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 350000 U.S. business establishments
- Single stage, fixed-size, stratified sampling design
- Strata-indexed probabilities assigned by employment size $\equiv$ pps
- $100000 - 150000$ establishments report employment changes
- $\overset{(d=4)\times 1}{\mathbf{x}}$ = (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, $\mathbf{X}_U = X_1, \ldots, X_N \sim P_\theta$.
- 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 $\mathbf{X}_U$
- $P_\theta(\mathbf{X}_S) \neq P_\theta(\mathbf{X}_U)$
- Want to estimate outliers from $P_\theta(\mathbf{X}_U)$ using $\mathbf{X}_S$.
- Use $\tilde{w}_i \propto 1/\pi_i$

Mixture / Cluster Model for Outlier Detection

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

$$
\begin{aligned}
\overset{d\times 1}{\mathbf{x}_i} | s_i, \mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K_{\max}})', \sigma^2, \tilde{w}_i &\overset{\text{ind}}{\sim} \mathcal{N}_d\left(\boldsymbol{\mu}_{s_i}, \sigma^2 \mathbb{I}_d\right)^{\tilde{w}_i} \\
s_i | \boldsymbol{\tau} &\overset{\text{iid}}{\sim} \mathcal{M}\left(1, \tau_1, \ldots, \tau_{K_{\max}}\right) \\
\boldsymbol{\mu}_p | G_0 &\overset{\text{iid}}{\sim} G_0 := \mathcal{N}_d\left(\mathbf{0}, \rho^2 \mathbb{I}_d\right) \\
\tau_1, \ldots, \tau_{K_{\max}} &\sim \mathcal{D}\left(\alpha/K_{\max}, \ldots, \alpha/K_{\max}\right)
\end{aligned}
$$

Sampling-weighted Pseudo Posterior

- Pseudo Posterior $\propto$ Weighted Likelihood $\times$ Priors
- Marginalize out τ from the joint prior, $f(\mathbf{s}, \boldsymbol{\tau}|\boldsymbol{\alpha}) = f(\mathbf{s}|\boldsymbol{\tau}) f(\boldsymbol{\tau}|\alpha)$
- $\underset{K\times d}{\mathbf{M}} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$
- $n_p = \sum_{i=1}^{n} \mathbf{1}(s_i = p) \equiv$ number of establishments assigned to cluster, p

$$f(\mathbf{s}, \mathbf{M}|\mathbf{X}, \tilde{\mathbf{w}}) \propto f(\mathbf{X}, \mathbf{s}, \mathbf{M}|\tilde{\mathbf{w}}) = \prod_{p=1}^{K} \prod_{i:s_i=p} \mathcal{N}_d\left(\mathbf{x}_i|\boldsymbol{\mu}_p, \sigma^2 \mathbb{I}_d\right)^{\tilde{w}_i}$$

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

$$\prod_{p=1}^{K} \mathcal{N}_d\left(\boldsymbol{\mu}_p|\mathbf{0}, \rho^2 \mathbb{I}_d\right).$$

Approximate MAP as $\sigma^2 \downarrow 0$

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

$$
\begin{aligned}
-2\sigma^2 \times \log f\left(\mathbf{X}, \mathbf{s}, \mathbf{M}, \tilde{\mathbf{w}}\right) = & \sum_{p=1}^{K} \sum_{i:s_i=p} \left[-2\sigma^2 \times \mathcal{O}\left(\log \sigma^2\right) + \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2\right] \\
& + K\lambda - 2\sigma^2 \times \mathcal{O}\left(1\right) \\
& - 2\sigma^2 \times \ \mathcal{O}\left(1\right),
\end{aligned}
$$

Approximate MAP Optimization

$$\underset{K,\mathbf{s},\mathbf{M}}{\operatorname{argmin}} \sum_{p=1}^{K} \sum_{i:s_i=p} \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2 + K\lambda,$$

