Optimal Prediction of the Number of Unseen Species

Orlitsky, Alon, Ananda Theertha Suresh, and Yihong Wu. "Optimal prediction of the number of unseen species." *Proceedings of the National Academy of Sciences* 113.47 (2016): 13283-13288.

Presenter: Jason Vega

Wednesday, December 11, 2024

Overview

1. Background
2. Proposed Estimator
3. Theory

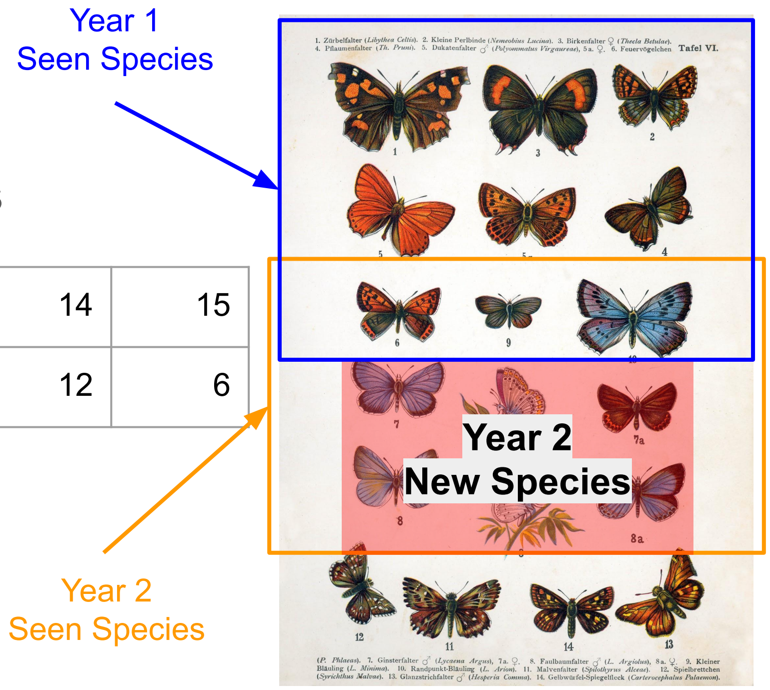
Background

Unseen Species Problem

Data from 1 year of trapping butterflies

Frequency	1	2	3	4	5	…	14	15
Species	118	74	44	24	29	…	12	6

e.g., for 118 species, only 1 specimen was captured!



Problem: How many **new species** from **another year** of trapping?

Unseen Species Problem

Year 1: *n* i.i.d. samples $X^n = (X_1, X_2, \ldots, X_n)$

Year 2: *m* i.i.d. samples $X_{n+1}^{n+m} = (X_{n+1}, X_{n+2}, \ldots, X_{n+m})$

Year 1 unique species: $\{X^n\}$

Year 2 unique species: $\{X_{n+1}^{n+m}\}$

Year 2 new species: $U = |\{X_{n+1}^{n+m}\} \setminus \{X^n\}|$

Prior Work: Good-Toulmin Estimator

Φ_i: # of species w/ freq. *i* (“**prevalence of i**”)

Frequency	1	2	3	4	5	…	14	15
Species	118	74	44	24	29	…	12	6

Φ_1 Φ_2 Φ_3 Φ_4 Φ_5 Φ_{14} Φ_{15}

Let **t = m/n** (yr. 2 to yr 1. sample size ratio)

Good-Toulmin Estimator: $U^{\mathrm{GT}} = -\sum_{i=1}^{\infty} (-t)^i \Phi_i$

e.g., for table data w/ t=1, U^{GT} = 118 - 74 + 44 - 24 + … - 12 + 6 = 75

Background: Bias, Variance and MSE

Given sample X, want to estimate some quantity Y.

