Sparse Linear Regression via Generalized Orthogonal Least-Squares

Abolfazl Hashemi and Haris Vikalo
University of Texas at Austin, Austin, TX, USA

abolfazl@utexas.edu, hvikalo@ece.utexas.edu

IEEE Global Conference on Signal and Information Processing

Greater Washington, D.C., December 7, 2016

Introduction

- Sparse linear regression $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
- Unknown sparse signal $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\|_0 \leq k$
- Vector of observations $\mathbf{y} \in \mathbb{R}^n$
- Full rank coefficient matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, $n \leq m$
- Observation noise vector $\mathbf{e} \in \mathbb{R}^n$
- Sparse linear regression as an optimization task

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_0 \leq k.$$

- A non-convex NP-hard program
- Efficient approximations:
  - Convex relaxation vs greedy methods

Convex Relaxation Methods

- Replacing ℓ_0-norm constraint problem with a ℓ_1- norm optimization

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon$$

- A related formulation: Least Absolute Shrinkage and Selection Operator (LASSO)

$$\underset{\mathbf{x}}{\text{minimize}} \, \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_1$$

- $\mathbf{A}$ Having near orthonormal columns guarantees perfect reconstruction with high probability [Candes et. al., 2006]
  - Sampling complexity $n = \mathcal{O}(k \log m)$
- Often computationally challenging in practice

Greedy Methods

- Successively identifying columns of $\mathbf{A}$ which correspond to non-zero components of $\mathbf{x}$
- Popular methods: Orthogonal Matching Pursuit (OMP) and its variants, e.g., stage-wise OMP and subspace pursuit
- Selection criterion relies on correlation with a residual vector $\mathbf{r} \in \mathbb{R}^n$

$$j_s = \mathrm{argmax}_{j \in \mathcal{I}} \left| \mathbf{r}^\top \mathbf{a}_j \right|$$

- $\mathbf{A}$ Having near orthonormal columns guarantees perfect reconstruction with high probability [Tropp et. Al., 2007]
  - Sampling complexity $n = \mathcal{O}\left(k \log m\right)$

Orthogonal Least-Squares (OLS)

- Dates back to 1980, but recent in compressed sensing
- Selection criterion relies on minimizing approximation error

$$j_s = \underset{j \in \mathcal{I}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathbf{P}_{\mathcal{S}_{i-1} \cup \{j\}} \mathbf{y} \right\|_2$$

- Empirically shown to outperform L_1 and OMP for an $\mathbf{A}$ with correlated columns [Soussen et. Al., 2013]
- More complex than OMP and more challenging to analyze

Contribution

- Sufficient condition on recovery properties of OLS from random linear measurements
- Improved OLS-based algorithms

Sampling Complexity of OLS

Theorem

For $\mathbf{A} \sim \mathcal{N}(0, 1/n)$ or $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm\frac{1}{\sqrt{n}})$, OLS can recover $\mathbf{x}$ in k iterations from $n = \mathcal{O}\left(k \log m/\delta\right)$ noiseless measurements with probability of success exceeding $1 - \delta^2$.

Toward Improved OLS

1. Reducing complexity of OLS
   - $\mathbf{a}$: Selected column in current iteration
   - $\mathbf{P}_i^{\perp}$: Projection onto span of previously selected columns
   - A recursion for $\mathbf{P}_i^{\perp}$

$$\mathbf{P}_{i+1}^{\perp} = \mathbf{P}_i^{\perp} - \frac{\mathbf{P}_i^{\perp}\mathbf{a}\mathbf{a}^{\top}\mathbf{P}_i^{\perp}}{\left\|\mathbf{P}_i^{\perp}\mathbf{a}\right\|_2^2}$$

   - Equivalent selection criterion

$$j_s = \mathrm{argmax}_{j\in\mathcal{I}} \left|\mathbf{y}^{\top}\frac{\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j}{\left\|\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\right\|_2}\right|$$

