L_1 Regularized L_1-norm PCA

Xiao Ling

Virginia Commonwealth University

July 17, 2020

Outline

- Classical subspace estimation
  ▸ With Outliers ▸ L_1PCA
- Proposed L_1-norm PCA
- Proposed L_1 Regularized L_1-norm PCA
- Implementation on Spark
- Computational performance
- Results

Subspace Estimation Background

- Factorization method
  - Computationally expensive
  - not amenable to adaptive situation
- Manifold Optimization
  - no generic tools for theoretical analysis
- Nonparametric

L_2-norm based Subspace Estimation

Finding a best-fit subspace for n points can be formulated as following optimization problem

$$\min_{V,\alpha_i} \sum_{i=1}^{n} \|x_i - V\alpha_i\|_2^2$$

PCA disadvantages

- Sensitive to outliers
- Batch Mode
  - Impracticable for Big Data
  - Impracticable for Streaming

PCA disadvantages

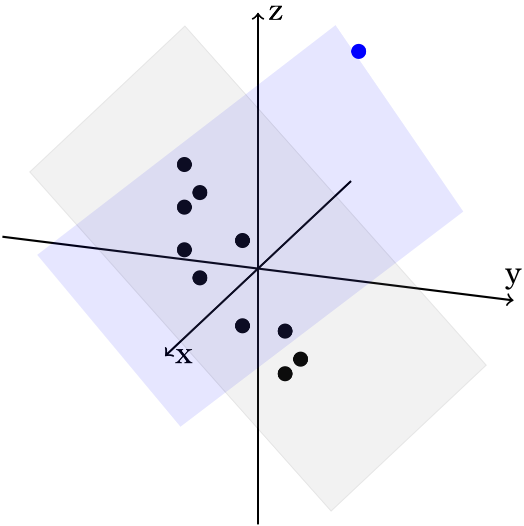

Figure: Data set for m=3 with points on a plane and a leverage point as outlier

The L_1-Norm Best-Fit Subspace Problem

L_1-norm subspace estimation is the essence of L_1-norm principal component analysis (PCA). Finding L_1-Norm best-fit subspace for points x_i, $i = 1, \ldots, n$ in m dimensions, we can solve

$$\min_{V,\alpha_i} \sum_{i=1}^{n} \|x_i - V\alpha_i\|_1$$

n = number of points, m = number of dimensions, q = dimension of subspace.

L_1-Norm Milestones

1. $q = 2$: Finding the best-fit line can be solved in polynomial time (Singleton 1940, Karst 1958, Guwirtz 1990, Megiddo and Tamir 1993).
2. $q = m - 1$: The best-fit hyperplane can be found by solving a small number of LPs (Martini and Schöbel 1998, Brooks and Dulá 2013).
3. $q = 1$: Finding the best-fit line is NP-hard (Gillis and Vavasis 2015).

L_1 Projection and Linear Programming

Given a point $x \in \mathbb{R}^m$ and a subspace $S = \{V\alpha | \alpha \in \mathbb{R}^q\}$, a projection of x onto S is given by an optimal solution to the following optimization problem:

$$\min_{V,\alpha} \|x - V\|_1$$

This problem can be converted into the following linear program:

$$LP(S,x) = \min_{V,\alpha} \sum_{j=1}^{m} (\lambda_j^+ + \lambda_j^-)$$

subject to

$$\begin{array}{rcl} V\alpha + \lambda^+ - \lambda^- & = & x \\ \lambda^+, \lambda^- & \geq & 0 \end{array}$$

The L_1-Norm Best-Fit Line Problem

$$\min_{v,\alpha_i} \sum_{i=1}^{n} \|x_i - v\alpha_i\|_1$$

The vector v determines the line through the origin. The α_i are the scaling factors that locate the projection of each point on the fitted line.

Estimator for the L_1-Norm Best-Fit Line Problem

Idea:

- Fix one of the m dimensions which will be "preserved" in all points' projections. Let this be $\hat{\jmath}$-th dimension.
- Without restricting the line that will be defined by the vector v, we can set $v_{\hat{\jmath}} = 1$.

From this we get $\alpha_i = x_{i\hat{\jmath}}$ for $i = 1, \ldots, n$ in the constraints and the formulation becomes:

$$z_{\hat{\jmath}} = \min_{\substack{v \in \Re^m, v_{\hat{\jmath}}=1, \\ \lambda^+, \lambda^- \in (\Re^{m \times n})^+}} \sum_{i=1}^{n} \sum_{j=1}^{m} (\lambda_{ij}^+ + \lambda_{ij}^-),$$

subject to:

$$v_j x_{i\hat{\jmath}} + \lambda_{ij}^+ - \lambda_{ij}^- \quad = \quad x_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, m; \; j \neq \hat{\jmath}.$$

A Solution to the Estimator LP

Proposition

A solution to the LP for a given component $\hat{\jmath}$ is:

$$v_j^* = \frac{\tilde{x}_j}{\tilde{x}_{\hat{\jmath}}},$$

for $j = 1, \dots, m$, where $\tilde{x}_j$ is j^{th} coordinate of a given point satisfying

$$\sum_{i:(i)>(\tilde{\imath}_j)} |x_{i\hat{\jmath}}| < \frac{1}{2}\sum_{i=1}^{n} |x_{i\hat{\jmath}}|,$$

and

$$\sum_{i:(i)<(\tilde{\imath}_j)} |x_{i\hat{\jmath}}| \leq \frac{1}{2}\sum_{i=1}^{n} |x_{i\hat{\jmath}}|,$$

for $j = 1, \dots, m$, where (i) is the place of $\frac{x_{ij}}{x_{i\hat{\jmath}}}$ in the ordered list $\left(\frac{x_{(1)j}}{x_{(1)\hat{\jmath}}}, \frac{x_{(2)j}}{x_{(2)\hat{\jmath}}}, \dots, \frac{x_{(n)j}}{x_{(n)\hat{\jmath}}}\right)$ and $\tilde{\imath}_j$ is the index of $\tilde{x}_j$.

