RandomizedNLA for GLMs with Big Datasets

Master Thesis - Barcelona Graduate School of Economics

Robert T. Lange
Supervisors: Prof. Ioannis Kosmidis
Prof. Omiros Papaspiliopoulos

April 13, 2017

Outline

1. **RandNLA for Least Squares**
   Problem | Structural Requirements | RSampling Algorithm
2. **RandNLA for GLM**
   Problem | IWLS Scheme | Quality of Approximation
3. **Questions | Next Steps**

Problem Formulation: LS

- Notation: $y, \beta \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$ and $\Pi \in \mathbb{R}^{r \times n}$ where $r << n$

$$RSS_{min} = \min_{\beta \in \mathbb{R}^d} ||y - X\beta||_2^2 \iff \widetilde{RSS}_{min} = \min_{\beta \in \mathbb{R}^d} ||\Pi(y - X\beta)||_2^2$$

- Interpretation as weighted least squares problem
- Requirements for a "good" approximation:

$$\textit{Solution Vector: } \tilde{\beta}_{LS} \approx \beta_{LS}$$

$$\textit{Objective Function Value: } \widetilde{RSS}_{min} \approx RSS_{min}$$

- Approaches: Random Sampling and Random Projection

Structural Conditions for a "good" Approximation

$$\textit{Rotation: } \sigma^2_{min}(\Pi U^{(X)}) = \lambda_{min}(U^{(X)T}\Pi^T\Pi U^{(X)}) \geq \frac{1}{\sqrt{2}}$$

- Lower bound on singular values of $\Pi U^{(X)}$, where $U^{(X)}$ denotes the orthonormal basis of span(X) (from SVD or QR).

$$\textit{Subspace Embedding:} ||U^{(X)}\Pi^T\Pi y^{\perp}||_2^2 \leq \frac{\epsilon}{2} RSS_{min}$$

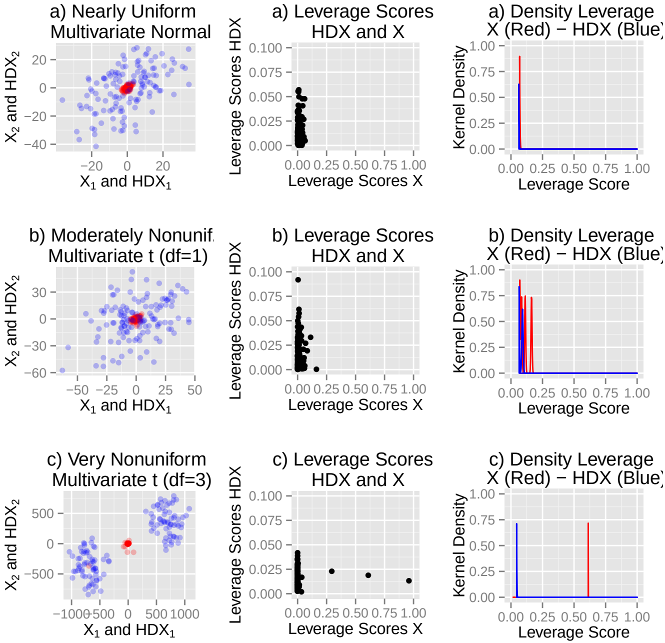
- Orthogonal projection property of LS approximately "survives" the transformation
- $\Pi y^{\perp}$ has to be approximately orthogonal to $\Pi U^{(X)}$

Fast Algorithms for LS I

- Rely on Fast Fourier/Hadamard preprocessing of input matrix (Ailon & Chazelle, 2006)

$$\Pi = PHD$$

- $P \in \mathbb{R}^{r \times n}$ - Sparse JL matrix/Uniform sampling (Achlioptas, 2003; Frankl & Maehara, 1988)
- $H \in \mathbb{R}^{n \times n}$ - DFT/Normalized Hadamard Matrix
- $D \in \mathbb{R}^{n \times n}$ - random $\{\pm 1\}$ matrix: Preprocesses bad cases
- Spreads out energy and flattens "spiky" vector in terms of sup-norm. Makes spare JL or uniform sampling effective.
- If r is appropriately large, Π is going to be ϵ-FJLT with high probability $\rightarrow$ compute ΠX in $O(ndlog(r))$ time.


