Subsampling least squares and elemental estimation

Keith Knight
Department of Statistical Sciences, University of Toronto
keith@utstat.toronto.edu

Least squares and leveraging

Given observations $\{(\boldsymbol{x}_i, y_i) : i = 1, \cdots, n\}$, define the OLS estimate $\widehat{\boldsymbol{\beta}}$ as the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2.$$

We will assume here that n and $p = \dim(\boldsymbol{\beta})$ are both very large with $n \gg p$. $\widehat{\boldsymbol{\beta}}$ can be computed in $O(np^2)$ time – prohibitive when n and p are very large.

To reduce the computational burden, we can take a random subsample of $m \ll n$ observations and define $\widehat{\boldsymbol{\beta}}_{ss}$ to minimize

$$\sum_{i\in\mathcal{A}} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

where $\mathcal{A}$ is the set of sampled indices and $\{w_i\}$ are some weights. The goal is to determine $\mathcal{A}$ and $\{w_i\}$ so that $\|\widehat{\boldsymbol{\beta}}_{ss} - \widehat{\boldsymbol{\beta}}\|$ is small with high probability. (One may also want to examine the statistical properties of $\widehat{\boldsymbol{\beta}}_{ss}$ but that will not be pursued here.)

Algorithmic leveraging samples observations with probabilities proportional to the diagonals $h_{11}, \cdots, h_{nn}$ of the **hat matrix** $H = X(X^TX)^{-1}X^T$. This tends to produce a more informative subsample than simple random sampling as illustrated by the plots given below in Figure 1 (for simple linear regression).

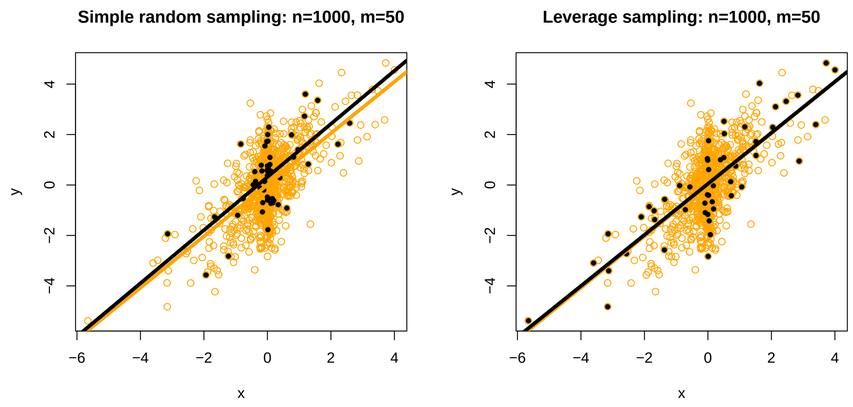

Figure 1: Subsamples drawn using simple random sampling and leverage samples; subsampled points are black. The orange and black lines represent, respectively, the LS estimate using all the data and the LS estimate using only the subsampled (black) points.

However, exact computation of $\{h_{ii}\}$ has a similar computational complexity to that of computing the OLS estimate $\widehat{\boldsymbol{\beta}}$. The feasibility of leveraging relies on the fact that $\{h_{ii}\}$ can be approximated well (i.e. with $\pm\epsilon$ relative error) in $o(np^2)$ time (Drineas *et al.*, 2011; Drineas *et al.*, 2012) using the Fast Johnson-Lindenstrauss Transform (Ailon & Chazelle, 2009).

Leveraging gives us a multinomial vector

$$(m_1, \cdots, m_n) \sim \text{Mult}\left(m; \frac{h_{11}}{p}, \cdots, \frac{h_{nn}}{p}\right)$$

representing the number of occurrences of each observation in the subsample; thus $\mathcal{A} = \{i : m_i \geq 1\}$. Two types of weighting are generally used:

- **Unweighted leveraging:** $w_i = m_i$ for $i \in \mathcal{A}$;
- **Weighted leveraging:** $w_i = m_i/h_{ii}$ for $i \in \mathcal{A}$.

The rationale for weighted leveraging is that it makes the objective function unbiased. Hybrid methods (e.g. convex combinations) can also be used.

Past research: Monte Carlo studies by Ma *et al.* (2015) and Ma & Sun (2015) indicate that unweighted leveraging tends to outperform weighted leveraging, particularly when the design contains high leverage points (i.e. the distribution of $\{h_{ii}\}$ is highly skewed to the right).

Leveraging and elemental estimates

Elemental estimates are estimates of $\boldsymbol{\beta}$ defined on subsets of p observations: If $s = \{i_1, \cdots, i_p\}$ is a subset of $\{1, \cdots, n\}$ then the elemental estimate $\widehat{\boldsymbol{\beta}}_s$ satisfies

$$\boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_s = y_j \text{ for all } j \in s.$$

Both the OLS and leveraging estimates can be written as expected values of elemental estimates with respect to probability distributions defined on all subsets of p observations.

