Matrix Sketching for Alternating Direction Method of Moments Optimization

Daniel J. McDonald
Indiana University, Bloomington
mypage.iu.edu/~dajmcdon

Nonlinear Dimension Reduction (SDSS)
17 May 2018

EXPLICIT+IMPLICIT DIMENSION REDUCTION

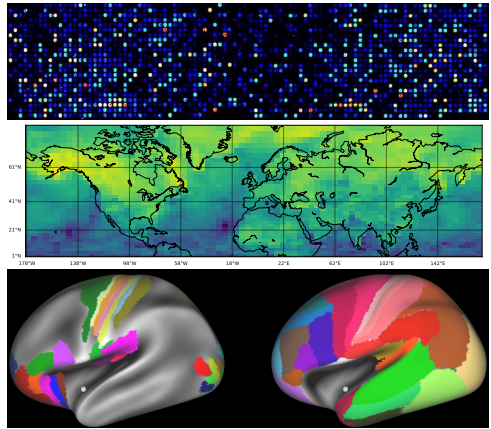
Modern statistical applications — genomics, neural image analysis, text analysis, weather prediction — have large numbers of covariates p

Also frequently have lots of observations n.

Need algorithms which can handle these kinds of data sets. With good statistical properties

MOTIVATING EXAMPLES

1. Localizing groups of genes that predict disease
2. Finding global temperature trends using satellite imagery
3. Detecting outliers in fMRI scans



1. Sparse PCR (BT04, PBHT08, DM17)

$$\widehat{V} = \underset{V \in \mathcal{F}^d}{\operatorname{argmin}} -\frac{1}{n} \operatorname{tr}(X^\top X V) + \lambda \sum_{ij} |V_{ij}|$$

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\| Y - X\widehat{V}\theta \right\|_2^2.$$

2. ℓ_1-trend filtering (KKBG09, TT12, T14, MK18)

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} -\mathcal{L}(Y \mid \theta) + \lambda \left\| D\theta \right\|_1$$

3. PCA leverage (MNECL16, MMD18)

$$\widehat{U} = \underset{U \in \mathcal{F}^d}{\operatorname{argmin}} -\frac{1}{n} \operatorname{tr}\left(X X^\top U \right) + \lambda \left\| D \sum_j |V_{ij}| \right\|_1$$

GENERIC CONVEX OPTIMIZATION

Many estimators have the form:

$$\min_x f(x) + g(x)$$

Consider $f(x)$ as the negative log-likelihood and $g(x)$ as some kind of penalty that preferences useful structure.

- The negative likelihood is convex and differentiable.
- The penalty may be neither.
- Sometimes relax the penalty to something convex to get approximate structure:

Example:

$$\min_\beta \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \longrightarrow \min_\beta \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

ALTERNATING DIRECTION METHOD OF MULTIPLIERS

One way to solve optimization problems like this is to restate the problem

Original	Equivalent
$\min_{x} \quad f(x) + g(x)$	$\min_{x,z} \quad f(x) + g(z)$ s.t. $\quad x - z = 0$

Then, iterate the following with $\rho > 0$

$$
\begin{aligned}
x &\leftarrow \operatorname*{argmin}_{x} f(x) + \frac{\rho}{2} \|x - z + u\|_2^2 \\
z &\leftarrow \operatorname*{argmin}_{z} g(z) + \frac{\rho}{2} \|x - z + u\|_2^2 \\
u &\leftarrow u + x - z
\end{aligned}
$$

WHY WOULD YOU DO THIS?

- It decouples f and g: this can be easier
- If f and g have the right structure, the individual updates can be parallelized
- The algorithm converges under very general conditions
- There are often many ways to decouple a problem

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- The individual minimizations don't have to be solved in closed form

Example:

$$\begin{aligned}
\beta &\leftarrow (X^\top X + \rho I)^{-1}(X^\top Y + \rho(\alpha - u)) \\
\alpha &\leftarrow \mathcal{S}_{\lambda/\rho}(\beta + u) \\
u &\leftarrow u + \beta - \alpha
\end{aligned}$$

$$[\mathcal{S}_a(b)]_k = \text{sgn}(b_k)(|b_k| - a)_+$$

CONDITIONS FOR CONVERGENCE

- When the updates are exact (as with lasso), all you need for convergence is
  1. f, g are convex, extended real valued.
  2. $f(x) + g(z) + u^\top (x - z)$ has a saddle point.
- The convergence rate is not well understood.
- It turns out, you can solve the minimizations approximately.

$$\sum_{k=1}^{\infty} \left\| \Pi(y^k) - \widetilde{\Pi}(y^k) \right\|_2 < \infty$$

Source: Eckstein and Bertsekas 1992

Why approximate?

