A Decomposition Algorithm for the Sparse Generalized Eigenvalue Problem

Ganzhao Yuan[1,3], Li Shen[2], Wei-Shi Zheng[3]

[1] Center for Quantum Computing, Peng Cheng Laboratory
[2] Tencent AI Lab [3] Sun Yat-sen University

Apirl. 2, 2019

Outline of this talk

- The Sparse Generalized Eigenvalue Problem
- The Proposed Decomposition Algorithm
- Existing Sparse Optimization Methods
- Theoretical Analysis
- Experiments

The Sparse Generalized Eigenvalue Problem

The Sparse Generalized Eigenvalue Problem

$$\min_{\mathbf{x}\neq \mathbf{0}, \|\mathbf{x}\|_0 \leq s} f(\mathbf{x}) \triangleq \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{C} \mathbf{x}}.$$

- Statistical learning models
  1. Principle Component Analysis (PCA)

     $$\min_{\mathbf{x}\neq 0} \frac{-\mathbf{x}^T \Sigma \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

  2. Fisher Discriminant Analysis (FDA)

     $$\min_{\mathbf{x}\neq 0} \frac{-\mathbf{x}^T ((\boldsymbol{\mu}_{(1)} - \boldsymbol{\mu}_{(2)})(\boldsymbol{\mu}_{(1)} - \boldsymbol{\mu}_{(2)})^T)\mathbf{x}}{\mathbf{x}^T (\Sigma_{(1)} + \Sigma_{(2)})\mathbf{x}}$$

  3. Canonical Correlation Analysis (CCA)

     $$\min_{\mathbf{x}} \frac{-\mathbf{x}^T \begin{pmatrix} \mathbf{0} & \Sigma_{xy} \\ \Sigma_{yx} & \mathbf{0} \end{pmatrix} \mathbf{x}}{\mathbf{x}^T \begin{pmatrix} \Sigma_{xx} & \mathbf{0} \\ \mathbf{0} & \Sigma_{yy} \end{pmatrix} \mathbf{x}}$$

- Applications: object recognition, object dection, visual tracking, pixel selection, text summarization

The Proposed Algorithm

The Concept of Block-k Optimality

- Optimization Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; F(\mathbf{x})$$

- Variable Splitting: $B \in \mathbb{N}^k$ is a subset of $\{1, 2, ..., n\}$. $N \triangleq \{1, ..., n\} \setminus B$, $\mathbf{x} = \mathbf{I}\mathbf{x} = (\mathbf{U}_B\mathbf{U}_B^T + \mathbf{U}_N\mathbf{U}_N^T)\mathbf{x} = \mathbf{U}_B\mathbf{x}_B + \mathbf{U}_N\mathbf{x}_N$.
- Block-k Optimal Solution

$$\mathcal{P}(B, \mathbf{x}) \triangleq \arg\min_{\mathbf{x}_B} F(\mathbf{U}_B\mathbf{x}_B + \mathbf{U}_N\mathbf{x}_N)$$

$$\{\bar{\mathbf{x}}_B = \mathcal{P}(B, \bar{\mathbf{x}}) \text{ for all } |B| = k\} \;\Leftrightarrow\; \bar{\mathbf{x}} \text{ is the block-k optimal solution}$$

- Block-k Optimality Measure [$\{\mathcal{B}_{(i)}\}_{i=1}^{C_n^k}$ denotes all the possible combinations]

$$\mathcal{M}(\mathbf{x}) \triangleq \frac{1}{C_n^k} \sum_{i=1}^{C_n^k} \|\mathcal{P}(\mathcal{B}_{(i)},\ \mathbf{x}) - \mathbf{x}_{\mathcal{B}_{(i)}}\|_2^2$$

$$\mathcal{M}(\bar{\mathbf{x}}) = 0 \Leftrightarrow \bar{\mathbf{x}} \text{ is the block-k optimal solution}$$

.

Optimization Problem:

$$\min_{\mathbf{x}\neq\mathbf{0},\ \|\mathbf{x}\|_0\leq s} f(\mathbf{x}) \triangleq \frac{\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}}{\frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}}$$

We define $h(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$, $\;g(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}$.

$$\begin{aligned} h(\mathbf{x}_B,\mathbf{x}_N) &= \tfrac{1}{2}\mathbf{x}_B^T\mathbf{A}_{BB}\mathbf{x}_B + \tfrac{1}{2}\mathbf{x}_N^T\mathbf{A}_{NN}\mathbf{x}_N + \langle \mathbf{x}_B, \mathbf{A}_{BN}\mathbf{x}_N\rangle, \\ g(\mathbf{x}_B,\mathbf{x}_N) &= \tfrac{1}{2}\mathbf{x}_B^T\mathbf{C}_{BB}\mathbf{x}_B + \tfrac{1}{2}\mathbf{x}_N^T\mathbf{C}_{NN}\mathbf{x}_N + \langle \mathbf{x}_B, \mathbf{C}_{BN}\mathbf{x}_N\rangle. \end{aligned}$$

The Proposed Decomposition Algorithm

Input: k, θ, a feasible solution $\mathbf{x}^0$, $t = 0$

while *not converge* **do**

S1 Find a working set B of size k. Denote $N \triangleq \{1, ..., n\} \setminus B$.

