Discussion of sampling approach in big data

Big data discussion group at MSCS of UIC

Outline

Mainly based on
Ping Ma, Michael Mahoney, Bin Yu (2015), *A statistical perspective on algorithmic leveraging*, Journal of Machine Learning Research, 16, 861-911

Sampling in big data analysis

- One popular approach
- Choose a small portion of full data
- One possible way: uniform random sampling
- “Worst-case” may perform poorly

Leveraging approach

- Data-dependent sampling process
- Least-square regression (Avron et al. 2010, Meng et al. 2014)
- Least absolute deviation and quantile regression (Clarkson et al. 2013, Yang et al. 2013)
- Low-rank matrix approximation (Mahoney and Drineas, 2009)
- Leveraging provides uniformly superior worst-case algorithmic result
- No work addresses the statistical aspects

Summary of the results

- Based on linear model
- Analytic framework for evaluating sampling approaches
- Use Taylor expansion to approximate the subsampling estimator

Uniform approach vs leveraging approach

- Compare the biases and variance, both conditional and not unconditional
- Both are unbiased to leading order
- Leveraging approach improve the “size-scale” of the variance but may inflate the variance with small leverage scores
- Neither leveraging nor uniform approach dominates each other

New approaches

- Shrinkage Leveraging Estimator (SLEV): a convex combination of leveraging sampling probability and uniform probability
- Unweighted leveraging Estimator (LEVUNW): leveraging sampling approach with unweighted LS estimation
- Both approaches have some improvements

Outline

Linear Model

$$y = X\beta_0 + \epsilon$$

- X is $n \times p$ matrix
- β_0 is $p \times p$
- $\epsilon \sim N(0, \sigma^2)$
- Least-squared estimator: $\hat{\beta}_{ols} = (X^T X)^{-1} X^T y$

About $\hat{\beta}_{ols}$

- Computation time $O(np^2)$
- Can be written as $V\Delta^{-1}U^Ty$, where $X = U\Delta V^T$ (thin SVD)
- Can be solved approximately with computation time $o(np^2)$ with error bounded by ϵ

University of Illinois at Chicago

Leverage

- Consider $\hat{y} = Hy$, where $H = X(X^TX)^{-1}X^T$
- The ith diagonal element, $h_{ii} = x_i^T(X^TX)^{-1}x_i$, called the statistical leverage of the ith observation.
- $Var(e_i) = (1 - h_{ii})\sigma^2$
- Student residual: $\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$
- h_{ii} has been used to qualify for the influential observations

Leverage

- $h_{ii} = \sum_{j=1}^{p} U_{ij}^2$
- Exact computation time: $O(np^2)$
- Approximate computation time: $o(np^2)$

Sampling algorithm

- $\{\pi_i\}_{i=1}^n$ is a sampling distribution
- Randomly sample $r > p$ rows of X and the corresponding elements of y, using $\{\pi_i\}_{i=1}^n$
- Rescale each sampled row/element by $\frac{1}{(r\sqrt{\pi_i})}$ to form a weighted LS subproblem
- Solve the weighted LS subproblem, the solution denoted as $\tilde{\beta}_{wls}$

Weighted LS subproblem

- Let S_X^T $(r \times n)$ be the sampling matrix indicating the selected samples
- Let D $(r \times r)$ be the diagonal matrix with the ith element being $\frac{1}{\sqrt{r\pi_k}}$ if the kth data is chosen
- The weighted LS estimator is

$$argmin_{\beta}||DS_X^T y - DS_X^T X\beta||$$

Weighted sampling estimators

$$\tilde{\beta}_W = (X^T W X)^{-1} X^T W y$$

with $W = S_X D^2 S_X^T$ ($n \times n$ diagonal random matrix). W is a random matrix with $E(W_{ii}) = 1$.

