Unified
Maximum Likelihood Estimation
of
Symmetric Distribution Properties

Jayadev Acharya Hirakendu Das Alon Orlitsky Ananda Suresh

Cornell **Yahoo** **UCSD** **Google**

Frontiers in Distribution Testing Workshop

FOCS 2017

Special Thanks to…

I don’t smile

I only smile

Symmetric properties

$\mathcal{P} = \{ (p_1, \dots p_k) \}$ - collection of distributions over {1,…,k}

Distribution property: $f: \mathcal{P} \to \mathbb{R}$

f is symmetric if unchanged under input permutations

Entropy $H(p) \triangleq \sum p_i log \frac{1}{p_i}$

Support size $S(p) \triangleq \sum_i \mathbb{I}_{\{p_i > 0\}}$

Rényi entropy, support coverage, distance to uniformity, …

Determined by the probability multiset $\{p_1, p_2, \dots, p_k\}$

Symmetric properties

$\boldsymbol{p = (p_1, \ldots p_k)}$, ($\boldsymbol{k}$ finite or infinite) - discrete distribution

$\boldsymbol{f(p)}$, a property of $\boldsymbol{p}$

$\boldsymbol{f(p)}$ is symmetric if unchanged under input permutations

Entropy $\boldsymbol{H(p) \triangleq \sum p_i log \frac{1}{p_i}}$

Support size $\boldsymbol{S(p) \triangleq \sum_i \mathbb{I}_{\{p_i > 0\}}}$

Rényi entropy, support coverage, distance to uniformity, …

Determined by the probability multiset $\boldsymbol{\{p_1, p_2, \ldots, p_k\}}$

Property estimation

$\boldsymbol{p}$ unknown distribution in $\mathcal{P}$

Given independent samples $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n} \sim \mathbf{p}$

Estimate $\boldsymbol{f(p)}$

Sample complexity $\boldsymbol{S(f, k, \varepsilon, \delta)}$

Minimum n necessary to

Estimate $\boldsymbol{f(p) \pm \varepsilon}$

With error probability $< \boldsymbol{\delta}$

Plug-in estimation

Use $X_1, X_2, \ldots, X_n \sim p$ to find an estimate $\hat{p}$ of p

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

How to estimate p?

Sequence Maximum Likelihood (SML)

$$p^{\mathrm{sml}} = \arg\max_{p} \; p(x^n) = \arg\max_{p(x)} \; \Pi_i \, p(x_i)$$

$$x^3 = h, h, t$$

$$p^{sml}_{h,h,t} = \mathrm{argmax} \; p^2(h) \cdot p(t)$$

$p^{sml}_{h,h,t}$(h)=2/3 $p^{sml}_{h,h,t}$(t)=1/3

Same as empirical-frequency distribution

Multiplicity N_x - # times x appears in x^n

$$p^{\mathrm{sml}}(x) = \frac{N_x}{n}$$

Prior Work

For several important properties

Empirical-frequency plugin requires $\Theta(k)$ samples

New complex (non-plugin) estimators need $\Theta\left(\frac{k}{\log k}\right)$ samples

Property	Notation	SML	Optimal	References
Entropy	$H(p)$	$\frac{k}{\varepsilon}$	$\frac{k}{\log k}\frac{1}{\varepsilon}$	(Valiant & Valiant, 2011a; Wu & Yang, 2016; Jiao et al., 2015)
Support size	$\frac{S(p)}{k}$	$k \log \frac{1}{\varepsilon}$	$\frac{k}{\log k} \log^2 \frac{1}{\varepsilon}$	(Wu & Yang, 2015)
Support coverage	$\frac{S_m(p)}{m}$	m	$\frac{m}{\log m} \log \frac{1}{\varepsilon}$	(Orlitsky et al., 2016)
Distance to u	$\|p-u\|_1$	$\frac{k}{\varepsilon^2}$	$\frac{k}{\log k}\frac{1}{\varepsilon^2}$	(Valiant & Valiant, 2011b; Jiao et al., 2016)

Different estimator for each property

Use sophisticated approximation theory results

Entropy estimation

SML estimate of entropy $= \sum_x \frac{N_x}{n} \log \frac{n}{N_x}$

Sample complexity: $\Theta(k/\varepsilon)$

Various corrections proposed: Miller-Maddow, Jackknifed estimator, Coverage adjusted, …

Sample complexity: $\Omega(k)$ for all the above estimators

Entropy estimation

[Paninski'03]: $o(k)$ sample complexity (existential)

[ValiantValiant'11a]: Constructive LP based methods: $\Theta_{\varepsilon}\left(\frac{k}{\log k}\right)$

[ValiantValiant11b, WuYang'14, HanJiaoVenkatWeissman'14]:

Simplified algorithms, and growth rate: $\Theta\left(\frac{k}{\varepsilon \log k}\right)$

New (as of August) Results

Unified, simple, sample-optimal

approach for all above problems

Plug-in estimator,
replace **sequence** maximum likelihood
with **profile** maximum likelihood

Profiles

h, h, t or h, t, h or t, h, t → same estimate

One element appeared once, on appeared twice

Profile: Multi-set of multiplicities: $\Phi(X_1^n) = \{N_x : x \in X_1^n\}$

$$\Phi(h, h, t) = \Phi(t, h, t) = \{1, 2\}$$

$$\Phi(\alpha, \gamma, \beta, \gamma) = \{1, 1, 2\}$$

Sufficient statistic for symmetric properties

Profile maximum likelihood [+SVZ'04]

Profile probability

$$p(\Phi) = \sum_{\Phi(x_1^n)=\Phi} p(x_1^n)$$

Maximize the profile probability

$$p_\Phi^{pml} = \arg\max_p p(\Phi(X_1^n))$$

See "On estimating the probability multiset", Orlitsky, Santhanam, Viswanathan, Zhang for a detailed treatment, and an argument for competitive distribution estimation.