Estimator: $\hat{Y} = f(X)$

Bias: $\mathbb{E}[\hat{Y} - Y]$

Variance: $\mathrm{Var}(\hat{Y} - Y)$ (Def. used in this paper)

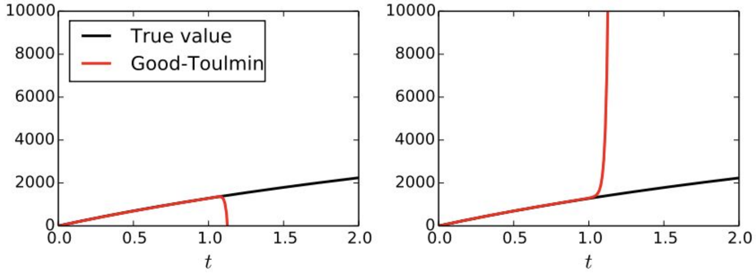
MSE: $\mathbb{E}[(\hat{Y} - Y)^2]$

Note that Variance ≤ MSE

Ideally: low bias and variance!

Good-Toulmin Performance

- For $t \leq 1$:
  - **Bias:** nearly unbiased!
  - **MSE:** $O(nt^2)$
- For $t > 1$: **high variance!**



Example for a Zipf distribution

$$U^{\mathrm{GT}} = -\sum_{i=1}^{\infty}(-t)^i \Phi_i$$

For $t > 1$, $|(-t)^i|$ explodes as i grows!

As $t \to \infty$, U^{GT} dominated by $(-t)^i\Phi_i$ for largest i s.t. $\Phi_i > 0$

Problem: How to reduce variance for $t > 1$?

Prior Work: Efron-Thisted Estimator

Intuition: Counteract exploding $(-t)^i$ with something that *decays* with i!

Efron-Thisted Estimator: $$U^{\mathrm{ET}} = -\sum_{i=1}^{n} (-t)^i \underbrace{\mathbb{P}(\mathrm{Bin}(k, \tfrac{1}{1+t}) \geq i)}_{\text{Tail probability of Binomial distribution}} \Phi_i$$

- Derived by truncating Euler transform of U^{GT} after k terms… **complicated!**
  - Later: U^{ET} derived easily through a probabilistic interpretation!
- **Performance:** Good empirical performance!
- **Problem:** No theoretical guarantees… (until Orlitsky et al., 2016!)

Proposed Estimator

Initial Attempt: Truncated Good-Toulmin Estimator

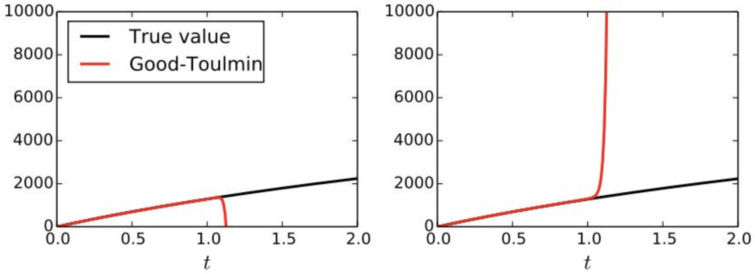
- What if we did something even simpler than U^{ET}?
- Idea: Since high variance due to $(-t)^i\Phi_i$ for large i, truncate U^{GT} after ℓ terms

Truncated Good-Toulmin Estimator: $$U^{\ell} = -\sum_{i=1}^{\ell}(-t)^i\Phi_i$$

- **Problem:** for t > 1, in the worst case, bias is still large! (will show later)

Smoothed Good-Toulmin Estimator

- U^ℓ may have positive or negative bias, depending on sign of dominant $(-t)^i\Phi_i$:



- **Idea:** Average over many U^ℓ so that biases "cancel" out!

Smoothed Good-Toulmin Estimator: $$U^L = \mathbb{E}_L\left[-\sum_{i=1}^{L}(-t)^i\Phi_i\right]$$

- The **smoothing distribution** of L can control the MSE behavior!

Probabilistic Interpretation of Efron-Thisted

Smoothed Good-Toulmin:

$$U^L = \mathbb{E}_L\left[-\sum_{i=1}^{L}(-t)^i\Phi_i\right]$$

$$= \mathbb{E}_L\left[-\sum_{i=1}^{\infty}(-t)^i\Phi_i\mathbb{I}\{i \le L\}\right]$$