Smapling approaches

- Uniform: $\pi_i = 1/n$, for all i; Uniform sampling estimator (UNIF)
- Leverage-based: $\pi_i = \frac{h_{ii}}{\sum_i^n h_{ii}} = h_{ii}/p$; Leveraging Estimator (LEV)
- Shrinkage: $\pi_i = \alpha\pi_i^{Lev} + (1-\alpha)\pi_i^{Unif}$; Shrinkage leveraging estimator (SLEV)
- Unweighted leveraging: with π_i^{Lev} solving

$$argmin_\beta ||S_X^T y - S_X^T X\beta||$$

Outline

Lemma 1

A Taylor expansion of $\tilde{\beta}_W$ around the point $E(W) = 1$ yields

$$\tilde{\beta}_W = \hat{\beta}_{ols} + (X^T X)^{-1} X^T Diag\{\hat{e}\}(w-1) + R_w$$

where $\hat{e} = y - X\hat{\beta}_{ols}$ and R_w is the Taylor expansion reminder

Remark: (1) when Taylor expansion is valid when $R_W = o_p(||W-1||)$. No theoretical justification when it holds.
(2) the formula does not apply to LEVUNW

Lemma 2

$$
\begin{aligned}
E_W\left[\tilde{\beta}_W|y\right] =& \hat{\beta}_{ols} + E_W\left[R_w\right] \\
Var_W\left[\tilde{\beta}_W|y\right] =& (X^TX)^{-1}\left[Diag\{\hat{e}\}Diag\{\frac{1}{r\pi}\}Diag\{\hat{e}\}\right]X(X^TX)^{-1} \\
& + Var_W\left[R_w\right]
\end{aligned}
$$

Remark: when $E_W\left[\tilde{\beta}_W|y\right]$ is negligible, $\tilde{\beta}_W$ is approximately unbiased relative to full sample estimate $\hat{\beta}_{ols}$. The variance is inversely proportional to subsample size r.

Lemma 2

$$
\begin{aligned}
E\left[\tilde{\beta}_W\right] =& \beta_0 \\
Var\left[\tilde{\beta}_W\right] =& \sigma^2(X^T X)^{-1} + \frac{\sigma^2}{r}(X^T X)^{-1} Diag\{\frac{(1-h_{ii})^2}{\pi_i}\} X (X^T X)^{-1} \\
& + Var\left[R_W\right]
\end{aligned}
$$

Remark: $\tilde{\beta}_W$ is unbiased to true value β_0. The variance depends on leverage and sampling probability, and is inversely proportional to subsample size r.

UNIF

$$
\begin{aligned}
E_W\left[\tilde{\beta}_{UNIF}|y\right] =& \hat{\beta}_{ols} + E_W\left[R_{UNIF}\right] \\
Var_W\left[\tilde{\beta}_{UNIF}|y\right] =& \frac{n}{r}(X^TX)^{-1}\left[Diag\{\hat{e}\}Diag\{\hat{e}\}\right]X(X^TX)^{-1} \\
&+ Var_W\left[R_{UNIF}\right] \\
E\left[\tilde{\beta}_{UNIF}\right] =& \beta_0 \\
Var\left[\tilde{\beta}_{UNIF}\right] =& \sigma^2(X^TX)^{-1} + \frac{n}{r}(X^TX)^{-1}Diag\{(1-h_{ii})^2\}X(X^TX)^{-1} \\
&+ Var\left[R_{UNIF}\right]
\end{aligned}
$$

Remark: (1) The variance depends on $\frac{n}{r}$, could be very large unless r is closed to n; (2) The sandwich-type expression will not be inflated by small h_{ii}.