2. Selecting L indices in each iteration

Generalized OLS Algorithm

A. Initialize $\mathcal{S}_0 = \emptyset$, $\mathbf{P}_0^{\perp} = \mathbf{I}$, $\mathcal{I} = \{1, 2, \ldots, m\}$

B. Repeat for $i = 1 \text{ to } \min\{k, \lfloor \frac{n}{L} \rfloor\}$

1. $\{i_1, \ldots, i_L\} = \arg_L \max_{j \in \mathcal{I}} \left| \mathbf{y}^\top \frac{\mathbf{P}_{i-1}^{\perp} \mathbf{a}_j}{\left\| \mathbf{P}_{i-1}^{\perp} \mathbf{a}_j \right\|_2} \right|$

2. Update set of selected indices $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{i_1, \ldots, i_L\}$, $\mathcal{I} = \mathcal{I} \backslash \mathcal{S}_i$

3. Update the projection matrix $\mathbf{P}_i^{\perp}$ using recently selected indices

$$\mathbf{P}_{i+1}^{\perp} = \mathbf{P}_{i_L}^{\perp}, \mathbf{P}_{i_{l+1}}^{\perp} = \mathbf{P}_{i_l}^{\perp} - \frac{\mathbf{P}_i^{\perp} \mathbf{a}_{i_l} \mathbf{a}_{i_l}^\top \mathbf{P}_{i_l}^{\perp}}{\left\| \mathbf{P}_{i_l}^{\perp} \mathbf{a} i_l \right\|_2^2}, \mathbf{P}_{i_1}^{\perp} = \mathbf{P}_i^{\perp}$$

C. Find the recovered signal $\hat{\mathbf{x}}_k = \mathbf{A}_{\mathcal{S}_k}^{\dagger} \mathbf{y}$

Computational Complexity

- Cost per iteration
  - Step 1: $\{i_1, \ldots, i_L\} = \arg_L \max_{j \in \mathcal{I}} \left| \mathbf{y}^\top \frac{\mathbf{P}_{i-1}^\perp \mathbf{a}_j}{\left\| \mathbf{P}_{i-1}^\perp \mathbf{a}_j \right\|_2} \right|$

    total cost $\mathcal{O}\left(mn^2\right)$
  - Step 3: $\mathbf{P}_{i+1}^\perp = \mathbf{P}_{i_L}^\perp, \mathbf{P}_{i_{l+1}}^\perp = \mathbf{P}_{i_l}^\perp - \frac{\mathbf{P}_i^\perp \mathbf{a}_{i_l} \mathbf{a}_{i_l}^\top \mathbf{P}_{i_l}^\perp}{\left\| \mathbf{P}_{i_l}^\perp \mathbf{a} i_l \right\|_2^2}, \mathbf{P}_{i_1}^\perp = \mathbf{P}_i^\perp$

    total cost $\mathcal{O}\left(Ln^2\right)$
- Worst case complexity $\mathcal{O}\left(kmn^2\right)$ Assuming $k = \mathcal{O}(n/L)$
- In practice terminates much sooner than reaching the predetermined maximum number of iterations

Accelerated OLS

- Accelerated recursion and selection criterion

$$j_s = \arg\max_{j \in \mathcal{I} \backslash \mathcal{S}_i} \left\| \mathbf{q}_j \right\|_2$$