S2 Solve the following subproblem *globally*:

$$\mathbf{x}_B^{t+1} \Leftarrow \arg\min_{\mathbf{x}_B} \frac{h(\mathbf{x}_B, \mathbf{x}_N^t) + \frac{\theta}{2}\|\mathbf{x}_B - \mathbf{x}_B^t\|_2^2}{g(\mathbf{x}_B, \mathbf{x}_N^t)} \quad s.t. \ \|\mathbf{x}_B\|_0 + \|\mathbf{x}_N^t\|_0 \leq s \tag{1}$$

S2 Increment t by 1

end

Algorithm 1: The Proposed Decomposition Algorithm

Remarks on the Decomposition Algorithm

1. A new proximal strategy (only applied to the numerator) $\Rightarrow$ sufficient descent and global convergence
2. When $k = n$, the subproblem reduces to the original problem.
3. Finding the working set
   1. **Random strategy**. Select one combination (which contains k coordinates) from the whole working set $\{\mathcal{B}_{(i)}\}_{i=1}^{C_n^k}$ uniformly.
   2. **Swapping strategy**. Pick the top pairs of coordinates that lead to the greatest descent by measuring $\mathbf{D} \in \mathbb{R}^{|\mathcal{S}(\mathbf{x})| \times |\mathcal{Z}(\mathbf{x})|}$:

$$\mathbf{D}_{i,j} = \min_{\beta} \; f(\mathbf{x}^t + \beta \mathbf{e}_i - \mathbf{x}_j^t \mathbf{e}_j) - f(\mathbf{x}^t).$$

with $\mathcal{S}(\mathbf{x}) \triangleq \{i \mid \mathbf{x}_i \neq 0\}$ and $\mathcal{Z}(\mathbf{x}) \triangleq \{j \mid \mathbf{x}_j = 0\}$.

Remarks on the Decomposition Algorithm

4. The subproblem reduces to the following problem:

$$\min_{\mathbf{z}\in\mathbb{R}^k,\ \|\mathbf{z}\|_0\leq q} \ p(\mathbf{z}) \triangleq \frac{\frac{1}{2}\mathbf{z}^T\bar{\mathbf{Q}}\mathbf{z}+\bar{\mathbf{p}}^T\mathbf{z}+\bar{w}}{\frac{1}{2}\mathbf{z}^T\bar{\mathbf{R}}\mathbf{z}+\bar{\mathbf{c}}^T\mathbf{z}+\bar{v}}$$

Our solution:

- We consider the following problem:

$$\min_{\mathbf{z}\in\mathbb{R}^k} \ p(\mathbf{z}),\ \ s.t.\ \mathbf{z}_K = \mathbf{0}$$

  where K has $\sum_{i=0}^{q} C_k^i$ possible choices.
- It reduces to the following problem:

$$\min_{\mathbf{y}} \ \mathcal{L}(\mathbf{y}) \triangleq \frac{\frac{1}{2}\mathbf{y}^T\mathbf{Q}\mathbf{y}+\mathbf{p}^T\mathbf{y}+w}{\frac{1}{2}\mathbf{y}^T\mathbf{R}\mathbf{y}+\mathbf{c}^T\mathbf{y}+v}$$

$$\min_{\mathbf{y}} \ \mathcal{L}(\mathbf{y}) \triangleq \frac{\frac{1}{2}\mathbf{y}^T\mathbf{Q}\mathbf{y} + \mathbf{p}^T\mathbf{y} + w}{\frac{1}{2}\mathbf{y}^T\mathbf{R}\mathbf{y} + \mathbf{c}^T\mathbf{y} + v}$$

$$\mathcal{J}(\alpha) = 0, \ \text{with } \mathcal{J}(\alpha) \triangleq \min_{\mathbf{y}} \ u(\mathbf{y}) - \alpha q(\mathbf{y})$$

We have the following results.

1. It holds that: $\lambda_{\min}(\mathbf{Z}) \leq \min_{\mathbf{y}} \ \mathcal{L}(\mathbf{y}) < \lambda_{\min}(\mathbf{O})$.
2. The dual equivalent function $\mathcal{J}(\alpha)$ is monotonically decreasing on the range $\lambda_{\min}(\mathbf{Z}) \leq \alpha < \lambda_{\min}(\mathbf{O})$.
3. The optimal solution $\mathbf{y}^* = \mathbf{R}^{-1/2}(\mathbf{u}^* - \mathbf{R}^{-1/2}\mathbf{c})$, with $\mathbf{u}^* = -(\mathbf{O} - \alpha^*\mathbf{I})^{-1}\mathbf{g}$ and α^* being the unique root of the equation $\mathcal{J}(\alpha) = 0$ on the range $\lambda_{\min}(\mathbf{Z}) \leq \alpha < \lambda_{\min}(\mathbf{O})$.