LEV

$$
\begin{aligned}
E_W\left[\tilde{\beta}_{LEV}|y\right] =& \hat{\beta}_{ols} + E_W\left[R_{LEV}\right] \\
Var_W\left[\tilde{\beta}_{LEV}|y\right] =& \frac{p}{r}(X^TX)^{-1}\left[Diag\{\hat{e}\}Diag\{\frac{1}{h_{ii}}\}Diag\{\hat{e}\}\right]X(X^TX)^{-1} \\
&+ Var_W\left[R_{LEV}\right] \\
E\left[\tilde{\beta}_{LEV}\right] =& \beta_0 \\
Var\left[\tilde{\beta}_{LEV}\right] =& \sigma^2(X^TX)^{-1} + \frac{p\sigma^2}{r}(X^TX)^{-1}Diag\{\frac{(1-h_{ii})^2}{h_{ii}}\}X(X^TX) \\
&+ Var\left[R_{LEV}\right]
\end{aligned}
$$

Remark: (1) The variance depends on $\frac{p}{r}$, not sample size n; (2) The sandwich-type expression can be inflated by small h_{ii}.

SLEV

- $\pi_i = \alpha \pi_i^{Lev} + (1-\alpha)\pi_i^{Unif}$
- Lemma 2 still holds
- If $(1-\alpha)$ is not small, variance of the SLEV does not get inflated too much
- If $(1-\alpha)$ is not large, variance of the SLEV has a scale of p/r
- Not only increase the small scores, but also shrinkage on large scores

LEVUNW

A Taylor expansion of $\tilde{\beta}_W$ around the point $E(W) = r\pi$ yields

$$\tilde{\beta}_{LEVUNW} = \hat{\beta}_{wls} + (X^T X)^{-1} X^T Diag\{\hat{e}_W\}(W - r\pi) + R_{LEVUNW}$$

where $\hat{\beta}_{wls} = (X^T W_0 X)^{-1} X W_0 y$ and $\hat{e}_W = y - X\hat{\beta}_{wls}$, $W_0 = Diag\{rh_{ii}/p\}$

LEVUNW

$$
\begin{aligned}
E_W\left[\tilde{\beta}_{LEVUNW}|y\right] =& \hat{\beta}_{wls} + E_W\left[R_{LEVUNW}\right] \\
Var_W\left[\tilde{\beta}_{LEVUNW}|y\right] =& (X^T W_0 X)^{-1} Diag\{\hat{e}_W\} W_0 Diag\{\hat{e}_W\} X (X^T W_0 X)^{-} \\
& + Var_W\left[R_{LEVUNW}\right]
\end{aligned}
$$

Remark: for a given data set, $\tilde{\beta}_{LEVUNW}$ is approximately unbiased to $\hat{\beta}_{wls}$, but not $\hat{\beta}_{ols}$.

LEVUNW

$$
\begin{aligned}
E_W\left[\tilde{\beta}_{LEVUNW}\right] &= \beta_0 \\
Var_W\left[\tilde{\beta}_{LEVUNW}\right] &= \sigma^2 (X^T W_0 X)^{-1} X^T W_0^2 X (X^T W_0 X)^{-1} \\
&+ (X^T W_0 X)^{-1} X^T Diag\{I - P_{X,W_0}\} W_0 Diag\{I - P_{X,W_0}\} X (X^T W_0 X)^{-1} \\
&+ Var_W\left[R_{LEVUNW}\right]
\end{aligned}
$$

Remark: $\tilde{\beta}_{LEVUNW}$ is unbiased to β_0 and the variance is not inflated by small leverage

Outline

Approximate computation

Based on Drineas et al. (2012)

- Generate an $r_1 \times n$ random matrix $\prod_1$
- Generate an $p \times r_2$ random matrix $\prod_2$
- Compute R, where R is the thin SVD of $\prod_1 X = QR$
- Return the leverage score of $XR^{-1} \prod_2$

Computation time

For approximate choices of r_1 and r_2, if one chooses $\prod_1$ to be a Hadamard-based random matrix, the the computation time is $o(np^2)$