- Bayesian motivation for K-means clustering
- Higher value for λ 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, $\boldsymbol{\mu}_p$
- 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 local clustering of $L_{\max}$ possible clusters in industry, j.
- Local cluster, c, in industry, j, connected to global cluster center, $\boldsymbol{\mu}_p$
- For $p \in (1, \ldots, K_{\max})$ possible global clusters
- Local clusters across industries may share a common global cluster
- $s_i^j \equiv$ global cluster assignment for establishment, i, in industry, j

Hierarchical Clustering Optimization

$$\underset{K,\mathbf{s},\mathbf{M}}{\text{argmin}} \sum_{p=1}^{K} \sum_{j=1}^{J} \sum_{i:s_i^j=p} \tilde{w}_i^j \|\mathbf{x}_i^j - \boldsymbol{\mu}_p\|^2 + K\lambda_K + L\lambda_L,$$

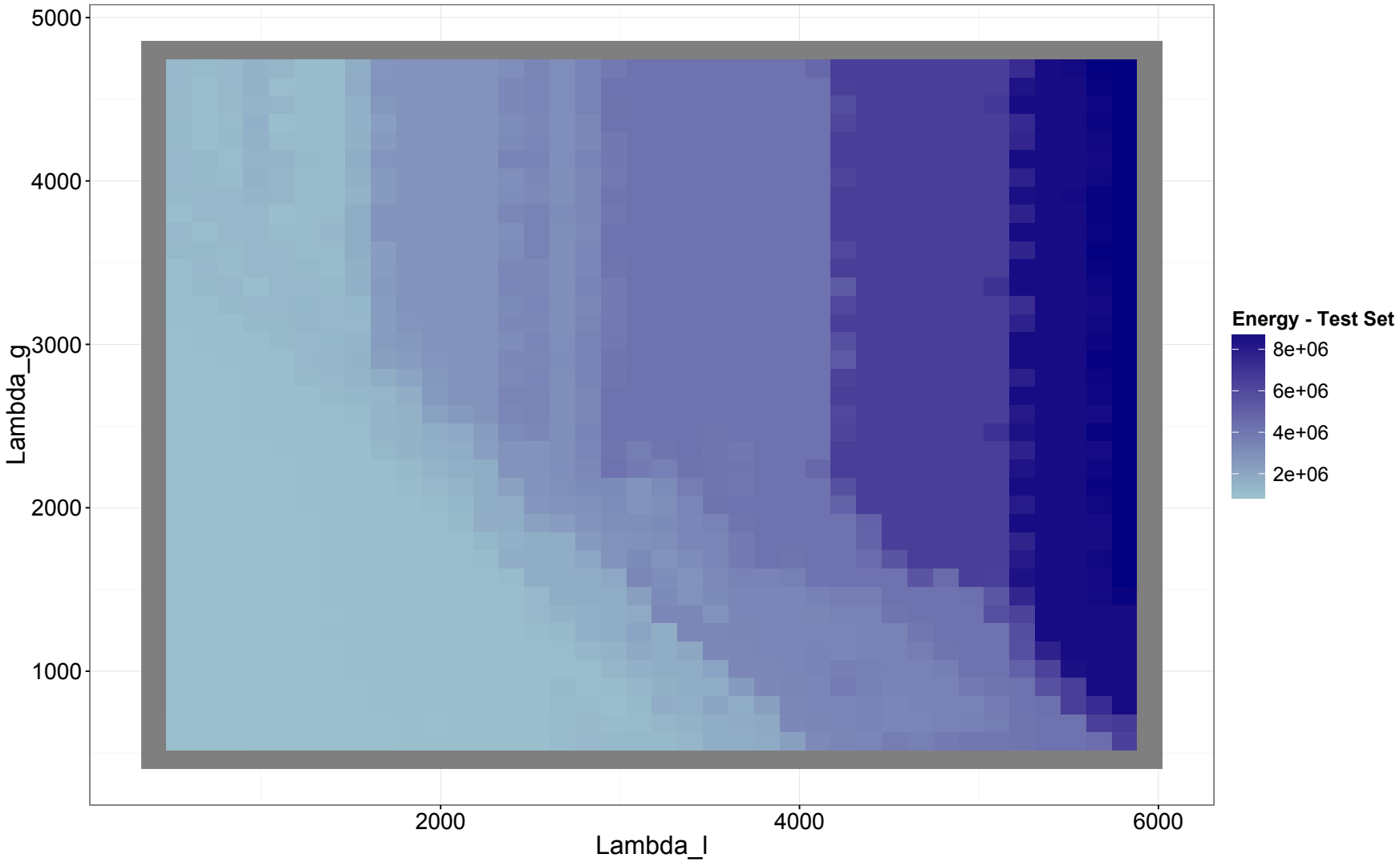
- $L = \sum_{j=1}^{J} L_j$ denotes the total number of local clusters
- L_j denotes the number local clusters estimated for data set, $j = 1, \ldots, J$
- K denotes the number of estimated global clusters
- λ_K denotes penalty on number of global clusters estimated
- λ_L denotes penalty on number of local clusters estimated
- $\tilde{w}_i^j$ is the sampling weight for establishment, i, in industry, j