a) Nearly Uniform Multivariate Normal
X2 and HDX2
X1 and HDX1
a) Leverage Scores HDX and X
Leverage Scores HDX
Leverage Scores X
a) Density Leverage X (Red) − HDX (Blue)
Kernel Density
Leverage Score
b) Moderately Nonunif. Multivariate t (df=1)
X2 and HDX2
X1 and HDX1
b) Leverage Scores HDX and X
Leverage Scores HDX
Leverage Scores X
b) Density Leverage X (Red) − HDX (Blue)
Kernel Density
Leverage Score
c) Very Nonuniform Multivariate t (df=3)
X2 and HDX2
X1 and HDX1
c) Leverage Scores HDX and X
Leverage Scores HDX
Leverage Scores X
c) Density Leverage X (Red) − HDX (Blue)
Kernel Density
Leverage Score

Fast Algorithms for LS II

- Approaches:
  (i) Random Sampling: Use FJLT to compute approximate leverage scores and sample accordingly
  (ii) Random Projection: Uniformize leverage scores and sample uniformly/use sparse projection matrix

Theorem 1 (Least Squares Quality of Approximation)

$\tilde{\beta}$ is such that with probability at least 0.8

Solution Certificate: $||\tilde{\beta} - \beta_{LS}||_2 \leq \sqrt{\epsilon}\kappa(X)\sqrt{\gamma^{-2} - 1}||\beta_{LS}||_2$

Objective Function: $||X\tilde{\beta} - y||_2 \leq (1 + \epsilon)||X\beta_{LS} - y||_2$

$$\text{where } \gamma = \frac{||U^{(X)}U^{(X)T}y||_2}{||y||_2}$$

Fast Leverage Score Approximations

Algorithm 1 Mahoney (2016, 81) - FJLT approximation for leverage scores

Input: $X \in \mathbb{R}^{n \times d}$ with SVD $X = U^{(X)} \Sigma V^T$ and an ϵ-level

Output: Approximate leverage scores, $\tilde{l}_i, i = 1, ..., n$

1: Let $\Pi_1 \in \mathbb{R}^{r_1 \times n}$ be an ϵ-FJLT for $U^{(X)}$ with $r_1 = \Omega\left(\frac{dlog(n)}{\epsilon^2} log\left(\frac{dlog(n)}{\epsilon^2}\right)\right)$.

2: Compute $\Pi_1 X$ and its SVD/QR where $R = \Sigma V^T$.

3: View rows of $XR^{-1} \in \mathbb{R}^{n \times d}$ as n vectors in $\mathbb{R}^d$. Let $\Pi_2 \in \mathbb{R}^{d \times r_2}$ be an ϵ-JLT for n^2 vectors, with $r_2 = O(\frac{log(n)}{\epsilon^2})$.

4: **return** $\tilde{l}_i = ||(XR^{-1}\Pi_2)_i||_2^2$, an ϵ-approximation of l_i.