Empirical studies

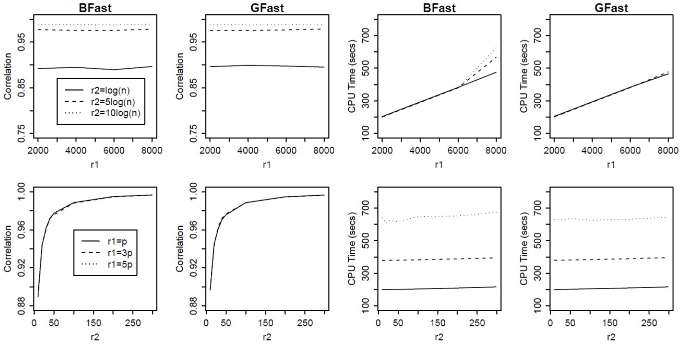
$n = 20,000$ and $p = 1,000$

- BFast: each element of $\prod_1$ and $\prod_2$ is generated i.i.d from $\{$-1,1$\}$ with equal sampling
- GFast: each element of $\prod_1$ and $\prod_2$ is generated i.i.d from $N(0, \frac{1}{n})$ and $N(0, \frac{1}{p})$
- $n = 20,000$ and $p = 1,000$
- $r_1 = p, 1.5p, 2p, 3p, 4p, 5p$ and $r_2 = klog(n)$ with $k = 1, 2, \ldots, 20$

Empirical studies: choose of r_1 and r_2

- With the increase of r_1, the correlation are not sensitive but the running time increase linearly
- With the increase of r_2, the correlation increase rapidly but the running time not sensitive
- Choose small r_1 and large r_2

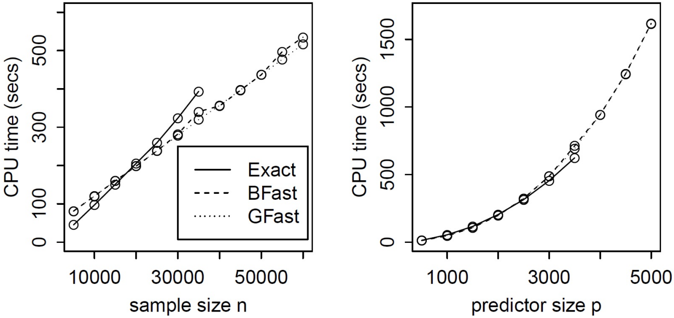
Empirical studies: choose of r_1 and r_2



Empirical studies: computation time

- When $n \leq 20,000$, exact method takes less time
- When $n > 20,000$, the approximate approach has some advantage

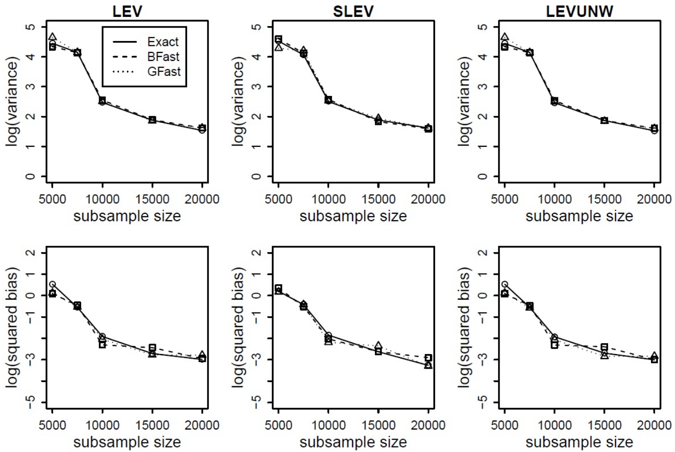
Empirical studies: computation time



Empirical studies: estimation comparision

- Compare the bias and variance of LEV, SLEV, and LEVUNW using exact, BFast, and GFast
- The results are almost identical

Empirical studies: estimation comparision



Outline

Plan

- Unconditional bias and variance for LEV and UNIF
- Unconditional bias and variance for SLEV and LEVUNW
- Conditional bias and variance of SLEV and LEVUNW
- Real data application

