

Scalable Approximate Bayesian Inference for Outlier Detection under Informative Sampling

Terrance D. Savitsky

SAVITSKY.TERRANCE@BLS.GOV

U. S. Bureau of Labor Statistics

Office of Survey Methods Research

Washington, DC 20212, USA

Editor: Kevin Murphy

Abstract

Government surveys of business establishments receive a large volume of submissions where a small subset contain errors. Analysts need a fast-computing algorithm to flag this subset due to a short time window between collection and reporting. We offer a computationally-scalable optimization method based on non-parametric mixtures of hierarchical Dirichlet processes that allows discovery of multiple industry-indexed local partitions linked to a set of global cluster centers. Outliers are nominated as those clusters containing few observations. We extend an existing approach with a new “merge” step that reduces sensitivity to hyperparameter settings. Survey data are typically acquired under an informative sampling design where the probability of inclusion depends on the surveyed response such that the distribution for the observed sample is different from the population. We extend the derivation of a penalized objective function to use a pseudo-posterior that incorporates sampling weights that “undo” the informative design. We provide a simulation study to demonstrate that our approach produces unbiased estimation for the outlying cluster under informative sampling. The method is applied for outlier nomination for the Current Employment Statistics survey conducted by the Bureau of Labor Statistics.

Key words: survey sampling, hierarchical dirichlet process, clustering, bayesian hierarchical models, optimization

1. Introduction

1.1 Outlier Detection and Informative Sampling

The U.S. Bureau of Labor Statistics (BLS) administers the Current Employment Statistics (CES) survey to over 350000 non-farm, public and private business establishments across the U.S. on a monthly basis, receiving approximately 270000 submitted responses in each month. Estimated total employment is published for local, state and national geographies in the U.S., as well as for domains defined by establishment size and industry categories within each geography. The BLS conducts a quality check to discover and correct establishment submission errors, particularly among those establishments whose entries are influential in the overall published domain-level estimates. Of the 270000 submissions, approximately 100000 – 150000 of those include employment changes from the prior to the current submission month. The CES maintains a short lag time of approximately 7 days between receipt of establishment submissions at the end of a month and subsequent publication of employment estimates for that month, such that investigations of submission data quality must be done quickly. The relatively large number of submissions with non-zero changes in employment levels, coupled with the rapid publication schedule, require use of quick-executing, automated data analysis tools that output a relatively small, outlying set of the submissions for further investigation and correction by BLS analysts.

The CES survey utilizes a stratified sampling design with strata constructed by combinations of state, broad industry grouping, and employment size (divided into 8 categories). Business establishments are sampled by their unique unemployment insurance (UI) tax identification numbers, which may contain a cluster of multiple individual sites. If a business establishment is selected for inclusion at the UI level, all of the associated sites in that cluster are also included. Stratum-indexed inclusion probabilities are set to be proportional to average employment size for member establishments of that stratum, which is

done because larger establishments compose a higher percentage of the published employment statistics. The correlation between establishment employment levels and inclusion probabilities induces informativeness into the sampling design, meaning that the probabilities of inclusion are correlated with the response variable of interest. The distribution for the resulting observed sample will be different from that for the population (because the sample emphasizes relatively larger establishments), such that inference made (e.g. about outliers) with the former distribution will be biased for the latter.

1.2 Methodologies for Outlier Detection

The Dirichlet process (DP) (Blackwell and MacQueen 1973) and more generally, species sampling formulations (Ishwaran and James 2003), induce a prior over partitions due to their almost surely discrete construction. Convolving a discrete distribution under a DP prior with a continuous likelihood in a mixture formulation is increasingly used for outlier detection (Quintana and Iglesias 2003), where one supposes that each observation is realized from one of multiple generation processes (Shotwell and Slate 2011). Each cluster collects an ellipse cloud of observations that are centered on the cluster mean.

In this article, we improve and extend sequential, penalized partition (clustering) algorithms of Kulis and Jordan (2011); Broderick et al. (2012), that each approximately estimate a maximum a posteriori (MAP) partition and associated cluster centers or means, to account for data acquired under an informative sampling design under which the distribution for the observed sample is different from that for the population (Bonnry et al. 2012). We incorporate first order sampling weights into our model for partitions that “undo” the sampling design so that our application nominates outliers with respect to the population generating (mixture) distribution. The algorithm of Kulis and Jordan (2011) is of a class of approaches (see also Wang and Dunson 2011; Shotwell and Slate 2011) that each produce

an approximation for the MAP from a mixture of DPs posterior distribution (where the (unknown) mixing measure over the parameters that define the data likelihood is drawn from a DP).

We follow Kulis and Jordan (2011) and straightforwardly extend our formulation from a DP prior imposed on the mixing distribution to a hierarchical Dirichlet process (HDP) prior (Teh et al. 2006), which allows our estimation of a distinct “local” partition (set of clusters) for each subset of establishments in the CES binned to one of $J = 24$ industry groupings. The local clusters of the industry-indexed partitions may share cluster centers (means) from a global clustering partition. The hierarchical construction of local clusters that draw from or share global cluster means permits the estimation of a dependence structure among the local partitions and encourages a parsimonious discovery of global clusters as compared to running separate global clustering models on each industry group. We refer to this model as hierarchical clustering.

We apply our (sampling-weighted) global and hierarchical clustering models, derived from the MAP approximations, for the nomination of outliers based on identification of the subset of clusters that together collect relatively few observations (with “few” defined heuristically). Quintana and Iglesias (2003) implement a mixture of DPs, for the purpose of outlier detection, where a partition is sampled at each step of an MCMC under an algorithm that minimizes a loss function. The loss function penalizes the size of the partition (or total number of clusters discovered), which will tend to assign most observations into relatively few clusters, while only a small number of observations will lie outside these clusters, a useful feature for outlier detection. We extend small variance asymptotic result of Broderick et al. (2012) for our global and hierarchical clustering formulations to produce objective functions that penalize the number of estimated clusters. We further devise a new “merge” step among all pairs of discovered clusters in each iteration of our estimation algorithms.

Both the penalized likelihood and the merge step encourage parsimony in the number of clusters discovered.

We next derive our sampling-weighted global and hierarchical clustering algorithms in Section 2, followed by a simulation study that demonstrates the importance of correcting estimation for an informative sampling design for outlier detection in Section 3. The simulation study also demonstrates the efficacy of outlier detection performance for the sampling-weighted hierarchical formulation estimated on synthetic data generated in a nested fashion that is similar to the structure of CES survey data. We apply the sampling-weighted hierarchical algorithm to discover local, by-industry partitions and to nominate observations for outliers for our CES survey application in Section 4. We conclude the paper with a discussion in Section 5.

2. Estimation of Partitions from Asymptotic Non-parametric Mixtures

We begin by specifying a Dirichlet process mixtures of Gaussians model to which we apply small variance asymptotics to derive a penalized K -means optimization expression and an associated fast-computing estimation algorithm. We include the sampling weights in our mixture formulation that asymptotically corrects for an informative sampling design (Savitsky and Toth 2015). We continue and extend this MAP approximation approach to a hierarchical Dirichlet process mixtures of Gaussians model to achieve a hard, point estimate sampling-weighted hierarchical clustering algorithm that permits specification of a set of local, industry indexed, clusters linked to a common set of global clusters.

We implement both algorithms we present, below, in the `growclusters` package for **R** (R Core Team 2014), which is written in **C++** for fast computation and available from the authors on request.

2.1 Asymptotic Sampling-weighted Mixtures of Dirichlet Processes

We specify a probability model for a mixture of K_{\max} allowed components to introduce Bayesian estimation of partitions of observations into clusters and the associated cluster centers. The mixtures of DPs will be achieved in the limit of this probability model as K_{\max} increases to infinity.

$$\mathbf{x}_i^{d \times 1} | s_i, \mathbf{M} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{K_{\max}})', \sigma^2, \tilde{w}_i \stackrel{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}_{s_i}, \sigma^2 \mathbb{I}_d)^{\tilde{w}_i} \quad (1a)$$

$$s_i | \boldsymbol{\tau} \stackrel{\text{iid}}{\sim} \mathcal{M}(1, \tau_1, \dots, \tau_{K_{\max}}) \quad (1b)$$

$$\boldsymbol{\mu}_p | G_0 \stackrel{\text{iid}}{\sim} G_0 := \mathcal{N}_d(\mathbf{0}, \rho^2 \mathbb{I}_d) \quad (1c)$$

$$\tau_1, \dots, \tau_{K_{\max}} \sim \mathcal{D}(\alpha/K_{\max}, \dots, \alpha/K_{\max}), \quad (1d)$$

for $i = 1, \dots, n$ sample establishments and $p = 1, \dots, K_{\max}$ allowable clusters. Each cluster candidate, p , indexes a unique value for the associated cluster center, $\boldsymbol{\mu}_p$. The set of cluster assignments of observations together compose a partition of observations, where partitions are indexed by $\mathbf{s} = (s_1, \dots, s_n)$, for $s_i \in (1, \dots, K)$. The prior distributions for the cluster centers and assignments induce a random distribution prior, G , for drawing a mean value for each \mathbf{x}_i for $i = 1, \dots, n$, by sampling unique values in the support of G from G_0 and assigning the mean value (that we interpret as a cluster) for each observation by drawing from the set of unique values with probabilities, $\boldsymbol{\tau}$, which are, in turn, randomly drawn from a Dirichlet distribution. As $K_{\max} \uparrow \infty$ in Equation 1d, such that number of possible clusters are countably infinite, this construction converges to a (non-parametric) mixture of Dirichlet processes (Neal 2000). We also note that α influences the number of clusters discovered. For any given K_{\max} , larger values for α will produce draws for $\boldsymbol{\tau}$ that assign larger probabilities to more of the $p \in (1, \dots, K_{\max})$ clusters, such that it is highly influential in the determination of the number of discovered clusters, $K < K_{\max}$.

The likelihood contribution for each observation, $i \in (1, \dots, n)$, in Equation 1a is up-lifted by a sampling weight, $\tilde{w}_i = n \times w_i / \sum_{i=1}^n w_i$ for $w_i \propto 1/\pi_i$, constructed to be inversely proportional to the known inclusion probability in the sample. This formulation for the sampling weight assigns importance for that observation likelihood based on the amount of information in the finite population represented by that observation. The weights sum to the sample size, n , so that an observation with a low inclusion probability in the sample will have a weight greater than 1, meaning that it “represents” more than a single unit in a re-balanced sample. The weighting serves to “undo” the sampling design by re-balancing the information in the observed sample to approximate that in the population. The sum of the weights directly impacts the estimation of posterior uncertainty such that the summation to n expresses the asymptotic amount of information in the observed sample. This construction is known as a “pseudo” likelihood because the distribution for the population is approximated by weighting each observation back to the population. Savitsky and Toth (2015) provide three conditions that define a class of sampling designs under which the pseudo posterior (defined from the pseudo likelihood) is guaranteed to contract in L_1 on the unknown true generating distribution, P_0 . The first condition states that the inclusion probability, π_i , for each unit in the finite population, $i \in (1, \dots, N)$, must be strictly greater than 0. This condition ensures that no portion of the population may be systematically excluded from the sample. The second condition on the sampling design requires the second order inclusion dependence for any two units, $i, j \in (1, \dots, N)$, be globally bounded from above by a constant; for example, a two-stage design in which units within selected clusters are sampled without replacement is within this class to the extent that the number of population units within each cluster asymptotically increases. The third condition requires the sampling fraction, n/N , to converge to a constant as the population size limits to infinity.

