Scalable Bayes Clustering for Outlier Detection Under Informative Sampling

Based on JMLR paper of T. D. Savitsky

Terrance D. Savitsky

Office of Survey Methods Research

FCSM  2018
March 7-9, 2018
Motivating Dataset

- Monthly survey of 350,000 U.S. business establishments
- Single stage, fixed-size, stratified sampling design
- Strata-indexed probabilities assigned by employment size $\equiv \text{pps}$
- 100,000 – 150,000 establishments report employment changes
- $(d=4) \times 1 = \text{(Employment, Production Workers, Payroll, Weekly Hours)}$
- Size variable is total employment, $z$
- $x \not\perp z$
- 7-day turnaround between submissions and publication
- Which establishment submissions contain reporting errors?
Estimating Population Model Under Informative Sample

- Finite $U = (1, \ldots, N)$
- With data, $X_U = X_1, \ldots, X_N \sim P_\theta$.
- Don’t fully observe the finite population.
- Draw a sample, $S = (1, \ldots, n \leq N)$.
- Inclusion probabilities, $P(\delta_i = 1) := \pi_i$ correlated with $X_U$
- $P_\theta(X_S) \neq P_\theta(X_U)$
- Want to estimate outliers from $P_\theta(X_U)$ using $X_S$.
- Use $\tilde{w}_i \propto 1/\pi_i$
Mixture / Cluster Model for Outlier Detection

- Mixture of Gaussians
- \( s_i \in (1, \ldots, K_{\text{max}}) \) indexes cluster memberships for \( i \in (1, \ldots, n) \)
- \( (\tau_1, \ldots, \tau_{K_{\text{max}}}) \) cluster assignment probabilities
- \( \alpha \uparrow \), number of \( \tau_p > 0 \), \( \uparrow \)
- Dirichlet Process mixing measure in the limit of \( K_{\text{max}} \)

\[
\begin{align*}
\mathbf{x}_i \mid s_i, \mathbf{M} &= (\mu_1, \ldots, \mu_{K_{\text{max}}})', \sigma^2, \tilde{w}_i \sim \mathcal{N}_d \left( \mu_{s_i}, \sigma^2 I_d \right) \tilde{w}_i \\
\mathbf{s}_i \mid \mathbf{\tau} &\sim \mathcal{M} (1, \tau_1, \ldots, \tau_{K_{\text{max}}}) \\
\mu_p \mid G_0 &\sim G_0 := \mathcal{N}_d (0, \rho^2 I_d) \\
\tau_1, \ldots, \tau_{K_{\text{max}}} &\sim \mathcal{D} \left( \alpha/K_{\text{max}}, \ldots, \alpha/K_{\text{max}} \right)
\end{align*}
\]
Sampling-weighted Pseudo Posterior

- Pseudo Posterior $\propto$ Weighted Likelihood $\times$ Priors
- Marginalize out $\tau$ from the joint prior, $f(s, \tau|\alpha) = f(s|\tau) f(\tau|\alpha)$
- $M = (\mu_1, \ldots, \mu_K)$
- $n_p = \sum_{i=1}^{n} 1(s_i = p) \equiv$ number of establishments assigned to cluster, $p$

\[
f(s, M|X, \tilde{w}) \propto f(X, s, M|\tilde{w}) = \prod_{p=1}^{K} \prod_{i:s_i=p} \mathcal{N}_d(x_i|\mu_p, \sigma^2 I_d) \tilde{w}_i
\]

\[
\alpha^K \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + n)} \prod_{p=1}^{K} (n_p - 1)!
\]

\[
\prod_{p=1}^{K} \mathcal{N}_d(\mu_p|0, \rho^2 I_d).
\]
Approximate MAP as $\sigma^2 \downarrow 0$

- Each observation assigned to its own cluster as $\sigma^2 \downarrow 0$
- Define a constant $\lambda$ and set $\alpha = \exp (-\lambda / (2\sigma^2))$
- Produces $\alpha \downarrow 0$ as $\sigma^2 \downarrow 0$
- $\lambda$ hyperparameter controls the size of the partition as $\sigma^2 \downarrow 0$

$$-2\sigma^2 \times \log f (X, s, M, \tilde{w}) = \sum_{p=1}^{K} \sum_{i:s_i=p} \left[ -2\sigma^2 \times O (\log \sigma^2) + \tilde{w}_i \|x_i - \mu_p\|^2 \right]$$

$$+ K\lambda - 2\sigma^2 \times O (1)$$

$$- 2\sigma^2 \times O (1),$$
Approximate MAP Optimization

\[ \arg\min_{K,s,M} \sum_{p=1}^{K} \sum_{i:s_i=p} \tilde{w}_i \| \mathbf{x}_i - \mu_p \|^2 + K \lambda, \]

- Bayesian motivation for K-means clustering
- Higher value for \( \lambda \) reduces number of estimated clusters
- Goal to minimize energy expression
Add *Merge Step* to Algorithm

- Test all pairs of clusters and merge those that *reduce energy*
- Collapse 2 clusters by assigning establishments from both to single cluster
- Recompute cluster center, $\mu_p$
- Encourages fewer clusters, which supports outlier detection
- Reduces sensitivity to initial values
Weighted Hierarchical Clustering - Set-up