Selecting Penalty Parameters, (λ_K, λ_L)

- Synthetic data, $L_j = 5$ local clusters for $j = 1, \ldots, (J = 3)$ industries
- Sharing $K = 7$ global clusters
- $\mathbf{X}_j^{N_j \times (d=15)}$
- $(N_j = 15000, n_j = 2500)$ 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 (λ_K, λ_L)

- Estimate clustering on training data and compute energy on test data



Use Calinski-Harabasz (C) criterion

- Cohesion within each cluster, $WGSS$
- Separation between clusters, $BGSS$

$$WGSS = \sum_{p=1}^{K} \sum_{i:s_i^v=k} \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2$$

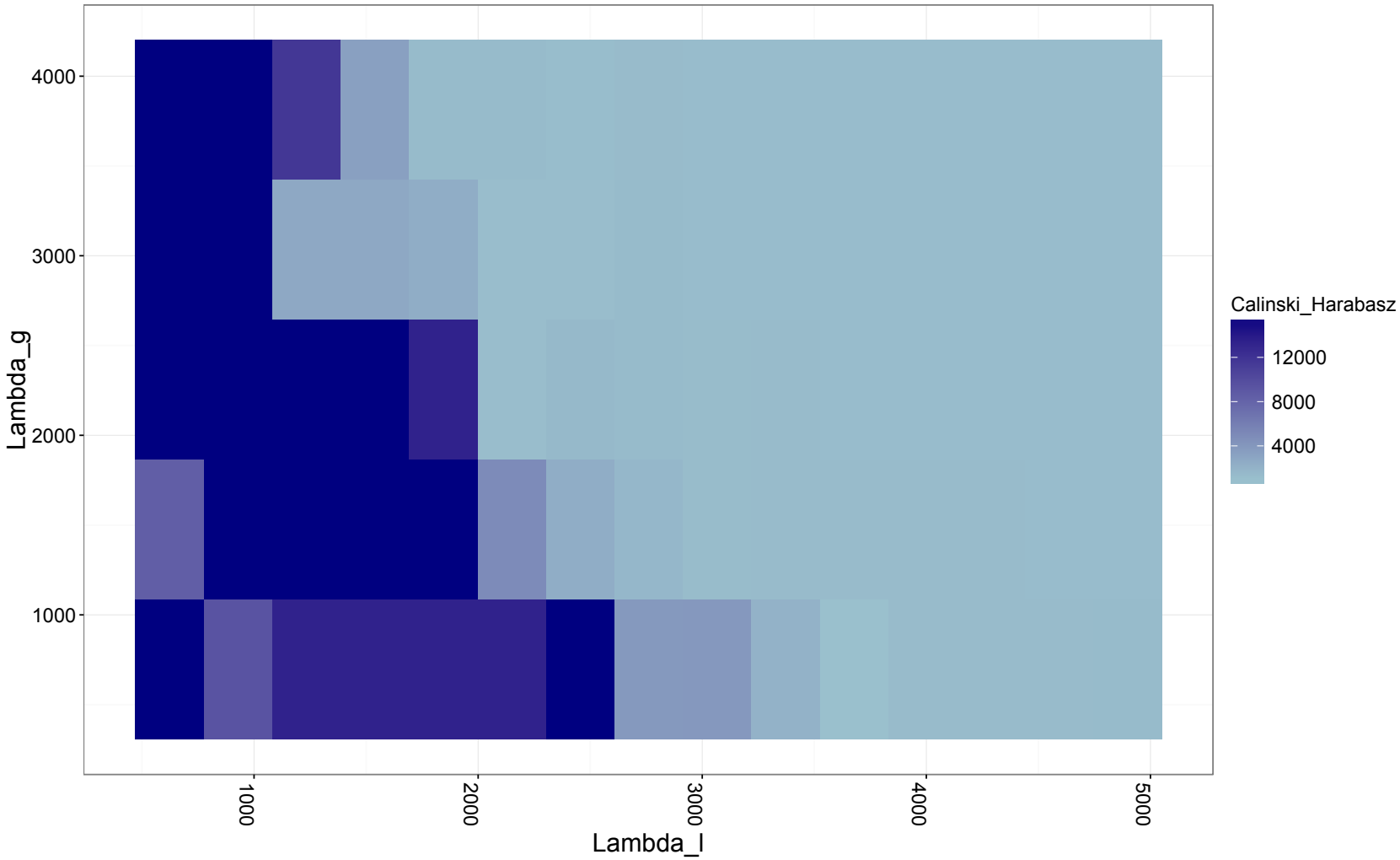
$$BGSS = \sum_{p=1}^{K} n_p \|\boldsymbol{\mu}_p - \boldsymbol{\mu}^G\|^2$$

$$C = \frac{n-K}{K-1} \frac{BGSS}{WGSS}$$

- $\boldsymbol{\mu}^G = \frac{\sum_{i=1}^{n} \tilde{w}_i \mathbf{x}_i}{\sum_{i=1}^{n} \tilde{w}_i}$
- K is number of global clusters