Samples may be drawn from the joint posterior distribution under Equation 1 using Markov chain Monte Carlo (MCMC), though the computation will not scale well with increasing sample size as a partition is formed through sequential assignments of observations to clusters (including to possibly new, unpopulated clusters) on each MCMC iteration. We, instead, derive an optimization expression that allows for scalable approximation of the single MAP partition by taking the limit of the joint posterior distribution as the likelihood variance, $\sigma^2 \downarrow 0$.

We marginalize out τ from the joint prior, $f(\mathbf{s}, \boldsymbol{\tau} | \boldsymbol{\alpha}) = f(\mathbf{s} | \boldsymbol{\tau}) f(\boldsymbol{\tau} | \boldsymbol{\alpha})$, by using the Pólya urn scheme of Blackwell and MacQueen (1973), which produces,

$$f(s_1, \dots, s_n | \boldsymbol{\alpha}) = \alpha^K \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + n)} \prod_{p=1}^K (n_p - 1)!,$$

where $n_p = \sum_{i=1}^n \mathbf{1}(s_i = p)$ is the number of observations (e.g. establishments) assigned to cluster, p and K denotes the number of estimated clusters. We generally follow Broderick et al. (2012) and derive our optimization expression from the joint pseudo posterior distribution,

$$\begin{aligned} f(\mathbf{s}, \mathbf{M} | \mathbf{X}, \tilde{\mathbf{w}}) &\propto f(\mathbf{X}, \mathbf{s}, \mathbf{M} | \tilde{\mathbf{w}}) = \prod_{p=1}^K \prod_{i:s_i=p} \mathcal{N}_d(\mathbf{x}_i | \boldsymbol{\mu}_p, \sigma^2 \mathbb{I}_d)^{\tilde{w}_i} \\ &\alpha^K \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + n)} \prod_{p=1}^K (n_p - 1)! \quad (2) \\ &\prod_{p=1}^K \mathcal{N}_d(\boldsymbol{\mu}_p | \mathbf{0}, \rho^2 \mathbb{I}_d). \end{aligned}$$

Broderick et al. (2012) point out that if one limits $\sigma^2 \downarrow 0$ in Equation 2, each observation will be allocated to its own cluster since the prior allows a countably infinite number of clusters. To avoid this degenerate outcome, define a constant λ and set $\alpha = \exp(-\lambda / (2\sigma^2))$, which

produces $\alpha \downarrow 0$ as $\sigma^2 \downarrow 0$. This functional form for α achieves the result and avoids the degenerate outcome, where the λ hyperparameter controls the size of the partition as $\sigma^2 \downarrow 0$. Commonly-used approaches that also specify an approximate MAP of a mixture of DPs also restrict α ; Shotwell and Slate (2011) fix the value for α and Wang and Dunson (2011) restrict the values of α to a discrete grid. Plugging in this expression for α into Equation 2, and taking -1 times the logarithm of both sides results in,

$$\begin{aligned} -\log f(\mathbf{X}, \mathbf{s}, \mathbf{M}, \tilde{\mathbf{w}}) &= \sum_{p=1}^K \sum_{i:s_i=p} \left[\mathcal{O}(\log \sigma^2) + \frac{\tilde{w}_i}{2\sigma^2} \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2 \right] \\ &+ K \frac{\lambda}{2\sigma^2} + \mathcal{O}(1) \\ &+ \mathcal{O}(1), \end{aligned} \tag{3}$$

where the expression in the first line derives from the log-kernel of a multivariate Gaussian distribution and $f(\sigma^2) = \mathcal{O}(h(\sigma^2))$ denotes that there exist constants, $c_1, c_2 > 0$, such that $|f(\sigma^2)| \leq c_1 |h(\sigma^2)|$, for $\sigma^2 < c_2$. Multiplying both sides of Equation 3 by $2\sigma^2$ and taking the limit as $\sigma^2 \downarrow 0$ results in the asymptotically equivalent sampling-weighted global clustering optimization problem,

$$\operatorname{argmin}_{K, \mathbf{s}, \mathbf{M}} \sum_{p=1}^K \sum_{i:s_i=p} \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2 + K\lambda, \tag{4}$$

whose solution is a MAP approximation for the sample-adjusted mixtures of DPs of Equation 1. We minimize a sampling-weighted distance in Equation 4, which will approximate a MAP partition and associated cluster centers with respect to the distribution for the *population*, which is our interest, rather than that for the realized sample.

Although we derive the sampling-weighted approximate MAP by letting the variance around a partition, σ^2 , limit to 0, this constructive derivation offers no comment on whether

the distribution over the space of partitions for a particular data set expresses a low or high variance. Our purpose is to extract a hard, point estimate that computationally scales to a large number of multivariate observations (that we employ for the purpose of nominating outliers). To the extent that computational considerations allow, one would obtain richer inference on both an asymptotically exact MAP of the marginal distribution over the space of partitions and an associated measure of uncertainty by using the fully Bayesian nonparametric mixture model.

Shotwell and Slate (2011) note that employing a MAP approximation (which selects the partition assigned the largest approximated value for the posterior mass) implicitly uses a 0 – 1 loss function. While other choices for the loss function may produce useful results, we prefer use of the MAP approximation as we would expect concentration of the estimated joint posterior distribution from the model of Equation 1 around the MAP due to the large sample sizes for our CES application.

2.2 Sampling-weighted Global Clustering Algorithm

We specify an estimation algorithm for MAP approximation, below, which is guaranteed to converge to a local optimum, since each step reduces (or doesn't increase) the objective function of Equation 4 and there are only a finite number of possible partitions. The algorithm operates, sequentially, on the observations and specifies successive assignment and cluster center estimation steps.

We introduce a new merge step, that tests all unique cluster pairs for merging and conducts a merge of any pair of clusters if such reduces the objective function, which reduces the sensitivity of the estimated partition to the specified values of the penalty parameter.

The observations are each weighted using the sampling weights in the computations of distances to cluster centers and the computation of cluster centers. Observations with higher weights contribute relatively more information to the computation of cluster centers. We demonstrate in the simulation study to follow that this procedure produces nearly unbiased population estimates for cluster centers, conditioned on the generation of the finite population from disjoint clusters. Since lower-weighted observations (less representative of the population from which they were drawn) contribute relatively less to the objective function, the weighting will also impact the number of clusters discovered and assignment of the multivariate observations to those clusters.

The algorithm for estimating a MAP partition and associated cluster centers is specified with,

1. **Input:** $n \times d$ data matrix, \mathbf{X} ; $n \times 1$ vector of sampling weights, $\tilde{\mathbf{w}}$, for observed units and cluster penalty parameter, λ .
2. **Output:** $K \times d$ matrix of cluster centers, $\mathbf{M} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)'$, and $n \times 1$ vector of cluster assignments, \mathbf{s} , where $s_i \in (1, \dots, K)$ for units, $i = 1, \dots, n$.
3. **Initialize:** $s_i = 1, \forall i$. $d \times 1$ cluster center, $\boldsymbol{\mu}_1 = \sum_{i=1}^n (\mathbf{x}_i \tilde{w}_i) / \sum_{i=1}^n \tilde{w}_i$.
4. Repeat the following steps until convergence (when the decrease in energy is below

$$e \leftarrow \sum_{p=1}^K \sum_{i \in \{i: s_i = p\}} w_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2 + \lambda K$$

a set threshold).

5. **Assign units to clusters for each unit, $i = 1, \dots, n$:**
 - (a) Compute distance, $d_{ip} = \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2$ for $p = 1, \dots, K$.
 - (b) If $\min_p d_{ip} > \lambda$, create new cluster to which unit i is assigned; $K \leftarrow K + 1$, $s_i \leftarrow K$, $\boldsymbol{\mu}_K \leftarrow \mathbf{x}_i$; else $s_i \leftarrow \operatorname{argmin}_p d_{ip}$.
6. **Re-compute centers for clusters, $p = 1, \dots, K$:** Let $\mathbf{S}_p = \{i : s_i = p\}$ and $\boldsymbol{\mu}_p = \sum_{i \in \mathbf{S}_p} (\mathbf{x}_i w_i) / \sum_{i \in \mathbf{S}_p} \tilde{w}_i$.

7. Assess merge of unique cluster pairs, $(p \in (1, \dots, K), p' \in (1, \dots, K))$:

- (a) Perform test merge of each pair of clusters:
- (b) Set matrix of cluster centers for virtual step, $\mathbf{M}^* = \mathbf{M}$.
- (c) Compose index set of observations assigned to clusters p' or p ; $\mathbf{S}_p^* = \{i : s_i = p \text{ or } s_i = p'\}$.
- (d) Compute (weighted average) merged cluster centers, $\boldsymbol{\mu}_p^* = \left(\sum_{i \in \mathbf{S}_p^*} \mathbf{x}_i \tilde{w}_i \right) / \sum_{i \in \mathbf{S}_p^*} \tilde{w}_i$ and set rows p and p' of \mathbf{M}^* equal to this value.
- (e) Compute energy under the tested move, $e^* \leftarrow \sum_{p=1}^K \sum_{i \in \mathbf{S}_p^*} \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p^*\|^2 + \lambda(K - 1)$. If $e^* < e$, execute a merge of p and p' :
- (f) Re-assign units linked to p' to p , $\mathbf{s}_{i:s_i=p'} \leftarrow p$.
- (g) Re-set cluster labels, p in \mathbf{s} to be contiguous after removing p' , such that $p \leftarrow p - 1, \forall p > p'$, leaving $s_i \in (1, \dots, K - 1)$. Delete row, p' , from \mathbf{M}^* and set $\mathbf{M} \leftarrow \mathbf{M}^*$.