$$= -\sum_{i=1}^{\infty}(-t)^i\Phi_i\mathbb{E}_L[\mathbb{I}\{i \le L\}]$$

$$= -\sum_{i=1}^{\infty}(-t)^i\Phi_i\mathbb{P}(L \ge i)$$

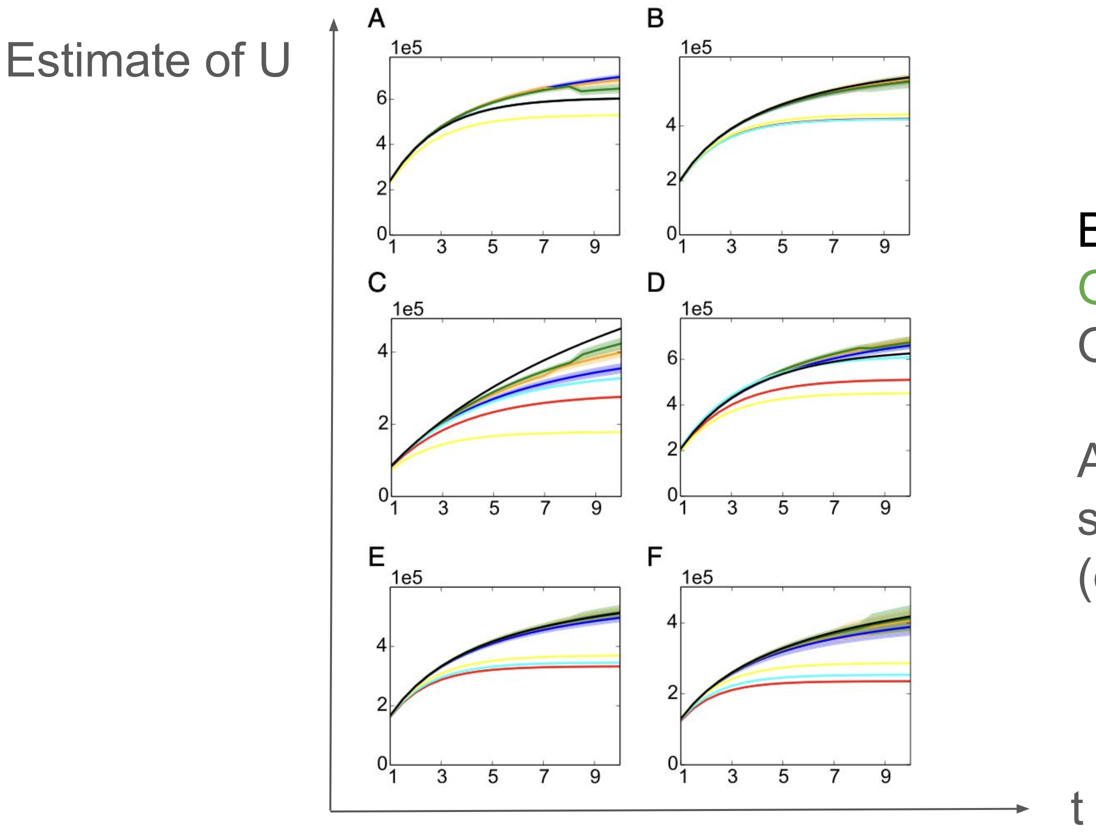
Efron-Thisted:

$$U^{\mathrm{ET}} = -\sum_{i=1}^{n}(-t)^i\mathbb{P}(\mathrm{Bin}(k, \tfrac{1}{1+t}) \ge i)\Phi_i$$

Efron-Thisted is U^L with

$$L \sim Bin(k, \tfrac{1}{1+t})$$

Experimental Results (Synthetic Data)

