Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance

Yanjun Han (Stanford EE)

Joint work with:

Jiantao Jiao	Stanford EE
Tsachy Weissman	Stanford EE

COLT 2018, Stockholm, Sweden

Our Problem

Target

Given n independent samples from $P = (p_1, \cdots, p_k)$, estimate the distribution vector P up to permutation.

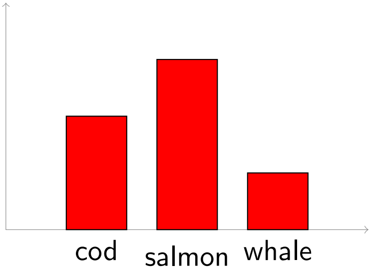
Our Problem

Target

Given n independent samples from $P = (p_1, \cdots, p_k)$, estimate the distribution vector P up to permutation.

Example

Observation for fish species in the ocean: {salmon, cod, whale, salmon, cod, salmon}



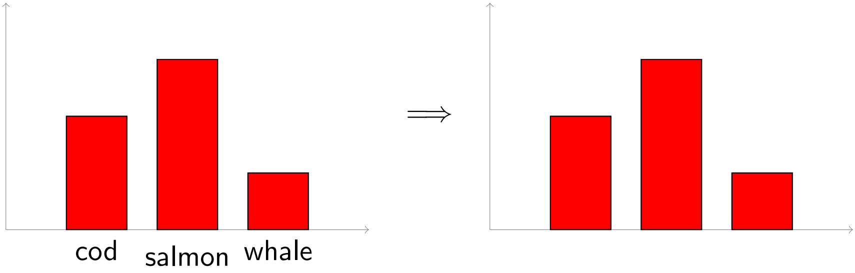
Our Problem

Target

Given n independent samples from $P = (p_1, \cdots, p_k)$, estimate the distribution vector P up to permutation.

Example

Observation for fish species in the ocean: {salmon, cod, whale, salmon, cod, salmon}



Loss Criterion for Our Problem

Let $P_< = (p_{(1)}, p_{(2)}, \cdots, p_{(k)})$ with $p_{(1)} \le p_{(2)} \le \cdots \le p_{(k)}$ be the sorted version of P. We would like to minimize the sorted ℓ_1 loss:

Minimax Sorted ℓ_1 Risk

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P \|\hat{P} - P_<\|_1$$

Motivation

Why distribution up to permutation (sorted distribution)?

Motivation

Why distribution up to permutation (sorted distribution)?

- shape of the distribution: light-tail or heavy-tail, etc

Motivation

Why distribution up to permutation (sorted distribution)?

- shape of the distribution: light-tail or heavy-tail, etc
- significantly easier to learn compared to P itself

Motivation

Why distribution up to permutation (sorted distribution)?

- shape of the distribution: light-tail or heavy-tail, etc
- significantly easier to learn compared to P itself
- provide insights into P learning via a two-step procedure: first learn P up to permutation, and then the labeling

Motivation

Why distribution up to permutation (sorted distribution)?

- shape of the distribution: light-tail or heavy-tail, etc
- significantly easier to learn compared to P itself
- provide insights into P learning via a two-step procedure: first learn P up to permutation, and then the labeling
- general insights of learning parameters up to group transformations (method of the invariant)

Motivation

Why distribution up to permutation (sorted distribution)?

- shape of the distribution: light-tail or heavy-tail, etc
- significantly easier to learn compared to P itself
- provide insights into P learning via a two-step procedure: first learn P up to permutation, and then the labeling
- general insights of learning parameters up to group transformations (method of the invariant)
- symmetric functional of the distribution: can be plugged into general functionals such that $F(P) = F(P_\pi)$

Symmetric Functional Estimation

Estimate functionals of the form $F(P) = \sum_{i=1}^{k} f(p_i)$

- Interesting regime: non-smooth f and $k \gtrsim n$
- Examples: Shannon entropy $\sum_{i=1}^{k} -p_i \log p_i$, power sum $\sum_{i=1}^{k} p_i^{\alpha}$, support size $\sum_{i=1}^{k} \mathbb{1}(p_i \neq 0)$, distance to uniformity $\sum_{i=1}^{k} |p_i - k^{-1}|$, etc