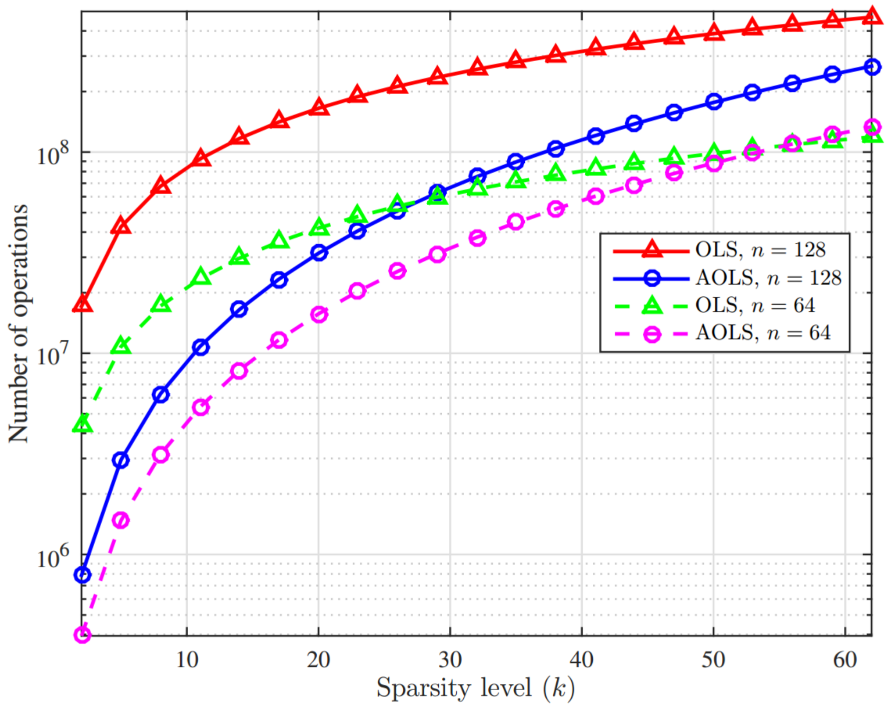
where

$$\mathbf{q}_j = \frac{\mathbf{a}_j^\top \mathbf{r}_i}{\mathbf{a}_j^\top \mathbf{t}} \mathbf{t}, \qquad \mathbf{t} = \mathbf{a}_j - \sum_{l=1}^{i} \frac{\mathbf{a}_j^\top \mathbf{u}_l}{\left\| \mathbf{u}_l \right\|_2^2} \mathbf{u}_l$$

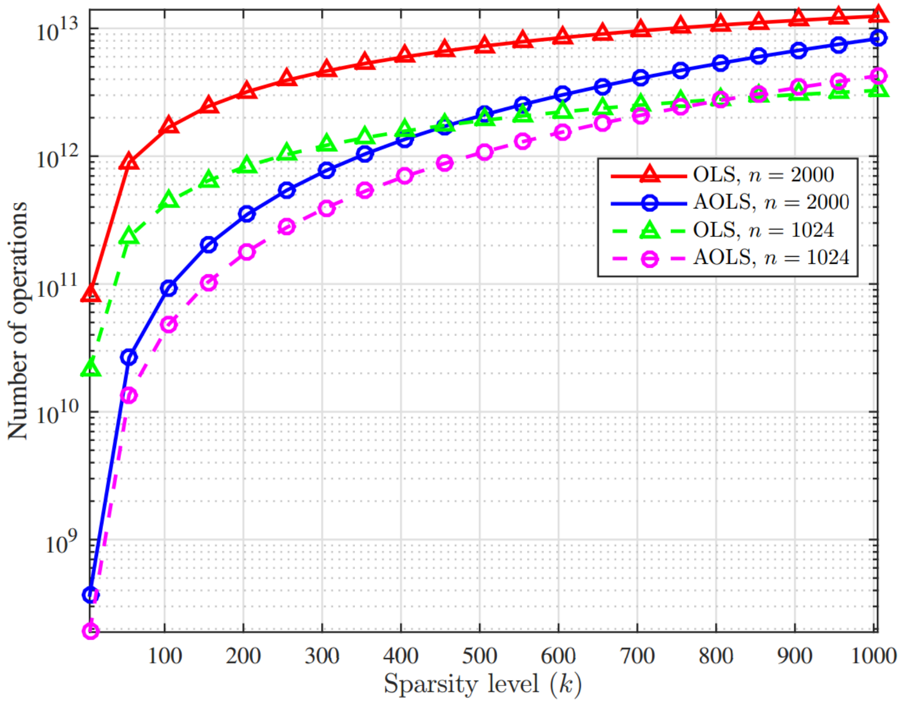
$$\mathbf{u}_{i+1} = \mathbf{q}_{j_s}, \qquad \mathbf{r}_{i+1} = \mathbf{r}_i - \mathbf{u}_{i+1}$$

- Worst case complexity $\mathcal{O}\left(k^2 mn\right)$ vs $\mathcal{O}\left(kmn^2\right)$