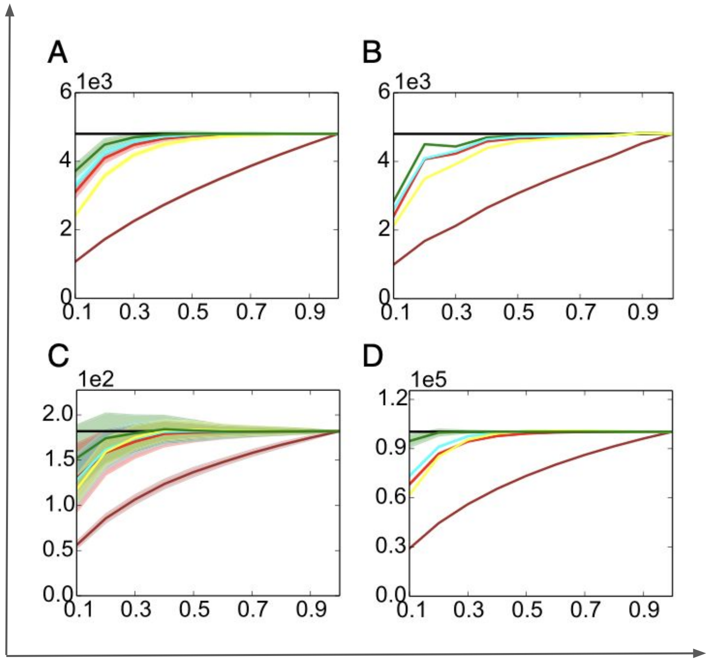


Black = True value
Green = SGT
Others = Other baselines

A-F denote various synthetic distributions (e.g. uniform, Zipf)

Experimental Results: Support Size Estimation (Real Data)

Estimated Support Size



Fraction of seen data

Black = True value
Green = SGT
Others = Other baselines

A-D denote various real data distributions (e.g. Shakespearean vocabulary, last names from Census data)

Est. Support Size = (# already seen species) + (U^E for m = # remaining data)

Theory

Evaluation Metric

Note: given n and t, U is at most m = nt

Worst-Case Normalized MSE of Estimator U^E:

$$\mathcal{E}_{n,t}(U^E) = \sup_p \mathbb{E}_p \left[\left(\frac{U^E - U}{nt} \right)^2 \right]$$

Recall that Variance ≤ MSE

Theorem 1: Performance of SGT

for t ≥ 1

Table 1. NMSE of SGT estimators for three smoothing distributions

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^{L})$
Poisson (r)	$r = \frac{1}{2t}\log_e \frac{n(t+1)^2}{t-1}$	$O(n^{-1/t})$
U^{ET} → Binomial (k, q)	$k = \left\lfloor \frac{1}{2}\log_2 \frac{nt^2}{t-1} \right\rfloor,\ q = \frac{1}{t+1}$	$O(n^{-\log_2(1+1/t)})$
Binomial (k, q)	$k = \left\lfloor \frac{1}{2}\log_3 \frac{nt^2}{t-1} \right\rfloor,\ q = \frac{2}{t+2}$	$O(n^{-\log_3(1+2/t)})$ ← Best!

for $t > 1$: $-\log_3(1+2/t) \le -\log_2(1+1/t) \le -1/t$

- Principled method of selecting k for Efron-Thisted with performance guarantees
- Slight modification to q beats original U^{ET}

Theorem 2: Best-Case Performance

∃ c' s.t. for any n and any estimator U^E,

$$\mathcal{E}_{n,t}(U^E) = \Omega\left(n^{-c'/t}\right)$$

Table 1. NMSE of SGT estimators for three smoothing distributions

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^L)$
Poisson (r)	$r=\frac{1}{2t}\log_e\frac{n(t+1)^2}{t-1}$	$O(n^{-1/t})$
Binomial (k,q)	$k=\left\lfloor\frac{1}{2}\log_2\frac{nt^2}{t-1}\right\rfloor,\ q=\frac{1}{t+1}$	$O(n^{-\log_2(1+1/t)})$
Binomial (k,q)	$k=\left\lfloor\frac{1}{2}\log_3\frac{nt^2}{t-1}\right\rfloor,\ q=\frac{2}{t+2}$	$O(n^{-\log_3(1+2/t)})$

SGT estimators are near-optimal!