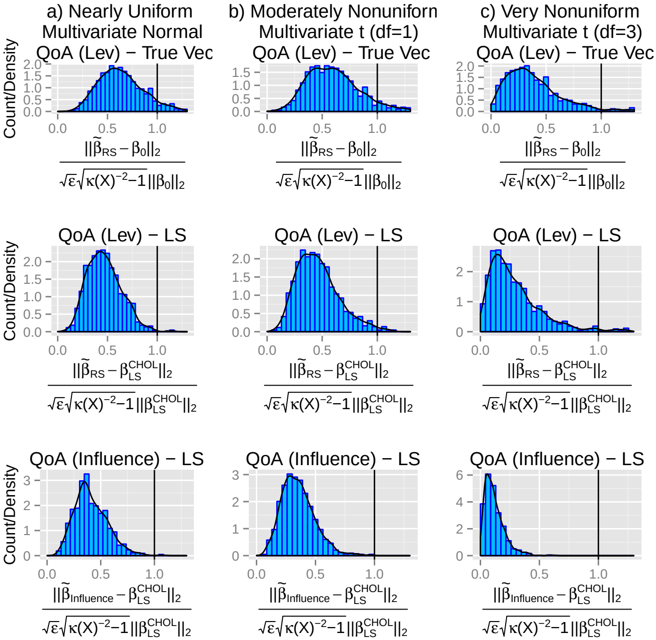
- $O(ndlog(d/\epsilon) + nd\epsilon^{-2}log(n) + d^3\epsilon^{-2}log(n)log(d\epsilon^{-1}))$

Fast Random Sampling Algorithm for LS

Algorithm 2 Drineas *et al.* (2012, 3451) - "Fast" Random Sampling Algorithm for LS

Input: LS problem with $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$ and an ϵ-level
Output: Approximate LS solution, $\tilde{\beta}$

1. Let $\{\tilde{l}_i\}_{i=1}^n$ be an $1 \pm \epsilon$ approximation to the leverage score computed using Algorithm 2.
2. Randomly sample $r = O(\frac{dlog(d)}{\epsilon})$ rows of X and y with probability depending on $\tilde{l}_i$, rescale them by $\frac{1}{\sqrt{rp_i}}$ and form $\tilde{X} \in R^{r\times d}, \tilde{y} \in R^r$.
3. Solve $(\tilde{X}'\tilde{X})\tilde{\beta} = \tilde{X}'\tilde{y}$ by SVD/QR/Cholesky.
4. **return** $\tilde{\beta}$, an ϵ-approximation of β_{LS} in $O(ndlog(r))$ time.



a) Nearly Uniform
Multivariate Normal
b) Moderately Nonuniform
Multivariate t (df=1)
c) Very Nonuniform
Multivariate t (df=3)
QoA (Lev) − True Vec
QoA (Lev) − LS
QoA (Influence) − LS
Count/Density
||β̃RS − β0||2 / √ε√κ(X)−2−1 ||β0||2
||β̃RS − βLS CHOL||2 / √ε√κ(X)−2−1 ||βLS CHOL||2
||β̃Influence − βLS CHOL||2 / √ε√κ(X)−2−1 ||βLS CHOL||2

Problem Formulation: IWLS

- IWLS: $\beta_{(k+1)} = (X^T W_{(k)} X)^{-1} X^T W_{(k)} z_{(k)}$
  - $\rightarrow$ $\eta = X\beta$ and $\mu = (\mu_1, ..., \mu_n)$ where $\mu_i = \mathbb{E}(y_i)$
  - $\rightarrow$ $z_{(k)} = X\beta_{(k-1)} + (y - \mu) diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$
  - $\rightarrow$ $W_{(k)} = diag(w_{1(k)}, ..., w_{n(k)})$ with $w_{i(k)} = Var(\mu_i)^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$
- Problem formulation:

$$\left((\Pi X)^T \tilde{W}_{(k)} \Pi X\right) \beta_{(k+1)} = \left((\Pi X)^T \tilde{W}_{(k)} \Pi z_{(k)}\right)$$
$$\tilde{X}^T \tilde{W}_{(k)} \tilde{X} \tilde{\beta}_{(k+1)} = \tilde{X}^T \tilde{W}_{(k)} \tilde{z}_{(k)}$$

- Solve weighted normal equation problem at each iteration.

RandIWLS: Sampling Schemes

Definition 2 (Weighted Influence Scores - Jinzhu Jia (2014))

Given an orthonormal basis $U^{(X)}$ for $span(X)$ the weighted leverage scores of X are defined as $WL_i = ||w_i U_i^{(X)}||_2^2$.

