Nonconvex Sparse Spectral Clustering by Alternating Direction Method of Multipliers and Its Convergence Analysis

Canyi Lu[1], Jiashi Feng[1], Zhouchen Lin[2], Shuicheng Yan[1]. [1]National University of Singapore, [2]Peking University

Convex Sparse Spectral Clustering

- **Data Clustering Task:** Given n data points $X = [x_1, \cdots, x_n] = [X_1, \cdots, X_k] \in \mathbb{R}^{d\times n}$, the task is to group them into k clusters.
- **Spectral Clustering** (Ng et al., NIPS, 2002)
  1. Compute the affinity matrix $W \in \mathbb{R}^{n\times n}$;
  2. Compute the normalized Laplacian matrix $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, where D is a diagonal matrix with its diagonal element $d_{ii} = \sum_{i=1}^{n} w_{ij}$;
  3. Compute $U \in \mathbb{R}^{n\times k}$ by solving
$$\min_{U\in\mathbb{R}^{n\times k}} \langle UU^\top, L\rangle, \text{ s.t. } U^\top U = I;$$
  4. Form $\hat{U} \in \mathbb{R}^{n\times k}$ by normalizing each row of U;
  5. Treat each row of $\hat{U}$ as a point in $\mathbb{R}^k$, and cluster them into k groups by k-means.
- **Convex Sparse Spectral Clustering** (C. Lu, et al., TIP, 2016)
  - If W is block diagonal in the ideal case, $UU^\top$ is block diagonal and thus sparse.
$$\min_{U\in\mathbb{R}^{n\times k}} \left\langle L, UU^\top\right\rangle + \beta\|UU^\top\|_0, \text{ s.t. } U^\top U = I.$$
  - Convex relaxation
$$\min_{P\in\mathbb{R}^{n\times n}} \langle P, L\rangle + \beta\|P\|_1, \text{ s.t. } 0 \preceq P \preceq I, \ Tr(P) = k.$$
$$\min_{U\in\mathbb{R}^{n\times k}} \|P^* - UU^\top\|, \text{ s.t. } U^\top U = I.$$
  - Limitation: The convex relaxation may not lead to sparse $U^\top U$.

Nonconvex Sparse Spectral Clustering

- Our goal is to solve the following nonconvex Sparse Spectral Clustering problem
$$\min_{U\in\mathbb{R}^{n\times k}} \left\langle L, UU^\top\right\rangle + g(UU^\top), \text{ s.t. } U^\top U = I,$$
where $g : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a sparse regularizer.
- Problem reformulation
$$\min_{P\in\mathbb{R}^{n\times n}, U\in\mathbb{R}^{n\times k}} \left\langle L, UU^\top\right\rangle + g(P), \text{s.t. } P = UU^\top, \ U^\top U = I.$$
- Partial augmented Lagrangian function
$$\mathcal{L}(P, U, Y, \mu) = \left\langle L, UU^\top\right\rangle + g(P) + \left\langle Y, P - UU^\top\right\rangle + \frac{\mu}{2}\left\|P - UU^\top\right\|^2,$$
where Y is the dual variable and $\mu > 0$.
- **Alternating Direction Method of Multipliers (ADMM)**
  1. Fix $P = P_k$ and update U by
$$\begin{aligned} U_{k+1} &= \operatorname*{argmin}_{U\in\mathbb{R}^{n\times k}} \mathcal{L}(P_k, U, Y_k, \mu_k), \text{ s.t. } U^\top U = I. \\ &= \operatorname*{argmin}_{U} \left\|UU^\top - P_k + (L - Y_k)/\mu_k\right\|^2, \text{s.t. } U^\top U = I. \end{aligned}$$
  2. Fix $U = U_{k+1}$ and update P by
$$P_{k+1} = \operatorname*{argmin}_{P} \mathcal{L}(P, U_{k+1}, Y_k, \mu_k) = \operatorname*{argmin}_{P} g(P) + \frac{\mu_k}{2}\left\|P - U_{k+1}U_{k+1}^\top + Y_k/\mu_k\right\|^2.$$
  3. $Y_{k+1} = Y_k + \mu_k(P_{k+1} - U_{k+1}U_{k+1}^\top)$.
  4. $\mu_{k+1} = \min(\mu_{\max}, \rho\mu_k), \ \rho > 1$.

