Robust subspace recovery by geodesically convex optimization

Teng Zhang

University of Minnesota, Institute of Mathematics and its Applications

2012 SIAM Annual Meeting (AN12)

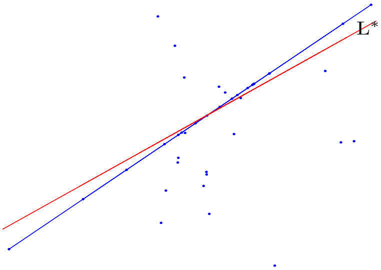
Jul 11, 2012

Outline

- Background: Robust Principal Components Analysis (PCA)
- Tyler's M-estimator and its properties
- Theory for exact recovery of the subspace
- Experiments

Problem Formulation

- Given: a linear subspace L^* and a data set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$, which contains some points sampled from L^* (we call them inliers) and outliers sampled from $\mathbb{R}^D \setminus \mathrm{L}^*$.
- Goal: recover L^* using $\mathcal{X}$.
- Fact: PCA is sensitive to outliers:



History

- Covariance estimators in robust statistics community: M-estimator, S-estimator, MVD (minimum volume ellipsoid) estimator, MCD (minimum covariance determinant) estimator, Stahel-Donoho estimator. See review by Maronna et al. (06)
- Projection Pursuit: Li & Chen (85), Ammann (93), McCoy & Tropp (10)
- Outlier detection and removal: Torre & Black(01), Xu et al. (10)

History

Some recent algorithms provide conditions for the exact recovery of the subspace L^*:

- Convex optimization based on nuclear norm: Xu et al. (10), McCoy & Tropp (11)
- Convex optimization based on l_1 distance: Zhang & Lerman (11), Lerman et al. (12).
- SSC algorithm based on sparse representation: Soltanolkotabi & Candès (11).

Motivation of Tyler's M-estimator for covariance

- Goal: robust covariance.
- Empirical covariance is also the MLE estimator when data points are drawn from Gaussian distribution:

$$\hat{\Sigma} = \arg\min_{\Sigma} \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}) + \frac{1}{2} \log\det(\boldsymbol{\Sigma}).$$

- For more general distribution

$$C(\rho) e^{-\rho(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})} / \sqrt{\det(\boldsymbol{\Sigma})}, \tag{1}$$

  the MLE estimator is

$$\hat{\Sigma} = \arg\min_{\Sigma} \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} \rho(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}) + \frac{1}{2} \log\det(\boldsymbol{\Sigma}). \tag{2}$$

- Tyler's M-estimator is defined for $\rho(x) = \frac{D}{2} \log(x)$, which corresponds to the MLE estimator for multivariate student distribution when $\nu \to 0$, or for angular Gaussian distribution (Gaussian distribution normalized to unit sphere).

Formulation

- (Tyler, 1987) Tyler's M-estimator for covariance is defined by

$$\mathbf{\Sigma}_* = \operatorname*{arg\,min}_{\mathrm{tr}(\mathbf{\Sigma})=1, \mathbf{\Sigma}=\mathbf{\Sigma}^T, \mathbf{\Sigma}\in \mathrm{S}_{++}(D)} F(\mathbf{\Sigma}), \quad \text{where} \tag{3}$$

$$F(\mathbf{\Sigma}) = \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{X}} \log(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}) + \frac{1}{D}\log\det(\mathbf{\Sigma}),$$

- Fix $\mathrm{tr}(\mathbf{\Sigma}) = 1$ because of scale-invariance: $F(\mathbf{\Sigma}) = F(c\mathbf{\Sigma})$.
- (Tyler, 1987) Use the limit of the iterative procedure to find Σ_*:

$$\mathbf{\Sigma}^{(k+1)} = \sum_{\mathbf{x}\in\mathcal{X}} \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{\Sigma}^{(k)\,-1}\mathbf{x}} / \,\mathrm{tr}\Big(\sum_{\mathbf{x}\in\mathcal{X}} \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{\Sigma}^{(k)\,-1}\mathbf{x}}\Big). \tag{4}$$

Property of formulation

- (Wiesel, 2012; Zhang, 2012) $F(\boldsymbol{\Sigma})$ is geodesically convex:

$$F(\boldsymbol{\Sigma}_1) + F(\boldsymbol{\Sigma}_2) \geq 2F(\boldsymbol{\Sigma}_1^{\frac{1}{2}}(\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_1^{\frac{1}{2}}). \tag{5}$$

- (Zhang 2012) When $\mathrm{Sp}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} = \mathbb{R}^D$, the equality in (5) holds if and only if $\boldsymbol{\Sigma}_1 = c\boldsymbol{\Sigma}_2$.
- Since $\mathrm{tr}(\boldsymbol{\Sigma})$ is fixed, we have strict convexity and uniqueness of the solution.

Geometry of positive definite matrices

- We call this property "geodesically convex" since $\mathbf{\Sigma}_1^{\frac{1}{2}}(\mathbf{\Sigma}_1^{-\frac{1}{2}}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{\Sigma}_1^{\frac{1}{2}}$ is the mean of the geodesic line connecting Σ_1 and Σ_2.
- In this geometry, $\mathrm{dist}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \|\log(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2)\|_F$, and $\mathbf{\Sigma}_1^{\frac{1}{2}}(\mathbf{\Sigma}_1^{-\frac{1}{2}}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-\frac{1}{2}})^t\mathbf{\Sigma}_1^{\frac{1}{2}}$ $(0 \leq t \leq 1)$ parametrizes the geodesic line connecting $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.
- This geometry can be obtained by differential geometry for the manifold of positive definite matrices, or by information geometry (Fisher's metric) for all multivariate Gaussian distributions with mean 0.