Corollary 1: Limits of Prediction Accuracy

$\forall \delta > 0,$

$$\max\{t : \exists U^E \, \mathcal{E}_{n,t}(U^E) < \delta\} = \Theta\left(\frac{\log n}{\log \frac{1}{\delta}}\right)$$

At best, an estimator will be accurate for new sample sizes up to $m \propto n * \log(n)$

Corollary 1: Limits of Prediction Accuracy

Paper claims SGT achieves corollary 1 limit

Rough idea to support claim (note: some abuse of notation)

(from Theorem 1) $\mathcal{E}_{n,t}(U^L) = O(n^{-1/t}) < \delta$

$$\Rightarrow \frac{1}{\delta} < n^{1/t}$$

$$\Rightarrow \log_n \frac{1}{\delta} < \frac{1}{t}$$

$$\Rightarrow \frac{\log \frac{1}{\delta}}{\log n} < \frac{1}{t}$$

$$\Rightarrow t < \frac{\log n}{\log \frac{1}{\delta}}$$

Analysis of Linear Estimators

- All estimators shown so far are linear
- Consider an arbitrary linear estimator:

$$U^h = \sum_{i=1}^{\infty} h_i \Phi_i$$

- Note that the series h_i can form the derivatives at 0 for some function h through the Taylor expansion:

$$h(y) = \sum_{i=1}^{\infty} \frac{h_i y^i}{i!}$$ (assuming h(0) = 0)

Analysis of Linear Estimators (Lemma 1)

- Let p_x be probability of observing species x, and let $\lambda_x = np_x$. The bias is then:

$$\mathbb{E}[U^h - U] = \sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right)$$

Want to approximate g well at points $\{\lambda_x\}$

- And the variance satisfies:

$$\mathrm{Var}(U^h - U) \le \mathbb{E}\left[\sum_{i=1}^{\infty} \Phi_i\right] \cdot \sup_{i \ge 1} h_i^2 + \mathbb{E}[U]$$

Want small derivatives at 0

- We can thus reinterpret constructing a linear estimator as **function approximation of g(y) = 1-e^{-ty}**

Analysis of Good-Toulmin

Lemma 1:

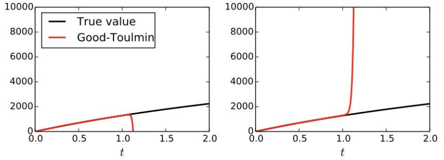
$$\mathbb{E}[U^h - U] = \sum_x e^{-\lambda_x}(h(\lambda_x) - (1 - e^{-t\lambda_x}))$$

$$\mathrm{Var}(U^h - U) \le \mathbb{E}[\sum_{i=1}^{\infty} \Phi_i] \cdot \sup_{i\ge 1} h_i^2 + \mathbb{E}[U]$$

- Recall Good-Toulmin estimator:

$$U^{\mathrm{GT}} = -\sum_{i=1}^{\infty}(-t)^i \Phi_i$$

- Note that its $\{h_i\}$ are the derivatives at 0 for $h(y) = 1-e^{-ty} = g(y)$!
    - For $h(y) = 1-e^{-ty}$, the i^{th} derivative is $-(-t)^i e^{-ty}$, so the derivatives at 0 are $-(-t)^i$
- According to Lemma 1, this means that Good-Toulmin is unbiased
- However, for $t > 1$, $|-(-t)^i| \to \infty$ as $i \to \infty$, so the variance can **explode**!



Analysis of Truncated Good-Toulmin

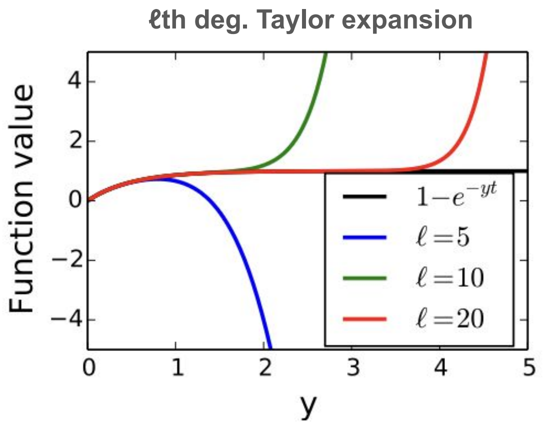
Lemma 1:

$$\mathbb{E}[U^h - U] = \sum_x e^{-\lambda_x}(h(\lambda_x) - (1 - e^{-t\lambda_x}))$$
$$\mathrm{Var}(U^h - U) \leq \mathbb{E}[\sum_{i=1}^{\infty} \Phi_i] \cdot \sup_{i \geq 1} h_i^2 + \mathbb{E}[U]$$