Convergence Analysis

Assumptions

A1. L is positive semi-definite.

A2. $g : \mathbb{R}^{n\times n} \to R$ is lower bounded, differential and ∇g is Lipschitz continuous, i.e., there exists $l > 0$ such that
$$\|\nabla g(X) - \nabla g(Y)\| \le l\|X - Y\|, \ \forall X, Y \in \mathbb{R}^{n\times n}.$$

A3. The stepsize μ_k is chosen large enough such that
(1) The P-subproblem is strongly convex with modulus γ_k.
(2) $\mu_k\gamma_k > l^2(\rho + 1)$ and $\mu_k \ge l$.

Convergence Results

Theorem. Under assumptions A1-A3, the generated sequences $\{P_k, U_k, Y_k\}$ satisfy

(a) $\mathcal{L}(P_k, U_k, Y_k, \mu_k)$ is monotonically decreasing, i.e.,
$$\mathcal{L}(P_{k+1}, U_{k+1}, Y_{k+1}, \mu_{k+1}) - \mathcal{L}(P_k, U_k, Y_k, \mu_k) \le -\left(\frac{\gamma_k}{2} - \frac{l^2(\rho+1)}{2\mu_k}\right)\|P_{k+1} - P_k\|^2.$$

(b) $\lim_{k\to+\infty} \mathcal{L}(P_k, U_k, Y_k, \mu_k) = \mathcal{L}^*$ for some constant $\mathcal{L}^*$.

(c) When $k \to +\infty$, $P_{k+1} - P_k \to 0$, $Y_{k+1} - Y_k \to 0$ and $P_k - U_kU_k^\top \to 0$.

(d) The sequences $\{P_k\}$, $\{U_k\}$ and $\{Y_k\}$ are bounded.

(e) There exists $G = [G_P \ G_U \ G_Y]$, where
$$\begin{aligned} G_P &= \partial_P\mathcal{L}(P_{k+1}, U_{k+1}, Y_{k+1}, \mu_k), \\ G_U &\in \partial_U\mathcal{L}(P_{k+1}, U_{k+1}, Y_{k+1}, \mu_k) + \partial_U\iota_{\mathcal{O}}(U_{k+1}), \\ G_Y &= \partial_Y\mathcal{L}(P_{k+1}, U_{k+1}, Y_{k+1}, \mu_k), \end{aligned}$$
such that
$$\|G\|^2 \le (8d + 1 + \frac{1}{\mu_0^2})\|Y_k - Y_{k+1}\|^2 + 8d\mu_{\max}^2\|P_k - P_{k+1}\|^2.$$

(f) Let (P^*, U^*, Y^*) denotes any limit point of the sequence $\{P_k, U_k, Y_k\}$. Then the limit point is a stationary point, i.e.,
$$\begin{aligned} 0 &\in \partial_U\mathcal{L}(P^*, U^*, Y^*, \mu^*) + \partial_U\iota_{\mathcal{O}}(U^*), \\ 0 &= \partial_P\mathcal{L}(P^*, U^*, Y^*, \mu^*), \\ 0 &= \partial_Y\mathcal{L}(P^*, U^*, Y^*, \mu^*) = P^* - U^*U^{*\top}. \end{aligned}$$