Comparison on required number of operations



$$m = 256$$



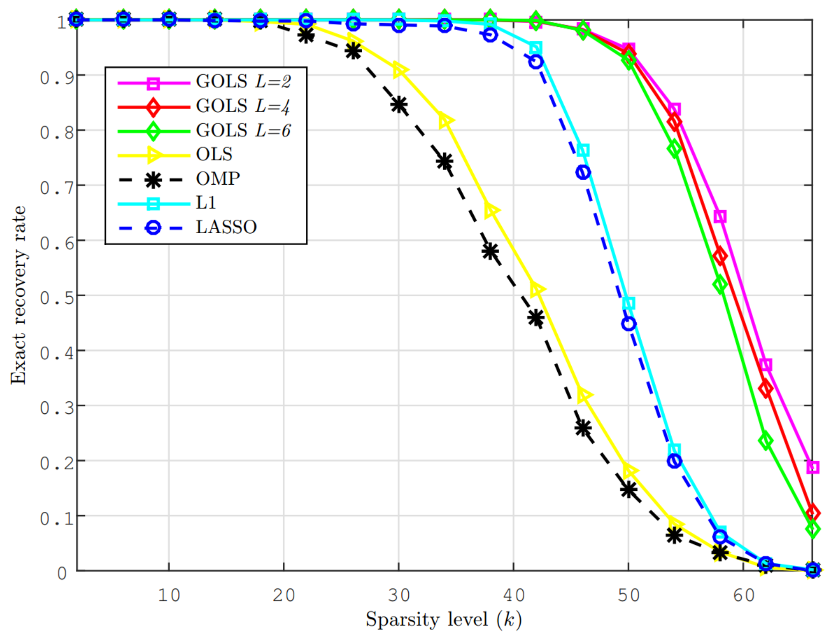
$$m = 2048$$

Results

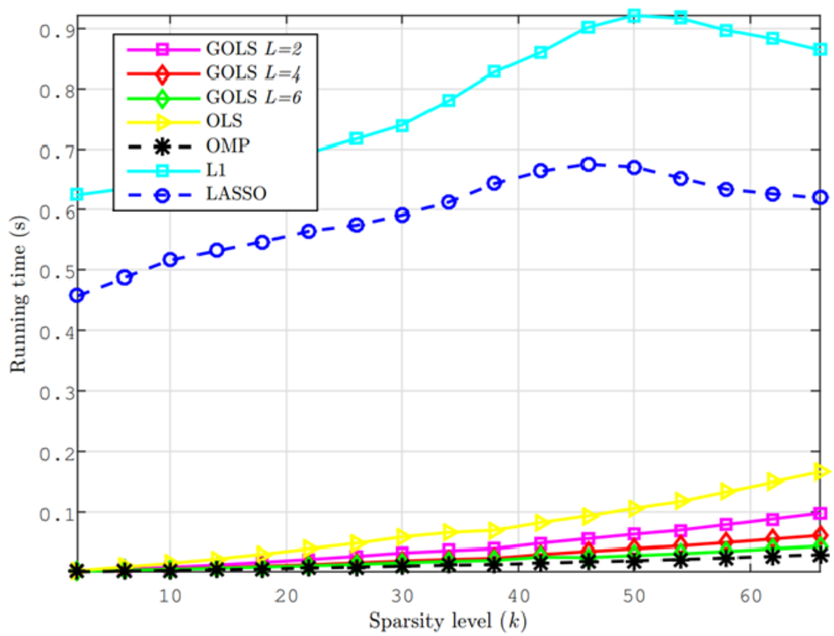
- Setting
  - Number of noiseless measurements $n = 128$
  - Dimension of unknown vector $m = 256$
  - Coefficient matrix $\mathbf{A} \sim \mathcal{N}(0, 1/n)$

- Benchmarking methods
  - OMP
  - OLS
  - LASSO
  - ℓ_1-Minimization via CVX
  - Generalized OLS with $L = 2, 4, 6$