- **OLS**: $\widehat{\boldsymbol{\beta}} = \sum_s \mathcal{P}(s)\widehat{\boldsymbol{\beta}}_s$ where $\mathcal{P}(s) = \left|\begin{pmatrix} h_{i_1i_1} & h_{i_1i_2} & \cdots & h_{i_1i_p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{i_pi_1} & h_{i_pi_2} & \cdots & h_{i_pi_p}\end{pmatrix}\right|$ for $s = \{i_1, \cdots, i_p\}$ where $\{h_{ij} : i, j = 1, \cdots, n\}$ are the elements of the hat matrix (Subrahmanyam, 1972).
- **Leveraging:** $\widehat{\boldsymbol{\beta}}_{ss} = \sum_s \mathcal{Q}(s)\widehat{\boldsymbol{\beta}}_s$ where $\mathcal{Q}(s) \propto \mathcal{P}(s)\prod_{j=1}^{p} w_{i_j}$ for $s = \{i_1, \cdots, i_p\} \subset \mathcal{A}$ with $\mathcal{Q}(s) = 0$ otherwise.

Proposition: *For a given set $\mathcal{A}$, the total variation (TV) distance $d_{tv}(\mathcal{Q}, \mathcal{P}) = \frac{1}{2}\sum_s |\mathcal{Q}(s) - \mathcal{P}(s)|$ is minimized for $\mathcal{Q}$ satisfying $\mathcal{Q}(s) = \lambda(s)\mathcal{P}(s)$ with $\lambda(s) \geq 1$ when $s \subset \mathcal{A}$ and $\lambda(s) = 0$ otherwise. The minimum TV distance is $1 - \gamma(\mathcal{A})$ where*

$$\gamma(\mathcal{A}) = \left|\begin{pmatrix} 1 - h_{i_1i_1} & -h_{i_1i_2} & \cdots & -h_{i_1i_\ell} \\ -h_{i_2i_1} & 1 - h_{i_2i_2} & \cdots & -h_{i_2i_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{i_\ell i_1} & -h_{i_\ell i_2} & \cdots & 1 - h_{i_\ell i_\ell}\end{pmatrix}\right| \approx \exp\left(-\sum_{j=1}^{\ell} h_{i_ji_j} - \frac{1}{2}\sum_{j=1}^{\ell}\sum_{k=1}^{\ell} h_{i_ji_k}^2\right)$$

where $\{i_1, \cdots, i_\ell\} = \mathcal{A}^c$ with $\ell = \sum_{i=1}^{n} I(m_i = 0) \approx \sum_{i=1}^{n}\exp(-mh_{ii}/p)$.

Corollary: *A leveraging estimate attains the lower bound on the TV distance if, and only if,*

$$\prod_{j\in s} w_j \geq \gamma(\mathcal{A})\sum_{u\subset\mathcal{A}}\left\{\frac{\mathcal{P}(u)}{\gamma(\mathcal{A})}\prod_{j\in u} w_j\right\} = \gamma(\mathcal{A})\sum_{u\subset\mathcal{A}}\left\{\mathcal{P}_{\mathcal{A}}(u)\prod_{j\in u} w_j\right\}$$

for all $s \subset \mathcal{A}$ where $\mathcal{P}_{\mathcal{A}} = \mathcal{P}/\gamma(\mathcal{A})$ is a probability distribution on subsets of $\mathcal{A}$.

Some notes

1. $\gamma(\mathcal{A})$ is monotone increasing in $\mathcal{A}$: If $\mathcal{A}_1 \subset \mathcal{A}_2$ then $\gamma(\mathcal{A}_1) \leq \gamma(\mathcal{A}_2)$.
   - We can use $m^* = n \times \{\gamma(\mathcal{A})\}^{1/p}$ as a (crude) measure of the effective subsample size.
   - Is it worthwhile modifying the sampling procedure to try to maximize $\gamma(\mathcal{A})$ for a fixed subsample size m?
2. For general $\{w_i\}$, the lower bound on the TV distance is attained provided that $\prod_{j\in s} w_j$ is not too variable for $s \subset \mathcal{A}$.
3. For weighted leveraging ($w_i = m_i/h_{ii}$), the lower bound is not necessarily attained, especially if the design contains high leverage points (i.e. some $h_{ii} \gg p/n$).
4. The lower bound is always attained when $w_i = 1$ (for $i \in \mathcal{A}$); this is similar to unweighted leveraging particularly if m is small compared to n.
   - For larger m, simulations indicate that taking $w_i = 1$ may be advantageous relative to other options.