L_1-Norm Subspace Estimation based on L_1-Norm Line Fitting

Given data $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$.

```
for q in 1, 2, ..., m − 1 do
    for j in 1, ..., m do
        Find the best-fit line using directions except for j.
    end for
    Project data into a m − q-dimensional space
    orthogonal to the first q principal components.
    Project out the directions already covered by the
    first m − 1 components.
end for
```

Sparsity

- Sparsity can be quantified as L_0-norm
- Easy to compressed and so require less memory resources
- Cut execution times

Sparse L_1-PCA Best-Fit line problem

Consider the optimization problem to find an L_1-norm regularized L_1-norm best-fit one-dimensional subspace:

$$\min \sum_{i \in N} \|x_i - v\alpha_i\|_1 + \lambda \|v\|_1, \tag{1}$$

The problem is separable into $|M|$ sub-problems, one for each column j. Each can be formulated as below linear programming

Sparse L_1-PCA Best-Fit line problem

$$\min_{\substack{v_j,\epsilon^+,\epsilon^-,\\ v^+,v^-}} \sum_{i\in N} |x_{i\hat{\jmath}}|(\epsilon_i^+ + \epsilon_i^-) + \lambda(v^+ + v^-), \tag{2}$$

$$s.t. v_j + \epsilon_i^+ - \epsilon_i^- = \frac{x_{ij}}{x_{i\hat{\jmath}}}, i \in N, \tag{3}$$

$$v_j + v^+ - v^- = 0, \tag{4}$$

$$\epsilon_i^+, \epsilon_i^- \geq 0, i \in N, \tag{5}$$

$$v^+, v^- \geq 0. \tag{6}$$

Sparse L_1-PCA Best-Fit line problem

Theorem

For data $x_i \in \mathbb{R}^m$, $i \in N$ and for a given penalty $\lambda > 0$, an optimal solution to (**??**) *can be constructed as follows. If $x_{i\hat{j}} = 0$ for all i, then set $v = 0$. Otherwise, for each $j \neq \hat{j}$,*

1. *Take points x_i, $i \in N$ such that $x_{i\hat{j}} \neq 0$ and sort the ratios $\frac{x_{ij}}{x_{i\hat{j}}}$ in increasing order.*
2. *If there is an $\tilde{i}$ where*

$$\left| sgn\left(\frac{x_{\tilde{i}j}}{x_{\tilde{i}\hat{j}}}\right)\lambda + \sum_{\substack{i \in N: \\ i < \tilde{i}}} |x_{i\hat{j}}| - \sum_{\substack{i \in N: \\ i > \tilde{i}}} |x_{i\hat{j}}| \right| \leq |x_{\tilde{i}\hat{j}}|, \tag{7}$$

   then set $v_j = \frac{x_{\tilde{i}j}}{x_{\tilde{i}\hat{j}}}$.

3. *If no such $\tilde{i}$ exists, then set $v_j = 0$.*

Algorithm

Input: $X \in \mathbb{R}^{m \times n}$
Output: A Vector Defining a one-dimensional subspace,v$\in \mathbb{R}^m$

for $\hat{j} \in M$ **do**
  for $j \neq \hat{j}$ **do**
    Sorting $\frac{x_{ij}}{x_{i\hat{j}}}$
    if $(\frac{x_{\tilde{i}j}}{x_{\tilde{i}\hat{j}}})\lambda + \sum_{i:i<\tilde{i}} |x_{i\hat{j}}| - \sum_{i:i>\tilde{i}} |x_{i\hat{j}}| \leq |x_{\tilde{i}\hat{j}}|$ **then**
      $v_j = \frac{x_{\tilde{i}j}}{x_{\tilde{i}\hat{j}}}$
    else
      $v_j = 0$
    end if
  end for
end for
return v

Algorithm

Input: $X \in \mathbb{R}^{m \times n}$, and the dimension of the fitted subspace, q.
Output: A matrix whose columns form an orthonormal basis of a fitted subspace,V$\in \mathbb{R}^m \times q$ and a matrix whose rows indicate the location of projections in the subspace A$\in \mathbb{R}^n \times q$.

Set $X^1 = X$.
for $k = 1, \cdots, q$ **do**
  Set v to be the estimate of the best-fit one-dimensional subspace derived using Algorithm 1 for data X^k
  if k=1 **then**
    $V^1 = v$
  else
    Set $v^k = v - V^{k-1}(V^{k-1})^T v$
    $V^k = [V^{k-1} v^k]$
    if $k < q$ **then**
      Set $X^{k+1} = X^k - X^k V^k (V^k)^T$
      $V = V^q$
    end if
  end if
end for

Choice of λ

The idea is based on interval

$$\sum_{i>\tilde{i}} |x_{i\hat{j}}| - \sum_{i\leq\tilde{i}} |x_{i\hat{j}}| \leq sgn(\frac{x_{\tilde{i}j}}{x_{\tilde{i}\hat{j}}})\lambda \leq \sum_{i\geq\tilde{i}} |x_{i\hat{j}}| - \sum_{i<\tilde{i}} |x_{i\hat{j}}| \qquad (8)$$

Distributed Implementation

Spark Implementation

Future Work

- Find optimal interval for λ
- Distributed Implementation
- Experimental Comparison to other L_1-PCA
- Derivations of regularization