Three Approaches towards Optimal Property Estimation and Testing

Jiantao Jiao (Stanford EE)

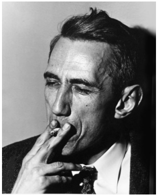
Joint work with: Yanjun Han, Dmitri Pavlichin, Kartik Venkat, Tsachy Weissman

Frontiers in Distribution Testing Workshop, FOCS 2017

Oct. 14th, 2017

Statistical properties

Disclaimer: Throughout this talk, n refers to the number of samples, S refer to the alphabet size of a distribution.



1. Shannon entropy: $H(P) \triangleq \sum_{i=1}^{S} -p_i \ln p_i$.
2. $F_\alpha(P)$: $F_\alpha(P) \triangleq \sum_{i=1}^{S} p_i^\alpha, \alpha > 0$.
3. KL divergence, χ^2 divergence, L_1 distance, Hellinger distance $F(P, Q) \triangleq \sum_{i=1}^{S} f(p_i, q_i)$ for $f(x, y) = x \ln(x/y), (x-y)^2/x, |x-y|, (\sqrt{x} - \sqrt{y})^2$.

Tolerant testing/learning/estimation

We focus on the question: how many samples are needed to achieve accuracy ϵ for estimating these properties from empirical data?

- Example: $L_1(P, U_S)$, $U_S = (1/S, 1/S, \ldots, 1/S)$, observe n i.i.d. samples from P;
- (VV'11, VV'11): exist approach whose error is $\sqrt{\frac{S}{n \ln n}}$ when $\frac{S}{\ln S} \lesssim n \lesssim S$; no consistent estimator when $n \lesssim \frac{S}{\ln S}$;
- The MLE plug-in $L_1(\hat{P}_n, U_S)$ achieves error $\sqrt{\frac{S}{n}}$ when $n \gtrsim S$.

Tolerant testing/learning/estimation

We focus on the question: how many samples are needed to achieve accuracy ϵ for estimating these properties from empirical data?

- Example: $L_1(P, U_S)$, $U_S = (1/S, 1/S, \ldots, 1/S)$, observe n i.i.d. samples from P;
- (VV'11, VV'11): exist approach whose error is $\sqrt{\frac{S}{n \ln n}}$ when $\frac{S}{\ln S} \lesssim n \lesssim S$; no consistent estimator when $n \lesssim \frac{S}{\ln S}$;
- The MLE plug-in $L_1(\hat{P}_n, U_S)$ achieves error $\sqrt{\frac{S}{n}}$ when $n \gtrsim S$.

Effective sample size enlargement

Minimax rate-optimal with n samples $\Longleftrightarrow$ MLE with $n \ln n$ samples

- Similar results also hold for Shannon entropy (VV'11, VV'11, VV'13, WY'16, JVHW'15), power sum functional (JVHW'15), Rényi entropy estimation (AOST'14), χ^2, Hellinger, and KL-divergence estimation (HJW'16, BZLV'16), L_r norm estimation under Gaussian white noise model (HJMW'17), L_1 distance estimation (JHW'16), etc. except for support size (WY'16)

Effective sample size enlargement

$$R_{\text{minmax}}(F, \mathcal{P}, n) = \inf_{\hat{F}(X_1,\ldots,X_n)} \sup_{P\in\mathcal{P}} \mathbb{E}|\hat{F} - F(P)|$$

$$R_{\text{plug-in}}(F, \mathcal{P}, n) = \sup_{P\in\mathcal{P}} \mathbb{E}|F(\hat{P}_n) - F(P)|.$$