2.3 Asymptotic Sampling-weighted Mixtures of Hierarchical Dirichlet
Processes

The CES survey is administered to establishments across a diverse group of industries that comprise the U.S. economy. CES survey analysts examine month-over-month reporting changes in the average employment and production worker variables within each of 24 NAICS industry categories (outlined in Section 4 that analyzes a data set of CES responses). The by-industry focus reflects the experience that both reporting patterns and establishment processes tend to express more similarity within than between industries. Yet, there are also commonalities in the reporting processes and types of reporting errors committed among the establishments due to overlapping skill sets of reporting analysts, similarities in organization structures, as well as the use of a single BLS-suggested reporting process.

So we want to perform separate estimations of partitions within each industry group, but allow those partitions to express similarities or dependencies across industries. We next

use the result of Kulis and Jordan (2011) that generalizes the approximate MAP algorithm obtained from mixtures of Dirichlet processes to mixtures of hierarchical Dirichlet processes. The hierarchical Dirichlet process (HDP) of Teh et al. (2006) specifies a DP partition for each industry that we refer to as a local (to industry) partition or clustering. These local partitions draw their cluster center values from a set of “global” clusters, such that local clusters may be connected across industry partitions by their sharing of common global cluster centers (from a single global partition) in a hierarchical manner.

We next specify a hierarchical mixture model that will converge to a sampling-weighted mixtures of HDPs in the limit of the *allowable* maximum number of global and local clusters as a generalization of our earlier DP specification,

$$\mathbf{x}_i^j | s_i^j, \mathbf{M} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{K_{\max}})', \sigma^2, \tilde{w}_i \stackrel{\text{ind}}{\sim} \mathcal{N}_d \left(\boldsymbol{\mu}_{s_i^j}, \sigma^2 \mathbb{I}_d \right)^{\tilde{w}_i}, \quad i = 1, \dots, n_j \quad (5a)$$

$$v_c^j | \boldsymbol{\tau} \stackrel{\text{iid}}{\sim} \mathcal{M}(1, \tau_1, \dots, \tau_{K_{\max}}), \quad c = 1, \dots, L_{\max} \quad (5b)$$

$$z_i^j | \boldsymbol{\pi}^j \stackrel{\text{iid}}{\sim} \mathcal{M}(1, \pi_1^j, \dots, \pi_{L_{\max}}^j) \quad (5c)$$

$$s_i^j = v_{z_i^j}^j \quad (5d)$$

$$\boldsymbol{\mu}_p | G_0 \stackrel{\text{iid}}{\sim} G_0 := \mathcal{N}_d(\mathbf{0}, \rho^2 \mathbb{I}_d) \quad (5e)$$

$$\tau_1, \dots, \tau_{K_{\max}} \sim \mathcal{D}(\alpha/K_{\max}, \dots, \alpha/K_{\max}) \quad (5f)$$

$$\pi_1^j, \dots, \pi_{L_{\max}}^j \sim \mathcal{D}(\gamma/L_{\max}, \dots, \gamma/L_{\max}), \quad (5g)$$

for each industry partition, $j = 1, \dots, (J = 24)$, where K_{\max} denotes the allowable number of global clusters and L_{\max} denotes the maximum allowable number of local clusters over all J partitions. An $n_j \times d$ data matrix, \mathbf{X}^j , includes the $d \times 1$ responses, (\mathbf{x}_i^j) , for establishments in industry, j . The $L_{\max} \times 1$, vector, \mathbf{v}^j , indexes the global cluster assignment for each local cluster, $c \in (1, \dots, L)$ for industry, j . The $n_j \times 1$ vector, \mathbf{z}^j , assigns each individual establishment, $i \in (1, \dots, n_j)$, to a local cluster, $c \in 1, \dots, L_{\max}$. So, z_i^j indexes

the local cluster assignment for establishment, i , in industry, j and v_c^j indexes the global assignment for all of the establishments assigned to local cluster, c , in industry, j . We may chain together the local cluster assignment for observation, i , and the global cluster assignment for the local cluster containing i to produce the $n_j \times 1$ index, \mathbf{s}^j , that holds the global cluster assignment for each establishment in industry, j . The vectors of probabilities for local cluster assignments of establishments, $(\boldsymbol{\pi}^j)_{j=1,\dots,J}$, and for global cluster assignments of local clusters, $\boldsymbol{\tau}$, are random probability vectors drawn from Dirichlet distributions. As before, Equation 5 converges to a sampling-weighted mixture of HDPs in the limit to infinity of K_{\max} and L_{\max} (Teh et al. 2006).

We achieve the following optimization algorithm by following a similar asymptotic derivation in the limit of $\sigma^2 \downarrow 0$, as earlier,

$$\operatorname{argmin}_{K,\mathbf{s},\mathbf{M}} \sum_{p=1}^K \sum_{j=1}^J \sum_{i:s_i^j=p} \tilde{w}_i^j \|\mathbf{x}_i^j - \boldsymbol{\mu}_p\|^2 + K\lambda_K + L\lambda_L, \quad (6)$$

where $L = \sum_{j=1}^J L_j$ denotes the total number of estimated local clusters and L_j denotes the number local clusters estimated for data set, $j = 1, \dots, J$. As before, K denotes the number of estimated global clusters.

2.4 Sampling-weighted Hierarchical Clustering Algorithm

We sketch a summary overview of the MAP approximation algorithm derived from sampling-weighted mixtures of HDPs. A more detailed, step-by-step, enumeration of the algorithm is offered in Appendix A. As with the sampling-weighted global clustering estimation, the algorithm will return a $K \times d$ matrix of global cluster centers, \mathbf{M} , and a vector for each industry, \mathbf{s}^j , $j \in (1, \dots, J)$, that holds the global cluster assignments, $s_i^j \in (1, \dots, K)$, for the set of n_j observations in industry, j . Each vector \mathbf{s}^j conveys additional information

about the local partitions of the hierarchical clustering. The estimation algorithm allows a distinct number of local clusters, L_j , to be discovered for each industry. The number of unique global cluster values, $p \in (1, \dots, K)$, that are assigned across all the establishments in industry j defines the structure - number of local clusters and establishments assigned to each cluster - of the local partition for that industry.

The algorithm contains an assignment step to directly assign each establishment, i , in industry, j , to a global cluster, as with the mixtures of DPs algorithm. For a new global and local cluster to be created, however, the minimum distance to the global cluster centers must be greater than the sum of local and global cluster penalty parameters, $\lambda_L + \lambda_K$. An observation is assigned to a global cluster by first assigning it to a local cluster that is, in turn, next assigned to that global cluster. A new local cluster will be created if an observation in an industry is assigned to a global cluster, p , that is not currently assigned to any local cluster for that industry. It is typical for only a subset of the K global clusters to be used in each local partition across the J industries. To the extent that the industry partitions share global clusters, then $K < \sum_{j=1}^J L_j$, and there will be a dependence among those partitions.

A second assignment step performs assignments to global clusters for *groups* of establishments in industry, j , who are assigned to the same local cluster, c . All of the establishments in local cluster, c , in industry, j , may have their assignments changed from global cluster, p , to global cluster, p' . This group assignment of establishments to global clusters contrasts with only changing global cluster assignments for establishments on an individual basis. This is a feature of the HDP and helps mitigate order dependence in the estimation of the MAP approximation. We specify a new “shed” step (that is not included in the algorithm of Kulis and Jordan (2011)) to remove any previously defined

global clusters that may not be linked to any local clusters across the J industries after this second assignment step of local-to-global clusters.

We continue to employ a merge step that tests each unique pair of global clusters for merging. We find in runs performed on synthetic data that the merge step helps reduce (but does not eliminate) the sensitivity (in number of clusters formed) to specifications of (λ_K, λ_L) . The number of merges increase under lower values for these penalty parameters. If global cluster, p' , is merged into p and deleted, then so are the local clusters, \mathbf{c} , assigned to p' across the J industries and establishments that were previously assigned to each $c \in \mathbf{c}$ are now re-assigned to existing local clusters linked to global cluster, p .

3. Simulation Study

We conduct a two-part Monte Carlo simulation study that assesses the effectiveness to detect an outlying cluster with respect to the population from an informative sample, where an outlying cluster is defined to contain a small percentage of the total number of establishments. Both parts of the simulation study generate a synthetic finite “population” from which repeated samples are drawn under a sampling design where the probabilities of selecting each establishment are a function of their response values, such that we produce informative samples. The informative samples are presented to the partition estimation models. The first part of the simulation study compares the accuracy of outlier detection under inclusion of sampling weights, used to correct for informativeness, versus excluding the sampling weights (where both employ the global (hard) clustering model using the estimation algorithm specified in Section 2.2. The sampling weights are set to a vector of 1’s in the case that excludes sampling weights). The second part of the simulation study compares the outlier estimation performance between the sampling-weighted global clustering model, on the one hand, and the sampling-weighted hierarchical clustering model,

on the other hand, for data generated from local partitions that share global cluster centers. We first devise an approach to select the penalty parameters for our estimation models, which determine the number of clusters discovered, that we will need for our two-part study.

3.1 Selection of Penalty Parameters

It is difficult to find a principled way to specify the penalty parameter, λ , for the global clustering model or (λ_L, λ_K) , for the hierarchical clustering model. Their specification is critical because these penalty parameters determine the number of discovered clusters. Relatively larger values estimate fewer clusters. Both cross-validation and randomly separating the data into distinct training and test sets (in the case the analyst possesses a sufficient number of observations) tends towards assignment of each observation to its own cluster by selecting nearly 0 values for the penalty parameters as poor fit overwhelms the penalty.

We demonstrate this tendency with a simulation study using the sampling-weighted hierarchical clustering model. We generate a population with $J = 3$ data sets, each of size $N_j = 15000$, with the local partition structure in each data set composed of $L_j = 5$, $j = 1, \dots, J$ clusters randomly selecting from $K = 7$ global cluster centers. Establishments are randomly assigned to the $L_j = 5$ local clusters for each data set under a skewed distribution of $(0.6, 0.25, 0.1, 0.025, 0.025)$ to roughly mimic the structure we find in the CES data set analysed in Section 4. Data observations are generated from a multivariate Gaussian,

$$\mathbf{x}_i^j \stackrel{\text{ind}}{\sim} \mathcal{N}_d \left(\boldsymbol{\mu}_{s_i^j}, \sigma^2 \mathbb{I}_d \right),$$

centered on the local cluster mean of each observation and σ is set equal to 1.5 times the average over the $J = 3$ data sets, $N_j = 15000$ establishments and $d = 15$ dimensions of the assignment-weighted centers, $\mathbf{M}_{s^j, 1:d}$, the mean values of \mathbf{X}_j .

