Least Squares Approximation for a Distributed System

Zhu et.al

August 14, 2020

Introduction

The common wisdom for addressing a distributed statistical problem can be classified into two categories.

- 'one-shot' or 'embarrassingly parallel' approach, which requires only one round of communication. It might not achieve the best efficiency in statistical estimation.
- iterative algorithms, which requires multiple iterations to be taken so that the estimation efficiency can be refined to match the global estimator.

Introduction

- The sparse learning problem using ℓ_1 shrinkage estimation.
- Ensure the model selection consistency and establish a criterion for consistent tuning parameter selection.
- The data possessed by different workers are allowed to be heterogeneous but share the same regression relationship.

Models and Notations

- Total N observations, which are indexed as $i = 1, 2, \ldots, N$, define $\mathcal{S} = \{1, 2, \ldots, N\}$
- The ith observation is denoted by $Z_i = (X_i^T, Y_i)^T \in \mathbb{R}^{p+1}$.
- The observations are distributed across K local workers, $\mathcal{S}_k$ collects observations distributed to kth worker and $\mathcal{S} = \cup_{k=1}^{K} \mathcal{S}_k$.
- Define $n = N/K$. Assume that $|\mathcal{S}_k| = n_k$ and that all n_k diverge in the same order $O(n)$.
- Due to the data storing strategy, the data in different workers could be quite heterogeneous, e.g., they might be collected according to spatial regions.
- Despite the heterogeneity here, we assume they share the same regression relationship, the parameter $\theta_0 \in \mathbb{R}^p$.

Models and Notations

Let $\mathcal{L}(\theta; Z)$ be a plausible twice differentiable loss function. Define the global loss function as $\mathcal{L}(\theta) = N^{-1} \sum_{i=1}^{N} \mathcal{L}(\theta, Z_i)$.

The global estimator is $\hat{\theta} = \arg\min \mathcal{L}(\theta)$ and the true value is θ_0.

It is assumed that $\hat{\theta}$ admits the following asymptotic rule

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma \in \mathbb{R}^{p \times p}$ is positive definite.

Models and Notations

Define the local loss function in the k th worker as

$$\mathcal{L}_k(\theta) = n_k^{-1} \sum_{i \in \mathcal{S}_k} \mathcal{L}(\theta; Z_i)$$

The local minimizer is

$$\hat{\theta}_k = \arg\min \mathcal{L}_k(\theta)$$

We assume that

$$\sqrt{n_k}(\hat{\theta}_k - \theta_0) \xrightarrow{d} N(0, \Sigma_k)$$

Least Squares Approximation

Approximate the global loss function using Taylor's expansion

$$
\begin{aligned}
\mathcal{L}(\theta) &= N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} \mathcal{L}(\theta; Z_i) \\
&= N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} \left\{ \mathcal{L}(\theta; Z_i) - \mathcal{L}\left(\widehat{\theta}_k; Z_i\right) \right\} + C_1 \\
&\approx N^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} \left(\theta - \widehat{\theta}_k\right)^\top \ddot{\mathcal{L}}\left(\widehat{\theta}_k; Z_i\right) \left(\theta - \widehat{\theta}_k\right) + C_2
\end{aligned}
$$

The last equation uses the fact that $\dot{\mathcal{L}}_k(\hat{\theta}_k) = 0$.