$F(P)$	$\mathcal{P}$	$R_{\text{minmax}}(F,\mathcal{P},n)$	$R_{\text{plug-in}}(F,\mathcal{P},n)$
$\sum_{i=1}^{S} p_i \log\left(\frac{1}{p_i}\right)$	$\mathcal{M}_S$	$\frac{S}{n\log(n)} + \frac{\log(S)}{\sqrt{n}}$	$\frac{S}{n} + \frac{\log(S)}{\sqrt{n}}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \quad 0 < \alpha \le \frac{1}{2}$	$\mathcal{M}_S$	$\frac{S}{(n\log(n))^\alpha}$	$\frac{S}{n^\alpha}$
$F_\alpha(P), \quad \frac{1}{2} < \alpha < 1$	$\mathcal{M}_S$	$\frac{S}{(n\log(n))^\alpha} + \frac{S^{1-\alpha}}{\sqrt{n}}$	$\frac{S}{n^\alpha} + \frac{S^{1-\alpha}}{\sqrt{n}}$
$F_\alpha(P), \quad 1 < \alpha < \frac{3}{2}$	$\mathcal{M}_S$	$(n\log(n))^{-(\alpha-1)}$	$n^{-(\alpha-1)}$
$F_\alpha(P), \quad \alpha \ge \frac{3}{2}$	$\mathcal{M}_S$	$\frac{1}{\sqrt{n}}$	$\frac{1}{\sqrt{n}}$
$\sum_{i=1}^{S} \mathbf{1}(p_i \neq 0)$	$\{P : \min_i p_i \ge \frac{1}{S}\}$	$Se^{-\Theta\left(\max\left\{\sqrt{\frac{n\log(n)}{S}}, \frac{n}{S}\right\}\right)}$	$Se^{-\Theta\left(\frac{n}{S}\right)}$
$\sum_{i=1}^{S} \|p_i - q_i\|$	$\mathcal{M}_S$	$\sum_{i=1}^{S} q_i \wedge \sqrt{\frac{q_i}{n\ln n}}$	$\sum_{i=1}^{S} q_i \wedge \sqrt{\frac{q_i}{n}}$

Effective sample size enlargement

Divergence functions: here $P, Q \in \mathcal{M}_S$ where we have m samples from p and n samples from q. For the Kullback-Leibler and χ^2 divergence estimators we only consider $(P, Q) \in \{(P, Q) | P, Q \in \mathcal{M}_S, \frac{P_i}{Q_i} \leq u(S)\}$ where $u(S)$ is some function of S.

$F(P, Q)$	$R_{\text{minmax}}(F, \mathcal{P}, m, n)$	$R_{\text{plug-in}}(F, \mathcal{P}, m, n)$
$\sum_{i=1}^{S} \lvert p_i - q_i \rvert$	$\sqrt{\frac{S}{\min\{m, n\} \log(\min\{m, n\})}} +$	$\sqrt{\frac{S}{\min\{m, n\}}}$
$\frac{1}{2} \sum_{i=1}^{S} (\sqrt{p_i} - \sqrt{q_i})^2$	$\sqrt{\frac{S}{\min\{m, n\} \log(\min\{m, n\})}}$	$\sqrt{\frac{S}{\min\{m, n\}}}$
$D(P \Vert Q) = \sum_{i=1}^{S} p_i \log\left(\frac{p_i}{q_i}\right)$	$\frac{S}{m \log(m)} + \frac{Su(S)}{n \log(n)} + \frac{\log(u(S))}{\sqrt{m}} + \frac{\sqrt{u(S)}}{\sqrt{n}}$	$\frac{S}{m} + \frac{Su(S)}{n} + \frac{\log(u(S))}{\sqrt{m}} + \frac{\sqrt{u(S)}}{\sqrt{n}}$
$\chi^2(P \Vert Q) = \sum_{i=1}^{S} \frac{p_i^2}{q_i} - 1$	$\frac{Su(S)^2}{n \log(n)} + \frac{u(S)}{\sqrt{m}} + \frac{u(S)^{3/2}}{\sqrt{n}}$	$\frac{Su(S)^2}{n} + \frac{u(S)}{\sqrt{m}} + \frac{u(S)^{3/2}}{\sqrt{n}}$

Goal of this talk

Understand the mechanism behind the logarithmic sample size enlargement.

- For what functionals do we have this phenomenon?
- What concrete algorithms achieve this phenomenon?
- If there exist multiple approaches, what are their relative advantages and disadvantages?

First approach: Approximation methodology

Question

Is the enlargement phenomenon caused by the fact that the functionals are permutation invariant (symmetric)?

First approach: Approximation methodology

Question

Is the enlargement phenomenon caused by the fact that the functionals are permutation invariant (symmetric)?

Answer

Nope. :)

Literature on approximation methodology

VV'11 (linear estimator), WY'16, WY'16 JVHW'15, AOST'14, HJW'16, BZLV'16, HJMW'16, JHW'16

Example: L_1 distance estimation

Given $Q = (q_1, q_2, \ldots, q_S)$, we estimate $L_1(P, Q)$ given i.i.d. samples from P.