- Establishments, \( i = 1, \ldots, n \), binned to \( j = 1, \ldots, J \) industry groups
- Estimate a \textit{local} clustering of \( L_{\text{max}} \) possible clusters in industry, \( j \).
- Local cluster, \( c \), in industry, \( j \), connected to \textit{global} cluster center, \( \mu_p \)
- For \( p \in (1, \ldots, K_{\text{max}}) \) possible \textit{global} clusters
- Local clusters across industries may share a common \textit{global} cluster
- \( s_{i}^{j} \equiv \textit{global} \) cluster assignment for establishment, \( i \), in industry, \( j \)
Hierarchical Clustering Optimization

\[
\arg\min_{K,s,M} \sum_{p=1}^{K} \sum_{j=1}^{J} \sum_{i:s_i^j = p} \tilde{w}_i^j \| x_i^j - \mu_p \|^2 + K\lambda_K + L\lambda_L,
\]

- \( L = \sum_{j=1}^{J} L_j \) denotes the total number of local clusters
- \( L_j \) denotes the number of local clusters estimated for data set, \( j = 1, \ldots, J \)
- \( K \) denotes the number of estimated global clusters
- \( \lambda_K \) denotes penalty on number of global clusters estimated
- \( \lambda_L \) denotes penalty on number of local clusters estimated
- \( \tilde{w}_i^j \) is the sampling weight for establishment, \( i \), in industry, \( j \).
Selecting Penalty Parameters, \((\lambda_K, \lambda_L)\)

- Synthetic data, \(L_j = 5\) local clusters for \(j = 1, \ldots, (J = 3)\) industries
- Sharing \(K = 7\) global clusters
- \(X_j^{N_j \times (d=15)}\)
- \((N_j = 15000, n_j = 2500)\) establishments in (population/sample)
- Randomly allocated to \(L_j = 5\) in skewed distribution, 
  
  \(0.6, 0.25, 0.1, 0.025, 0.025)\)
- Evenly divide data into training and test sets
- Estimate clustering on training data and compute energy on test data
Energy steadily decreases with lower \((\lambda_K, \lambda_L)\)

- Estimate clustering on training data and compute energy on test data
Use Calinski-Harabasz \((C)\) criterion

- Cohesion \textit{within} each cluster, \(WGSS\)
- Separation \textit{between} clusters, \(BGSS\)

\[
WGSS = \sum_{p=1}^{K} \sum_{i:s_i^p = k} \tilde{w}_i \| x_i - \mu_p \|^2
\]

\[
BGSS = \sum_{p=1}^{K} n_p \| \mu_p - \mu^G \|^2
\]

\[
C = \frac{n - K \cdot BGSS}{K - 1 \cdot WGSS}
\]

- \(\mu^G = \frac{\sum_{i=1}^{n} \tilde{w}_i x_i}{\sum_{i=1}^{n} \tilde{w}_i}\)
- \(K\) is number of global clusters
$C$ finds an optimum

- chose the values of $(\lambda_L = 1232, \lambda_K = 2254)$
Correct Clusterings Estimated

- Each panel presents a local clustering for industry, $j \in (1, \ldots, (J = 3))$.
- We see $L_j = 5$ with correct skewed allocation
- Sharing $K = 7$ global clusters
Merges Increase at lower values for $(\lambda_K, \lambda_L)$

- Higher number of merges for lower values of $(\lambda_K, \lambda_L)$
Outlier Detection Simulation Study Design

- $J = 8$ local populations, $X^j$ with $N_j = 25000$
- $L_j = 2$ local clusters, one an outlier, sharing $K = 5$ global clusters

\[
\begin{align*}
\mu_1 &= (1, 1.5, 2.0, \ldots, 7.5, 8) \\
\mu_2 &= (8, 7.5, \ldots, 1) \\
\mu_3 &= (1, \ldots, 7, 8, 7, \ldots, 1) \\
\mu_4 &= \text{Sampling from } (1, \ldots, 8) \text{ with replacement, } d = 15 \text{ times} \\
\mu_5 &= \text{Sampling from } (-2, \ldots, 6) \text{ with replacement, } d = 15 \text{ times},
\end{align*}
\]

- mean $\mu_5$ is assigned 150 observations
- Stratified design of $H = 10$ strata assign $\pi^j_h \propto \text{variance of, } X^j_h$
- $B = 100$ Monte Carlo draws
Outlier Detection Accuracy

- **True positive** $\equiv$ # of true outliers discovered / total # of true outliers
- **False positive** $\equiv$ # of false discoveries / total # nominated
- True positives measure **effectiveness**, False positives measure **efficiency**

![Bar chart showing true positives and false positives for different estimation types.](chart.png)
Estimation Bias of Outlier Center, $\mu_5$

- For each $d = 15$ dimensions
- Dashed line presents true values.
Take Aways

- Fast hierarchical clustering captures dependencies among industry clusterings.
- Incorporating sampling weights better detects outliers from the population.
- Implemented in growclusters in R.
CONTACT INFORMATION

Savitsky.Terrance@bls.gov