1.006	1.006	1
JOHNSON	0.810	1.816	2
WILLIAMS	0.699	2.515	3
JONES	0.621	3.136	4
BROWN	0.621	3.757	5
DAVIS	0.480	4.237	6
MILLER	0.424	4.660	7
WILSON	0.339	5.000	8
MOORE	0.312	5.312	9
TAYLOR	0.311	5.623	10

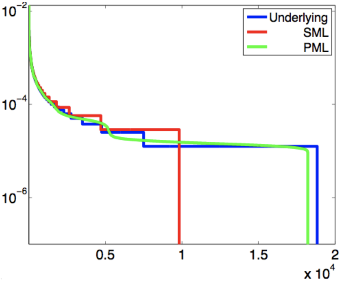
AMEND	0.001	77.478	18835
ALPHIN	0.001	77.478	18836
ALLBRIGHT	0.001	77.479	18837
AIKIN	0.001	77.479	18838
ACRES	0.001	77.480	18839
ZUPAN	0.000	77.480	18840
ZUCHOWSKI	0.000	77.481	18841
ZEOLLA	0.000	77.481	18842

18,839 names
77.48% population
~230 million

1990 Census - Last names

18,839 last names based on ~230 million

35,000 samples, observed 9,813 names



Coverage (# new symbols)

Zipf distribution over 15K elements

Sample 30K times

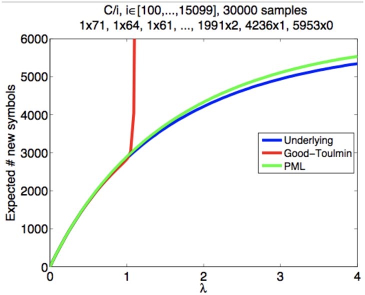
Estimate: # new symbols in sample of size λ * 30K

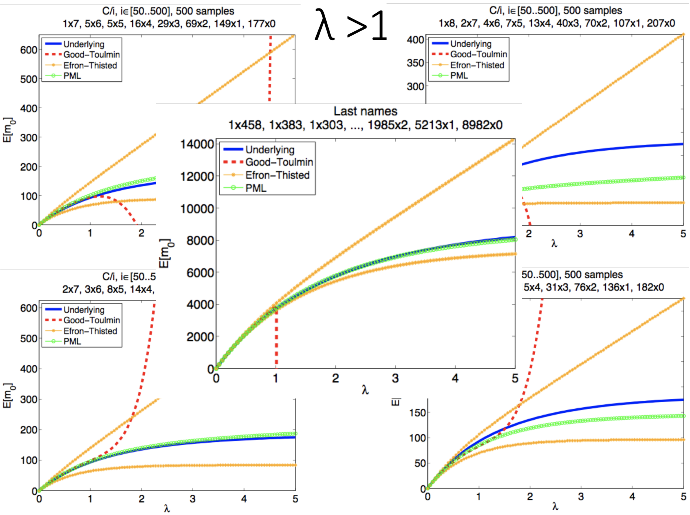
Good-Toulmin: λ < 1 λ > 1

Estimate PML & predict

Extends to λ > 1

Applies to other properties




λ >1
C/i, i∈[50..500], 500 samples
1x7, 5x6, 5x5, 16x4, 29x3, 69x2, 149x1, 177x0
C/i, i∈[50..500], 500 samples
1x8, 2x7, 4x6, 7x5, 13x4, 40x3, 70x2, 107x1, 207x0
Last names
1x458, 1x383, 1x303, ..., 1985x2, 5213x1, 8982x0
Underlying
Good–Toulmin
Efron–Thisted
PML
E[m0]
λ
C/i, i∈[50..5
2x7, 3x6, 8x5, 14x4,
50..500], 500 samples
5x4, 31x3, 76x2, 136x1, 182x0

Proof Elements

Upper bound probability of observing unlikely outcomes

p: distribution over $\mathcal{Z}$

$\delta > 0$

$z \in \mathcal{Z}$ is δ-*unlikely* if $p(z) \leq \delta$

$\Pr(\text{observing a } \delta - \text{ unlikely outcome}) = \sum_{z \in \mathcal{Z}_{\leq \delta}} p(z) \leq \sum_{z \in \mathcal{Z}_{\leq \delta}} \delta = \delta \cdot |\mathcal{Z}_{\leq \delta}|$.