Profile maximum likelihood (PML) [+SVZ '04]

Profile probability

$$p(\Phi) = \sum_{x^n:\, \Phi(x_1^n)=\Phi} p(x^n)$$

Distribution Maximizing the profile probability

$$p_\Phi^{pml} = \arg\max_p p(\Phi)$$

PML competitive for distribution estimation

PML example

$X^3 = \mathrm{h}, \mathrm{h}, \mathrm{t}$

$p^{sml}_{h,h,t}(\mathrm{h})=2/3 \quad p^{sml}_{h,h,t}(\mathrm{t})=1/3$

$\Phi(\mathrm{h}, \mathrm{h}, \mathrm{t}) = \{1,2\}$

$p(\Phi = \{1,2\}) = p(s,s,d) + p(s,d,s) + p(d,s,s)$

$= 3p(s,s,d)$

$= \binom{3}{1}\left(\sum_{x \neq y} p^2(x)p(y)\right)$

$$p^{sml}(\{1\,,2\}) = \binom{3}{1}\left(\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)\right) = \frac{18}{27} = \frac{2}{3}$$

PML of {1,2}

P({1,2}) = p(s,s,d)+p(s,d,s)+p(d,s,s) = 3 p(s,s,d)

(1/2,1/2) ➞ p(s,s,d) = 1/8 + 1/8 = 1/4

$$\begin{aligned} \mathrm{p}(\mathrm{s},\mathrm{s},\mathrm{d}) &= \Sigma_{\mathrm{x}\neq\mathrm{y}} p^2(x)p(y) \\ &= \Sigma_{\mathrm{x}} p^2(x)\big(1-p(x)\big) \\ &\leq ¼\, \Sigma_{\mathrm{x}} p(x) = \frac{1}{4} \end{aligned}$$

p^{pml} (s,s,d) = ¼

p^{pml} ({1,2}) = ¾ Recall: $p^{sml}(\{1\ ,2\})$=2/3

PML({1,1,2})

$\Phi(\alpha, \gamma, \beta, \gamma) = \{1,1,2\}$

$p^{pml}(\{1,1,2\}) = U[5]$

PML can predict existence of new symbols

Profile maximum likelihood

PML of {1,2} is {½, ½ }

$$p^{pml}(\{1,2\}) = \binom{3}{1}\left(\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)\right) = \frac{3}{4} > \frac{18}{27}$$

$$\Sigma_{x\neq y} p^2(x)p(y) = \Sigma_x p^2(x)\big(1-p(x)\big) \leq \tfrac{1}{4}\,\Sigma_x p(x) = \frac{1}{4}$$

$$X^n = \alpha,\gamma,\beta,\gamma, \quad \Phi(X^n) = \{1,1,2\}$$

$$p^{pml}(\{1,1,2\}) = U[5]$$

PML can predict existence of new symbols

PML Plug-in

To estimate a symmetric property f

Find $p^{pml}(\Phi(X^n))$

Output $f(p^{pml})$

Simple

Unified

No tuning parameters

Some experimental results (c. 2009)