Definition 3 (Working Variate Influence Score)

At iteration k of the Randomized IWLS scheme the working variate influence scores are defined as the leverage scores of the concatenated matrix $(X, z_{(k)}) \in \mathbb{R}^{n\times(d+1)}$.

$$wvI_i = ||U_i^{(X,z_{(k)})}||_2^2$$

- We propose 3 potential sampling schemes:
  1. Sample according to fast approximate leverage scores of X
  2. Sample according to weighted leverage scores
  3. Sample according to working variate influence scores

Algorithmic Leveraging for GLM

Algorithm 3 Random Sampling Algorithm for IWLS

Input: GLM problem with $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, initial $\beta_{(0)} \in \mathbb{R}^d$

Output: Approximate GLM solution, $\tilde{\beta}_{RandGLM}$

1: **while** $||\beta_{(k+1)} - \beta_{(k)}||_2^2 > \delta$ **do**

2: $z_{(k)} = X\beta_{(k)} + (y-\mu) diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$ with $\eta = X\beta_{(k)}$, $\mu = \mathbb{E}(\eta)$

3: $W_{(k)} = diag\left(Var(\mu_i)^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$ with $\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

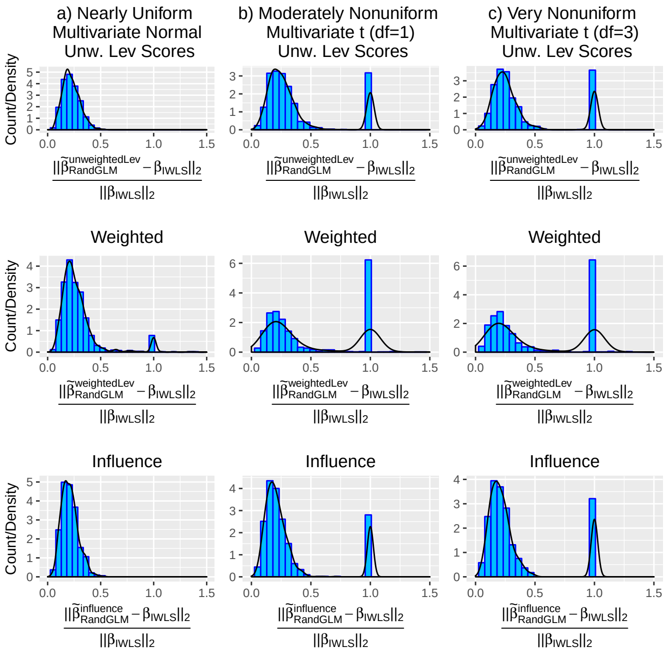
4: Form a normalized sampling distribution according to one of the three proposed sampling schemes.

5: Randomly sample $r = O(\frac{dlog(d)}{\epsilon})$ rows of X, W and z. Rescale them by $\frac{1}{\sqrt{rp_i}}$, form $\tilde{X} \in R^{r\times d}$, $\tilde{W}_{(k)} \in \mathbb{R}^{r\times r}$, $\tilde{z}_{(k)} \in R^r$.

6: Solve $\beta_{(k)} = (\tilde{X}^T \tilde{W}_{(k)} \tilde{X})^{-1} \tilde{X}^T \tilde{W}_{(k)} \tilde{z}_{(k)}$.