Consider the problem of symmetric property estimation

$\boldsymbol{\Phi_n}$: collection of profiles associated with samples of size n

Lemma Suppose $\hat{f} : \Phi_n \to \mathbb{R}$ is such that for all $p \in \mathcal{P}_{\mathcal{X}}$,

$$\Pr_{\varphi \sim p}(|\hat{f}(\varphi) - f(p)| > \varepsilon) < \delta,$$

then the PML plug-in estimator satisfies [ADOS17]

$$\Pr_{\varphi \sim p}\left(|\boldsymbol{f(p_{\varphi}^{\mathsf{ml}})} - f(p)| > 2 \cdot \varepsilon\right) < \delta \cdot \exp(3\sqrt{n})$$

Proof: Consider any $p \in \mathcal{P}_{\mathcal{X}}$

$\Phi^n_{\geq\delta} := \{\varphi \in \Phi_n : p(\varphi) \geq \delta\}$

For $\varphi \in \Phi^n_{\geq\delta}$:

$|\hat{f}(\varphi) - f(p)| \leq \varepsilon$ (condition in the lemma)

$p^{\mathsf{ml}}_{\varphi}(\varphi) \geq p(\varphi) \geq \delta$, hence $|\hat{f}(\varphi) - f(p^{\mathsf{ml}}_{\varphi})| \leq \varepsilon$

Triangle inequality: $|f(p^{\mathsf{ml}}_{\varphi}) - f(p)| \leq 2\varepsilon$

Therefore,

$$\Pr_{\varphi\sim p}\left(|\boldsymbol{f}(\boldsymbol{p}^{\mathsf{ml}}_{\boldsymbol{\varphi}}) - f(p)| > 2\varepsilon\right) \leq \Pr_{\varphi\sim p}(\varphi \notin \Phi^n_{\geq\delta}) \leq \delta \cdot |\Phi_n|$$

Finally, $|\Phi_n|$ is exactly the number of partitions of integer n, which $\leq \exp(3\sqrt{n})$ by the well-known result* of Hardy and Ramanujan

*Hardy, G. H. and Ramanujan, S. "Asymptotic Formulae in Combinatory Analysis." Proc. London Math. Soc. 17, 75-115, 1918.

$\boldsymbol{p}$ an unknown distribution in $\mathcal{P}_{\mathcal{X}}$

Given an i.i.d. sample $X^n \sim p$

Estimate $f(p)$ by estimator $\hat{f}$

Min-max sample complexity $n_f(|\mathcal{X}|, \varepsilon, \delta)$

- minimum n necessary to
- ensure $|\hat{f}(X^n) - f(p)| \leq \varepsilon$ with probability $\geq 1 - \delta$
- for every $p \in \mathcal{P}_{\mathcal{X}}$

Equivalent to result in table

The Broad Optimality of PML [HO19a]

Profile maximum likelihood (PML) is a unified time-
and sample-optimal approach to four fundamental problems:
additive property estimation, Rényi entropy estimation,
uniformity testing, and sorted distribution estimation.

.

Hao, Y., & Orlitsky, A. (2019). The Broad Optimality of Profile Maximum Likelihood.

Theorem For every f in a broad class of symmetric additive properties, including all Lipschitz properties, any $\mathcal{X}$, $p \in \mathcal{P}_{\mathcal{X}}$, and $n \geq n_f(|\mathcal{X}|, \varepsilon, 1/3)$, if $\varepsilon \geq n^{-0.1}$,

$$\Pr\left(\left|\boldsymbol{f}\left(\boldsymbol{p}^{\mathsf{ml}}_{\boldsymbol{\varphi(X^{4n})}}\right) - f(p)\right| > 5\varepsilon\right) \leq \exp(-\sqrt{n}).$$

Can use APML [CSS19], approximating PML in near linear time.

Prior work either:

- Used different estimators for different properties
- Applied a plug-in estimator for only few properties

(A)PML apply to all additive Lipschitz properties and more

Essentially strengthens original table

Runs in near-linear time

α-Rényi entropy estimation

For integer $\alpha > 1$, PML plug-in has optimal $k^{1-1/\alpha}$ sample complexity

For non-integer $\alpha > 3/4$, (A)PML plug-in improves best-known results

Sorted distribution estimation

Under ℓ_1 distance, (A)PML yields optimal $\Theta(k/(\varepsilon^2 \log k))$ sample complexity for sorted distribution estimation

Uniformity testing: $p = p_u$ v.s. $|p - p_u| \geq \varepsilon$; complexity $\Theta(\sqrt{k}/\varepsilon^2)$

Tester below is sample-optimal up to logarithmic factors of k

Input: parameters $k, \varepsilon,$ and a sample $X^n \sim p$ with profile φ

If any symbol appears $\geq 3\max\{1, n/k\}\log k$ times, return 1

If $\|p_\varphi^{\text{ml}} - p_u\|_2 \geq 3\varepsilon/(4\sqrt{k})$, return 1; else, return 0

Thank You!