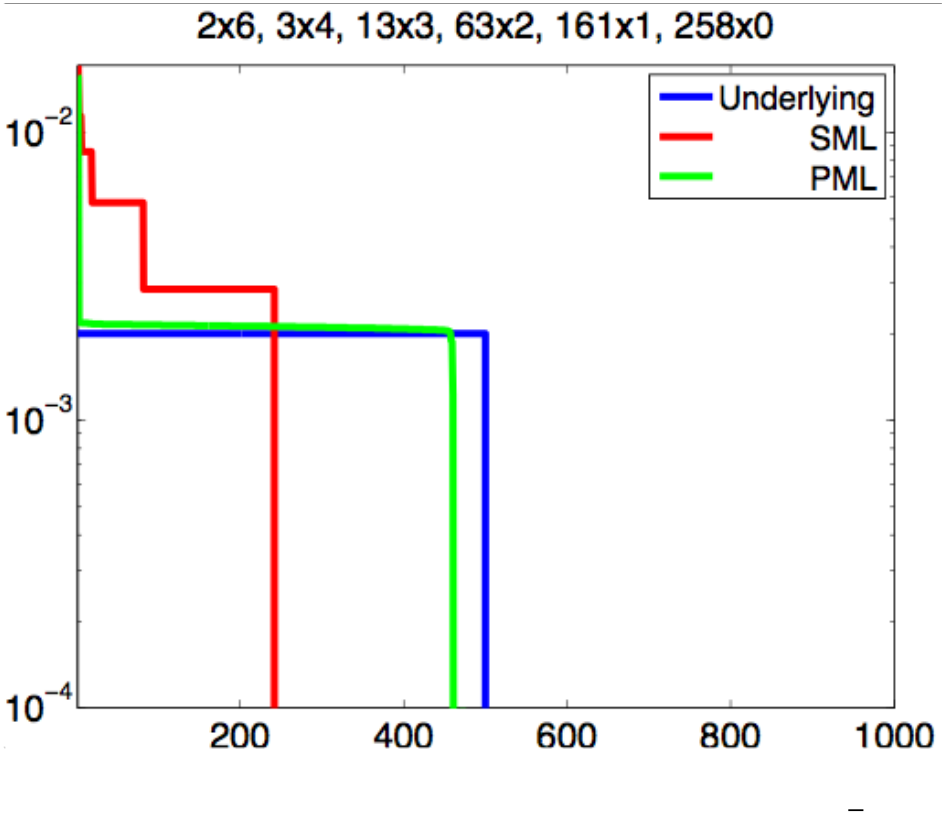
Uniform

500 symbols

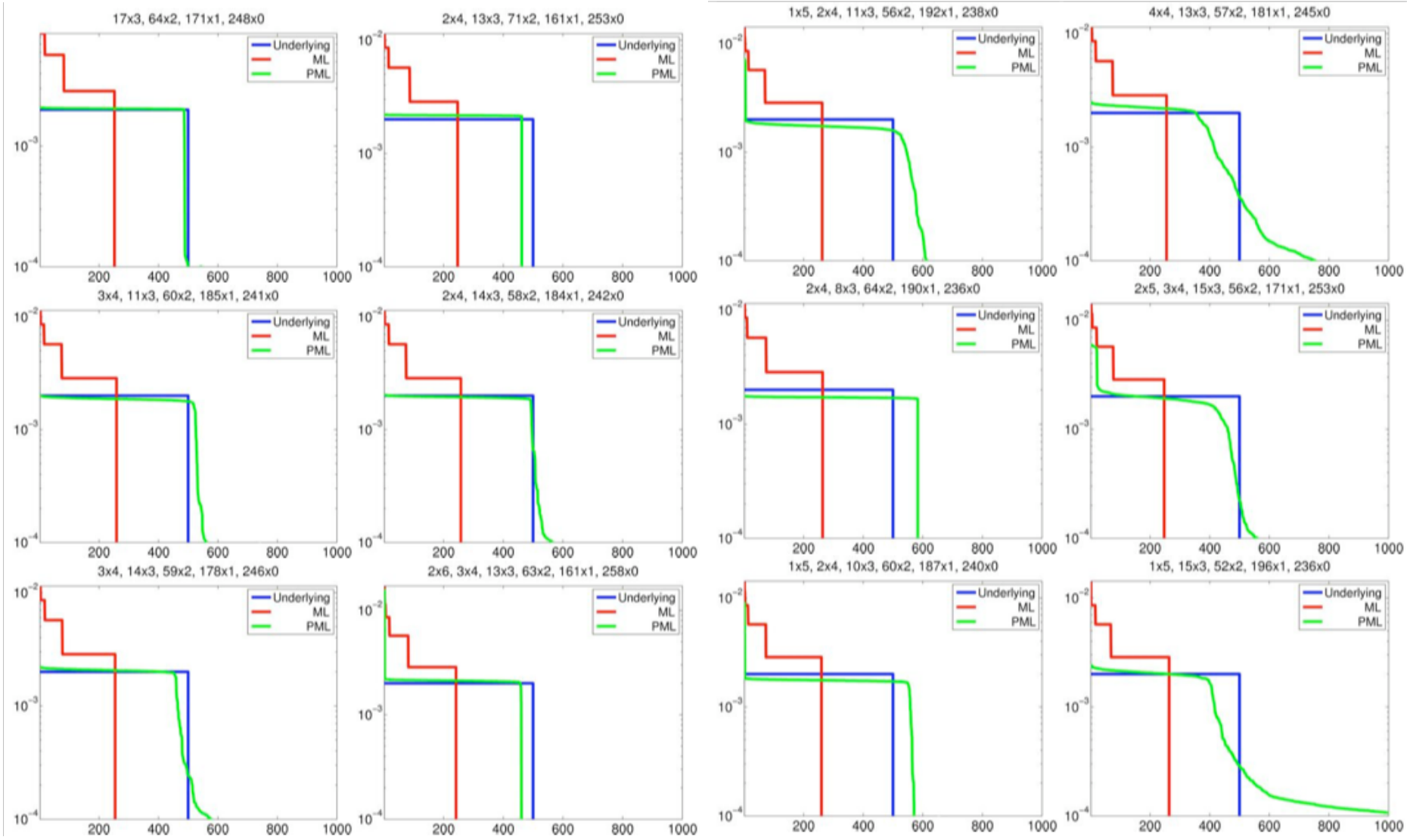
350 samples

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

242 appeared, 258 did not



U[500], 350x, 12 experiments



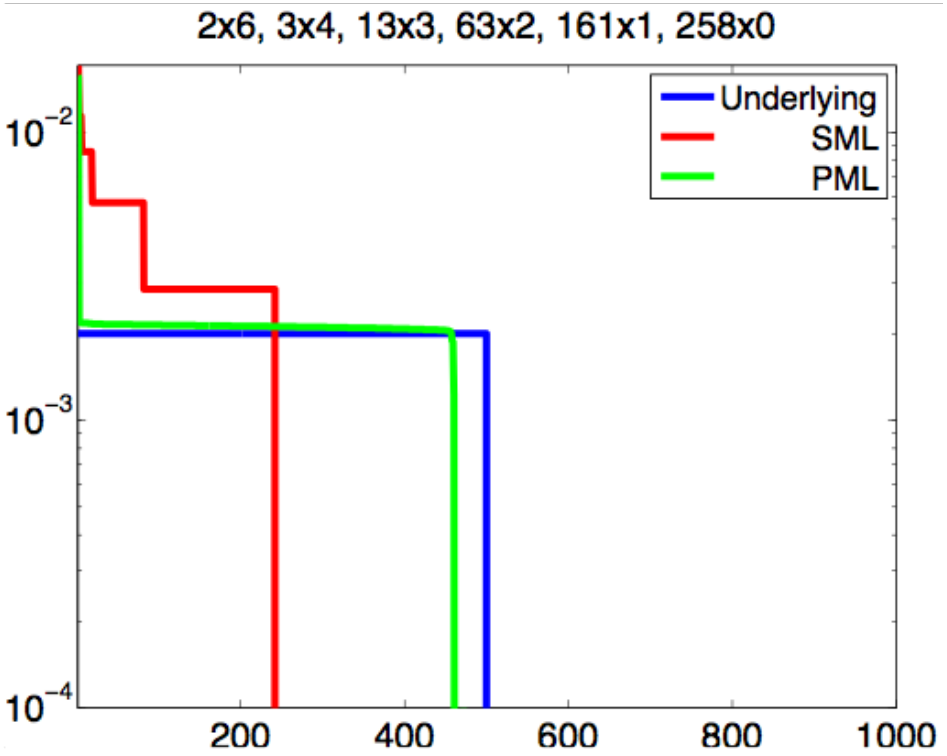
Uniform

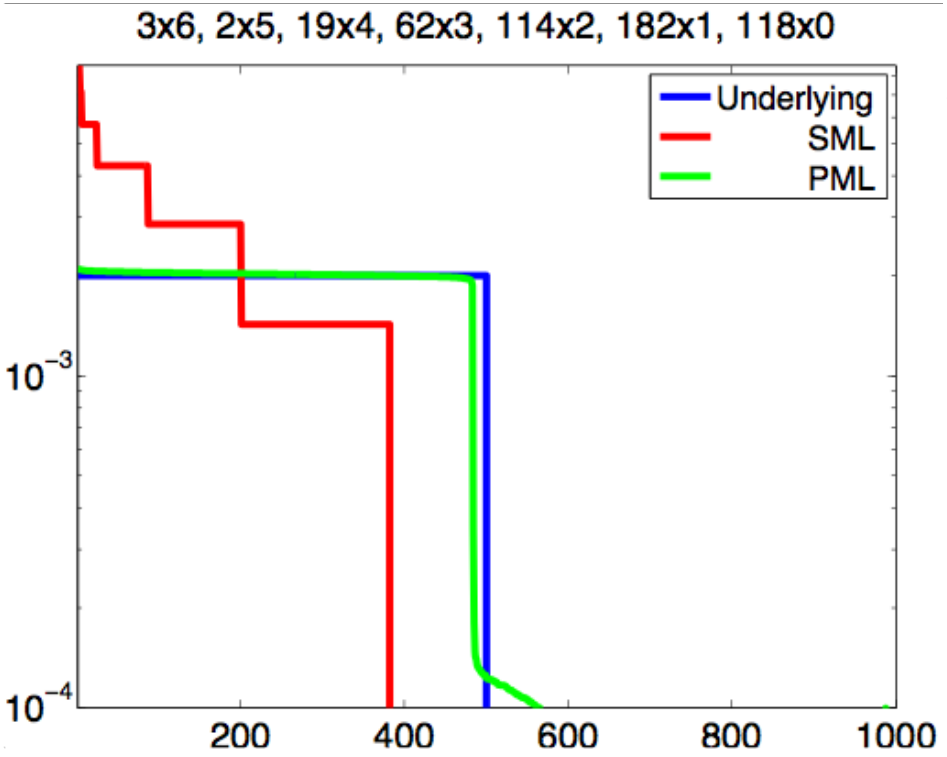
500 symbols

350 samples

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

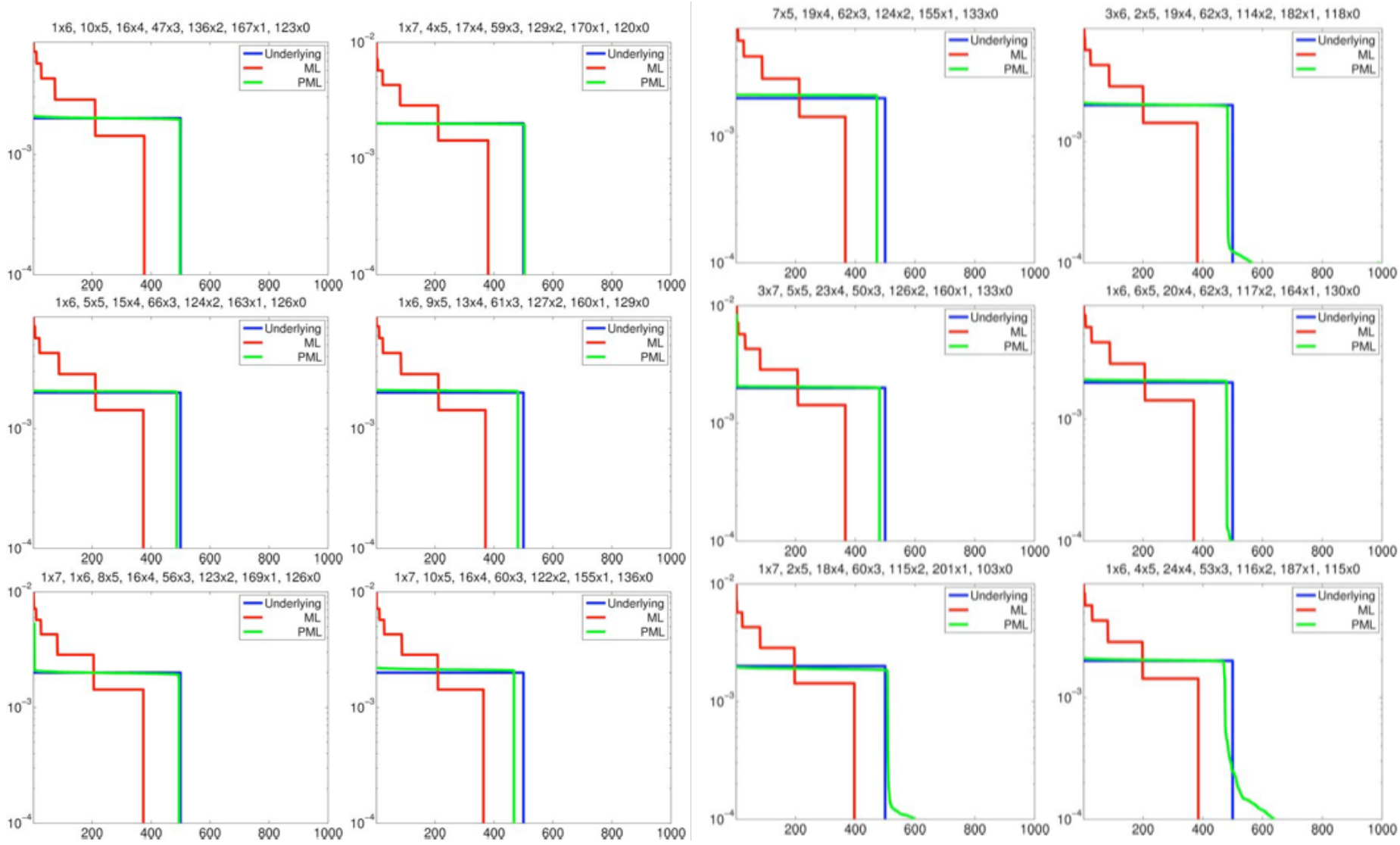
248 appeared, 258 did not