Property of iterative algorithm

- Recall the algorithm:

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{\mathbf{x}\in\mathcal{X}} \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\boldsymbol{\Sigma}^{(k)\,-1}\mathbf{x}} / \operatorname{tr}\Big(\sum_{\mathbf{x}\in\mathcal{X}} \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\boldsymbol{\Sigma}^{(k)\,-1}\mathbf{x}}\Big). \tag{6}$$

- (Wiesel, 2012; Zhang, 2012) This algorithm is monotone:

$$F(\boldsymbol{\Sigma}^{(k+1)}) \leq F(\boldsymbol{\Sigma}^{(k)})$$

- (Zhang 2012) If for any linear subspace L we have

$$\frac{|\mathcal{X}\cap L|}{N} < \frac{\dim(L)}{D}, \tag{7}$$

  then $\boldsymbol{\Sigma}_*$ exists and is unique, and $\lim_{k\to\infty}\boldsymbol{\Sigma}^{(k)} = \boldsymbol{\Sigma}_*$.
- Empirically it converges linearly.

Theoretical justification for exact subspace recovery

(Zhang 2012) If
(a)there exists a d-dimensional subspace L_* such that

$$\frac{|\mathcal{X} \cap L_*|}{|\mathcal{X}|} > \frac{d}{D}, \tag{8}$$

(b) the points in the set $\mathcal{Y}_1 = \{\mathbf{P}_{L_*}\mathbf{x} : \mathbf{x} \in \mathcal{X} \cap L_*\} \subset \mathbb{R}^d$ and $\mathcal{Y}_0 = \{\mathbf{P}_{L_*^\perp}\mathbf{x} : \mathbf{x} \in \mathcal{X} \setminus L_*\} \subset \mathbb{R}^{D-d}$ lie in general positions respectively (i.e., any k points in $\mathcal{Y}_1$ span a k-dimensional subspace for all $k \leq d$ and any k points in $\mathcal{Y}_0$ span a k-dimensional subspace for all $k \leq D - d$).
Then the sequence $\mathbf{\Sigma}^{(k)}$ converges to some $\hat{\mathbf{\Sigma}}$ such that $\mathrm{Im}(\hat{\mathbf{\Sigma}}) = L_*$.

Theoretical justification for exact subspace recovery

Properties of this theory:

- Condition (b) is weak: the theorem almost only depends on the ratio of the number of inliers/outliers.
- No probabilistic estimation involved.
- No incoherence condition of the data set involved.
- However, this theory tolerates less outliers than SCC algorithm when d/D is small, and inliers/outliers are drawn from gaussian distribution (with high probability).

Phase transition

If inliers/outliers lie in general position, then

- when
$$\frac{|\mathcal{X} \cap L_*|}{|\mathcal{X}|} > \frac{d}{D}, \tag{9}$$
we have $\mathrm{im}(\mathbf{\Sigma}_*) = L_*$.
- when
$$\frac{|\mathcal{X} \cap L_*|}{|\mathcal{X}|} < \frac{d}{D}, \tag{10}$$
we have $\mathrm{im}(\mathbf{\Sigma}_*) = \mathbb{R}^D$.

Other properties

- This method only depends on the directions of the data points: if we replace any $\mathbf{x} \in \mathcal{X}$ by $\mathbf{x}' = c\mathbf{x}$, then $\log(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})$ and $\log(\mathbf{x}'^T \boldsymbol{\Sigma}^{-1} \mathbf{x}')$ only differ by a constant of $2 \log c$, and the minimizer of $F(\boldsymbol{\Sigma})$ is unchanged.
- The algorithm is also independent of the magnitude of the data points.

Verification of exact recovery and phase transition

- In this example we let $D = 10$, $d = 5$, 100 outliers, and apply this algorithm for the case of different number of inliers.
- It turns out that we have exact recovery when the number of inliers is larger than 100.

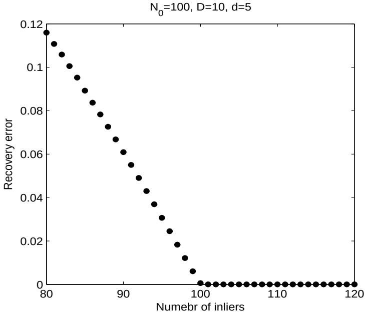


Figure: *The dependence on the number of inliers and recovery error: x-axis is the number of inlier and y-axis is the corresponding recovery error.*

Experiment

- 64 images of a single face under different illuminations from the Extended Yale Face database (used as inliers)
- 400 additional random images from the BACKGROUND/Google folder of the Caltech101 database (used as outliers)
- resolution downsampled to 20×20
- The face images lie on a nine-dimensional subspace (Basri & Jacobs, 03)
- Learn the subspace from a data set that contain 32 face images and 400 other random images.
- We recover the 9-dimensional subspace by the span of top 9 eigenvectors of $\boldsymbol{\Sigma}_*$.

Experiment

We compare Tyler's M-estimator with PCA, Reaper and S-reaper algorithms:

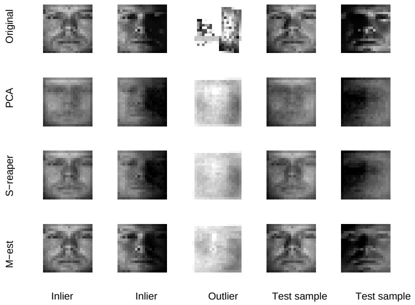


Figure: *The projection of images to the fitted subspace.*

Conclusions

- We analyze the properties of Tyler's M-estimator (geodesic convexity) and the convergence of the iterative algorithm.
- We provide a theory for robust subspace recovery, which almost only depends on the percentage of outliers.
- We verify its performance on real data set.