Least Squares Approximation

The weighted least squares objective function

$$
\begin{aligned}
\widetilde{\mathcal{L}}(\theta) &= N^{-1} \sum_k \left(\theta - \widehat{\theta}_k\right)^\top \left\{ \sum_{i \in \mathcal{S}_k} \ddot{\mathcal{L}}\left(\widehat{\theta}_k ; Z_i\right) \right\} \left(\theta - \widehat{\theta}_k\right) \\
&\stackrel{\text{def}}{=} \sum_k \left(\theta - \widehat{\theta}_k\right)^\top \alpha_k \widehat{\Sigma}_k^{-1} \left(\theta - \widehat{\theta}_k\right)
\end{aligned}
$$

where $\alpha_k = n_k / N$. The solution is (weighted least squares estimator(WLSE))

$$
\widetilde{\theta} = \arg\min_\theta \widetilde{\mathcal{L}}(\theta) = \left(\sum_k \alpha_k \widehat{\Sigma}_k^{-1} \right)^{-1} \left(\sum_k \alpha_k \widehat{\Sigma}_k^{-1} \widehat{\theta}_k \right).
$$

Remarks about WLSE

- The local worker sends $\hat{\theta}_k$ and $\hat{\Sigma}_k$ to the master node
- Then the master node produces WLSE by the above equation.
- The above WLSE requires only one round of communication.

Assumptions

(C1) The parameter space Θ is a compact and convex subset of $\mathbb{R}^p$. θ_0 lies in the interior of Θ.

(C2) Covariates $X_i (i \in \mathcal{S}_k)$ from kth worker are iid from $F_k(x)$.

(C3) For any $\delta > 0$, there exists $\varepsilon > 0$ such that

$$\lim_{n\to\infty} \inf P\left\{\inf_{\|\theta^* - \theta_0\| \geq \delta, 1\leq k \leq K} \left(\mathcal{L}_k\left(\theta^*\right) - \mathcal{L}_k\left(\theta_0\right)\right) \geq \epsilon\right\} = 1$$
$$\text{and } E\left\{ \left.\frac{\partial \mathcal{L}_k(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \right\} = 0$$

(C4) Define

$$\Omega_k(\theta) = E\left\{ \frac{\partial \mathcal{L}\left(\theta; Z_i\right)}{\partial \theta} \frac{\partial \mathcal{L}\left(\theta; Z_i\right)}{\partial \theta^\top} \mid i \in \mathcal{S}_k \right\}$$

Assume $\Omega_k(\theta)$ is nonsingular at θ_0. Let $\Sigma_k = \{\Omega_k(\theta_0)\}^{-1}$ and $\Sigma = \{\Sigma_k \alpha_k \Omega(\theta_0)\}^{-1}$.

Assumptions

(C5) Define $B(\delta) = \{\theta^* \in \Theta | ||\theta^* - \theta_0|| \leq \delta\}$. There exists function $M_{ijl}(Z)$ and $\delta > 0$ such that

$$\left| \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_l} \mathcal{L}\left(\theta^*; Z\right) \right| \leq M_{ijl}(Z), \quad \text{for all} \quad \theta^* \in B(\delta)$$

where
$E\left\{M_{ijl}\left(Z_m\right) \mid m \in \mathcal{S}_k\right\} < \infty$ for all $1 \leq i, j, l \leq p$ and $1 \leq k \leq K$.

Propsition 1

Assume Conditions (C1)-(C5). Then, we have

$$\sqrt{N}\left(\widetilde{\theta}-\theta_0\right)=V\left(\theta_0\right)+B\left(\theta_0\right)$$

with $\operatorname{cov}\left\{V\left(\theta_0\right)\right\}=\Sigma$ and $B\left(\theta_0\right)=O_p(K/\sqrt{N})$, where $\Sigma=\left(\sum_{k=1}^{K}\alpha_k\Sigma_k^{-1}\right)^{-1}$.

Theorem 1

Assume Conditions (C1)-(C5) and further assume $n/N^{1/2} \to \infty$. Then we have $\sqrt{N}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma)$, which achieves the same asymptotic normality as the global estimator $\hat{\theta}$.

Distributed Adaptive Lasso Estimation

How to conduct variable selection on a distributed system has not been sufficiently investigated.