700 samples

U[500], 700x, 12 experiments



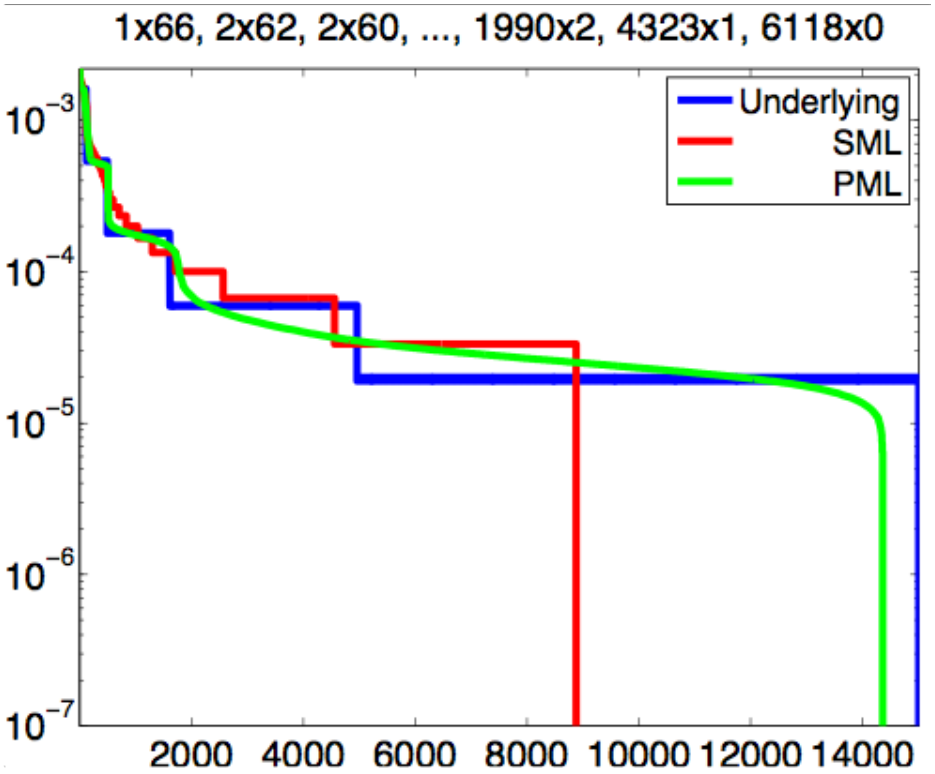
Staircase

15K elements, 5 steps, ~3x

30K samples

Observe 8,882 elts

6,118 missing



Zipf

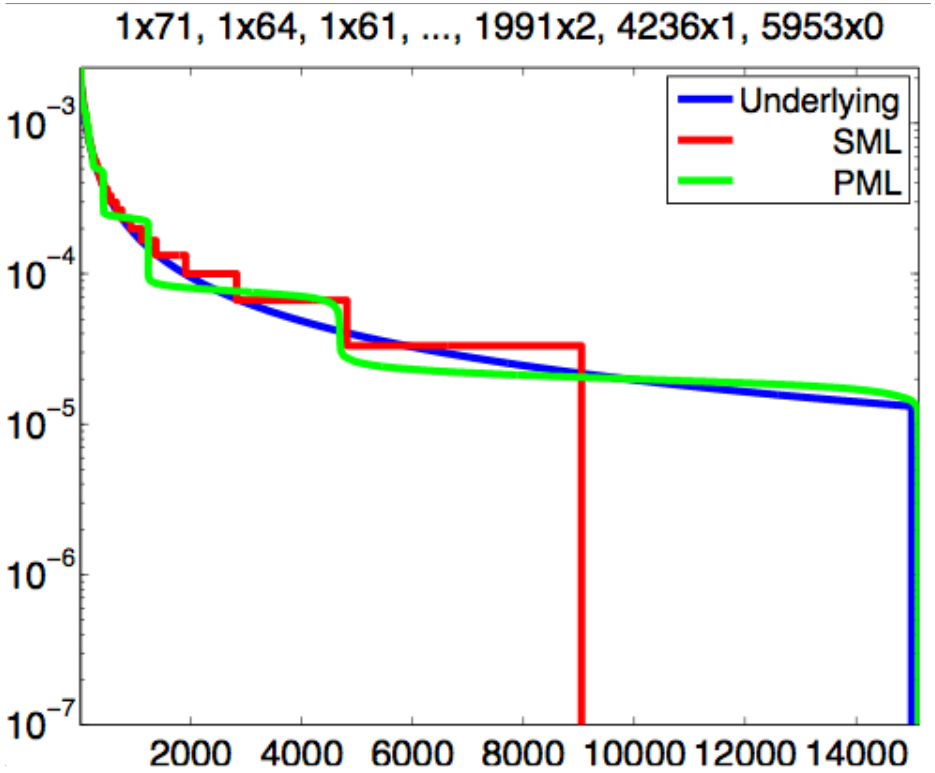
Underlies many natural phenomena

$p_i = C/i$, 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

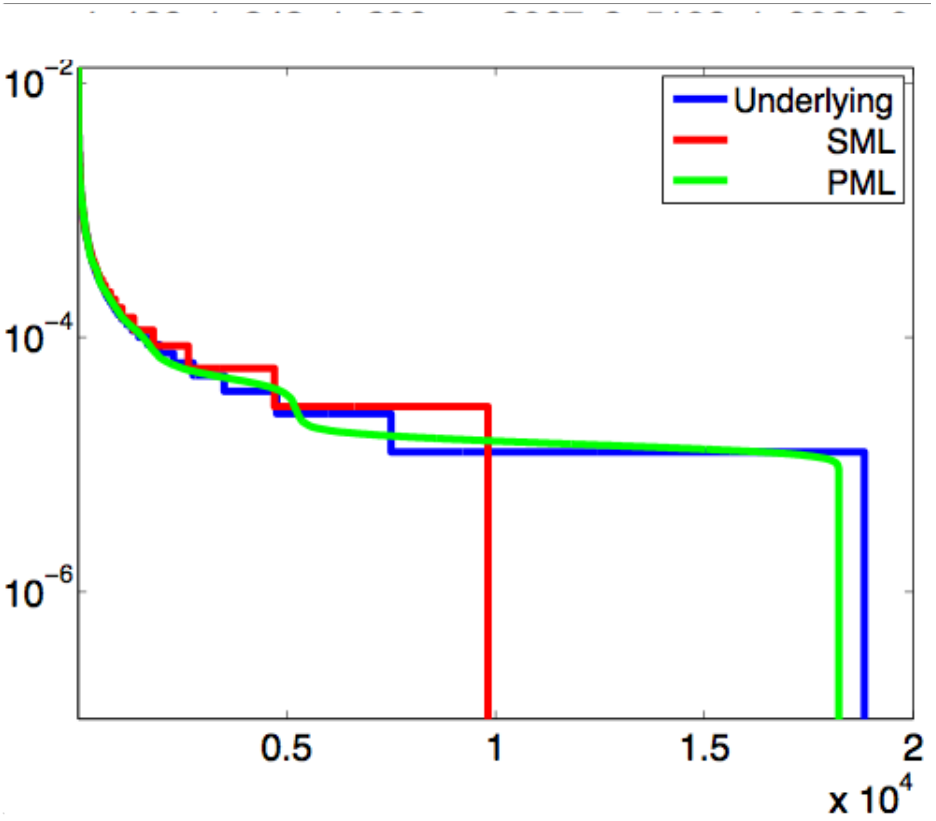
77.48% population

~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

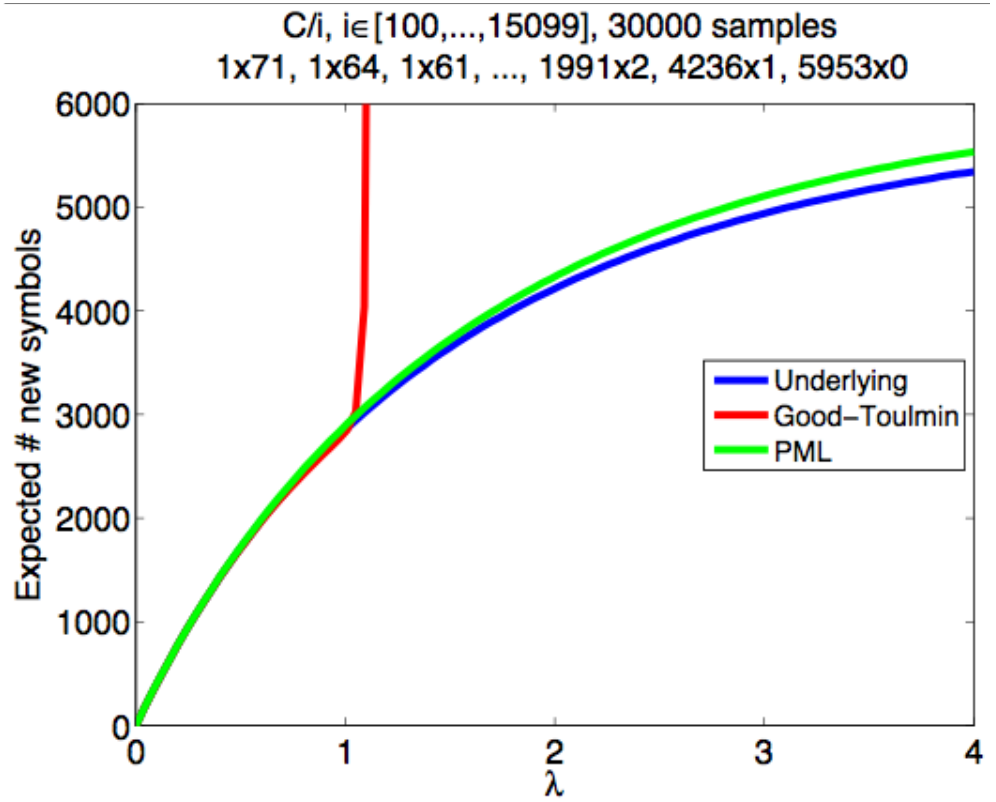