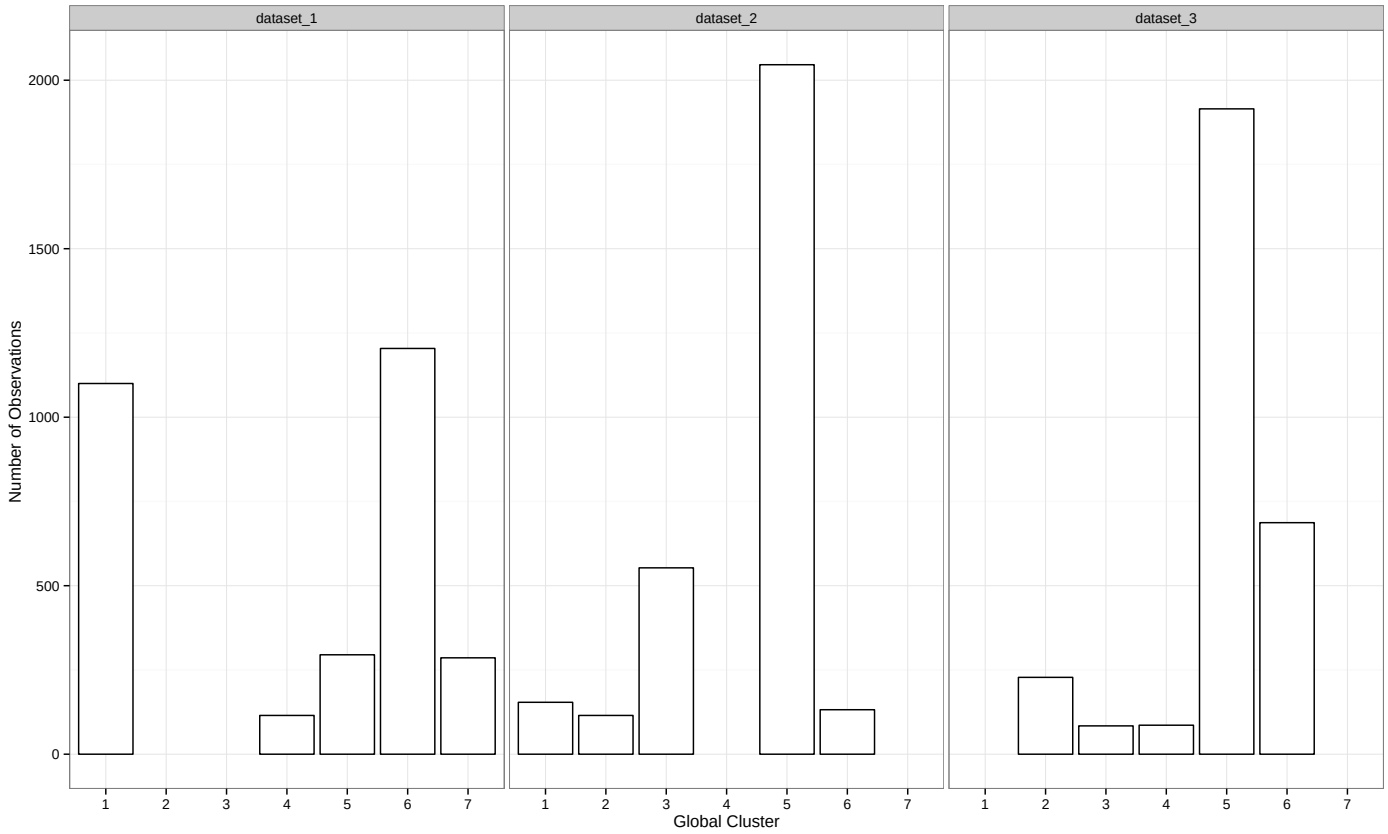
C finds an optimum

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



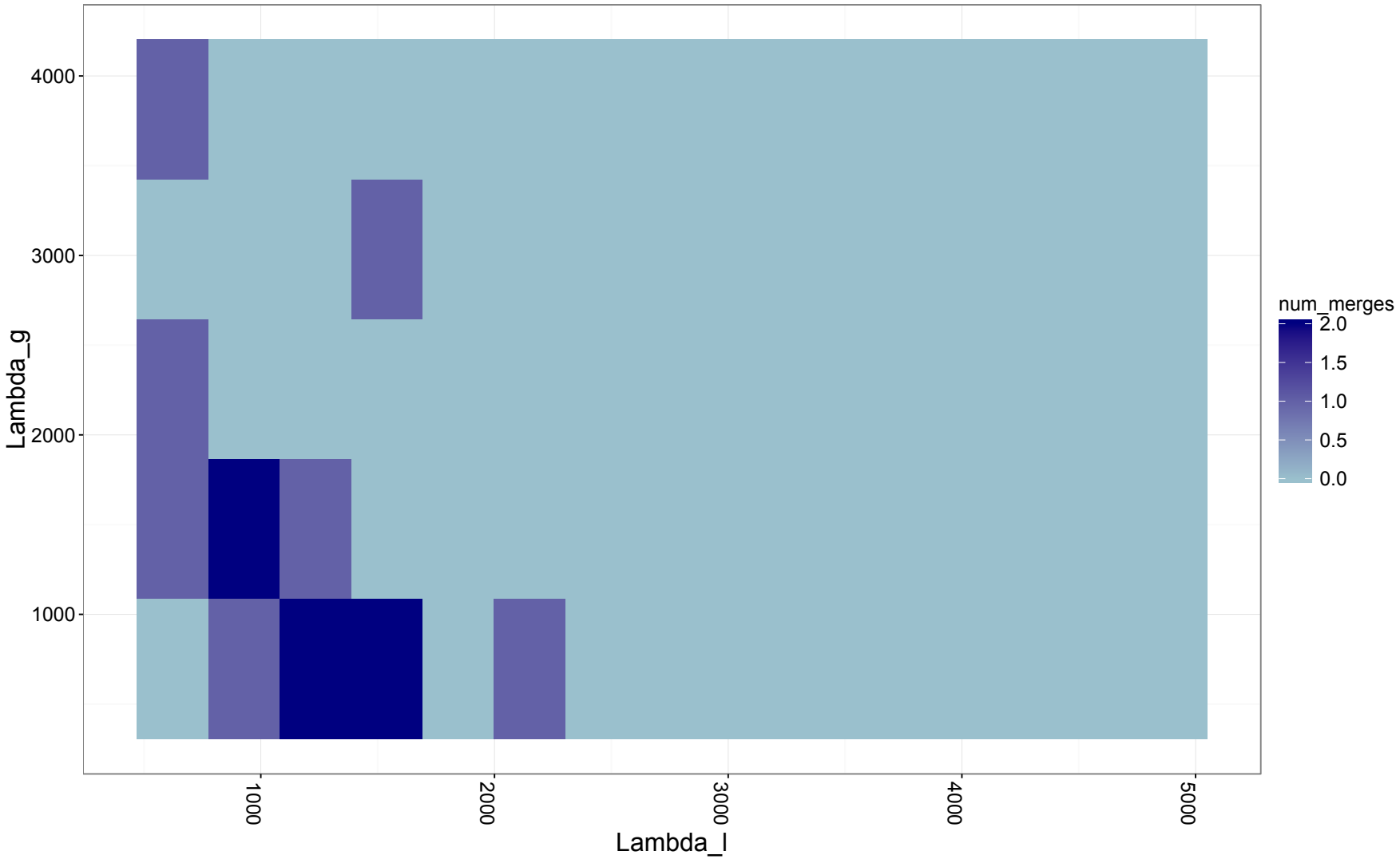
Correct Clusterings Estimated

- 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 (λ_K, λ_L)

- Higher number of merges for lower values of (λ_K, λ_L)