Symmetric Functional Estimation

Estimate functionals of the form $F(P) = \sum_{i=1}^{k} f(p_i)$

- Interesting regime: non-smooth f and $k \gtrsim n$
- Examples: Shannon entropy $\sum_{i=1}^{k} -p_i \log p_i$, power sum $\sum_{i=1}^{k} p_i^{\alpha}$, support size $\sum_{i=1}^{k} \mathbb{1}(p_i \neq 0)$, distance to uniformity $\sum_{i=1}^{k} |p_i - k^{-1}|$, etc

Key Idea

- Approximate f using low-degree polynomials
- n samples in optimal estimator essentially equivalent to $n \log n$ samples in plugging in the empirical distribution $F(\hat{P})$

Symmetric Functional Estimation: Examples

		minimax ℓ_1 risk	plug-in ℓ_1 risk
$\sum_{i=1}^k -p_i \log p_i$ (JVHW'15, WY'16)		$\frac{k}{n \log n} + \frac{\log k}{\sqrt{n}}$	$\frac{k}{n} + \frac{\log k}{\sqrt{n}}$
$\sum_{i=1}^k p_i^\alpha$ (JVHW'15)	$0 < \alpha \leq \frac{1}{2}$	$\frac{k}{(n \log n)^\alpha}$	$\frac{k}{n^\alpha}$
	$\frac{1}{2} < \alpha < 1$	$\frac{k}{(n \log n)^\alpha} + \frac{k^{1-\alpha}}{\sqrt{n}}$	$\frac{k}{n^\alpha} + \frac{k^{1-\alpha}}{\sqrt{n}}$
	$1 < \alpha < \frac{3}{2}$	$(n \log n)^{-(\alpha-1)}$	$n^{-(\alpha-1)}$
$\sum_{i=1}^k \mathbb{1}(p_i \neq 0)$ (WY'16)		$k \exp\left(-\sqrt{\frac{n \log n}{k}} - \frac{n}{k}\right)$	$k \exp\left(-\frac{n}{k}\right)$
$\sum_{i=1}^k \lvert p_i - k^{-1} \rvert$ (JHW'17)		$\sqrt{\frac{k}{n \log n}}$	$\sqrt{\frac{k}{n}}$

Symmetric Functional Estimation: Examples (Cont'd)

	minimax ℓ_1 risk	plug-in ℓ_1 risk
$\sum_{i=1}^k \lvert p_i - q_i\rvert$ (JHW'16)	$\sqrt{\frac{k}{m\log m}} + \sqrt{\frac{k}{n\log n}}$	$\sqrt{\frac{k}{m}} + \sqrt{\frac{k}{n}}$
$\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2$ (HJW'16)	$\sqrt{\frac{k}{m\log m}} + \sqrt{\frac{k}{n\log n}}$	$\sqrt{\frac{k}{m}} + \sqrt{\frac{k}{n}}$
$\sum_{i=1}^k p_i \log \frac{p_i}{q_i}$, $\max_i \frac{p_i}{q_i} \le g(k)$ (BZLV'16, HJW'16)	$\frac{k}{m\log m} + \frac{kg(k)}{n\log n} + \frac{\sqrt{k}}{\sqrt{m}} + \frac{\sqrt{g(k)}}{\sqrt{n}}$	$\frac{k}{m} + \frac{kg(k)}{n} + \frac{\sqrt{k}}{\sqrt{m}} + \frac{\sqrt{g(k)}}{\sqrt{n}}$
$\sum_{i=1}^k \frac{(p_i-q_i)^2}{q_i}$, $\max_i \frac{p_i}{q_i} \le g(k)$ (HJW'16)	$\frac{kg(k)^2}{n\log n} + \frac{g(k)}{\sqrt{m}} + \frac{g(k)^{3/2}}{\sqrt{n}}$	$\frac{kg(k)^2}{n} + \frac{g(k)}{\sqrt{m}} + \frac{g(k)^{3/2}}{\sqrt{n}}$
$\lVert f\rVert_r$, r non-even (HJMW'17)	$(n\log n)^{-\frac{s}{2s+d}}$	$n^{-\frac{s}{2s+d}}$
$\int -f(x)\log f(x)dx$ (HJWW'17)	$(n\log n)^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$	$n^{-\frac{s}{s+d}} + n^{-\frac{1}{2}}$

Is Plug-in Really Bad?