Normally Distributed Sparse Vector

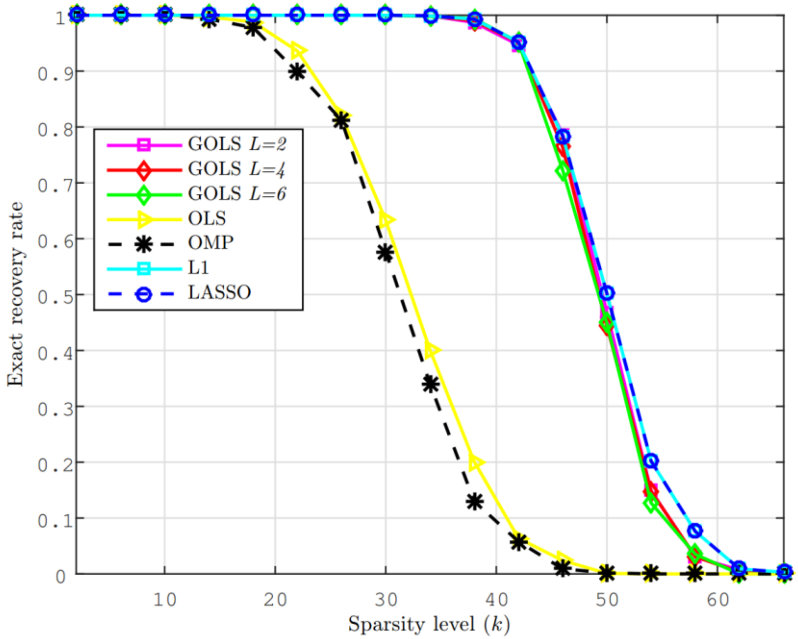


(a) Exact Recovery Rate

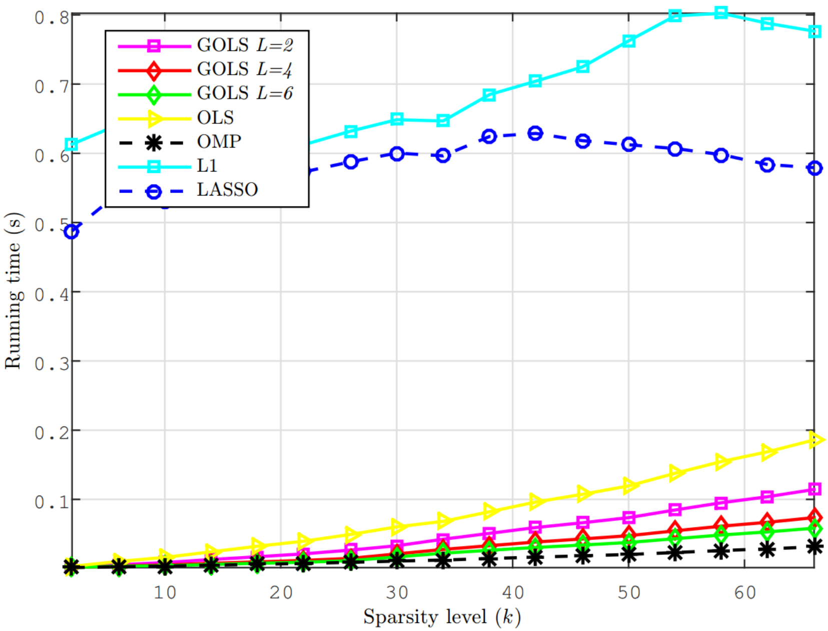


(b) Running time

$\{\pm 1, \pm 3\}$ -Valued Sparse Vector

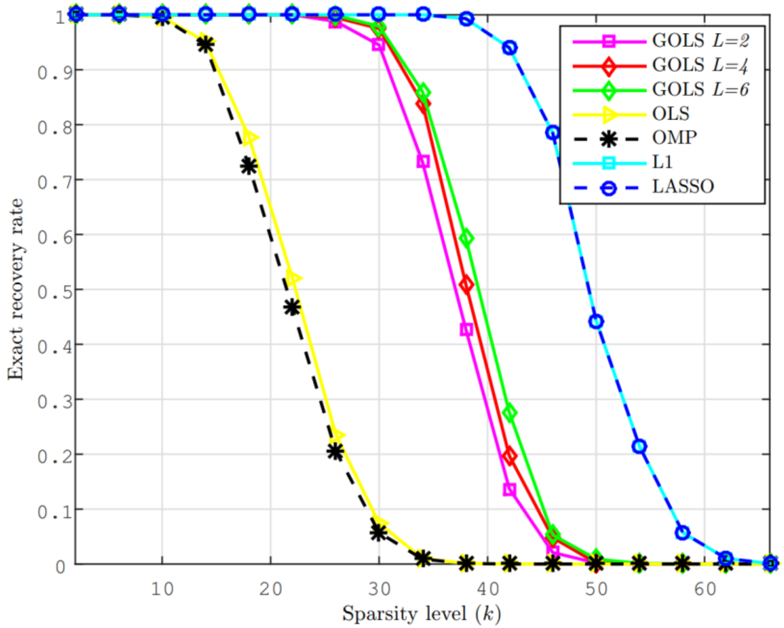


(a) Exact Recovery Rate

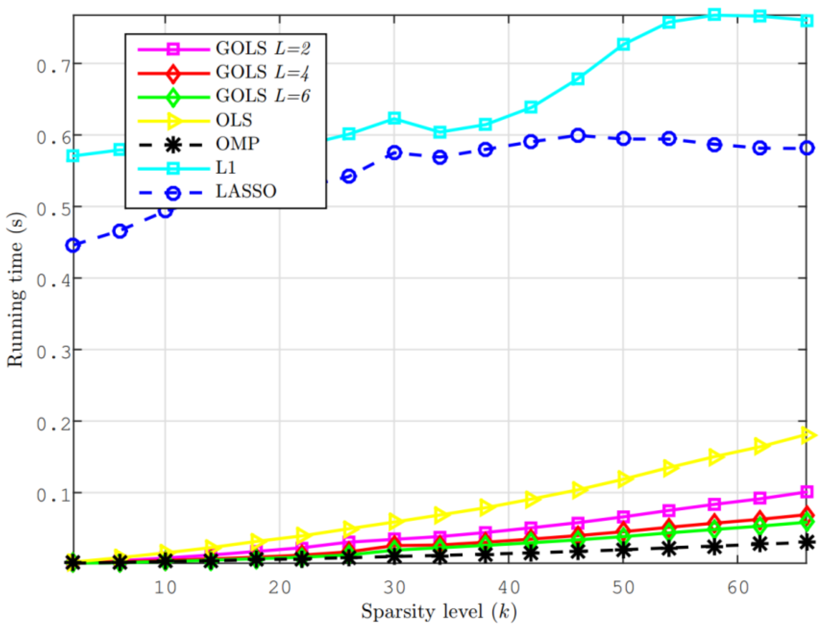


(b) Running time

$\{\pm 1\}$-Valued Sparse Vector



(a) Exact Recovery Rate



(b) Running time

Conclusion

- Sampling requirements of OLS for perfect recovery
- Improved OLS-based schemes
- Performance gain while being computationally more efficient than LASSO and L_1
- Exploring the case of correlated matrices

Thank you for your Attention!

Appendix Slides

Toward Improved OLS

- $\mathbf{B}_i$ the sub-matrix of $\mathbf{A}$ constructed by selecting of i its columns

- $\mathbf{B}_i^\dagger = \left(\mathbf{B}_i^\top \mathbf{B}_i\right)^{-1} \mathbf{B}_i^\top$ pseudo-inverse of $\mathbf{B}_i$

- $\mathbf{P}_i = \mathbf{B}_i \mathbf{B}_i^\dagger$ the projection matrix onto the span of the columns of $\mathbf{B}_i$, and $\mathbf{P}_i^\perp = \mathbf{I} - \mathbf{P}_i$