We next take a sample of $n_j = 2250$ from each of the $J = 3$ data sets under a fixed sample size, proportional-to-size design where inclusion probabilities are set to be proportional to the variances of the $(d = 15) \times 1$, $\{\mathbf{x}_i\}$, inducing informativeness. The taking of an informative sample addresses the class of data in which we are interested and mimics our motivating CES data application. Observations are next randomly allocated into two sets of equal size; one used to train the model and the other to evaluate the resultant energy. The training sample is presented to the sampling-weighted hierarchical clustering model. Figure 1 displays the sampling-weighted hierarchical objective function (energy) values, evaluated on the test set, over a grid of global and local cluster penalty parameters. We observe that the energy doesn't reach a balanced optimum, but decreases along with values of the penalty parameters. The rate of decrease declines for these synthetic data, possibly suggesting the identification of an "elbow" for selecting penalty parameters, though there is little-to-no sensitivity in the values for the global penalty parameter, λ_K , which prevents precise identification of a value. The rate of decrease in energy on the CES data does not decline, however. We see a similar result under 10– fold cross-validation.

As an alternative, we employ penalty parameter selection statistics that combine measures of cohesion within clusters and separation between clusters to select the number of clusters. We devise a sampling-weighted version of the Calinski Harabasz (C) criterion, which is based on within cluster sum-of-squares (*WGSS*) and between clusters sum-of-

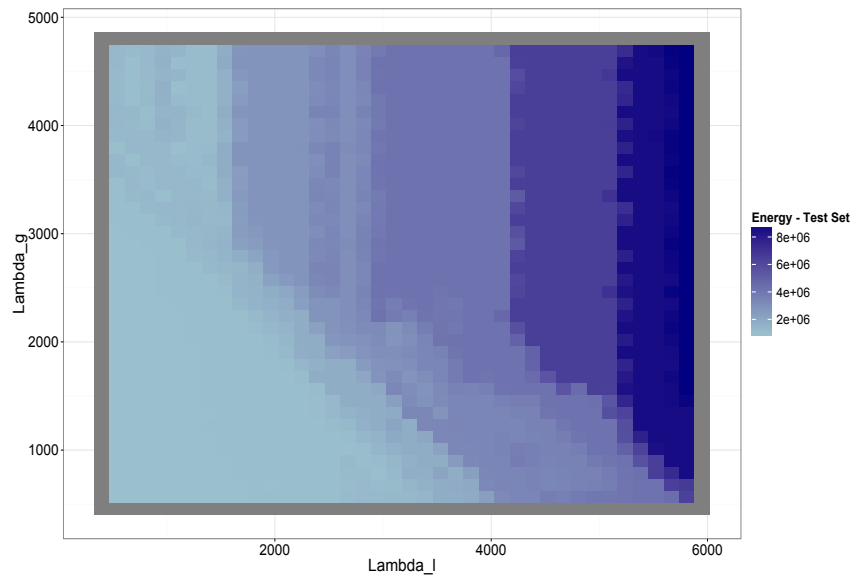


Figure 1: Heat map of (energy) values under the sampling-weighted hierarchical clustering model over grid of global and local cluster penalty parameters. The vertical axis represents the global cluster penalty, λ_K , and the horizontal axis, the local cluster penalty, λ_L . The energy is measured from assignment to a test set not used for training the model. The darker colors represent higher energy (optimization function) values.

squares (*BGSS*),

$$\begin{aligned}
 WGSS &= \sum_{p=1}^K \sum_{i:s_i^v=k} \tilde{w}_i \|\mathbf{x}_i - \boldsymbol{\mu}_p\|^2 \\
 BGSS &= \sum_{p=1}^K n_p \|\boldsymbol{\mu}_p - \boldsymbol{\mu}^G\|^2,
 \end{aligned}$$

where $\mathbf{s}^v = (\mathbf{s}^1, \dots, \mathbf{s}^J)'$ stacks the set of J , $\{\mathbf{s}^J\}$ into a vector, which retains information on the local partitions (based on the co-clustering relationships). The weighted total number of establishments linked to each cluster, $p \in (1, \dots, K)$ is denoted by $n_p = \sum_{i:s_i=p} \tilde{w}_i$ and $\boldsymbol{\mu}^G = \frac{\sum_{i=1}^n \tilde{w}_i \mathbf{x}_i}{\sum_{i=1}^n \tilde{w}_i}$. The C criterion is then formed as, $C = \frac{n-K}{K-1} \frac{BGSS}{WGSS}$, where K is determined by the number of unique values in \mathbf{s}^v , which is equal to the number of rows in the $K \times d$, $\mathbf{M} = (\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_K)$ (and d denotes the dimension of each observation, \mathbf{x}_i). Larger values for C are preferred.

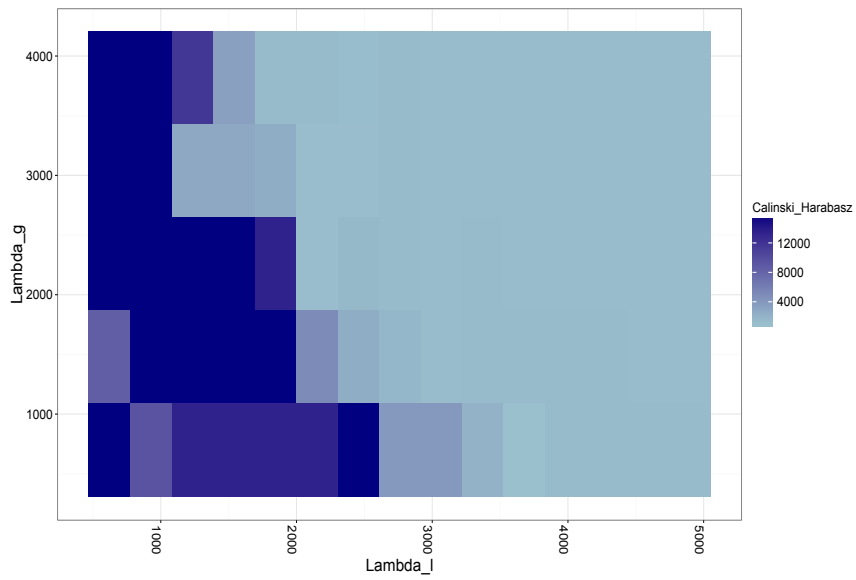


Figure 2: Heat map of Calinski Harabasz (C) index values under the mixtures of HDP model over grid of global and local cluster penalty parameters. The vertical axis represents the global cluster penalty, λ_K , and the horizontal axis, the local cluster penalty, λ_L . The darker colors represent higher C values. Global and local cluster penalty parameters with higher C values are preferred.

Our proposed procedure estimates a clustering on the sampled data over a grid of (local, global) penalty parameter values, where we compute the C statistic for each (λ_L, λ_K) in the grid, selecting that clustering which maximizes the C statistic. (Our `growclusters` package uses the farthest-first algorithm to convert user-specified minimum and maximum number of global and local clusters to the associated penalty parameters. The farthest first is a rough heuristic, so we are conservative in the specified range of $1 \geq \lambda_K \leq 30$, $1 \geq \lambda_L \leq 20$). We perform estimation using the sampling-weighted hierarchical clustering algorithm over a 5×15 grid of global and local clustering penalty parameters, respectively, on the full sample of $n_j = 2250$ for the $J = 3$ data sets. Figure 2 suggests selection of penalty parameters from the upper left-quadrant of the grid. We chose the values of $(\lambda_L = 1232, \lambda_K = 2254)$ that maximized the C index for our sampling-weighted hierarchical clustering model. Figure 3 presents the resulting estimated distribution of the n_j (informative sample) observations within clusters of each local partition (that index the $J = 3$ datasets), where the support of each local distribution indicates to which global cluster center each local cluster is linked. We note that the correct number ($K = 7$) of global clusters and local clusters ($L_j = 5$) is returned and that the skewed distribution of establishments within each local partition mimics that of the population (which is estimated from our informative sample). We use the Rand statistic $\in (0, 1)$, which measures the concordance of pairwise clustering assignments, to compare the true partition assigned in the the population versus the estimated partition, where the latter is estimated from the observed informative sample, rather than the population, itself. The computed Rand value is 1, indicating perfect agreement between the true and estimated partitions in their pairwise clustering assignments. (See Shotwell (2013) for a discussion of alternative statistics, including the Rand, for comparing the similarity of two partitions).

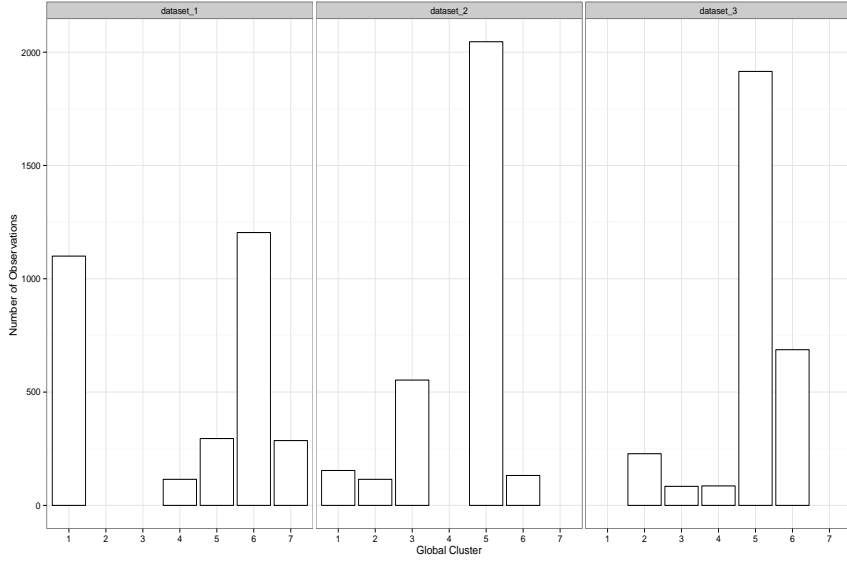


Figure 3: Estimated distribution of establishments in $J = 3$ local cluster partitions. The support of each distribution indicates the global cluster to which each local cluster is linked.

We additionally compute a sampling-weighted version of the silhouette index, $h_i = \frac{a_i - b_i}{\max(a_i, b_i)}$, where a_i is the average distance to observations co-clustered with i and b_i is the average distance to other observations in the nearest cluster to that holding i . The total index, $h^K = \frac{\sum_{i=1}^n \tilde{w}_i h_i}{\sum_{i=1}^n \tilde{w}_i} \in (0, 1)$ and, like the C index, prefers partitions where the clusters are compact and well-separated. These two indices select the same penalty parameter values from the grid of values evaluated on both the simulation and CES data. We prefer our sampling-weighted C since the computation is more scalable to a large number of observations. The C index was used to select the penalty parameters for all implementations of the mixtures of DPs and HDPs models.