Previous work: Lee et al., 2015; Battey et al., 2015; Wang et al., 2017a; Jordan et al., 2018

Notations:

- The first d_0 to be nonzero. i.e. $\theta_j \neq 0$ for $1 \leq j \leq d_0$. Denote $\mathcal{M}_T = \{1, 2, \ldots, d_0\}$ to be the true model.
- Let $\mathcal{M} = \{i_1, \ldots, i_d\}$ be an arbitrary candidate model.
- For an arbitrary vector v, define $v^{(\mathcal{M})} = (v_i : i \in \mathcal{M})^\top \in \mathbb{R}^{|\mathcal{M}|}$ and $v^{(-\mathcal{M})} = (v_i : i \notin \mathcal{M})^\top \in \mathbb{R}^{p-|\mathcal{M}|}$.
- For an arbitrary Matrix M, define $M^{(\mathcal{M})} = (m_{j_1 j_2} : j_1, j_2 \in \mathcal{M}) \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$.

Adaptive Lasso

Consider the adaptive Lasso objective function on the master

$$Q_\lambda(\theta) = \widetilde{\mathcal{L}}(\theta) + \sum_j \lambda_j \left|\theta_j\right|$$

Define $\tilde{\theta}_\lambda = \arg\min Q_\lambda(\theta)$.

Theorem 2

Assume the conditions (C1)-(C5). Let $a_\lambda = \max\{\lambda_j, j \leq d_0\}$ and $b_\lambda = \min\{\lambda_j, j > d_0\}$. Then the following results holds.

- If $\sqrt{N} a_\lambda \xrightarrow{p} 0$, then $\tilde{\theta}_\lambda - \theta = O_p(N^{-1/2})$.
- If $\sqrt{N} a_\lambda \xrightarrow{p} 0$ and $\sqrt{N} b_\lambda \xrightarrow{p} \infty$,

$$P\left(\widetilde{\theta}_\lambda^{(-\mathcal{M}_T)} = 0\right) \to 1.$$

Covariance Assumption

(C6) Define the global unpenalized estimator as $\hat{\theta}_M = \arg\min_{\left\{\theta \in \mathbb{R}^p : \theta_j = 0, \forall j \notin \mathcal{M}\right\}} \mathcal{L}(\theta)$. Assume for the global estimator $\hat{\theta}_M$ with $\mathcal{M} \supset \mathcal{M}_T$ that $\sqrt{N}\left(\widehat{\theta}_{\mathcal{M}}^{(\mathcal{M})} - \theta_{\mathcal{M}}^{(\mathcal{M})}\right) \to_d N\left(0, \Sigma_{\mathcal{M}}\right) = N\left(0, \Omega_{\mathcal{M}}^{-1}\right)$. Further assume for any $\mathcal{M} \supset \mathcal{M}_T$ that $\Omega_{\mathcal{M}} = \Omega_{\mathcal{M}_F}^{(\mathcal{M})}$, where $\mathcal{M}_F = \{1, 2, \ldots, p\}$ denotes the whole set.

Condition (C6) does not seem very intuitive. Nevertheless, it is a condition that is well satisifed by most maximum likelihood. estimators

Theorem 3

Assume Conditions (C1)-(C6). Let $\sqrt{N} a_\lambda \xrightarrow{p} 0$ and $\sqrt{N} b_\lambda \xrightarrow{p} \infty$, then it holds that

$$\sqrt{N}\left(\widetilde{\theta}_\lambda^{(\mathcal{M}_T)} - \theta^{(\mathcal{M}_T)}\right) \rightarrow_d N\left(0, \Sigma_{\mathcal{M}_T}\right).$$

Remarks about Theorem 3

- as long as the tuning parameters are approximately selected, the resulting estimator is selection consistent and as efficient as the oracle estimator.
- Specify $\lambda_j = \lambda_0 |\tilde{\theta}_j|^{-1}$.
- Since $\tilde{\theta}_j$ is $\sqrt{N}$ -consistent, then as long as as λ_0 satisfies the condition $\lambda_0 \sqrt{N} \to 0$ and $\lambda_0 N \to \infty$, then the conditions in Theorem 2 and Theorem 3 hold.