Experiments

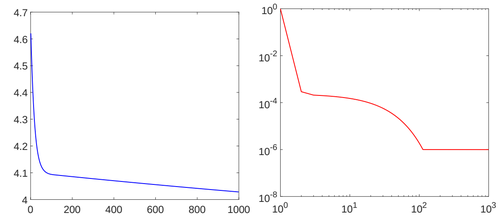
We use the following nonconvex SSC model in the experiment
$$\min_{P\in\mathbb{R}^{n\times n}, U\in\mathbb{R}^{n\times k}} \left\langle L, UU^\top\right\rangle + g_\sigma(P), \text{s.t. } P = UU^\top, \ U^\top U = I, \tag{1}$$
where g_σ is the smoothed ℓ_1-norm $\beta\|P\|_1$ with a smoothness parameter $\sigma > 0$ defined as follows
$$g_\sigma(P) = \max_Z \langle P, Z\rangle - \frac{\sigma}{2}\|Z\|^2, \text{ s.t. } \|Z\|_\infty \le \beta, \tag{2}$$
where $\|Z\|_\infty = \max_{ij}|z_{ij}|$.

Affinity matrix construction by the ℓ_1-graph

- Construct the *sparse* affinity matrix W by the ℓ_1-graph (Elhamifar and Vidal, 2013).
- Extended Yale B database is used. It consists of 2,414 face images of 38 subjects. Each subject has 64 faces.

Table: Clustering errors (%) on the Extended Yale B database based on the *sparse* affinity matrix W constructed by the ℓ_1-graph.

# of subjects	SC	SSC	SSC-PG	SSC-ADMM
2	1.56±2.95	1.80±2.89	1.37±3.15	**1.21±2.10**
3	3.26±7.69	3.36±7.76	3.12±6.23	**2.40±4.92**
5	6.33±5.36	6.61±5.93	5.65±4.33	**3.86±2.82**
8	8.93±6.11	4.98±4.00	4.95±3.36	**4.67±3.40**
10	9.94±4.57	**4.60±2.59**	5.91±4.52	5.84±3.43



(a) $\mathcal{L}(P_k, U_k, Y_k, \mu_k)$ (b) stopping criteria

Figure: Plots of (a) convergence of $\mathcal{L}(P_k, U_k, Y_k, \mu_k)$ v.s. k and (b) convergence of the stopping criteria $\max\{\|P_{k+1} - P_k\|_\infty, \|P_{k+1} - U_{k+1}U_{k+1}^\top\|_\infty\}$ v.s. k.

Affinity matrix construction by the Gaussian kernel

- Construct the *dense* affinity matrix W by the Gaussian kernel.
- Ten datasets of different sizes are used.

Table: Statistics of the used 10 datasets.

dataset	# samples	# features	# clusters
Wine	178	13	3
USPS	1,000	256	10
Glass	214	9	6
Letter	1,300	16	26
Vehicle	846	18	4
UMIST	564	1,024	20
PIE	1,428	1,024	68
COIL20	1,440	1,024	20
CSTR	476	1,000	4
AR	840	768	120

Table: Clustering accuracy on 10 datasets based on the *dense* affinity matrix constructed by the Gaussian kernel.

	k-means	NMF	SC	SSC	SSC-PG	SSC-ADMM
Wine	94.4	96.1	94.9	96.1	96.6	**97.2**
USPS	67.3	69.2	71.2	73.4	**76.8**	**76.8**
Glass	40.7	39.3	41.1	43.0	44.9	**45.3**
Letter	27.1	30.4	31.8	35.3	**36.4**	36.0
Vehicle	62.1	61.3	67.0	70.0	73.0	**73.4**
UMIST	52.1	63.8	63.3	64.2	65.1	**66.1**
PIE	35.4	37.9	42.0	46.7	46.8	**51.1**
COIL20	59.0	46.2	63.1	64.5	**69.1**	67.9
CSTR	65.0	70.0	68.9	72.7	71.0	**76.3**
AR	24.2	35.0	36.1	37.1	37.7	**39.0**

Conclusion

- We develop the first ADMM solver with the convergence guarantee for solving the nonconvex Sparse Spectral Clustering problem.