- Recall Truncated Good-Toulmin estimator:

$$U^\ell = -\sum_{i=1}^{\ell} (-t)^i \Phi_i$$

- Note that its $\{h_i\}$ are the derivatives at 0 for the ℓ^{th} degree Taylor expansion of g(y)
- Now we are guaranteed **finite variance**!
- Approximation quality degrades for y far from 0 => **large bias**



Analysis of Smoothed Good-Toulmin

Lemma 1:

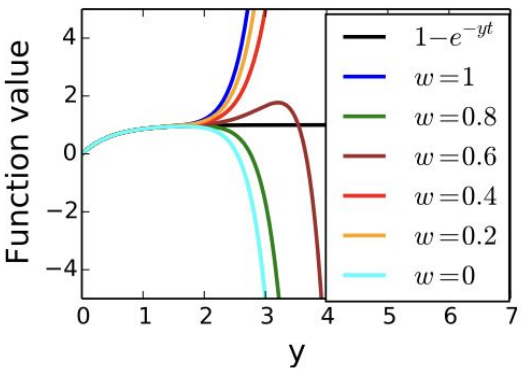
$$\mathbb{E}[U^h - U] = \sum_x e^{-\lambda_x}(h(\lambda_x) - (1 - e^{-t\lambda_x}))$$
$$\mathrm{Var}(U^h - U) \le \mathbb{E}[\sum_{i=1}^{\infty} \Phi_i] \cdot \sup_{i\ge 1} h_i^2 + \mathbb{E}[U]$$

- Recall Smoothed Good-Toulmin estimator:

$$U^L = \mathbb{E}_L\left[-\sum_{i=1}^{L}(-t)^i \Phi_i\right]$$

- Clearly, its $\{h_i\}$ are weighted averages of the derivatives at 0 for Taylor expansions of g(y) over all possible degrees
- Averaging over Taylor expansions gives overall better approximation of g(y) => **bias *can* be reduced**
- **Finite variance** possible for specific smoothing distributions

Avg. of 10th and 11th deg. Taylor expansion



Proof Sketch of Theorem 1 (Poisson)

for t ≥ 1

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^L)$
Poisson (r)	$r = \frac{1}{2t}\log_e \frac{n(t+1)^2}{t-1}$	$O(n^{-1/t})$

First, the following is proved (Theorem 3):

$$\mathbb{E}[(U^L - U)^2] \le \mathbb{E}\left[\sum_{i=1}^{\infty} \Phi_i\right] \mathbb{E}^2\left[t^L\right] + \left(\mathbb{E}\left[\sum_{i=1}^{\infty} \Phi_i\right] + \mathbb{E}[U]\right)^2 \xi_L(t)^2$$

where

$$\xi_L(t) = \max_{0 \le s < \infty} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| e^{-s/t}$$

Proof Sketch of Theorem 1 (Poisson)

for t ≥ 1

Smoothing distribution	Parameters	$\mathcal{E}_{n,t}(U^L)$
Poisson (r)	$r = \frac{1}{2t}\log_e \frac{n(t+1)^2}{t-1}$	$O(n^{-1/t})$

For Poisson distribution with parameter r,

$$\mathbb{E}[t^L] = e^{-r}\sum_{\ell=0}^{\infty}\frac{(rt)^\ell}{\ell!} = e^{r(t-1)}$$

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = e^{-r}\underbrace{\sum_{\ell=0}^{\infty}\frac{(-sr)^\ell}{(\ell!)^2}} \le e^{-r} \Longrightarrow \xi_L(t) \le e^{-r}$$

Bessel function, which has values in [-1, 1]

Substituting into Theorem 3 and optimizing over r yields the Theorem 1 bound.

Conclusion

- **Unseen species problem:** Estimates # unseen species in future sample given past data
- **Prior work:**
  - Good-Toulmin estimator works well for $t \leq 1$, but has large variance for $t > 1$
  - Efron-Thisted estimator empirically worked well for $t > 1$, but had no theoretical support
- Smoothed Good-Toulmin estimator proposed
  - Generalizes Efron-Thisted
  - Principled selection of parameters
  - Worst-case MSE performance guarantees
- Bias and variance of linear estimators can be analyzed via their Taylor expansions; used to show why truncated Good-Toulmin has high bias