Distributed Bayes Information Criterion

distributed Bayesian information criterion (DBIC)-based criterion

$$\mathrm{DBIC}_\lambda = \left(\widetilde{\theta}_\lambda - \widetilde{\theta}\right)^\top \widehat{\Sigma}^{-1} \left(\widetilde{\theta}_\lambda - \widetilde{\theta}\right) + \log N \times df_\lambda / N$$

where df_λ is the number of nonzero elements in $\tilde{\theta}_\lambda$.
Define the set of nonzero elements of $\hat{\theta}_\lambda$ by $\mathcal{M}_\lambda$. Define

$$\mathbb{R}_- = \{\lambda \in \mathbb{R}^p : \mathcal{M}_\lambda \not\supset \mathcal{M}_T\}, \mathbb{R}_0 = \{\lambda \in \mathbb{R}^p : \mathcal{M}_\lambda = \mathcal{M}_T\}$$
$$\mathbb{R}_+ = \{\lambda \in \mathbb{R}^p : \mathcal{M}_\lambda \supset \mathcal{M}_T, \mathcal{M}_\lambda \neq \mathcal{M}_T\}$$

where $\mathbb{R}_-$ denotes the under fitted model, and $\mathbb{R}_+$ denotes an over fitted model.

Theorem 4

Assume Conditions (C1)-(C6). Define a reference tuning parameter sequence $\{\lambda_N \in \mathbb{R}^p\}$, where the first d_0 elements of λ_N are $1/N$ and and the remaining elements are $\log N/N$. Then we have

$$P\left(\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} DBIC_\lambda > DBIC_{\lambda_N}\right) \to 1.$$

Simulation Models and Setting

For each model, we consider two typical settings to verify the numerical performance of the proposed method.

- The first strategy is to distribute data in a complete random manner. X_{ij} are sampled from the standard normal distribution N(0,1).
- The second strategy allows for covariate distribution on different workers to be heterogeneous. On the kth worker, the covariates are sampled from the multivariate normal distribution $N(\mu_k, \Sigma_k)$, where $\mu_k \sim U[-1, 1]$ and $\Sigma_k = (\rho_k^{|j_1 - j_2|})$ with $\rho_k \sim U[0.3, 0.4]$.

Simulation Models and Setting

Examples:

- Linear Regression $\theta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$
- Logistic Regression $\theta_0 = (3, 0, 0, 1.5, 0, 0, 2, 0)$
- Possion Regression $\theta_0 = (0.8, 0, 0, 1, 0, 0, -0.4, 0, 0)$
- Cox Model. We set the hazard function to be $h(t_i|x_i) = \exp(X_i^T\theta_0)$,where t_i is the survival time from the ith subject. $\theta_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0, 0)$. Censoring time is generated independently from an exponential distribution with a mean $\mu_i \exp(X_i^T\theta_0)$ where u_i sampled from a uniform distribution $U[1, 3]$.
- Ordered Probit Regression. The ordinal responses are independently generated as follows:

$$P\left(Y_i = l \mid X_i, \theta_0\right) = \left(\begin{array}{ll} \Phi\left(c_1 - X^\top\theta_0\right) & l = 1 \\ \Phi\left(c_l - X^\top\theta_0\right) - \Phi\left(c_{l-1} - X^\top\theta_0\right) & 2 \le l \le L - \\ 1 - \Phi\left(c_{L-1} - X^\top\theta_0\right) & l = L \end{array}\right.$$

where $\theta_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0, 0)$

Simulation Results : I

Page 42 -Page 46

Airline Data

- The dataset considered here is the U.S. Airline Dataset. It contains detailed flight information about U.S. airlines from 1987 to 2008.
- The task is to predict the delayed status of a fight given all other flight information.

The results are in Page 27.