Theorem (J., Han, Weissman'16)

Suppose $\ln S \lesssim \ln n \lesssim \ln\left(\sum_{i=1}^{S} \sqrt{q_i} \wedge q_i\sqrt{n \ln n}\right)$, $S \geq 2$. *Then,*

$$\inf_{\hat{L}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{L} - L_1(P, Q)| \asymp \sum_{i=1}^{S} q_i \wedge \sqrt{\frac{q_i}{n \ln n}}. \tag{1}$$

For the MLE, we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P |L_1(\hat{P}_n, Q) - L_1(P, Q)| \asymp \sum_{i=1}^{S} q_i \wedge \sqrt{\frac{q_i}{n}}. \tag{2}$$

Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$

0 ———————————————— 1

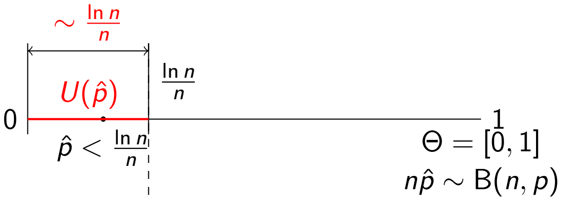
$\Theta = [0, 1]$

$n\hat{p} \sim \mathrm{B}(n, p)$

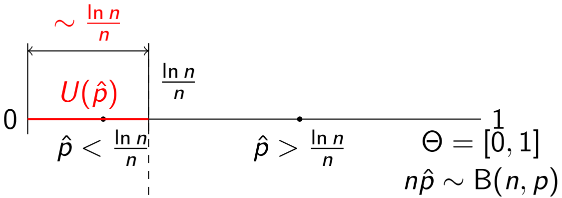
Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$

Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$

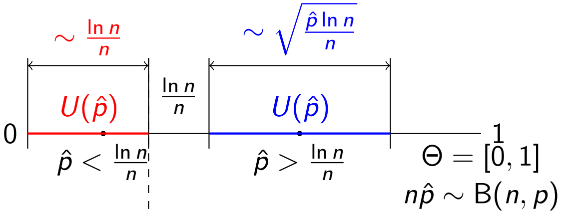
Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$



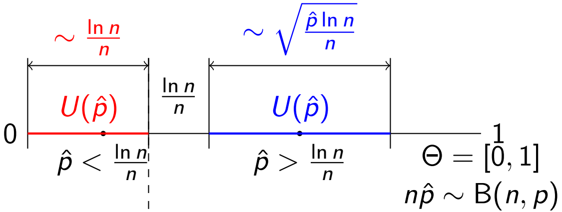
Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$



Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$



Confidence sets in binomial model: coverage probability $\asymp 1 - n^{-A}$



Theorem (J., Han, Weissman'16)

Partition $[0,1]$ *into finitely number of intervals* $I_i = [x_i, x_{i+1}]$, $x_0 = 0$, $x_1 \asymp \frac{\ln n}{n}$, $\sqrt{x_{i+1}} - \sqrt{x_i} \asymp \sqrt{\frac{\ln n}{n}}$. *Then,*

1. *if* $p \in I_i$, *then* $\hat{p} \in 2I_i$ *with probability* $1 - n^{-A}$;
2. *if* $\hat{p} \in I_i$, *then* $p \in 2I_i$ *with probability* $1 - n^{-A}$;
3. *Those intervals are of the shortest length.*

Algorithmic description of Approximation methodology

First conduct sampling splitting, get $\hat{p}_i, \hat{p}'_i$ i.i.d. with distribution $\frac{2}{n} \cdot \mathrm{B}(n/2, p_i)$.

Suppose $q_i \in I_j$. For each i do the following:

1. if $\hat{p}_i \in I_j$, compute best polynomial approximation in $2I_j$:

$$P_K(x; q_i) = \arg \min_{P \in \mathrm{Poly}_K} \max_{z \in 2I_j} ||z - q_i| - P(z)|, \tag{3}$$

   and then estimate $|p_i - q_i|$ by the unbiased estimator of $P_K(p_i; q_i)$ using $\hat{p}'_i$;

2. if $\hat{p}_i \notin I_j$, estimate $|p_i - q_i|$ by $|\hat{p}'_i - q_i|$;
3. sum everything up.

Why it works?

