Overlapping Clustering Models, and One (class) SVM to Bind Them All
Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti

One-class SVM solves ideal cone problem

\begin{align*}
\text{Primal:} & \quad \max b \\
& \text{s.t.} \quad \|w\| \leq 1, \ w^Ty_i \geq b, \ i \in I
\end{align*}

\begin{align*}
\text{Dual:} & \quad \min \ \frac{1}{2} \sum_{y,j} \beta_j y_j \\
& \text{s.t.} \quad \sum_{y} \beta_j = 1, \ \beta_j \geq 0, \ i \in I
\end{align*}

- One-class SVM works if the following condition is satisfied:
- Condition: The matrix $Y_P$ satisfies $(Y_P Y_P^T)^{-1} \geq 1 > 0$.

Empirical Cone Problem

Given the empirical matrix $Y$ with rows $y_i^T/\|y_i\|$.

- With $Z = MY \in \mathbb{R}^{r \times m}$, and
- $\max_i \|y_i^T - y_i^T\| \leq \epsilon$.
- Infer $M$.

One-class SVM still works if the following condition is satisfied:
- Condition: The matrix $Y_P$ satisfies $(Y_P Y_P^T)^{-1} \geq \eta I$ for some constant $\eta > 0$.

Algorithm

- **Step 1**: Express the rows of $Z$ by $l_2$ norm.
- **Step 2**: Infer one-class SVM to get support hyperplane.
- **Step 3**: Cluster all points close to this hyperplane.
- **Step 4**: Pick one point from each cluster to get near-corner set $C$.

Proof Sketch for Empirical Cone Problem

- **Step 1**: Show that SVM solution of empirical cone is nearly ideal, i.e., $(w, b) = (w, b)$.
- **Step 2**: Show that true corners of the cone are close to supporting hyperplane.
- **Step 3**: Also, all points close to support vectors are nearly corners.
- **Step 4**: Clustering the points that are close to support vectors yields exactly one corner for each cluster.

Per-node Consistency

- If $|y_i - \hat{z}_i| \leq \epsilon \rightarrow 0$ for all $i$, using SVM-cone will get $M$ consistently with ground truth $(\hat{m}_i = m_i)$.
- Holds with high probability for both DCMSMB-type models and topic models.

Per-node consistency guarantee for DCMSMB (informal)

- $\Theta = \text{Dirichlet}(\alpha), \ \alpha_i = \alpha^T$, under some conditions on $\alpha$ and $I$, w.h.p.,
- $|y_i^T(\Theta - \Theta^I)| = O(\frac{\max_j K_j^2 \min_j K_j (\alpha^T P)^j e^2 (1 + \alpha_j^T)}{\gamma_{\min} \|B\| \sqrt{m}})$

Solving the ideal cone problem with $Z = V$ gives us the pure nodes, from which community memberships can be inferred.

Ideal cone problem

- Given a matrix $Z$ that is known to be of the form $Z = MY$, where $M \in \mathbb{R}^{r \times m}$, and $\gamma_i \in \mathbb{R}$ are $\gamma$ (unknown) rows of $Z$, each scaled to unit $l_2$ norm.
- Infer $M$.

Simulation Experiments

- Comparing with GeoNMF (Mao et al. 2017), OCCAM (Zhang et al. 2014) and SAAC (Kaufmann et al. 2016).
- (a) Varying degree heterogeneity on DCMSB. (b) Varying sparsity on DCMSB. (c) Varying sparsity on OCCAM. (d) Varying sparsity on SBM.

Real-world Network Results

- Evaluation metric: Averaged Spearman rank correlation coefficients (RC) between $\Theta(\cdot, \cdot)$, $\alpha \in [K]$ and $\Theta(\cdot, \sigma(\cdot))$, where $\sigma$ is a permutation of $[K]$.
- $R_{\text{corr}}(\Theta, \Theta) = \frac{1}{K} \max_{i} \sum_j \text{RC}(\Theta(:, i), \Theta(:, \sigma(i))) \in [-1, 1]$. 

- Co-authorship networks (assortative).

- Author-paper bipartite networks (dissortative).

Topic Models Results

- $l_2$ reconstruction error and running time (log scale) for semi-synthetic data with number of documents set to 60,000.

Related work

Most algorithms (Mao et al., 2017a, 2017b; Jin et al., 2017; Panov et al., 2017; Rubin-Delanchy et al., 2017) use a two step method:
- Find the pure nodes by finding corners of a simplex.
- Estimate model parameters via regression.

For example, (Mao et al., 2017b) shows that for MMSB, the rows $v_i$ of $V$ (the top $K$ eigenvectors of $P$) lie on a simplex with corners as pure nodes of $Z$.

One can estimate community memberships by regressing $v_i$ on $Z_i$, with degree parameters, rows of $V$ are not in a simplex.

Eliminating the effects of degree parameters

- Rows of $V$ fall on a cone.
- One can estimate community memberships by regressing $v_i$ on $Z_i$ with degree parameters, rows of $V$ are not in a simplex.

Ideal cone problem

- Given a matrix $Z$ that is known to be of the form $Z = MY$, where $M \in \mathbb{R}^{r \times m}$, and $\gamma_i \in \mathbb{R}$ are $\gamma$ (unknown) rows of $Z$, each scaled to unit $l_2$ norm.
- Infer $M$.

Solving the ideal cone problem with $Z = V$ gives us the pure nodes, from which community memberships can be inferred.

Ideal cone problem

- Given a matrix $Z$ that is known to be of the form $Z = MYP \in \mathbb{R}^{r \times m}$, where $M \in \mathbb{R}^{r \times m}$, no row of $M$ is 0.
- $Y_i \in \mathbb{R}^{r \times m}$, $\gamma_i \in \mathbb{R}$ are $\gamma$ (unknown) rows of $Z$, each scaled to unit $l_2$ norm.
- Infer $M$.

Solving the ideal cone problem with $Z = V$ gives us the pure nodes, from which community memberships can be inferred.

Ideal cone problem

- Given a matrix $Z$ that is known to be of the form $Z = MY$, where $M \in \mathbb{R}^{r \times m}$, no row of $M$ is 0.
- $Y_i \in \mathbb{R}^{r \times m}$, $\gamma_i \in \mathbb{R}$ are $\gamma$ (unknown) rows of $Z$, each scaled to unit $l_2$ norm.
- Infer $M$.

Solving the ideal cone problem with $Z = V$ gives us the pure nodes, from which community memberships can be inferred.