Lastly, our procedure for selecting (λ_L, λ_K) by evaluating the Calinski Harabasz statistic over a grid generally selects the same clustering under inclusion or exclusion of the merge move on synthetic datasets. Figure 4 examines the numbers of merges that take place on

each run on the grid of penalty parameter values in the case we include the merge step, and reveals that the number of merges increases at relatively low values of the penalty parameters. In this case, and in others we tested on synthetic data, the merges took place outside of the range of nearly optimum penalty parameter values, though one may imagine the possibility of merges taking place in this range on a real dataset. There would be a reduction in the sensitivity for the number of clusters estimated to the values of the penalty parameter, which may have the effect of increasing the sub-space of nearly optimal penalty parameters (that produce the same clustering results). While we recommend and employ our selection procedure for penalty parameters in our simulations and application to follow, further simulations (not shown) demonstrate that inclusion of the merge step produces estimated clusters that are consistently closer to the truth as compared to excluding the merge step when the penalty parameters are specified by a heuristic procedure, like the farthest first algorithm employed in Kulis and Jordan (2011).

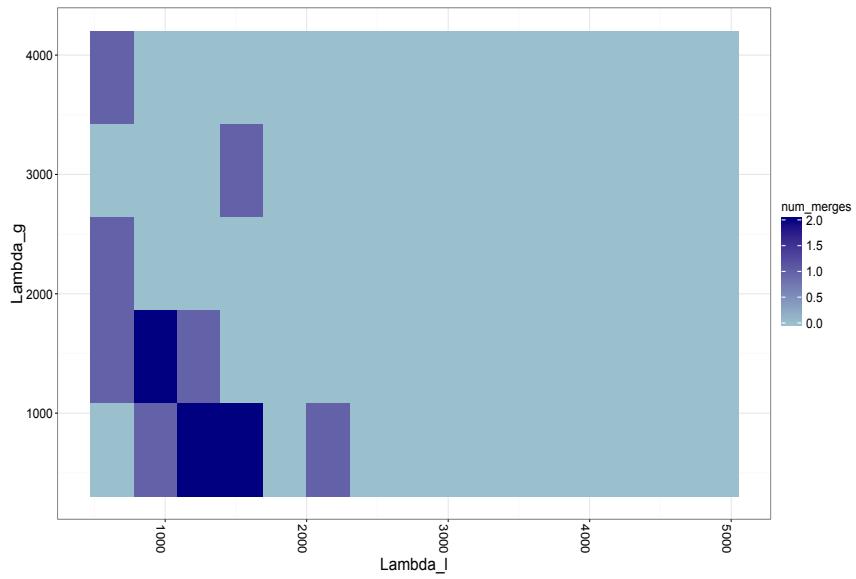


Figure 4: Heat map of the number of merges that took place under the mixtures of HDP estimation model over grid of global and local cluster penalty parameters. The vertical axis represents the global cluster penalty, λ_K , and the horizontal axis, the local cluster penalty, λ_L . The darker colors represent higher number of merges.

Our selection procedure that computes the Calinski Harabasz statistic over a range of global and local cluster penalty parameters defined on a grid and selects that clustering which produces the maximum value of this statistic will be used for every estimation run of the global and hierarchical clustering algorithms in the sequel.

3.2 Outlier Estimation Under Informative Sampling

Our next simulation study compares the outlier detection power when accounting for, versus ignoring, the informativeness in the observed sample by generating data under a simple, non-hierarchical global partition. We construct a set of $K = 5$ global cluster centers, each of dimensions, $d = 15$. Each cluster center expresses a distinct pattern over the $d = 15$ dimensions,

$$\begin{aligned}
 \mu_1^{(d=15) \times 1} &= (1, 1.5, 2.0, \dots, 7.5, 8) \\
 \mu_2 &= (8, 7.5, \dots, 1) \\
 \mu_3 &= (1, \dots, 7, 8, 7, \dots, 1) \\
 \mu_4 &= \text{Sampling from } (1, \dots, 8) \text{ under equal probability with replacement, } d = 15 \text{ times} \\
 \mu_5 &= \text{Sampling from } (-2, \dots, 6) \text{ under equal probability with replacement, } d = 15 \text{ times,}
 \end{aligned}$$

to generate $\mathbf{M} = (\mu_1, \dots, \mu_5)'$. Without loss of generality, the last cluster, μ_5 , is defined as the outlier cluster and includes negative values, though there is overlap with the support of the other four clusters for values > 0 . We create the $(N = 25000) \times 1$ cluster assignment vector, \mathbf{s} , by randomly assigning establishments in the (finite) population to clusters (with equal probabilities) such that the first 4 clusters are assigned equal numbers of observations, while the last (outlying) cluster (with mean μ_5 is assigned 150 observations, according with our earlier definition of an outlying cluster as having relatively few assigned observations).

We then generate our $N \times d$ matrix of population response values, \mathbf{X} , from the multivariate Gaussian distribution,

$$\mathbf{x}_i^{d \times 1} \stackrel{\text{ind}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}_{s_i}, \sigma^2 \mathbb{I}_d),$$

where the standard deviation, σ , is set equal to 1.5 times the average over the $N = 25000$ establishments and $d = 15$ dimensions of the assignment-weighted matrix, $\mathbf{M}_{\mathbf{s}, 1:d}$, the mean values of \mathbf{X} . We may think of this process for generating a finite population as taking a census of all establishments in the population. Establishments who report values in any of the first 4 clusters are drawn from the true population generating distribution, while those establishments who report values from cluster 5 commit errors such that the reported values are from a different (shifted) distribution that generates the population of errors. Our inferential interest is to uncover outlying values with respect to the population, rather than the sample, as an observation may not be outlying relative to the sample, but may relative to the population (or vice versa).

We assign the population establishments, evenly, to one of $H = 10$ strata, so there are $N_h = 2500$ establishments assigned to each stratum. Our sampling design employs simple random sampling of establishments within each of the H strata. The sample size taken from each stratum is set proportionally to the average of the by-establishment variances of the $(d = 15) \times 1$, $\{\mathbf{x}_i\}$ for establishments, $i = 1, \dots, N_h$, assigned to each stratum, $h = 1, \dots, H$. Each generated sample produces the by-stratum sample sizes, $\mathbf{n} = (45, 90, 136, 181, 227, 272, 318, 363, 409, 459)$, ordered according to variance quantile, from left-to-right, for a total sample size of $n = \sum_{h=1}^H n_h = 2500$. This sampling design assigns higher probabilities to larger variance strata (and all establishments in each stratum have an equal probability of selection), which is often done, in practice, because there is expected to be more information in these strata.

The population of establishments are assigned to clusters based on the mean values of \mathbf{x}_i , not the associated variances (as the variances in each cluster are roughly equal). We conversely use the variance of each \mathbf{x}_i , rather than the mean, in order to construct our sampling design with the goal to produce sampling inclusion probabilities that are nearly independent from the probabilities of assignment to the population clusters. So our model formulation doesn't (inadvertently) parameterize the sampling design, which is the most general set-up.

We generate a population and subsequently draw $B = 100$ samples under our informative single stage, stratified sampling design with inclusion probabilities of each stratum proportional to the average of the variances of member establishment response values (across the $d = 15$ dimensions), as described above. We run our sampling-weighted global clustering algorithm on each sample under two alternative configurations: 1. excluding the sampling weights, so that we do not correct for the informative sampling design; 2. including the sampling weights, such that we estimate outliers and cluster centers with respect to the clustering parameters for the population, asymptotically, conditioned on the generation of the finite population from disjoint clusters. We include two methods as comparators; firstly, we utilize the model-based clustering (MBC) algorithm of Fraley and Raftery (2002) that defines a finite mixture of Gaussians model of similar form as our (penalized) mixtures of DPs construction. Fraley and Raftery (2002) employ the EM algorithm to solve their mixture model (initialized by a hierarchical agglomeration algorithm) and select the number of clusters, K , using the Bayesian information criterion (BIC); secondly, we include the trimmed K-means method of Fritz et al. (2012), which intends to robustify the K-means algorithm by removing outlying data points that may induce mis-estimation of the partition. The authors note that these outlying points may be used to nominate outliers. Both implementations exclude employment of sampling weights to correct for informative

sampling. We also considered to include the algorithm of Shotwell and Slate (2011), but it did not computationally scale to the size of data we contemplate for our CES application.

The left-hand set of box plots in Figure 5 displays the distributions (within 95% confidence intervals under repeated sampling) of the true positive rate for identifying outlying observations, constructed as the number of true outliers discovered divided by the total number of true outliers, estimated on each Monte Carlo iteration for our three comparator models. The right-hand set of box plots display the false positive rate, which we define as the number of false discoveries divided by the total number of observations nominated as outliers. The inclusion of false positives permits assessment of the efficiency to detect outliers. Each set of box plots compares estimation under the global clustering algorithm *including* sampling weights, on the one hand, to a version of the global clustering algorithm, the MBC, and trimmed K-means, on the other hand, that all *exclude* the sampling weights. Outliers were detected in each simulation iteration, $b = 1, \dots, B$, based on selecting those clusters whose total observations (among the selected clusters) cumulatively summed to less than $C = 1.1$ times the total number of true outliers in the informative sample, which is how we would select outlying clusters on a real dataset where we don't know the truth. (We experimented with different values of $C \in [1, 1.5]$ and realized the same comparative results, as presented below).

Figure 5 reveals that failure to account for the informative sampling design induces a deterioration in outlier detection accuracy.

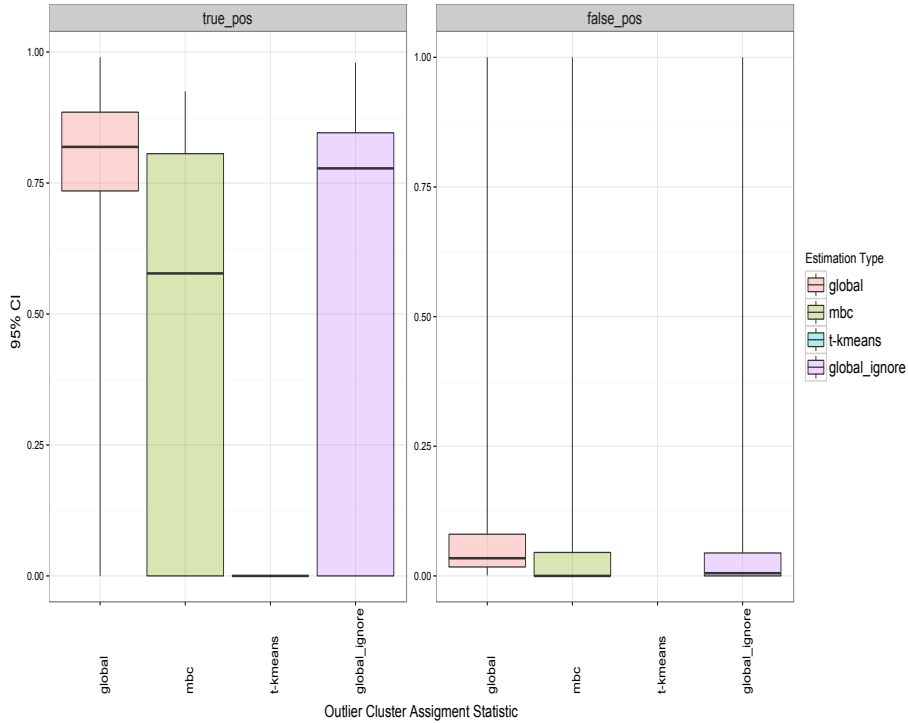


Figure 5: Accuracy of outlier detection under informative sampling: The left-hand plot panel presents the distributions of the true positive rates, and the right-hand panel presents the distribution for the false positive rates, both within 95% confidence intervals estimated from $B = 100$ Monte Carlo draws of informative samples. Each sample is of size, $n = 2500$, from a population of size, $N = 25000$, with a population cluster of outliers of size, $N_5 = 250$. The left-hand box plot within each plot panel is estimated from the sampling-weighted global clustering algorithm that accounts for the informative sampling design by including sampling weights, while the middle two box plots represent the model-based clustering (MBC) and trimmed K-means algorithms, respectively, that ignore the informative sampling design as does the right-hand box plot that represents the global clustering algorithm without inclusion of sampling weights.