- In our Example, the first step involved a matrix inversion $(X^\top X + \rho I)^{-1}$
- The same is true for the real data cases above: we need matrix decompositions/inversions.
- Focus on two methods of "approximate eigendecomposition"
  1. Nyström extension
  2. Column sampling

A QUICK SKETCH OF THE INTUITION

- Both methods fall into a larger class
- Suppose we want to approximate $S = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$
- S is symmetric and positive semi-definite
- Choose t and form a "sketching" matrix $\Phi \in \mathbb{R}^{p \times t}$
- Then write

$$S \approx (S\Phi)(\Phi^\top S\Phi)^\dagger (S\Phi)^\top$$

SPECIAL CASES

- Nyström and column sampling correspond to particular Φ
- But they are easy to implement without extra multiplications
- Randomly choose t entries in $\{1, \dots, p\}$ and
- Then partition the matrix so the selected portion is S_{11}

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

Nyström

$$S \approx \begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix} S_{11}^{\dagger} \begin{bmatrix} S_{11} & S_{12} \end{bmatrix}$$

Column sampling

$$S \approx U \left(\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix} \right) \Lambda \left(\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix} \right) U \left(\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix} \right)^{\top}$$

A SHORT LIST OF RELATED WORK

- Rokhlin, Tygert, (2008).
- Drineas, Mahoney, Muthukrishnan, Sarlós (2011).
- Halko, Martinsson, Tropp (2011).
- Gittens, Mahoney (2013).
- Woodruff (2014).
- Pourkamali (2014).
- Homrighausen, McDonald (2016).
- Wang, Gittens, Mahoney (2017)

ADMM FOR GENETICS

Goal is to find clusters of genes which predict the response.

The approach is semi-supervised: like PCR, but we assume that the eigenvectors are "row sparse".

1. This allows for consistent estimation when $p \gg n$.
2. Matches our assumption that only a few genes are predictive: $\|V_i\|_2 = 0 \Rightarrow \beta_i = 0$.

$$
\begin{aligned}
V &\leftarrow \Pi_{\mathcal{F}^d}\left(Y - U + \frac{1}{n\rho} X^\top X\right) \\
Y &\leftarrow \mathcal{S}_{\lambda/\rho}(V + U) \\
U &\leftarrow U + V - Y
\end{aligned}
$$

PROJECTING ONTO THE FANTOPE

Given an eigen decomposition of $A = \sum_i \gamma_i a_i a_i^\top$.

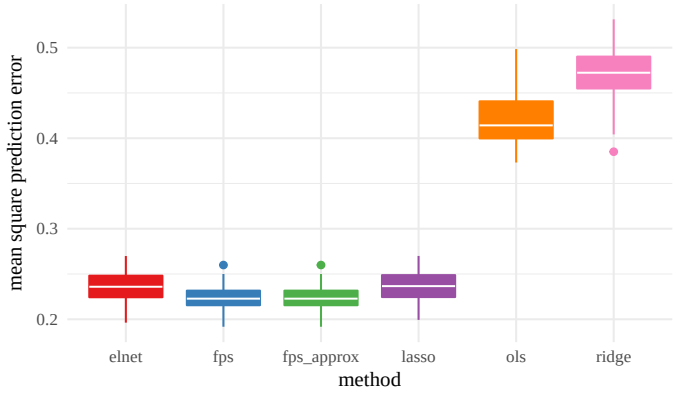
$$\Pi_{\mathcal{F}^d}(A) = \sum_i \gamma_i^+(\theta) a_i a_i^\top$$

$$\gamma_i^+(\theta) = \min(\max(\gamma_i - \theta,\ 0),\ 1), \qquad \theta \text{ s.t. } \sum_i \gamma_i^+(\theta) = d$$

- The γ-θ stuff solves a monotone, piecewise linear equation.
- For our data, S is $10^5 \times 10^5$.
- And we have to do the decomposition at every iteration.
- The fMRI outlier detection problem involves a similar step but the matrix is $n_{\text{voxels}} \times n_{\text{voxels}}$.

Source: Vu, Cho, Lei, and Rohe 2013

SIMULATION FOR GENES



$n = 1000,\ p = 2000$, 100 true genes, 3 principal components

A NOD TOWARD THEORY

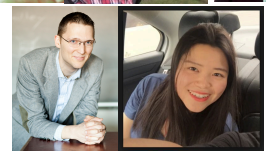
- At each iteration, we use column sampling with $t = 1000$
- Could also use "Nyström approximation"
- These approximations are accurate: something like $O(\epsilon^{-1})$ if $t = \Omega((1-\epsilon)^{-2})$
- Need $t \to p$ as $k \to \infty$ to guarantee convergence, though seems unnecessary in practice.

Source: See Alex's work as well as that of Mahoney, Woodroofe, Drineas, others

Conclusion

- This talk summarized some methodology for analyzing large data sets.
- Making these methods work requires computational approximations.
- These ideas combined algorithmic dimension reduction with nonlinear dimension reduction.
- Current work develops more detailed theoretical results for these methods.

COLLABORATORS AND FUNDING



Institute for
New Economic
Thinking