Outlier Detection Simulation Study Design

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

$$\overset{(d=15)\times 1}{\boldsymbol{\mu}_1} = (1, 1.5, 2.0, \ldots, 7.5, 8)$$

$$\boldsymbol{\mu}_2 = (8, 7.5, \ldots, 1)$$

$$\boldsymbol{\mu}_3 = (1, \ldots, 7, 8, 7, \ldots, 1)$$

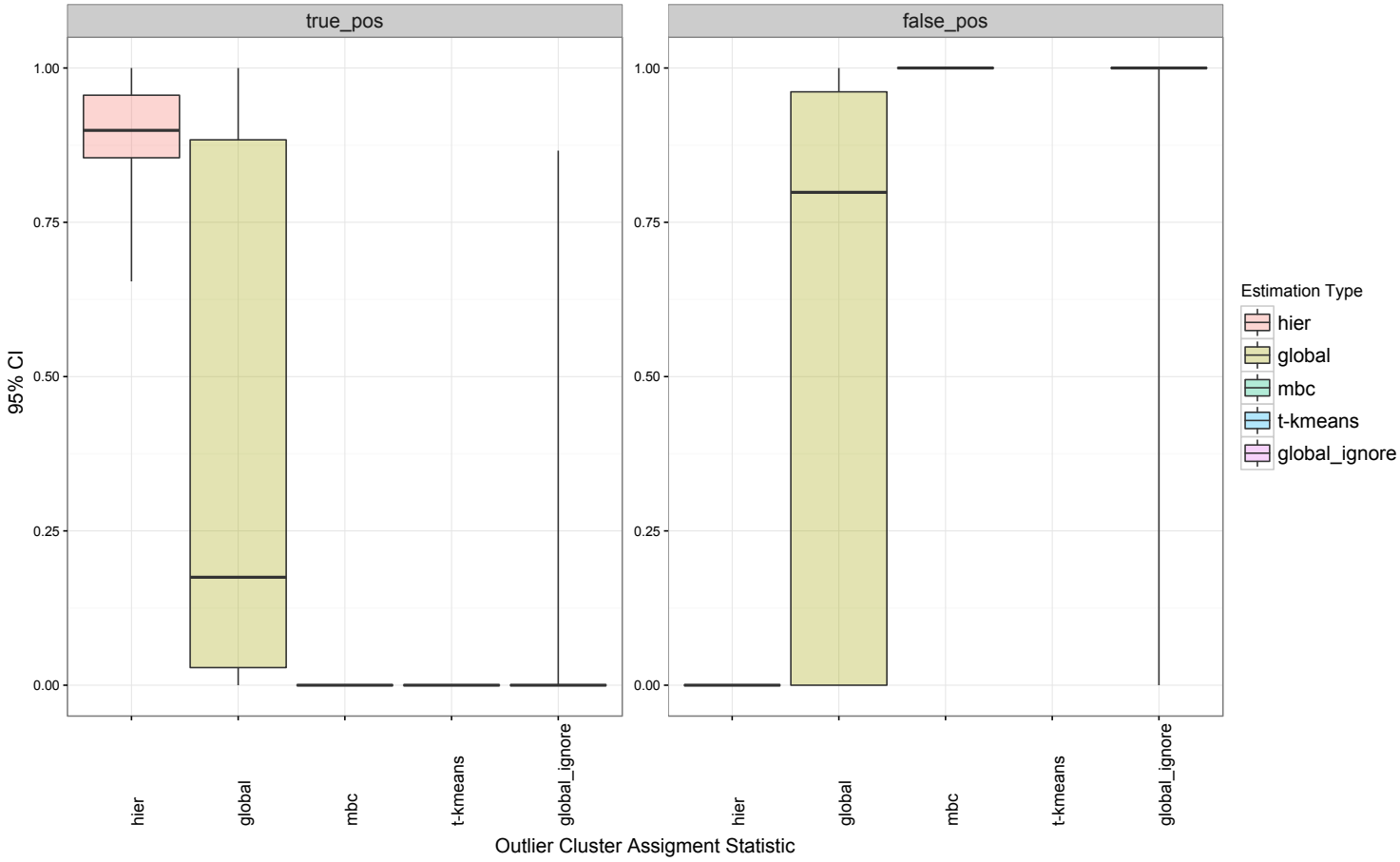
$$\boldsymbol{\mu}_4 = \text{Sampling from } (1, \ldots, 8) \text{ with replacement, } d = 15 \text{ times}$$

$$\boldsymbol{\mu}_5 = \text{Sampling from } (-2, \ldots, 6) \text{ with replacement, } d = 15 \text{ times,}$$

- mean $\boldsymbol{\mu}_5$ is assigned 150 observations
- Stratified design of $H = 10$ strata assign $\pi_h^j \propto$ variance of, $\mathbf{X}_h^j$
- $B = 100$ Monte Carlo draws