Synthetic data

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim N(0, 9I_n)$$

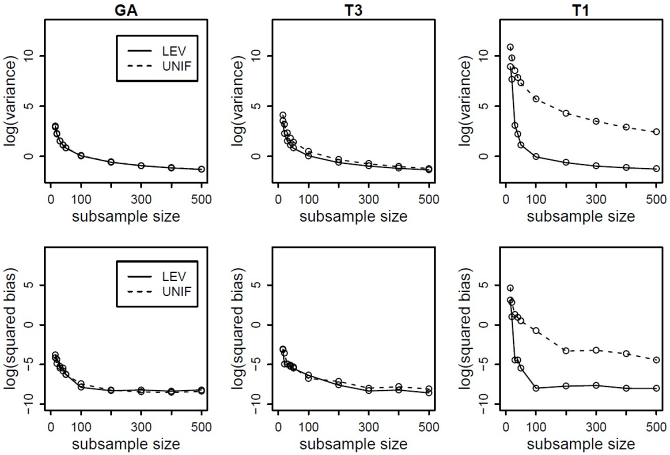
- Nearly uniform leverage scores (GA): $X \sim N(1_p, \Sigma)$, $\Sigma_{ij} = 2 \times 0.5^{|i-j|}$, and $\beta = (1_{10}, 0.11_{p-20}, 1_{10})$
- Moderately nonuniform leverage scores (T_3): X is from multivariate t-distribution with df=3
- Very nonuniform leverage scores (T_1): X is from multivariate t-distribution with df=1

LEV vs UNIF: square loss and variance

$n = 1000$, $p = 10, 50, 100$, and repeat sampling 1000 times

- Square loss is much smaller than variance
- Similarly for GA
- Less similarly for T_3
- Very different for T_1
- Both decrease as r increase, but slower for *UNIF*

LEV vs UNIF: square loss and variance

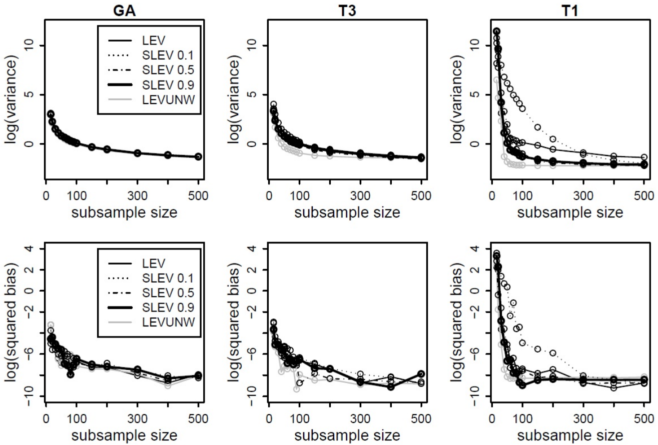


Improvements from SLEV and LEVUNW

$n = 1000$, $p = 10, 50, 100$, and repeat sampling 1000 times

- Similarly for GA
- Less similarly for T_3
- Different for T_1
- SLEV with $\alpha = 0.9$ and LEVUNW have better performance

Improvements from SLEV and LEVUNW

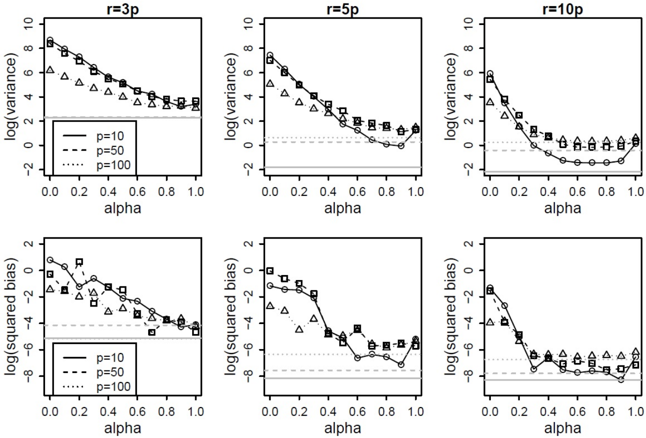


Choices of α in SLEV

$n = 1000$, $p = 10, 50, 100$, and repeat sampling 1000 times

- T_1 data
- $0.8 \leq \alpha \leq 0.9$ has beneficial effect
- Recommend $\alpha = 0.9$
- LEVUNW has better performance