Figure 6 presents the distribution over the number of discovered clusters for each of the three comparator models: 1. global clustering model, *including* sampling weights; 2. MBC, *excluding* sampling weights; 3. Trimmed K-means, *excluding* sampling weights; 4. global clustering model, *excluding* sampling weights. The dashed line at $K = 5$ clusters is the correct generating value. While the the models excluding the sampling weights (except for the trimmed K-means) estimate a higher number of clusters, such is not the primary reason for their reduced outlier detection accuracy, as we observe from Figure 5 that the false positive rates are slightly lower for these two models compared to the model that includes the sampling weights. The reduced accuracy is primarily driven by biased estimation of the $d \times 1$ cluster centers, $\{\boldsymbol{\mu}_p\}_{p=1,\dots,K}$, (relative to the population), whose estimation is performed together with assignment to clusters in a single iteration of the algorithm. We examine this bias in the next simulation study.

The trimmed K-means does relatively well in capturing the number of true clusters, absent the outlying cluster. The trimming is not isolating these points as outliers, however, but collapsing them into a larger cluster; hence, the trimmed k-means does not nominate any outliers.

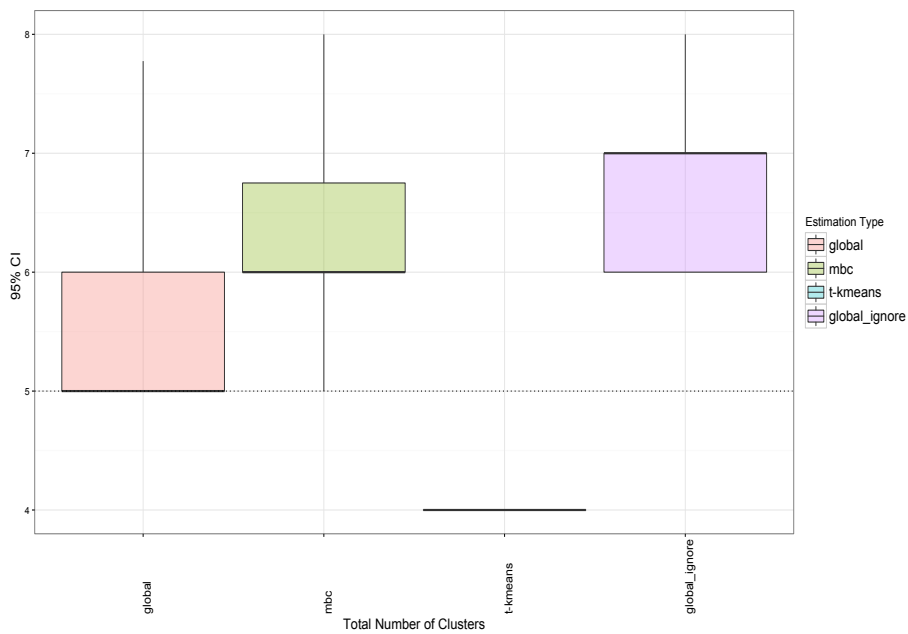


Figure 6: Comparison of distributions for number of estimated clusters, K , between the global clustering model including the sampling weights in the left-hand box plot, to the MBC in the middle box plot and the global clustering model in the right-hand box plot, where both exclude the sampling weights.

3.3 Comparison of Outlier Estimation between Hierarchical and Global Clustering under Informative Sampling

We now generate a set of local partitions hierarchically linked to a collection of global cluster centers. We generate $J = 8$ local populations, $(\mathbf{X}^j)_{j=1,\dots,J}$, each of size $N_j = 25000$, and associated local partitions, $(\mathbf{s}^j)_{j=1,\dots,J}$, where each local partition contains, $L_j = 2$ clusters, including one that is composed of 150 outliers. The set of J local partitions randomly select their local cluster centers from the same $K = 5$ global cluster centers that we earlier introduced, which induces a dependence structure among the set of local partitions. We conduct $B = 100$ Monte Carlo draws, where we take an informative sample within each of the $J = 8$ datasets on each draw, using the same stratified sampling informative design, described above. Estimation is conducted on the set of J informative samples produced in each draw using both the sampling-weighted global clustering algorithm and the hierarchical clustering algorithm outlined in Section 2.4. Our goal is to uncover the global cluster center for the outlier cluster and the set of observations in each local dataset that are assigned to the outlier cluster. We concatenate the set of informative samples generated over the J datasets when conducting estimation using the global clustering algorithm because our primary interest is in outlier detection, rather than inference on the local partitions. (The performance of the global clustering algorithm relative to the hierarchical clustering algorithm is worse than shown in the case we separately estimate a global clustering on each dataset).

Figure 7 is of the same format as Figure 5, with two panels displaying distributions over true positive and false positive rates, respectively, for our choice of models. The set of five box plots compare the following models: 1. hierarchical clustering model, including the sampling weights to both estimate global partitions linked to a global partition and control for informative sampling; 2. global clustering model, including sampling weights to

control for informative sampling; 3. model-based clustering, excluding sampling weights; 4. trimmed K-means robust clustering, excluding sampling weights; 5. global clustering model, excluding sampling weights. Results for the hierarchical clustering model, shown in the left-hand box plot of each panel, outperforms the global clustering model, where both include sampling weights, because the hierarchical model additionally estimates the dependent set of local clusters. The hierarchical clustering algorithm appears to do a better job of borrowing estimation information among the $J = 8$ local partitions.

Figure 8 presents distributions for each dimension, $d = 1, \dots, 15$, of the outlier cluster center, μ_5 , under four of the five comparison models (excluding the trimmed K-means) in the same order as displayed in Figure 7. The sampling-weighted hierarchical clustering model produces perfectly *unbiased* estimates for the population values of the global outlier cluster center, while the other 3 models (excluding the trimmed K-means) induce bias in proportion to their outlier detection performances. Unbiased estimation of the outlying cluster center as compared to the other clusters is important to detect an outlying cluster because the cluster centers encode the degree of separation of the outlying cluster(s) from the others. We see that the true positive detection rates across the methods shown in Figure 7 are roughly proportional to the levels of bias estimated in the dimensions of the outlying cluster center displayed in Figure 8. Since the trimmed K-means does not nominate any outliers, we cannot compute a outlying cluster center under this method.

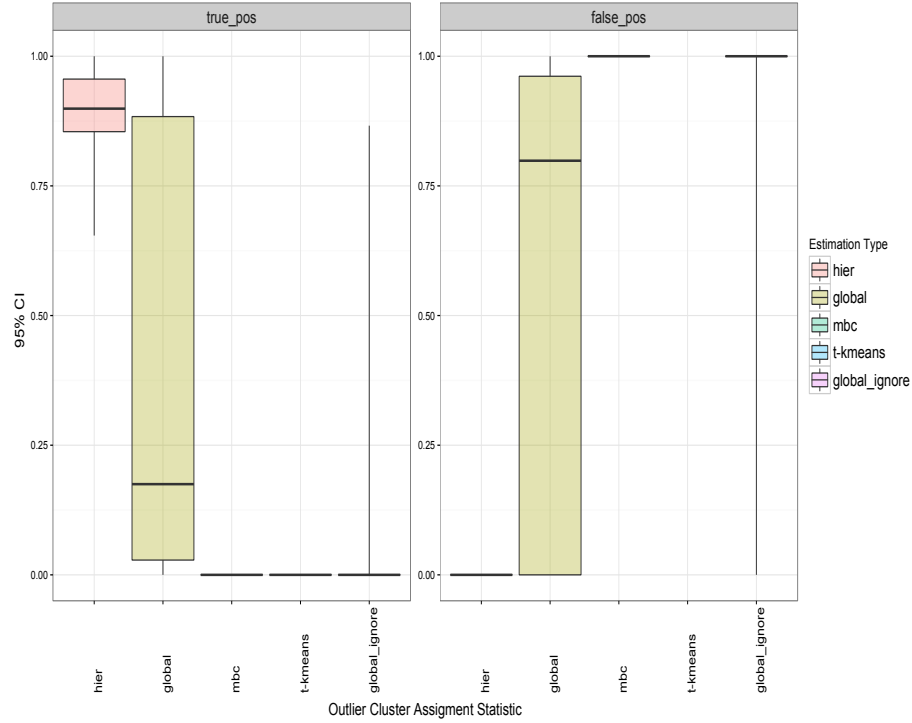


Figure 7: Accuracy of outlier detection under informative sampling for global vs hierarchical clustering model: The left-hand plot panel presents the distributions of the true positive rate, and the right-hand panel presents the distributions for the false positive rate, both within 95% confidence intervals estimated from $B = 100$ Monte Carlo draws of informative samples from each of $J = 8$ local populations. Each sample is of size, $n_j = 2500$, $j = 1, \dots, 8$, from a population of size, $N_j = 25000$, with a cluster of outliers of size, $N_{j,5} = 125$. The box plots represent the following models, from left-to-right: 1. hierarchical clustering model, including the sampling weights to both estimate global partitions linked to a global partition and control for informative sampling; 2. global clustering model, including sampling weights to control for informative sampling; 3. model-based clustering excluding sampling weights; 4. trimmed K-means, excluding sampling weights; 5. global clustering model excluding sampling weights.