Toward Improved OLS

$$
\begin{aligned}
\mathbf{P}_{i+1} &= \mathbf{B}_{i+1}\left(\mathbf{B}_{i+1}^\top \mathbf{B}_{i+1}\right)^{-1}\mathbf{B}_{i+1}^\top \\
&= \begin{bmatrix}\mathbf{B}_i & \mathbf{a}\end{bmatrix}\begin{bmatrix}\mathbf{B}_i^\top \mathbf{B}_i & \mathbf{B}_i^\top \mathbf{a} \\ \mathbf{a}^\top \mathbf{B}_i & \mathbf{a}^\top \mathbf{a}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{B}_i^\top \\ \mathbf{a}^\top\end{bmatrix} \\
&\overset{(a)}{=} \begin{bmatrix}\mathbf{B}_i & \mathbf{P}_i^\perp \mathbf{a}\end{bmatrix}\begin{bmatrix}\left(\mathbf{B}_i^\top \mathbf{B}_i\right)^{-1} & 0 \\ 0 & \left(\mathbf{a}^\top \mathbf{P}_i^\perp \mathbf{a}\right)^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{B}_i^\top \\ \mathbf{a}^\top \mathbf{P}_i^\perp\end{bmatrix} \\
&\overset{(b)}{=} \mathbf{P}_i + \frac{\mathbf{P}_i^\perp \mathbf{a}\mathbf{a}^\top \mathbf{P}_i^\perp}{\left\|\mathbf{P}_i^\perp \mathbf{a}\right\|_2^2}
\end{aligned}
$$