Choices of α in SLEV

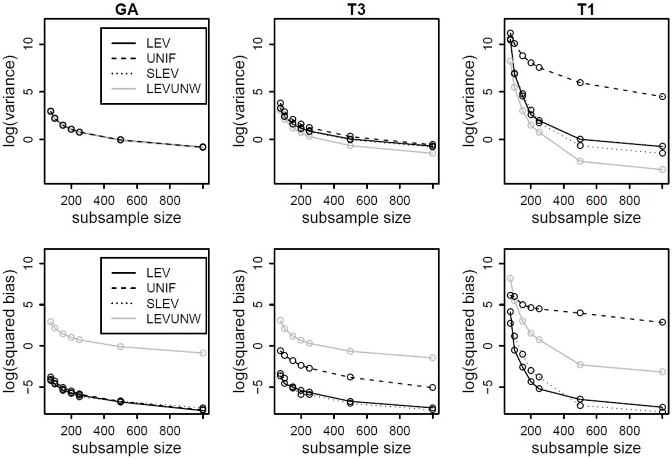


Conditional bias and variance

$n = 1000$, $p = 10, 50, 100$, and repeat sampling 1000 times

- LEVUNW is biased for $\hat{\beta}_{ols}$
- LEVUNW has smallest variance
- Recommend use SLEV with $\alpha = 0.9$

Conditional bias and variance



Real Data: RNA-SEQ data

$n = 51,751$ read counts from embryonic mouse stem cells

- n_{ij} denotes the counts of reads that are mapped to the genome starting at the jth nucleotide of the ith gene
- $y_{ij} = log(n_{ij} + 0.5)$
- Independent variables: 40 nucleotides denoted as $b_{ij,-20}$,, $b_{ij,-19}, \ldots, b_{ij,19}$.
- Linear model: $y_{ij} = \alpha + \sum_{k=-20}^{19} \sum_{h \in H} \beta_{kh} I(b_{ij,k} = h) + \epsilon_{ij}$, where $H = \{A, C, G\}$, T is used as baseline level.
- $p = 121$

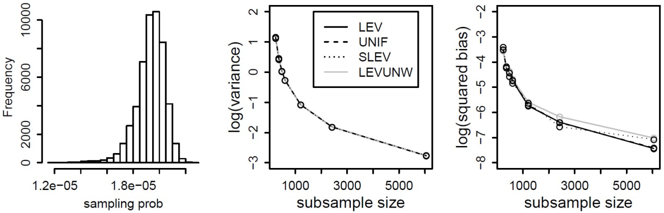
Sampling analysis

- UNIF, LEV, and SLEV
- $r = 2p, 3p, 4p, 5p, 10p, 20p, 50p$
- Compare sample bias (respect to $\hat{\beta}_{ols}$) and variance
- Sampling 100 times

Comparison

- Relatively uniform leverage scores
- Almost identical variances
- LEVUNW has slightly larger bias

Emprical resutls for real data I



Real Data: predicting gene expression of cancer patient

$n = 5,520$ genes for 46 patients.

- Randomly select one patient's gene expression as y and remaining patients' gene expressions as predictors ($p = 45$)
- Sample sizes from 100 to 5000
- UNIF, LEV, and SLEV

Comparison

- Relatively nonuniform leverage scores
- SLEV and LEV have smaller variances
- LEVUNW has the largest bias

Emprical resutls for real data II