Contributions

Contributions of this paper:

1. A decomposition algorithm, it finds stronger stationary points.
2. Two strategies to find the working set
3. Two methods to solve the subproblem
4. A convergence analysis for 'DEC'
5. 'DEC' consistently outperforms existing methods

Existing Sparse Optimization Methods

Existing Sparse Optimization Methods

Optimization Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}), \;\; s.t. \; \|\mathbf{x}\|_0 \leq s$$

- $f(\mathbf{x})$: smooth, convex, its gradient is L-Lipschitz continuous
- Existing methods
  1. Relaxed approximation method
  2. Greedy pursuit method
  3. Combinatorial search method
  4. Gradient projection method

Comparisons of the methods

1. Relaxed approximation method
    - convex: $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1$, $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_{\text{tok}-k}$
    - nonconvex: $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$, reweighted ℓ_1 norm

   $\Rightarrow$ our method directly controls the sparsity of the solution

2. Greedy pursuit method
    - the solution MUST be initialized to zero
    - $S = \emptyset$, $S = S \cup i_1$, $\min_{\mathbf{x}_S}\ f(\mathbf{x})$, $S = S \cup i_2$, $\min_{\mathbf{x}_S}\ f(\mathbf{x}), \ldots$

   $\Rightarrow$ our method is a greedy coordinate descent algorithm

Comparisons of the methods

3. Combinatorial search method: global optimization methods
   - cutting plane methods
   - branch-and-cut methods

   $\Rightarrow$ our method leverages the effectiveness of combinatorial search method methods

4. Gradient projection method
   - $\mathbf{x}^{k+1} = \text{Proj}_s(\mathbf{x}^k - \gamma \nabla f(\mathbf{x}^k))$
   - with $\text{Proj}_s(\mathbf{a}) = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{a}\|_2^2, \ s.t. \ \|\mathbf{x}\|_0 \leq s$

   $\Rightarrow$ our method significantly outperforms gradient projection method

Theoretical Analysis

Optimality Analysis

Basic Stationary Point

$\breve{\mathbf{x}}$ is called a basic stationary point if the following holds.
$\breve{\mathbf{x}} = \arg\min_{\mathbf{y}} \; f(\mathbf{y}), \; s.t. \; \mathbf{y}_Z = \mathbf{0}, \; |S| \leq s$, where $Z \triangleq \{i | \breve{\mathbf{x}}_i = 0\}$ and $S \triangleq \{i | \breve{\mathbf{x}}_i \neq 0\}$.

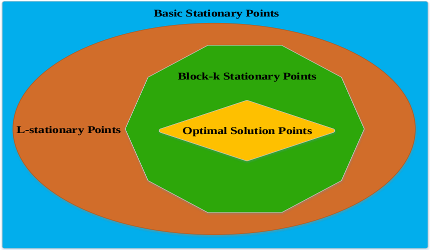
L-Stationary Point

A solution $\grave{\mathbf{x}}$ is an L-stationary point if it holds that:
$\grave{\mathbf{x}} = \mathrm{Proj}_s(\grave{\mathbf{x}} - \nabla f(\grave{\mathbf{x}})/L)$.

Block-k Stationary Point

A solution $\bar{\mathbf{x}}$ is a block-k stationary point if it holds that:
$\bar{\mathbf{x}} \in \arg\min_{\mathbf{z} \in \mathbb{R}^n} \; \mathcal{P}(\mathbf{z}; \bar{\mathbf{x}}, B) \triangleq \{F(\mathbf{z}), \; s.t. \; \mathbf{z}_N = \bar{\mathbf{x}}_N\}, \; \forall |B| = k, \; N \triangleq \{1, ..., n\} \setminus B$.

Optimality Hierarchy



Relations between the three types of stationary point.

We have the following optimality hierarchy [a]:

$\boxed{\text{Basic Stat. Point}} \overset{(1)}{\Leftarrow} \boxed{L\text{-Stat. Point}} \overset{(2)}{\Leftarrow} \boxed{\text{Block-2 Stat. Point}} \overset{(3)}{\Leftarrow} \boxed{\text{Block-3 Stat. Point}} \Leftarrow \ldots \Leftarrow \boxed{\text{Block-}n\text{ Stat. Point}} \overset{(4)}{\Leftrightarrow} \boxed{\text{Optimal Point}}$.