7: **end while**

8: **return** $\tilde{\beta}_{RandGLM}$, an approximation of β_{GLM}.


a) Nearly Uniform
Multivariate Normal
Unw. Lev Scores
b) Moderately Nonuniform
Multivariate t (df=1)
Unw. Lev Scores
c) Very Nonuniform
Multivariate t (df=3)
Unw. Lev Scores
Weighted
Weighted
Weighted
Influence
Influence
Influence
Count/Density
0.0
0.5
1.0
1.5
$\frac{\|\tilde{\beta}^{\text{unweightedLev}}_{\text{RandGLM}} - \beta_{\text{IWLS}}\|_2}{\|\beta_{\text{IWLS}}\|_2}$
$\frac{\|\tilde{\beta}^{\text{weightedLev}}_{\text{RandGLM}} - \beta_{\text{IWLS}}\|_2}{\|\beta_{\text{IWLS}}\|_2}$
$\frac{\|\tilde{\beta}^{\text{influence}}_{\text{RandGLM}} - \beta_{\text{IWLS}}\|_2}{\|\beta_{\text{IWLS}}\|_2}$

Following the Approximation Error

$$\hat{z}_{(k+1)} = X\tilde{\beta}_{(k)} + (y - \tilde{\mu})diag\left(\frac{\partial \tilde{\eta}_i}{\partial \tilde{\mu}_i}\right) = \prod_{j=1}^{k}(1 \pm \epsilon_j)z_{(k+1)}$$

$$\hat{W}_{(k+1)} = diag\left(Var(\tilde{\mu}_i)^{-1}\left(\frac{\partial \tilde{\mu}_i}{\partial \tilde{\eta}_i}\right)^2\right) = \prod_{j=1}^{k}(1 \pm \epsilon_j)W_{(k+1)}$$

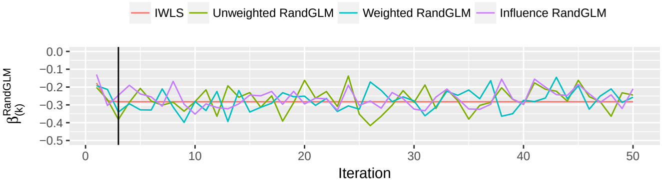
where

$$\tilde{\eta} = X\tilde{\beta}_{(k+1)} = \prod_{j=1}^{k}(1 \pm \epsilon_j)X\beta_{(k)}$$

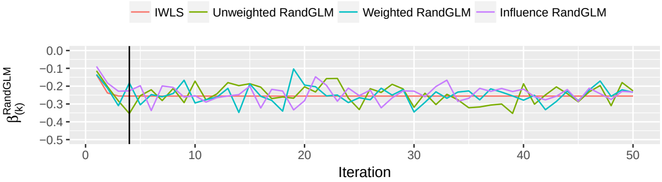
$$\tilde{\mu} = \mathbb{E}(X\tilde{\beta}_{(k)}) = \mathbb{E}(X\prod_{j=1}^{k}(1 \pm \epsilon_j)\beta_{(k)}) = \mathbb{E}(X\beta_{(k)})$$

- Given independence of the approximation errors $k \to \infty$, $\prod_{k=1}^{K}(1 \pm \epsilon_k) = (1 \pm \epsilon)^k \to 1$. $\Rightarrow$ Asymptotically consistent!

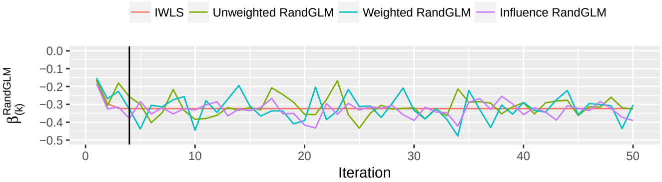
a) Nearly Uniform − Multivariate Normal



b) Moderately Nonuniform − Multivariate t (df=1)



c) Very Nonuniform − Multivariate t (df=2)



References I

Achlioptas, Dimitris. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, **66**(4), 671–687.

Ailon, Nir, & Chazelle, Bernard. 2006. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 557–563.

Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W, & Woodruff, David P. 2012. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, **13**(Dec), 3475–3506.

Frankl, Peter, & Maehara, Hiroshi. 1988. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, **44**(3), 355–362.

Mahoney, Michael W. 2016. Lecture Notes on Randomized Linear Algebra. *arXiv preprint arXiv:1608.04481*.