Is there a single estimator $\hat{P}$, such that:

- plugging-in the estimator into a large variety of symmetric functionals achieves the information theoretic limit;
- has clear correspondence with the approximation approach;
- achieves the minimax rate in estimating sorted distribution;
- is efficiently computable.

Is Plug-in Really Bad?

Is there a single estimator $\hat{P}$, such that:

- plugging-in the estimator into a large variety of symmetric functionals achieves the information theoretic limit;
- has clear correspondence with the approximation approach;
- achieves the minimax rate in estimating sorted distribution;
- is efficiently computable.

Past work:

- Linear Programming based approach (VV'11): optimal only for a small range of n
- Profile maximum likelihood approach (ADOS'17): hard to compute, cannot generalize to other models (e.g., Gaussian)

Main Result I: Distribution Estimation

Theorem

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{k}{n \ln n}} + \tilde{\mathcal{O}}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{k}{n}}\right).$$

Main Result I: Distribution Estimation

Theorem

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{k}{n \ln n}} + \tilde{\mathcal{O}}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{k}{n}} \right).$$

Corollary

Unlabeled distribution learning is possible if and only if $n \gg \frac{k}{\ln k}$.

Main Result I: Distribution Estimation

Theorem

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{k}{n \ln n}} + \tilde{\mathcal{O}}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{k}{n}}\right).$$

Corollary

Unlabeled distribution learning is possible if and only if $n \gg \frac{k}{\ln k}$.

Theorem

The sorted empirical distribution requires $n \gg k$ to learn the true sorted distribution.

Main Result I: Distribution Estimation

Theorem

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_k} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{k}{n \ln n}} + \tilde{\mathcal{O}}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{k}{n}}\right).$$

Corollary

Unlabeled distribution learning is possible if and only if $n \gg \frac{k}{\ln k}$.

Theorem

The sorted empirical distribution requires $n \gg k$ to learn the true sorted distribution.

Corollary

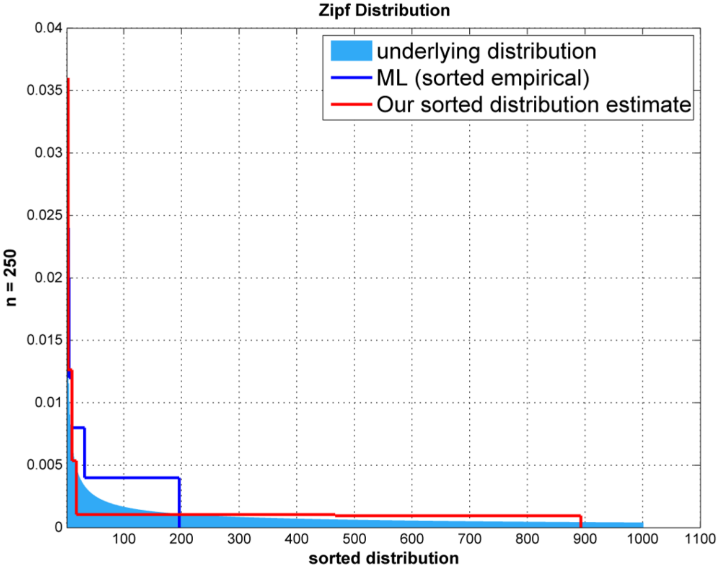
Sorted empirical distribution is improvable iff $k = \tilde{\Omega}(n^{\frac{1}{3}})$.

Main Result II: Symmetric Functional Estimation

Theorem

Plugging in the previous estimator $\hat{P}$ achieves the optimal phase transitions for ALL the permutation invariant 1D functionals mentioned before.

Ideas of LMM


Zipf Distribution
underlying distribution
ML (sorted empirical)
Our sorted distribution estimate
n = 250
sorted distribution
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
100
200
300
400
500
600
700
800
900
1000
1100

Properties of LMM

- Applies to a wide range of permutation invariant functionals
- Applies to a wide range of statistical models (Binomial, Poisson, Gaussian, etc)
- Polynomial complexity
- Agnostic to the support size k
- Implicit polynomial approximation

Properties of LMM

- Applies to a wide range of permutation invariant functionals
- Applies to a wide range of statistical models (Binomial, Poisson, Gaussian, etc)
- Polynomial complexity
- Agnostic to the support size k
- Implicit polynomial approximation

Thank you!