[a] Ganzhao Yuan, Li Shen, Wei-Shi Zheng. A Hybrid Method of Combinatorial Search and Coordinate Descent for Discrete Optimization. arXiv preprint, 2017. url: `https://arxiv.org/abs/1706.06493`.

A Running Example

Optimization Problems:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{x}^T\mathbf{p}, \ s.t. \ \|\mathbf{x}\|_0 \leq s$$

$n = 6$, $\mathbf{Q} = \mathbf{c}\mathbf{c}^T + \mathbf{I}$, $\mathbf{p} = \mathbf{1}$, $\mathbf{c} = [1 \ 2 \ 3 \ 4 \ 5 \ 6]^T$, $s = 4$.

Number of points satisfying optimality conditions.

Basic-Stat.	L-Stat.	Block-1 Stat.	Block-2 Stat.	Block-3 Stat.	Block-4 Stat.	Block-5 Stat.	Block-6 Stat.
57	56		3	1	1	1	1

Global Convergence Properties.

Assume that the subproblem is solved globally. We have the following results.

1. When the random strategy is used to find the working set, we have $\lim_{t\to\infty} \mathbb{E}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|] = 0$ and Algorithm 1 converges to the block-k stationary point in expectation.
2. When the swapping strategy is used to find the working set with $k \geq 2$, we have $\lim_{t\to\infty} \|\mathbf{x}^{t+1} - \mathbf{x}^t\| = 0$ and Algorithm 1 converges to the block-2 stationary point deterministically.

Experiments

Sparse Generalized Eigenvalue Problem

Optimization problems:

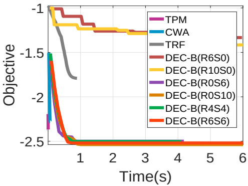
$$\min_{\mathbf{x}} \ \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{C}\mathbf{x}}, \ s.t. \ \|\mathbf{x}\|_0 \leq s$$

Compared Methods:

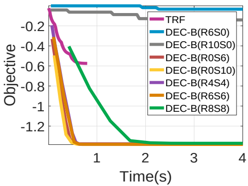
- Truncated Power Method (TPM) [Yuan & Zhang, JMLR2013]
- Coordinate-Wise Algorithm (CWA) [Beck & Vaisbourd, JOTA2016]
- Truncated Rayleigh Flow (TRF) [Tan, et al., JRSS2018]
- Quadratic Majorization Method (QMM) [Song, et al., TIP2015]
- Proposed Decomposition Method [1] (DEC-R*i*-G*j*) [This Paper]

[1]Selecting *i* coordinate using the **R**andom strategy and *j* coordinates using the **S**wapping strategy

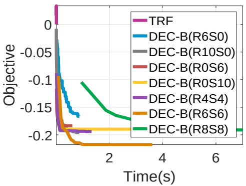
Convergence Behavior: i



(a) sparse PCA, 'rnd-2000'



(b) sparse FDA, 'rnd-2000'

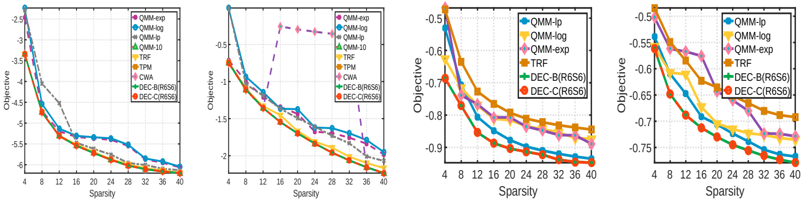


(c) sparse CCA, 'rnd-2000'

Conclusions:

1. {TPM,CWA,TRF} converge faster, but result in poor accuracy.
2. DEC-B(R10S0) achieves a lower objective value than DEC-B(R6S0). Larger k implies stronger stationary points.
3. The swapping strategy plays an indispensable role.

Experimental Results

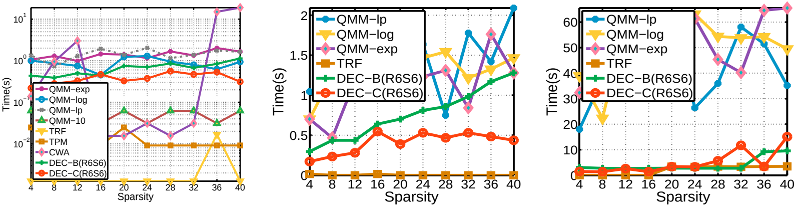


(a) a1a (b) w1a (c) a1a (d) w1a

Conclusions:

1. CWA is not stable (much worse results on 'w1a').
2. DEC achieves lowest objective values.
3. Both DEC-B and DEC-C perform similarly.

Computational Efficiency



(e) PCA (f) FDA (g) CCA

Conclusions:

1. DEC takes less than 15 seconds to converge in all our instances.
2. DEC is practical and it is much more efficient than QMM.

Thank You!