DPPs and GPs

August 6, 2019

1 8/5/2019

1.1 Distributed ridge via reweighting

Dobriban and Sheng (2019) give a distributed model where the ridge estimator is first taken on subsets of data (e.g. in a distributed setting)

$$\hat{\boldsymbol{\beta}}_S \coloneqq \boldsymbol{X}_S^\dagger \boldsymbol{y}_S$$

and the local estimates are averaged

$$\hat{\boldsymbol{\beta}} \coloneqq \sum_{S_i} w_i \hat{\boldsymbol{\beta}}_{S_i}$$

We propose an estimator consisting of averaged local estimates. Our method is unique in that we construct a globally regularized estimator from unregularized local estimators. This is achieved through exploiting the implicit regularization present in DPPs via importance reweighting.

Proposition 1. *Let $S \subset [n]$ and define the local estimate $\hat{\boldsymbol{\beta}}_S \coloneqq \boldsymbol{X}_S^\dagger \boldsymbol{y}_S$.*
For $|S| = k \geq d$, if we use weights

$$w_S = \frac{P(S)}{Q(S)}$$

where

$$P(S) = \frac{\det \lambda^{-1} \boldsymbol{X}_S^\top \boldsymbol{X}_S}{\sum_{|S|=k} \det \boldsymbol{X}_S^\top \boldsymbol{X}_S} = \frac{\det \lambda^{-1} \boldsymbol{X}_S^\top \boldsymbol{X}_S}{\binom{n-d}{k-d} \det \boldsymbol{X}^\top \boldsymbol{X}}$$

and $Q(S)$ is the probability that the subset $S \subset [n]$ is present on any given machine (assumed IID), then

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbb{E}\frac{1}{|I|} \sum_i w_{S_i} \hat{\boldsymbol{\beta}}_S = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{1}$$

where I is the index set over all machines.

Remark 1. This shows that the averaged estimator $\hat{\boldsymbol{\beta}}$ provides an unbiased estimator to the full ridge solution. An interesting property is that we are averaging ordinary least squares estimators; the regularization is implicit and obtained by the weights w_S.

Remark 2. By weighting estimates obtained on "larger" subsets $\boldsymbol{X}_S$, we give more importance to estimators using "larger" amounts of data.

Interpret as a form of variance reduction.

Remark 3. The proposition implies a trivial map-reduce algorithm which (assuming the data is already partitioned) can produce $\hat{\beta}$ in $O(k^3 + |I|D)$ time.

Proof.

$$\begin{aligned}\mathbb{E}\hat{\boldsymbol{\beta}} &= \frac{1}{|I|}\sum_i \mathbb{E} w_{S_i}\hat{\boldsymbol{\beta}}_{S_i} \\ &= \sum_{|S|=k} Q(S)\frac{P(S)}{Q(S)}\hat{\boldsymbol{\beta}}_S \\ &= \sum_{|S|=k} P(S)\hat{\boldsymbol{\beta}}_S\end{aligned}$$

Result follows by expectation lemma for regularized volume sampling, ■

1.2 Follow up from Michal's notes

Proof. (Proof of Lemma 1 from determinantal random projections) By the definition of Fredholm determinant (analogously, by L-ensemble normalizing constant result of Machi)

$$\det\left(\boldsymbol{I} + \boldsymbol{\Sigma}_\mu\right) = \sum_{k=0}^{d}\sum_{|S|=k}\det(\boldsymbol{\Sigma}_\mu)_S$$

where S ranges over all increasing subsequences of $[n]$ of length k. Hence, it suffices to show $\sum_{|S|=k}\det(\boldsymbol{\Sigma}_\mu)_S = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mu^k}[\det \boldsymbol{X}\boldsymbol{Y}^\top]$. Unpacking the summand, we see that

$$\sum_{|S|=k}\det(\boldsymbol{\Sigma}_\mu)_S = \sum_{|S|=k}\det\mathbb{E}[\boldsymbol{P}_S\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{P}_S^\top]$$

Here's where I get stuck

We would like to simplify the sum over S using Cauchy-Binet. However, this is only possible when the outer dimension is fixed and S is on the inside, i.e.

$$\sum_{|S|=k}\det \boldsymbol{X}\boldsymbol{P}_S\boldsymbol{P}_S^\top\boldsymbol{Y} = \binom{d-k}{k-k}\det \boldsymbol{X}\boldsymbol{Y}^\top$$

(when we take $\boldsymbol{X}\in\mathbb{R}^{k\times d}$). ■

Theorem 1 (Counterexample to Lemma 2 in determinantal random projections)**.** *Let μ be the uniform measure over a discrete set $[n]$ indexing the rows of $\boldsymbol{X}\in\mathbb{R}^{n\times d}$. Then if $S\sim DPP(L = \boldsymbol{X}\boldsymbol{X}^\top)$, we have*

$$\mathbb{E}[\boldsymbol{f}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S^\top)^{-1}\boldsymbol{g}_S] = \boldsymbol{f}^\top(\boldsymbol{I} + \boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{g} \tag{2}$$