1. Suppose $\hat{p}_i \in I_j$. No matter what we use to estimate, one can always assume that $p_i \in 2I_j$;
2. The bias of the MLE is approximately (Strukov and Timan'77)

$$\sup_{p_i \in 2I_j} \left| |p_i - q_i| - \mathbb{E}|\hat{p}_i - q_i| \right| \asymp q_i \wedge \sqrt{\frac{q_i}{n}}; \qquad (4)$$

3. The bias of the Approximation methodology is approximately (Ditzian and Totik'87)

$$\sup_{p_i \in 2I_j} \left| |p_i - q_i| - P_K(p_i; q_i) \right| \asymp q_i \wedge \sqrt{\frac{q_i}{n \ln n}}. \qquad (5)$$

4. Permutation invariance does not play a role since we are doing symbol by symbol bias correction;
5. The bias dominates in high dimensions (measure concentration phenomenon).

Properties of the Approximation Methodology

1. Applies to essentially any functional
2. Applies to a wide range of statistical models (binomial, Poisson, Gaussian, etc)
3. Near-linear complexity
4. Explicit polynomial approximation for each different functional
5. Need to tune parameters in practice

Second approach: Local moment matching methodology

Motivation

Does there exist a single plug-in estimator that can replace the Approximation methodology?

Second approach: Local moment matching methodology

Motivation

Does there exist a single plug-in estimator that can replace the Approximation methodology?

Answer

No. For any plug-in rule $\hat{P}$, there exists a fixed Q such that $L_1(\hat{P}, Q)$ requires $n \gg S$ samples to consistently estimate $L_1(P, Q)$, while the optimal method requires at most $n \gg \frac{S}{\ln S}$.

Second approach: Local moment matching methodology

Motivation

Does there exist a single plug-in estimator that can replace the Approximation methodology?

Answer

No. For any plug-in rule $\hat{P}$, there exists a fixed Q such that $L_1(\hat{P}, Q)$ requires $n \gg S$ samples to consistently estimate $L_1(P, Q)$, while the optimal method requires at most $n \gg \frac{S}{\ln S}$.

Weakened goal

What about we only consider permutation invariant functionals?

Literature on the local moment matching methodology

VV'11 (linear programming), HJW'17

Local moment matching methodology

Theorem (Han, J., Weissman'17)

There exists a single estimator $\hat{P}$, efficiently computable, and achieves the optimal phase transitions for ALL the permutation invariant functionals mentioned above.
In particular, it solves the minimax problem

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}\|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{S}{n \ln n}} + \left(\tilde{\mathcal{O}}(n^{-1/3}) \wedge \sqrt{\frac{S}{n}} \right), \qquad (6)$$

where $P_< = (p_{(1)}, p_{(2)}, \ldots, p_{(S)})$, $p_{(i)} \leq p_{(i+1)}$.

A simple example

Assume for all i, $p_i \le \frac{\ln n}{n}, \hat{p}_i \le \frac{\ln n}{n}$. Consider the Shannon entropy functional $H(P) = \sum_{i=1}^{S} f(p_i), f(x) = x\ln(1/x)$.

Theorem (VV'11, Wu and Yang'16, J. et al'15)

Optimal error in estimating H is $\frac{S}{n \ln n}$, while MLE error is $\frac{S}{n}$.

A simple example

Assume for all i, $p_i \le \frac{\ln n}{n}, \hat{p}_i \le \frac{\ln n}{n}$. Consider the Shannon entropy functional $H(P) = \sum_{i=1}^{S} f(p_i), f(x) = x \ln(1/x)$.

Theorem (VV'11, Wu and Yang'16, J. et al'15)

Optimal error in estimating H is $\frac{S}{n \ln n}$, while MLE error is $\frac{S}{n}$.

Suppose we use the plug-in rule $\sum_{i=1}^{S} f(q_i)$ to estimate $H(P)$, where $q_i \le \frac{\ln n}{n}$. Then, for any $P_K(x) \in \mathrm{Poly}_K, K = \ln n$,

$$
\begin{aligned}
H - \sum_i f(q_i) &= \sum_i (f(p_i) - P_K(p_i)) + \sum_i (P_K(p_i) - P_K(q_i)) \\
&\quad + \sum_i (P_K(q_i) - f(q_i)) \\
&\le 2S \cdot \inf_{P_K} \max_{x \in [0, \frac{\ln n}{n}]} |f(x) - P_K(x)| + \sum_i (P_K(p_i) - P_K(q_i)) \\
&\lesssim \frac{S}{n \ln n} + \sum_i (P_K(p_i) - P_K(q_i)).
\end{aligned}
$$

Local moment matching

We showed for any plug-in rule Q,

$$H - \sum_i f(q_i) \lesssim \frac{S}{n \ln n} + \sum_i (P_K(p_i) - P_K(q_i)). \tag{7}$$

Why MLE is bad?