Good-Toulmin: $\lambda < 1$

$\lambda > 1$

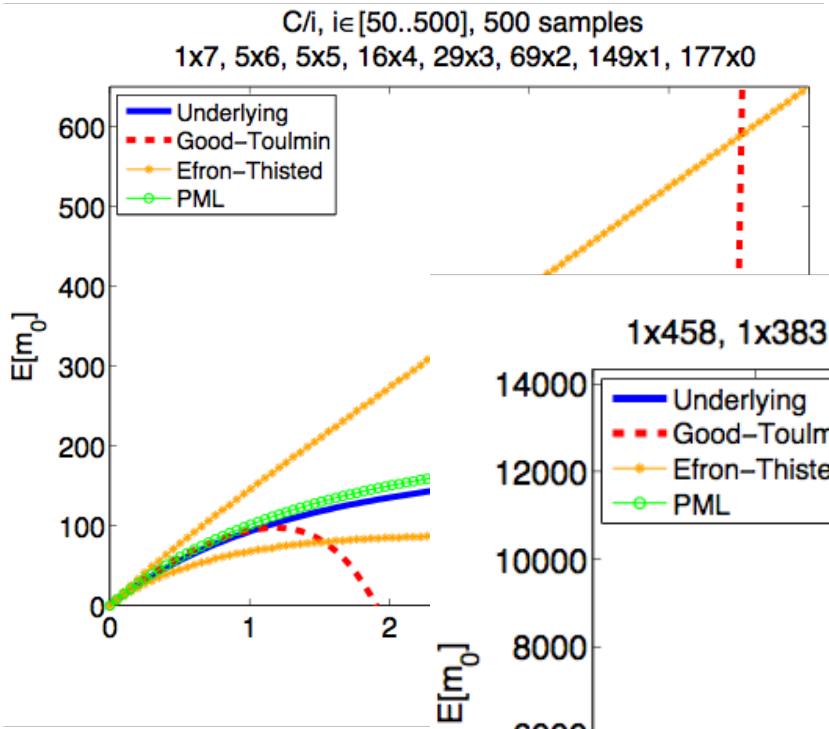
Estimate PML & predict

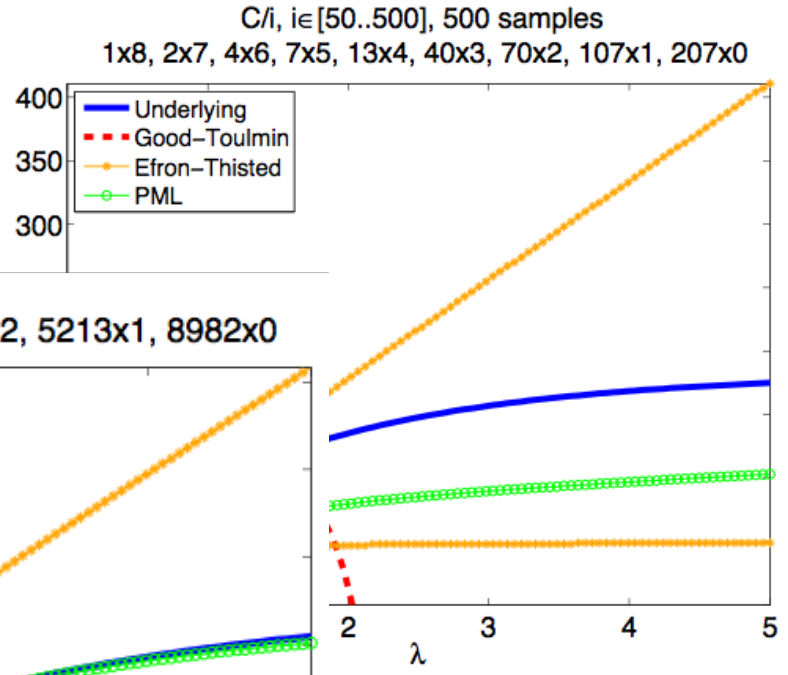
Extends to $\lambda > 1$

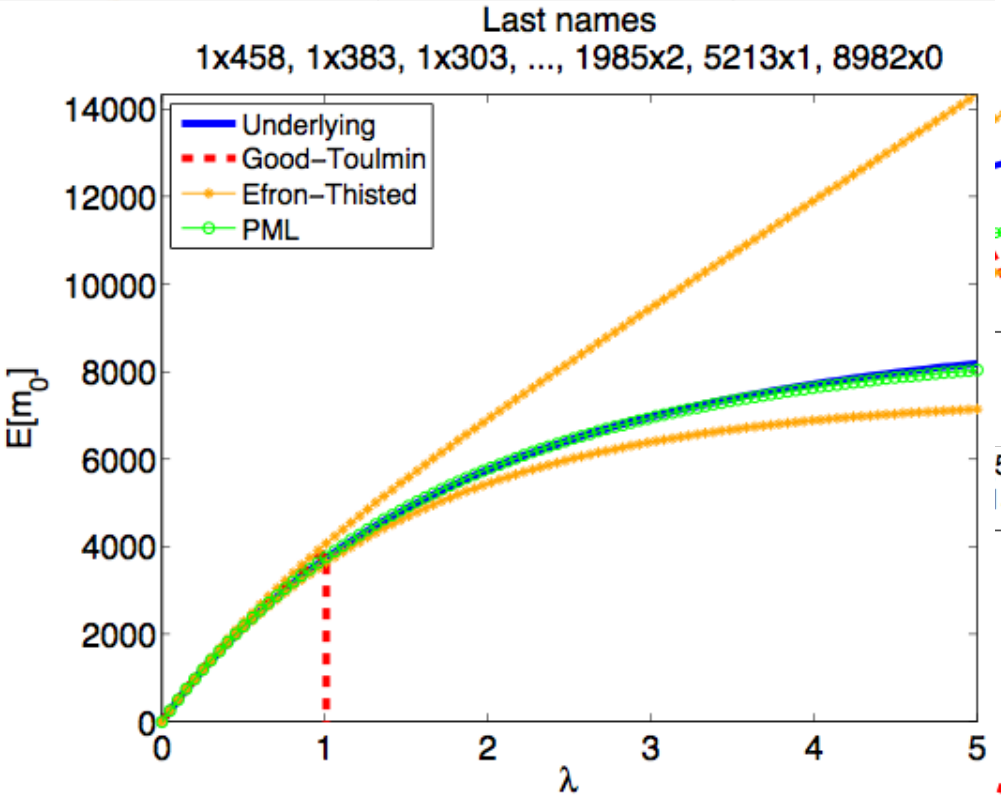
Applies to other properties

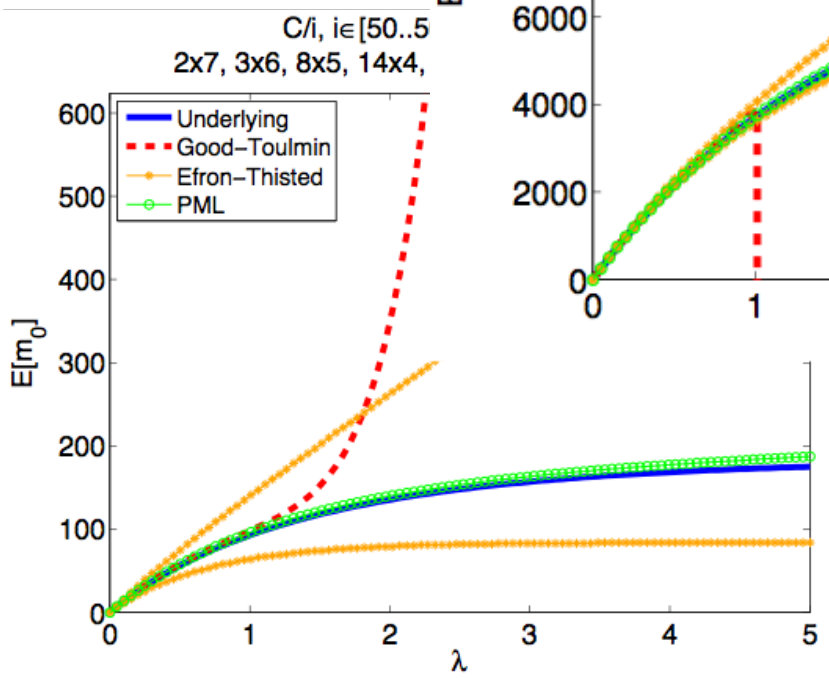


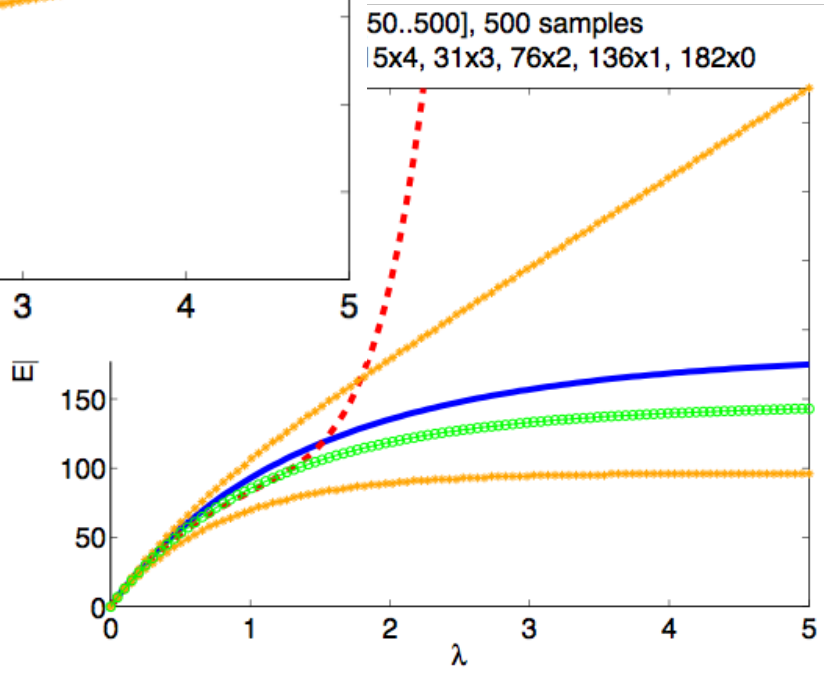
$\lambda > 1$


C/i, i∈[50..500], 500 samples
1x7, 5x6, 5x5, 16x4, 29x3, 69x2, 149x1, 177x0
Underlying
Good–Toulmin
Efron–Thisted
PML
E[m0]


C/i, i∈[50..500], 500 samples
1x8, 2x7, 4x6, 7x5, 13x4, 40x3, 70x2, 107x1, 207x0
Underlying
Good–Toulmin
Efron–Thisted
PML
λ


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,
Underlying
Good–Toulmin
Efron–Thisted
PML
E[m0]
λ


50..500], 500 samples
5x4, 31x3, 76x2, 136x1, 182x0
E|
λ

Finding the PML distribution

EM algorithm [+ Pan, Sajama, Santhanam, Viswanathan, Zhang ’05 - ’08]

Approximate PML via Bethe Permanents [Vontobel]

Extensions of Markov Chains [Vatedka, Vontobel]

No provable algorithms known

Motivated Valiant & Valiant

Maximum Likelihood Estimation Plugin

General property estimation technique

$\mathcal{P}$ - collection of distributions over domain $\mathcal{Z}$

$f: \mathcal{P} \to \mathbb{R}$ any property (say entropy)

MLE estimator

Given $z \in \mathcal{Z}$

Determine $p_z^{\mathrm{MLE}} \triangleq \arg\max_{p \in \mathcal{P}} p(z)$

Output $f(p_z^{\mathrm{MLE}})$

How good is MLE?