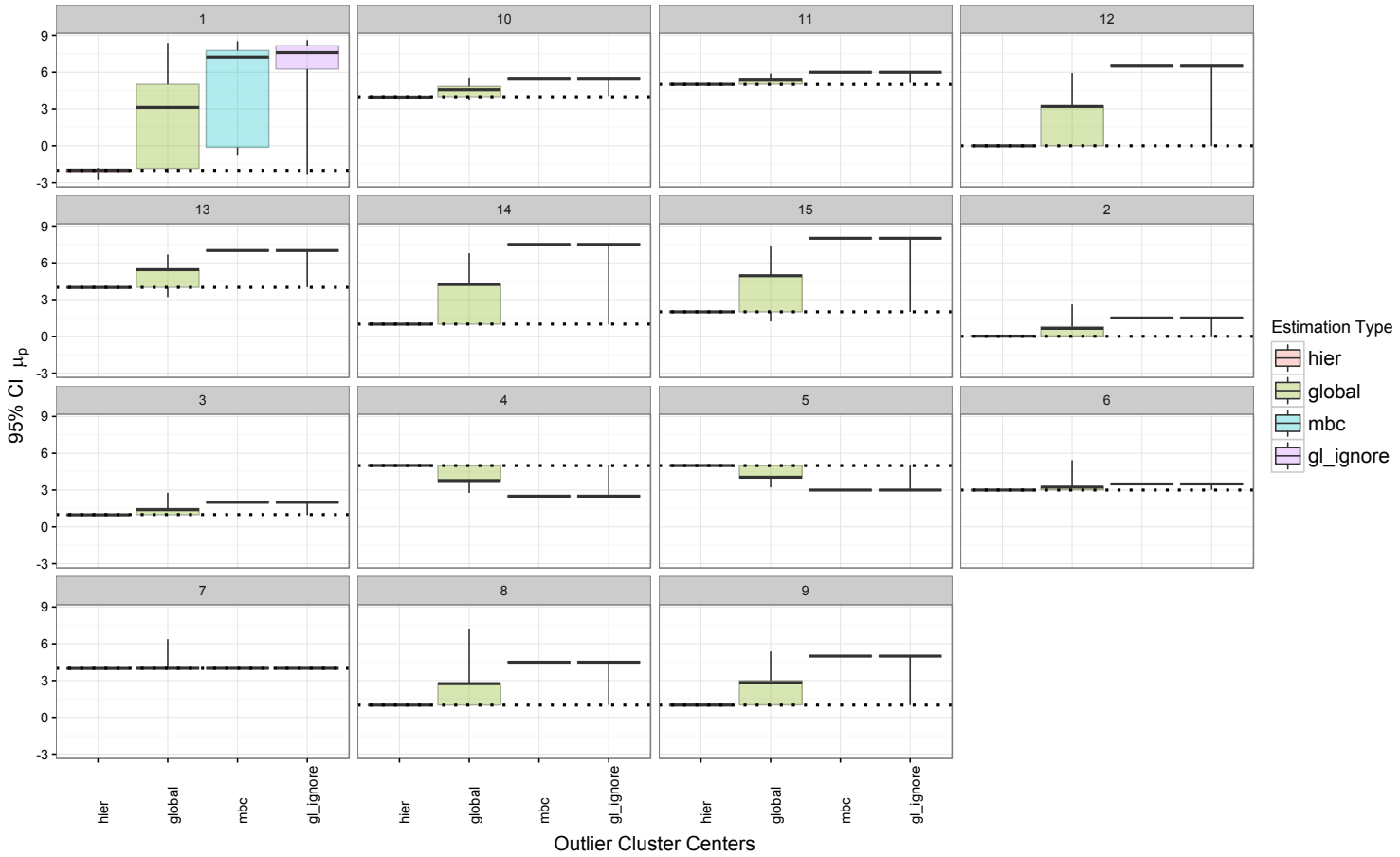
Outlier Detection Accuracy

- True positive ≡ # of true outliers discovered / total # of true outliers
- False positive ≡ # of false discoveries / total # nominated
- True positives measure effectiveness, False positives measure efficiency



Estimation Bias of Outlier Center, μ_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