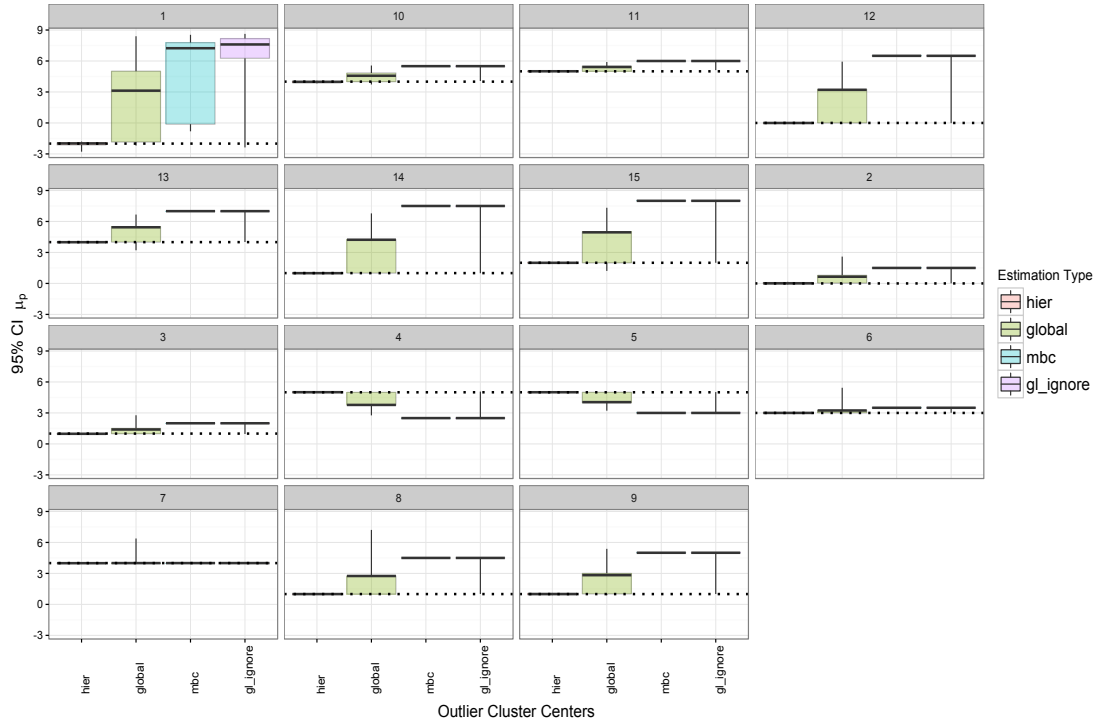


Figure 8: Comparison of estimation bias for each of $d = 15$ dimensions of the outlier global cluster mean, μ_5 , estimated from $J = 8$ local partitions under the following models, from left-to-right: 1. hierarchical clustering model, including the sampling weights to both estimate global partitions linked to a global partition and control for informative sampling; 2. global clustering model, including sampling weights to control for informative sampling; 3. model-based clustering, excluding sampling weights; 4. global clustering model, excluding sampling weights. The dashed line in each panel is the true value the center for each dimension, $d = 1, \dots, 15$

We also examined the effect of excluding the merge move for the hierarchical clustering algorithm in this simulation study. While excluding the merge move induced an overestimation of the the number of clusters (6 instead of the true 5), the outlier detection accuracy wasn't notably impacted, likely because employment of the C algorithm for selection of the number of local and global clustering penalty parameters protected against a large magnitude misestimation of the number of global and local clusters. We find a larger discrepancy between employment or not of the merge move for a *single* run of the clustering algorithm where the penalty parameters are set in an ad hoc fashion (e.g., through use of the farthest-first algorithm as in Kulis and Jordan (2011), a procedure that we do not recommend).

4. Application

We apply our hierarchical clustering algorithm to a data set of one month changes in CES survey responses for estimation of industry-indexed (local) partitions that may express a dependence structure across industries by potentially sharing global clusters. Outliers are nominated by selecting all establishment observations in any local cluster that holds a small percentage of observations (e.g. < 1%). Our CES survey application focuses on the set of 108017 establishments whose reported employment statistics differ between November and December, 2009 as a typical illustration. We are interested to flag those establishments whose responses express unusually large changes in employment statistics between November to December, relative to their November employment levels. So we normalize the observed statistics to,

$$\delta_{ijt} = \frac{x_{ijt}}{x_{ij(t-1)}}, \quad j = 1, \dots, d \quad (7)$$

which has support on the positive real line that we, in turn, cluster. The distribution of δ_{ijt} is highly right-skewed for all dimensions, $j \in (1, \dots, d)$, so we use a logarithm transform of the statistic to perform clustering in order to conform to the mixtures of Gaussians assumptions in Equation 5a. Alternatively, we could replace the squared Euclidean distance in Equation 6 with the Bregman divergence (which is uniquely specified for exponential family distributions) and replace the mixtures of Gaussian distributions with Exponential distributions, as suggested by Jiang et al. (2012) (which produces the same results on these data).

We focus on four employment variables of particular importance to CES (that also express high response rates); 1. “ae”, which is the total employment for each reporting establishment; 2. “pw”, which is production worker employment for each establishment; 3. “npr”, which is the total payroll dollars expended for employees of each establishment; 4. “nhr”, which is the average weekly hours worked for each establishment. So the dimension of our CES survey application is $d = 4$.

We select penalty parameters from the range, $(9 \leq \lambda_L \leq 829, 17 \leq \lambda_K \leq 187)$, on a 15×20 grid, using the Calinski Harabasz and silhouette statistics, which both choose $(\lambda_L = 19, \lambda_K = 145)$ from this grid. The selected penalty parameters did not change when including or excluding the merge step. Nevertheless, the number of clusters estimated without inclusion of the merge step was $K = 10$, while $K = 9$ was estimated when including the merge step. The resulting inference on the nomination of outliers, however, was unchanged so that we prefer the parsimonious model (in terms of number of clusters selected). The selected penalty parameters and resulting estimated clustering also did not change when we employed a finer grid. Results from inclusion of the merge step are presented in Table 1, where we discover $K = 9$ global clusters shared among the $J = 23$ local partitions, and each local partition holds between $L_j = 2 - 9$ clusters. Each column of Table 1 presents

the results for one of the $K = 9$ global clusters, from left-to-right in descending order of the number of establishments assigned to each global cluster. The first four rows (labeled “ae_ratio”, “pw_ratio”, “npr_ratio”, “nhr_ratio”) present the estimated global cluster centers, $\boldsymbol{\mu}_p$, for change ratio in ae, pw, npr, and nhr respectively, after reversing the logarithm transform. The fourth through final rows present the by-industry, local partitions, \mathbf{s}^j , and their links to the global cluster centers. Scanning each of the columns, we see that all of the $K = 9$ global clusters are shared among the local partitions, indicating a high degree of dependence among them.

The fifth row (labeled “ae_avg”) averages the total employment, ae, over all establishments in all industries linked to the associated global cluster as reported for the month of November. The second column from the right represents a cluster whose centers indicate unusually large increases in reported employment and payroll from November-to-December. The establishments linked to this cluster are of generally small-sized, with an average of 20 reported employees in November. Establishments with a relatively small number of employees will receive high sampling weights due to their low inclusion probabilities, such that their high magnitude shifts in reported employment levels may be influential in the estimation of the sampling-weighted total employment estimates for November and December. It might, however, be expected that the retail (and associated wholesale) hiring might dramatically increase in anticipation of holiday shopping. So we would nominate the remaining 470 establishments linked to this global cluster as outliers for further analyst investigation by the BLS. Previous analysis conducted within the BLS suggests that smaller establishments generally tend to commit submission errors at a higher rate. The result excluding the merge step splits this cluster into two in a manner that doesn’t change inference about outlier nominations.

The smallest-size (right-most) cluster has a mean of 0.01 in the monthly change ratio for the variable “npr”, indicating an unusually large magnitude decrease in the number of employees (and payroll dollars) reported from November-to-December. This cluster contains 113 relatively large-sized establishments with an average of 278 reported employees in November. The moderate-to-large size of establishments in this cluster will tend to receive smaller sampling weights, however, (because they have higher sample inclusion probabilities), and so are less influential. So BLS would generally place a lower priority for investigation on this cluster than the previous discussed. The small number of establishments in this cluster, coupled with the large magnitude decreases in reported employment variables, however, would likely prompt an investigation of the submissions for all establishments in this cluster. The two Retail industries show 17 establishments in this cluster of establishments expressing a large decrease in employment. The seasonal hiring in this industry, mentioned above, may suggest to place an especially high priority on investigation of these establishments.

5. Discussion

The BLS seeks to maintain a short lead time for the CES survey between receipt of monthly establishment responses and publication of estimates. Retaining estimation quality requires investigation of establishment responses for errors, which may prove influential in published estimates (by region and industry, for example). BLS analysts, therefore, require an automated, fast-computing tool that nominates as outliers for further investigation a small subset of the over 100000 establishments whose responses reflect month-over-month changes in employment levels.

We have extended the MAP approximation algorithm of Kulis and Jordan (2011) and Broderick et al. (2012) for estimation of partitions among a set of observations to appli-

Table 1: CES Mixtures of HDPs: Global cluster centers and local partitions

Variable Values, by cluster										
ae_ratio	0.961	0.98	1.157	0.829	0.657	2.008	0.156	19.212	0.093	
pw_ratio	0.951	0.973	1.199	0.797	0.596	2.586	0.189	11.95	0.146	
npr_ratio	0.939	0.963	1.283	0.423	0.465	4.106	0.143	16.617	0.010	
nhr_ratio	0.944	0.979	1.213	0.639	0.547	2.469	0.182	20.37	0.109	
Average Employment Count, by cluster										
ae_avg (November)	168	205	104	154	125	182	258	20	278	
Units in Super Sector Clusters										
Agriculture	0	81	0	0	16	0	0	0	0	
Mining	379	0	118	0	105	0	20	10	3	
Utilities	0	405	0	0	0	0	4	2	1	
Construction	2977	0	1111	0	1008	174	125	34	9	
Manufacturing(31)	1015	0	254	0	0	0	28	6	6	
Manufacturing(32)	1229	0	272	0	131	30	19	5	3	
Manufacturing(33)	0	2117	494	169	0	46	34	16	3	
Wholesale	1828	0	486	0	161	47	26	23	3	
Retail(44)	0	14648	4484	915	0	160	168	33	7	
Retail(45)	0	9381	5243	996	0	132	58	17	10	
Transportation(48)	1266	0	346	155	0	41	26	16	4	
Transportation(49)	0	1149	703	101	0	64	39	11	14	
Information	1852	0	456	198	0	30	22	30	1	
Finance	4947	0	1894	476	0	138	51	111	6	
Real Estate	1409	0	981	0	159	0	18	15	0	
Professional Services	2999	0	1151	349	0	149	70	47	5	
Management of Companies	0	881	185	64	0	0	6	6	1	
Waste Mgmt	3674	0	1301	0	602	144	103	40	11	
Education	666	0	229	0	0	23	19	6	0	
Health Care	0	6437	1068	368	0	126	63	23	14	
Arts-Entertainment	953	0	308	0	213	0	42	25	2	
Accommodation	11757	0	2541	822	0	181	100	49	7	
Other Services	1499	0	433	0	158	51	11	18	3	
	38450	35099	24058	4613	2553	1536	1052	543	113	108017

cations where the observations were acquired under an informative sampling design. We replace the usual likelihood in estimation of the joint posterior over partitions and associated cluster centers with a pseudo-likelihood that incorporates the first order sampling weights that serve to undo the sampling design. The resulting estimated parameters of the approximate MAP objective function are asymptotically unbiased with respect to the population, conditioned on the generation of the finite population from disjoint clusters. Our simulation study demonstrated that failure to correct for informative sampling reduces the outlier detection accuracy by inducing biased estimation of the cluster centers.