Competitiveness of MLE plugin

$\mathcal{P}$ - collection of distributions over domain $\mathcal{Z}$

$\hat{f}: \mathcal{Z} \to \mathbb{R}$ **any** estimaor such that $\forall\, p \in \mathcal{P},\ Z \sim p$

$$\Pr\left(\left|f(p) - \hat{f}(Z)\right| > \varepsilon\right) < \delta$$

MLE plugin error bounded by

$$\Pr\left(\left|f(p) - f(p_z^{\mathrm{MLE}})\right| > 2 \cdot \varepsilon\right) < \delta \cdot |\mathcal{Z}|$$

Simple, universal, competitive with any $\hat{f}$

Quiz: Probability of unlikely outcomes

6-sided die, p=$(p_1, p_2, \ldots, p_6)$

$p_i \geq 0$, and $\Sigma p_i = 1$, otherwise arbitrary

Z$\sim$p

$P_r(p_z \leq 1/6)$ Can be anything

(1,0,…,0) ➞ Pr($p_z \leq$ 1/6) = 0

(1/6, …, 1/6) ➞ Pr($p_z \leq 1/6$) = 1

$P_r(p_z \leq 0.01)$ ≤ 0.06

$P_r(p_z \leq 0.01) = \Sigma_{i:p_i \leq 0.01}\ p_i \leq 6 \cdot 0.01 = 0.06$

Competitiveness of MLE plugin - proof

$\hat{f}: \mathcal{Z} \to \mathbb{R}$: $\quad \forall\, p \in \mathcal{P},\ \ Z \sim p \ \rightarrow\ \Pr\left(\left|f(p) - \hat{f}(Z)\right| > \varepsilon\right) < \delta$

then $\Pr\left(\left|f(p) - f(p_Z^{\mathrm{MLE}})\right| > 2 \cdot \varepsilon\right) < \delta \cdot |\mathcal{Z}|$

For all z such that $p(z) \geq \delta$: 1) $\left|f(p) - \hat{f}(z)\right| \leq \varepsilon$

2) $p_Z^{\mathrm{MLE}}(z) \geq p(z) > \delta$, hence $\left|f(p_Z^{\mathrm{MLE}}) - \hat{f}(z)\right| \leq \varepsilon$

Triangle inequality: $\left|f(p_Z^{\mathrm{MLE}}) - f(p)\right| \leq 2\varepsilon$

If $\left|f(p_Z^{\mathrm{MLE}}) - f(p)\right| > 2\varepsilon$ then $p(z) < \delta$,

$$\Pr\left(\left|f\left(p_Z^{\mathrm{MLE}}\right) - f(p)\right| > 2\varepsilon\right) \leq \Pr(p(Z) < \delta) \leq \sum_{p(z)<\delta} p(z) \leq \delta \cdot |\mathcal{Z}|$$

PML performance bound

If $n = S(f, k, \varepsilon, \delta)$, then $S^{pml}(f, k, 2 \cdot \varepsilon, |\Phi^n| \cdot \delta) \leq n$

$|\Phi^n|$: number of profiles of length n

Profile of length n: partition of n

{3}, {1,2}, {1,1,1} ➞ 3, 2+1, 1+1+1

$|\Phi^n| = partition\ \#\ of\ n$

Hardy-Ramanujan: $|\Phi^n| < e^{3\sqrt{n}}$

Easy: $e^{\log n \cdot \sqrt{n}}$

If $\boldsymbol{n = S\left(f, k, \varepsilon, e^{-4\sqrt{n}}\right)}$, then $\boldsymbol{S^{pml}\left(f, k, 2\varepsilon, e^{-\sqrt{n}}\right) \leq n}$

Summary

Symmetric property estimation

PML plug-in approach

- Simple
- Universal
- Sample optimal for known sublinear properties

Future directions

- Provably efficient algorithms
- Independent proof technique

Thank You!

PML for symmetric f

For any symmetric property f,

if $n = S(f, k, \varepsilon, 0.1)$, then $S^{pml}(f, k, 2 \cdot \varepsilon, 0.1\ ) = O(n^2)$.

Proof. By median trick, $S(f, k, \varepsilon, e^{-m}) = O(n \cdot m)$.

Therefore,

$S^{pml}\left(f, k, 2 \cdot \varepsilon, e^{3\sqrt{n \cdot m} - m}\right) = O(n \cdot m)$,

Plugging in $m = C \cdot n$, gives the desired result.

Better error probabilities – warm up

Estimating a distribution p over $[k]$ to L1 distance ε w.p. > 0.9 requires $\Theta(k/\varepsilon^2)$ samples.

Proof: Exercise.

Estimating a distribution p over $[k]$ to L1 distance ε w.p. $1 - e^{-k}$ requires $\Theta(k/\varepsilon^2)$ samples.

Proof:

- Empirical estimator $\hat{p}$
- $|\hat{p} - p|$ has bounded difference constant (b.d.c.) $2/n$
- Apply McDiarmid's inequality

Better error probabilities

Recall

$$S(H, \varepsilon, k, 2/3\ ) = \Theta\left(\frac{k}{\varepsilon \cdot \log k}\right)$$

- Existing optimal estimators: high b.d.c.
- Modify them to have small b.d.c., and still be optimal
- In particular, can get b.d.c. = $n^{-0.95}$ (exponent close to 1)
- With twice the samples error drops **super-fast**

$$S\left(H, \varepsilon, k, e^{-n^{0.9}}\right) = \Theta\left(\frac{k}{\varepsilon \cdot \log k}\right)$$

Similar results for other properties

Even approximate PML works!

- Perhaps finding exact PML is hard.
- Even approximate PML works.

Find a distribution q such that

$$q\left(\Phi(X_1^n)\right) \geq e^{-n^{.8}} \cdot p^{pml}(\Phi(X_1^n))$$

Even this is optimal (for large k)

In Fisher's words …

Of course nobody has been able to prove that MLE is best under all circumstances. MLE computed with all the information available may turn out to be inconsistent. Throwing away a substantial part of the information may render them consistent.

R. A. Fisher

Proof of PML performance

If $\boldsymbol{n = S(f, k, \varepsilon, \delta)}$, then $\boldsymbol{S^{pml}(f, k, 2 \cdot \varepsilon, |\Phi^n| \cdot \delta) \leq n}$

$S(f, k, \varepsilon, \delta)$, achieved by an estimator $\hat{f}(\Phi(X^n))$

- Profiles $\Phi(X^n)$ such that $p\big(\Phi(X^n)\big) > \delta$,

$$p^{PML}(\Phi) \geq p(\Phi) > \delta$$

$$\left|f\left(p_{\Phi}^{PML}\right) - f(p)\right| \leq \left|f\left(p_{\Phi}^{PML}\right) - \hat{f}(\Phi)\right| + \left|\hat{f}(\Phi) - f(p)\right| < 2\varepsilon$$

- Profiles with $p\big(\Phi(X^n)\big) < \delta$,

$$p(p(\Phi(X^n) < \delta) < \delta \cdot |\Phi^n|$$