The MLE is bad because

$$\left| \mathbb{E} \left[\sum_i (P_K(p_i) - P_K(q_i)) \right] \right| \gtrsim \frac{S}{n}. \tag{8}$$

Solution

It suffices to reduce the bias of $P_K(q_i)$ in estimating $P_K(p_i)$.

Local moment matching

Ideal situation

Suppose for each $0 \le k \le \ln n$,

$$\sum_j p_j^k = \sum_j q_j^k, \tag{9}$$

we immediately have

$$\mathbb{E}\left[\sum_i (P_K(p_i) - P_K(q_i))\right] = 0. \tag{10}$$

Algorithmic description of local moment matching

For each interval I_j, collect $\mathcal{A} = \{i : \hat{p}_i \in I_j\}$. Then, for each $0 \le k \le \ln n$, we solve Q such that

$$\left| \sum_{i \in \mathcal{A}} q_i^k - \left(\text{unbiased estimates of } \sum_{i \in \mathcal{A}} p_i^k \right) \right| \lesssim n^{\epsilon} \cdot \sigma_{k,\mathcal{A}}, \qquad (11)$$

here

$$\sigma_{k,\mathcal{A}} = \text{standard deviation of unbiased estimates of } \sum_{i \in \mathcal{A}} p_i^k. \qquad (12)$$

Existence of solution

The solution exists with overwhelming probability since the true distribution P satisfies these inequalities with overwhelming probability.

Properties of the Local moment matching Methodology

1. Applies only to permutation invariant functionals
2. Applies to a wide range of statistical models (binomial, Poisson, Gaussian, etc)
3. Polynomial complexity
4. Implicit polynomial approximation, just need to compute once
5. Need to tune parameters in practice

Third approach: the profile maximum likelihood methodology (PML)

Properties	Approximation	Local MM	PML
Permutation invariant	No	Yes	Yes
Statistical model	Broad	Broad	(Conjectured) Broad
Complexity	Near-linear	Polynomial	Unclear
Functional dependent	Yes	No	No
Parameter tuning	Yes	Yes	No

Thank you!

Literature

- Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. "A unified maximum likelihood approach for optimal distribution property estimation", Proceedings of ICML, 2017.
- Jiantao Jiao, Yanjun Han, and Tsachy Weissman. "Minimax Estimation of the L_1 Distance", arXiv e-prints, May 2017
- Gregory Valiant and Paul Valiant. "A CLT and tight lower bounds for estimating entropy", Electronic Colloquium on Computational Complexity (ECCC), 2010
- Gregory Valiant and Paul Valiant. "Estimating the unseen: a sublinear-sample canonical estimator of distributions", Electronic Colloquium on Computational Complexity, 2010.
- Gregory Valiant and Paul Valiant, "Estimating the unseen: an $n/\log n$ sample estimator for entropy and support size, shown optimal via new clts", Proceedings of STOC, 2011.
- Gregory Valiant and Paul Valiant, "The power of linear estimators", Proceedings of FOCS, 2011.

Literature

- Yihong Wu and Pengkun Yang. "Minimax rates of entropy estimation on large alphabets via best polynomial approximation." IEEE Transactions on Information Theory 62.6 (2016): 3702-3720.
- Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. "Minimax estimation of functionals of discrete distributions." IEEE Transactions on Information Theory 61.5 (2015): 2835-2885.
- Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, Himanshu Tyagi. "The complexity of estimating Rnyi entropy." Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2014.
- Yanjun Han, Jiantao Jiao, and Tsachy Weissman. "Minimax Rate-Optimal Estimation of Divergences between Discrete Distributions. " arXiv preprint arXiv:1605.09124 (2016).
- Yuheng Bu, Shaofeng Zou, Yingbin Liang, Venugopal V. Veeravalli. "Estimation of KL Divergence: Optimal Minimax Rate." arXiv preprint arXiv:1607.02653 (2016).

Literature

- Yanjun Han, Jiantao Jiao, Rajarshi Mukherjee, and Tsachy Weissman. "On Estimation of L_r-Norms in Gaussian White Noise Models. " arXiv preprint arXiv:1710.03863 (2017).
- Yihong Wu and Pengkun Yang. "Chebyshev polynomials, moment matching, and optimal estimation of the unseen." arXiv preprint arXiv:1504.01227 (2015).
- Yanjun Han, Jiantao Jiao, Tsachy Weissman, "Local moment matching: a unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance", in preparation