Selective inference for clustering via k-means

July 8, 2021

1 Formulation

With two clusters $\widehat{C}_1$ and $\widehat{C}_2$ obtained from k-means algorithm on the realization $X = x \in \mathbb{R}^{n\times d}$, the hypothesis of interest takes the form

$$\mathcal{H}_0 : \mu_{\widehat{C}_1} = \mu_{\widehat{C}_2}. \tag{1.1}$$

Here, the number of clusters k is pre-specified by the algorithm and should not be taken into account by the selective inference framework. Since the clusters are determined by the original data, the chi-squared test ought to be corrected conditioning on the clustering result. Denote $\phi = \|\bar{X}_{\widehat{C}_1} - \bar{X}_{\widehat{C}_2}\|$, a general approach is developed as:

$$\begin{aligned} &\mathbb{P}\left(\phi \geq \|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\| \,\middle|\, X \text{ results in } \widehat{C}_1, \widehat{C}_2, \text{ others}\right) && (1.2)\\ =&\mathbb{P}\left(\phi \geq \|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\| \,\middle|\, x'(\phi) \text{ results in } \widehat{C}_1, \widehat{C}_2\right). && (1.3) \end{aligned}$$

The novel choice of $x'(\phi)$ explicit in x and ϕ is for simplifying the computation where

$$[x'(\phi)]_i = \begin{cases} x_i + \left(\frac{|\widehat{C}_2|}{|\widehat{C}_1|+|\widehat{C}_2|}\right)\left(\phi - \|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\|\right) \operatorname{dir}\left(\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\right), & i \in \widehat{C}_1,\\ x_i - \left(\frac{|\widehat{C}_1|}{|\widehat{C}_1|+|\widehat{C}_2|}\right)\left(\phi - \|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\|\right) \operatorname{dir}\left(\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\right), & i \in \widehat{C}_2,\\ x_i, & i \notin \widehat{C}_1 \cup \widehat{C}_2. \end{cases} \tag{1.4}$$

2 Computation of the conditional event

The selective inference approach can be viewed as the truncation of the chi-squared test function on the set $\mathcal{S} = \left\{\phi \geq 0 : \widehat{C}_1, \widehat{C}_2 \in C\left(x'(\phi)\right)\right\}$, where $C(\cdot)$ is the clustering map. Then, to ease the implementation, we consider the characterization of $\mathcal{S}$ based on the k-means algorithm. As $x'(\phi)$ results in the two clusters of interest, this event with respect to the specific k-means algorithm can be written as:

$$\begin{aligned} \mathcal{S} &= \left\{\phi \geq 0 : x'(\phi) \text{ results in } \widehat{C}_1, \widehat{C}_2\right\} && (2.1)\\ &= \left(\bigcap_{A\in C(x)\cap \widehat{C}_1^c} \left\{\phi \geq 0 : \forall i \in \widehat{C}_1, d\left([x'(\phi)]_i, \tilde{m}_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, \tilde{m}_A\right)\right\}\right) && (2.2)\\ &\bigcap \left(\bigcap_{A\in C(x)\cap \widehat{C}_2^c} \left\{\phi \geq 0 : \forall i \in \widehat{C}_2, d\left([x'(\phi)]_i, \tilde{m}_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, \tilde{m}_A\right)\right\}\right) && (2.3) \end{aligned}$$

$$= \left(\bigcap_{A \in C(x) \cap \widehat{C}_1^c \cap \widehat{C}_2^c} \bigcap_{i \in \widehat{C}_1} \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, m_A\right) \right\} \right) \tag{2.4}$$

$$\bigcap \left(\bigcap_{A \in C(x) \cap \widehat{C}_1^c \cap \widehat{C}_2^c} \bigcap_{i \in \widehat{C}_2} \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_2}\right) \leq d\left([x'(\phi)]_i, m_A\right) \right\} \right) \tag{2.5}$$

$$\bigcap \left(\bigcap_{i \in \widehat{C}_1} \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, \tilde{m}_{\widehat{C}_2}\right) \right\} \right) \tag{2.6}$$

$$\bigcap \left(\bigcap_{i \in \widehat{C}_2} \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_2}\right) \leq d\left([x'(\phi)]_i, \tilde{m}_{\widehat{C}_1}\right) \right\} \right) \tag{2.7}$$

$$\tag{2.8}$$

Here $\tilde{m}_A$ is the centroid for cluster A with $x'(\phi)$ and m_A is the centroid for cluster A with original data x. For $A \neq \widehat{C}_1, \widehat{C}_2$, $m_A = \tilde{m}_A$ and the distance-structure within each cluster remains the same after the perturbation.

WLOG, we consider $i \in \widehat{C}_1$, then there are two kinds of sets:

1. Cluster $A \neq \widehat{C}_1, \widehat{C}_2$. Denote $\mathcal{S}_{1,i} = \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, m_A\right) \right\}$. In $\mathcal{S}_{1,i}$, if we choose $d(\cdot,\cdot)$ to be the l_2 distance, $d\left(x_i, m_{\widehat{C}_1}\right)$ can be computed using the clustering result with x and is therefore treated as fixed. For the second term, $x'(\phi)$ is linear in ϕ, then $\|[x'(\phi)]_i - m_A\|^2$ is a quadratic function in ϕ for any $A \in C(x) \cap \widehat{C}_1^c \cap \widehat{C}_2^c$.

2. Cluster $A = \widehat{C}_2$. The other kind of set has the form $\mathcal{S}_{2,i} = \left\{ \phi \geq 0 : d\left(x_i, m_{\widehat{C}_1}\right) \leq d\left([x'(\phi)]_i, \tilde{m}_{\widehat{C}_2}\right) \right\}$. To compute $\tilde{m}_{\widehat{C}_2}$, $[x'(\phi)]_i = x_i - c_2\phi - a_2$ for $i \in \widehat{C}_2$, then $\tilde{m}_{\widehat{C}_2} = \frac{1}{|\widehat{C}_2|} \sum_{i \in \widehat{C}_2} [x'(\phi)]_i = m_{\widehat{C}_2} - c_2\phi - a_2$ that is linear in ϕ. As both $[x'(\phi)]_i$ and $\tilde{m}_{\widehat{C}_2}$ are linear in ϕ, the squared distance is also quadratic in ϕ similar with the first case.

From above, the restricted set $\mathcal{S}$ is the intersection of quadratic constraints for ϕ and can be computed explicitly with x, $x'(\phi)$ and the clustering results $C(x)$ together with the perturbed version $C(x'(\phi))$. To sum up, the selective p-value can be written as

$$\mathbb{P}\left(\phi \geq \|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\| \middle| \phi \in \bigcap_{j \in \mathcal{J}} \mathcal{A}_j \right) = 1 - F_{\bigcap_{j \in \mathcal{J}} \mathcal{A}_j}(\|\bar{x}_{\widehat{C}_1} - \bar{x}_{\widehat{C}_2}\|), \tag{2.9}$$

where $\mathcal{A}_j = \{\phi \geq 0 : \kappa_j^2\phi^2 - \lambda_j\phi + \theta_j \geq 0\}$ is the quadratic constraint for ϕ and $F_{\mathcal{S}}$ is the cdf of $\sigma\sqrt{\left(1/|\widehat{C}_1| + 1/|\widehat{C}_2|\right)}\chi_p$ truncated to $\mathcal{S}$.