(a) $$\begin{bmatrix}\mathbf{A} & \mathbf{E} \\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{I} & -\mathbf{A}^{-1}\mathbf{E} \\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I}\end{bmatrix}$$

$$\mathbf{A} = \mathbf{B}_i^\top \mathbf{B}_i,\ \mathbf{E} = \mathbf{B}_i^\top \mathbf{a},\ \mathbf{C} = \mathbf{a}^\top \mathbf{B}_i,\ \mathbf{D} = \mathbf{a}^\top \mathbf{a},\ \mathbf{\Delta} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{E}$$

(b) Idempotent property $\mathbf{P}_i^\perp = {\mathbf{P}_i^\perp}^\top = {\mathbf{P}_i^\perp}^2$

Selecting new indices

- Equivalently $\mathbf{P}_{i+1}^{\perp} = \mathbf{P}_{i}^{\perp} - \frac{\mathbf{P}_{i}^{\perp}\mathbf{a}\mathbf{a}^{\top}\mathbf{P}_{i}^{\perp}}{\left\|\mathbf{P}_{i}^{\perp}\mathbf{a}\right\|_{2}^{2}}$

- Following the recursive relation and idempotent property

$$
\begin{aligned}
j_s &= \operatorname*{argmin}_{j\in\mathcal{I}} \left\|\mathbf{y} - \mathbf{A}_{\mathcal{S}_{i-1}\cup\{j\}}\mathbf{A}^{\dagger}_{\mathcal{S}_{i-1}\cup\{j\}}\mathbf{y}\right\|_2 \\
&= \operatorname*{argmin}_{j\in\mathcal{I}} \left\|\left(\mathbf{I} - \mathbf{P}_i\right)\mathbf{y}\right\|_2^2 \\
&= \operatorname*{argmin}_{j\in\mathcal{I}} \mathbf{y}^{\top}\mathbf{y} - \mathbf{y}^{\top}\mathbf{P}_i\mathbf{y} - \mathbf{y}^{\top}\mathbf{P}_i^{\top}\mathbf{y} + \mathbf{y}^{\top}\mathbf{P}_i^{\top}\mathbf{P}_i\mathbf{y} \\
&= \operatorname*{argmin}_{j\in\mathcal{I}} \mathbf{y}^{\top}\mathbf{y} - \mathbf{y}^{\top}\mathbf{P}_i\mathbf{y} \\
&= \operatorname*{argmax}_{j\in\mathcal{I}} \mathbf{y}^{\top}\mathbf{P}_{i-1}\mathbf{y} + \mathbf{y}^{\top}\frac{\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\mathbf{a}_j^{\top}\mathbf{P}_{i-1}^{\perp}}{\left\|\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\right\|_2^2}\mathbf{y} \\
&= \operatorname*{argmax}_{j\in\mathcal{I}} \frac{\left\|\mathbf{y}^{\top}\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\right\|_2^2}{\left\|\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\right\|_2^2} = \operatorname*{argmax}_{j\in\mathcal{I}} \left|\mathbf{y}^{\top}\frac{\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j}{\left\|\mathbf{P}_{i-1}^{\perp}\mathbf{a}_j\right\|_2}\right|
\end{aligned}
$$

Accelerated OLS

Table I. Computational Complexity of OLS and Accelerated OLS

Algorithm	Number of arithmetic operations
OLS	$4n\left(km-\frac{k(k-1)}{2}\right)+\frac{5}{2}nk+2n^2\left(km-\frac{k(k-1)}{2}\right)+\frac{7}{2}n^2k$
Accelerated OLS	$5n\left(km-\frac{k(k-1)}{2}\right)+nk+2nk(k+1)(m+1)-\frac{2}{3}k(k+1)(2k+1)$