Proof. First notice that this L-ensemble is the same as the setting in Lemma 2, since for ordered sequences $S\subset[n]$ of size k we have

$$P(S)\propto\frac{1}{k!}\det\boldsymbol{X}_S\boldsymbol{X}_S^\top \propto \frac{1}{k!}\mu(S)\det\boldsymbol{X}_S\boldsymbol{X}_S^\top = \frac{1}{k!}\mathbb{E}_T\left[\mathbb{1}_{T=S}\det\boldsymbol{X}_T\boldsymbol{X}_T^\top\right] = \frac{1}{k!}\mathbb{E}_{\boldsymbol{Z}\sim\mu^k}\left[\mathbb{1}_{\{z_i\}_{i=1}^k=S}\det\boldsymbol{Z}\boldsymbol{Z}^\top\right] \tag{3}$$

Next, note that the resulting conclusion differs:

$$\det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top)\mathbb{E}\boldsymbol{f}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S)^{-1}\boldsymbol{g}_S = \sum_{\substack{S\subset[n]\\0\le|S|\le d}} \det(\boldsymbol{X}_S\boldsymbol{X}_S^\top)\boldsymbol{f}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S)^{-1}\boldsymbol{g}_S \tag{4}$$

$$= \sum_{\substack{S\subset[n]\\0\le|S|\le d}} \boldsymbol{f}_S^\top \operatorname{adj}(\boldsymbol{X}_S\boldsymbol{X}_S)^{-1}\boldsymbol{g}_S \tag{5}$$

$$= \sum_{\substack{S\subset[n]\\0\le|S|\le d}} \det\left(\boldsymbol{X}_S\boldsymbol{X}_S^\top + \boldsymbol{f}_S\boldsymbol{g}_S^\top\right) - \det \boldsymbol{X}_S\boldsymbol{X}_S^\top \tag{6}$$

$$= \left(\sum_{\substack{S\subset[n]\\0\le|S|\le d}} \det\left(\left[\boldsymbol{X},\boldsymbol{f}\right]_S\left[\boldsymbol{X},\boldsymbol{g}\right]_S^\top\right)\right) - \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top) \tag{7}$$

$$= \det\left(\left[\boldsymbol{X},\boldsymbol{f}\right]\left[\boldsymbol{X},\boldsymbol{g}\right]^\top\right) - \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top) \tag{8}$$

$$= \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top+\boldsymbol{f}\boldsymbol{g}^\top) - \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top) \tag{9}$$

$$= \boldsymbol{f}^\top \operatorname{adj}(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{g} + \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top) - \det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top) \tag{10}$$

$$= \boldsymbol{f}^\top \operatorname{adj}(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{g} \tag{11}$$

Divide both sides by $\det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{X}^\top)$ to get the desired result. ■

Proposition 2 (Sylvester's for operators)**.** *For trace class operators $A : H_1 \to H_2$ and $B : H_2 \to H_1$ on Hilbert spaces*

$$\det(\boldsymbol{I}+\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{I}+\boldsymbol{B}\boldsymbol{A}) \tag{12}$$

Proof.

$$\det(\boldsymbol{I}+\boldsymbol{A}\boldsymbol{B}) = \sum_{k=0}^{\infty} \operatorname{Tr}\wedge^k(\boldsymbol{A}\boldsymbol{B}) \tag{13}$$

$$= \sum_{k=0}^{\infty} \operatorname{Tr}\wedge^k(\boldsymbol{A})\wedge^k(\boldsymbol{B}) \tag{14}$$

$$= \sum_{k=0}^{\infty} \operatorname{Tr}\wedge^k(\boldsymbol{B})\wedge^k(\boldsymbol{A}) \tag{15}$$

$$= \det(\boldsymbol{I}+\boldsymbol{B}\boldsymbol{A}) \tag{16}$$

where trace class ensures that the sums manipulated are finite. ■

1.3 Connections to Jackknife

Wu et al. (1986) extend the analogy of the LSE being the mean of the volume sampling distribution

$$\hat{\boldsymbol{\beta}} = \sum_{|S|=k} \det(\boldsymbol{X}_S^\top \boldsymbol{X}_S)\hat{\boldsymbol{\beta}}_S = \mathbb{E}\hat{\boldsymbol{\beta}}_S \tag{17}$$

to second-order variance estimators and show (for $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2)$)

$$\mathbb{E}_{\boldsymbol{\xi}} \sum_{|S|=k} \det(\boldsymbol{X}_S^\top \boldsymbol{X}_S)(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}^\top) = \binom{n-d}{k-d+1} \det(\boldsymbol{X}^\top \boldsymbol{X})(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{18}$$

In other words, the variance estimator is unbiased for homoscedastic noise.

Bibliography

Dobriban, E. and Y. Sheng
2019. One-shot distributed ridge regression in high dimensions. *arXiv preprint arXiv:1903.09321*.

Wu, C.-F. J. et al.
1986. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295.