An illustration

Two designs with $p = 100$ and $n = 100000$ are considered:

- High leverage design: $\max h_{ii} \approx 40p/n$.
- Low leverage design: $\max h_{ii} \approx 1.7p/n$.

For a given subsampling estimate, we define the squared error SqErr $= \|\widehat{\boldsymbol{\beta}}_{ss} - \widehat{\beta}\|^2$. Figures 2 and 3 show the distribution (based on 1000 replications) of the ratio SqErr(unweighted)/SqErr(weighted) for $m = 200, 1000, 5000$ for the two designs.

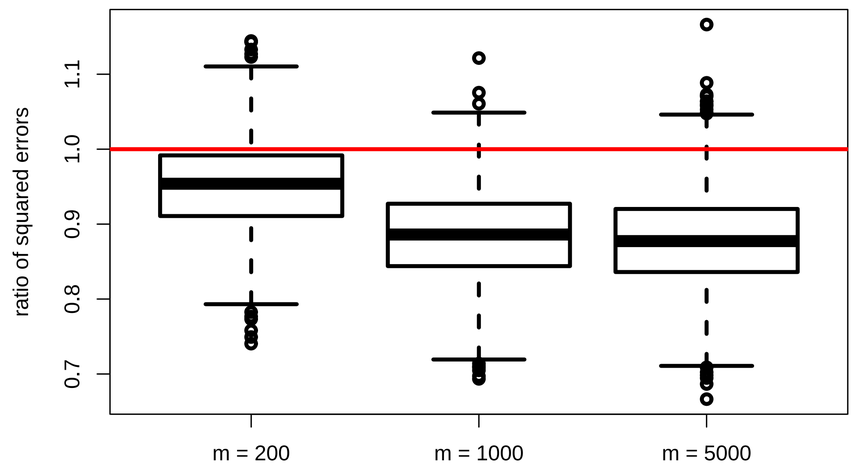

Figure 2: High leverage design: SqErr(unweighted)/SqErr(weighted) for $m = 200, 1000, 5000$.

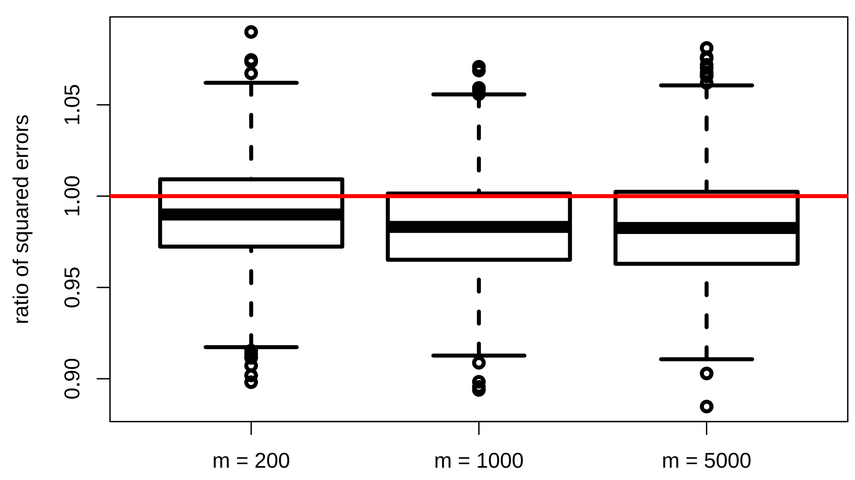

Figure 3: Low leverage design: SqErr(unweighted)/SqErr(weighted) for $m = 200, 1000, 5000$.

The simulation results confirm the theoretical results:

- For the high leverage design, unweighted leveraging is clearly superior with the advantage increasing with the subsample size.
- For the low leverage design, unweighted leveraging is still better although the advantage is less definitive.

References

[1] Ailon, N., Chazelle, B.: The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. SIAM Journal on Computing **39**, 302-322 (2009)

[2] Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.: Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research. **13**, 3475–3506 (2012)

[3] Drineas, P., Mahoney, M.W., Muthukrishnan, S., Sarlós, T.: Faster least squares approximation. Numerische Mathematik. **117**, 219–249 (2011)

[4] Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research. **16**, 861–911 (2015)

[5] Ma, P., Sun, X.: Leveraging for big data regression. Wiley Interdisciplinary Reviews: Computational Statistics. **7**, 70–76 (2015)

[6] Subrahmanyam, M.: A property of simple least squares estimates. Sankhya, Series B. **34**, 355-356 (1972)

Acknowledgements

Research supported by the Natural Sciences and Engineering Research Council of Canada.