Our use of sampling-weighted partition estimation algorithms focuses on outlier detection, so we incorporated a new merge step, which may increase robustness against estimation of local optima and encourage discovery of clusters containing large numbers of observations, which may be a feature for outlier detection where we nominate as outliers those observations in clusters containing relatively few observations. We additionally constructed the sampling-weighted C statistic that our simulation study demonstrated was very effective for selection of the local and global cluster penalty parameters, (λ_L, λ_K) , that together determine the numbers of local and global clusters.

The sampling-weighted hierarchical clustering algorithm permitted our estimation of industry-indexed local clusters, which fits well into the BLS view that industry groupings tend to collect establishments with similar employment patterns and reporting processes. We saw that all of the estimated $K = 9$ global clusters for percentage change in employment from November to December, 2009 were shared among the local by-industry clusters, which served to both sharpen estimation of the local partitions and global cluster centers and to discover small global clusters of potentially outlying observations.

Acknowledgments

The authors wish to thank Julie Gershunskaya, a Mathematical Statistician colleague at the Bureau of Labor Statistics, for her clever insights and thoughtful comments that improved the preparation of this paper.

References

- D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.
- Daniel Bonnry, F. Jay Breidt, and Francois Coquet. Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli*, 18(4):1361–1385, 11 2012. doi: 10.3150/11-BEJ369. URL <http://dx.doi.org/10.3150/11-BEJ369>.
- T. Broderick, B. Kulis, and M. I. Jordan. MAD-Bayes: MAP-based Asymptotic Derivations from Bayes. *ArXiv e-prints*, December 2012.
- Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.
- Heinrich Fritz, Luis A. García-Escudero, and Agustín Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012. URL <http://www.jstatsoft.org/v47/i12/>.
- Hemant Ishwaran and Lancelot F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1235, 2003.
- Ke Jiang, Brian Kulis, and Michael I. Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In F. Pereira, C.J.C. Burges, L. Bottou, and

- K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3158–3166. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4853-small-variance-asymptotics-for-exponential-family-dirichlet-process-mixture-models.pdf>.
- Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *CoRR*, abs/1111.0352, 2011. URL <http://arxiv.org/abs/1111.0352>.
- R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- Fernando A. Quintana and Pilar L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 65(2):557–574, 2003.
- R** Core Team. *R: A Language and Environment for Statistical Computing*. **R** Foundation for Statistical Computing, Vienna, Austria, 2014. URL <http://www.R-project.org/>.
- Terrance Savitsky and Daniell Toth. Convergence of pseudo posterior distributions under informative sampling. 2015. URL <http://arxiv.org/abs/1507.07050>.
- Matthew S. Shotwell. Profdpm: An R package for MAP estimation in a class of conjugate product partition models. *Journal of Statistical Software*, 53(8):1–18, 2013.
- Matthew S. Shotwell and Elizabeth H. Slate. Bayesian outlier detection with Dirichlet process mixtures. *Bayesian Analysis*, 6(4):665–690, 2011.
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Lianming Wang and D. B. Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20:196–216, 2011.

Appendix A. Hierarchical Clustering Algorithm

Loop over algorithm blocks, A.2 and A.3 until convergence.

Algorithm A.1: INITIALIZE LOCAL AND GLOBAL CLUSTER OBJECTS

Input: Set of $n_j \times d$ data matrices, $\{\mathbf{X}^j\}_{j=1,\dots,J}$, where each holds $1 \times d$ observations for n_j units in local dataset, j . Set of $n_j \times 1$ vectors, $\{\tilde{\mathbf{w}}^j\}$ of sampling weights for n_j units in local dataset, j . λ_L , local cluster penalty parameter. λ_K , global cluster penalty parameter.

Output: $K \times d$, $\mathbf{M} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)'$, a matrix of $1 \times d$ global cluster centers for $p = 1, \dots, K$ global clusters. A set of $n_j \times 1$ vectors, $\{\mathbf{s}^j\}_{j=1,\dots,J}$, where $s_i^j \in (1, 2, \dots, K)$, linking units in each dataset, \mathbf{X}^j , to a global cluster, $p \in (1, \dots, K)$. The number of unique values in \mathbf{s}^j , denoted by L_j , specifies the local partition structure for dataset, \mathbf{X}^j .

1 **Initialize:** Initialize number of global clusters, $K = 1$.

2 **for** $j \leftarrow 1$ **to** J **do**

3 $L_j \times 1$, \mathbf{v}^j records link of each local cluster, $c \in (1, \dots, L_j)$, to global cluster, $p \in (1, \dots, K)$. Initialize $L_j = 1$ and $v_{L_j}^j = 1$.

4 $n_j \times 1$, \mathbf{z}^j , records link of each unit, i in dataset, j , to a *local* cluster, $c \in (1, \dots, L_j)$. Initialize $z_i^j = 1, \forall i$.

5 $n_j \times 1$, \mathbf{s}^j records link of each unit, i , in dataset, j to a *global* cluster, $p \in (1, \dots, K)$. Initialize $s_i^j \leftarrow v_{z_i^j}^j, \forall i$.

6 Set $d \times 1$ cluster center, $\boldsymbol{\mu}_1 \leftarrow \left(\sum_{j=1}^J \sum_{i=1}^{n_j} \tilde{w}_i^j \mathbf{x}_i^j \right) / \left(\sum_{j=1}^J \sum_{i=1}^{n_j} \tilde{w}_i^j \right)$.

7 Compute energy,

$$e \leftarrow \sum_{p=1}^K \sum_{j=1}^J \sum_{i \in \{i: s_i^j = p\}} \tilde{w}_i^j \left\| \mathbf{X}_i^j - \boldsymbol{\mu}_p \right\|^2 + \lambda_K K + \lambda_L L, \text{ where } L = \sum_{j=1}^J L_j.$$

Algorithm A.2: BUILD LOCAL AND GLOBAL CLUSTERS

1 **Assignment of units to global clusters.**

2 **for** $j \leftarrow 1$ **to** J **and** $i \leftarrow 1$ **to** n_j **do**

3 Compute distance metric, $d_{ijp} = \tilde{w}_i^j \left\| \mathbf{X}_{i,1:d}^j - \mu_p \right\|^2$ for $p = 1, \dots, K$.

4 For those global clusters, p , not linked to any local cluster in j ,
 $\{p : v_c^j \neq p, \forall c \in (1, \dots, L_j)\}$,

5 $d_{ijp} \leftarrow d_{ijp} + \lambda_L$ (since must add local cluster if assign i to p).

6 **if** $\min_p d_{ijp} > \lambda_L + \lambda_K$ **then**

7 Create new local cluster linked to new global cluster.

8 $L_j \leftarrow L_j + 1$, $z_i^j = L_j$

9 $K \leftarrow K + 1$, $\mu_K \leftarrow \mathbf{X}_i^j$, $v_{L_j}^j \leftarrow K$

10 **else**

11 Let $\hat{p} \leftarrow \operatorname{argmin}_p d_{ijp}$.

12 **if** $v_c^j = \hat{p}$ for some $c \in (1, \dots, L_j)$ **then**

13 Assign unit i in j to local cluster c , $z_i^j \leftarrow c$ and $v_c^j = \hat{p}$.

14 **else**

15 Create new local cluster for unit i and link to global cluster, \hat{p} .

16 $L_j \leftarrow L_j + 1$ and $z_i^j = L_j$ and $v_{L_j}^j = \hat{p}$.

17 **Re-assignment of (all units in) local clusters to global clusters.**

18 **for** local partitions, $j \leftarrow 1$ **to** J and local cluster, $c \leftarrow 1$ **to** L_j **do**

19 Let $\mathbf{z}_c^j = \{i : \mathbf{z}_i^j = c\}$ and $1 \times d$, $\mu_{jc} = \left(\sum_{i \in \mathbf{z}_c^j} \tilde{w}_i^j \mathbf{x}_i^j \right) / \left(\sum_{i \in \mathbf{z}_c^j} \tilde{w}_i^j \right)$.

20 Compute sum of local-to-global cluster distances:

21 $d_{jcp} = \sum_{i \in \mathbf{z}_c^j} \tilde{w}_i^j \left\| \mathbf{X}_i^j - \mu_p \right\|^2$, for $p = 1, \dots, K$.

22 **if** $\min_p d_{jcp} > \lambda_K + \sum_{i \in \mathbf{z}_c^j} \tilde{w}_i^j \left\| \mathbf{X}_i^j - \mu_{jc} \right\|^2$ **then**

23 Set $K \leftarrow K + 1$, $v_c^j = K$, and $\mu_K = \mu_{jc}$

24 **else** $v_c^j \leftarrow \operatorname{argmin}_p d_{jcp}$

25 **Shed global clusters no longer assigned to any local clusters.**

26 **for** $p \leftarrow 1$ **to** K global clusters **do**

27 **if** $\mathbf{v}^j \neq p$, $\forall j \in (1, \dots, J)$ **then**

28 Recode global cluster labels in \mathbf{v}^j such that $p' \leftarrow p' - 1$, $\forall p' > p$

29 Set $K \leftarrow K - 1$

30 Delete cluster center for p , $\mathbf{M} \leftarrow \mathbf{M}_{-p,1:d}$.

31 **for** local partitions, $j \leftarrow 1$ **to** J and units, $i \leftarrow 1$ **to** n_j **do**

32 $s_i^j \leftarrow v_{z_i^j}^j$.

33 **Re-compute global cluster centers**

34 **for** $p \leftarrow 1$ **to** K global clusters **do**

35 **for** $j \leftarrow 1$ **to** J datasets **do**

36 Compute units in j assigned to global cluster p , $\mathbf{S}_{jp} = \{i : s_i^j = p\}$;

37 Compute $\mu_p = \left(\sum_{j=1}^J \sum_{i \in \mathbf{S}_{jp}} \mathbf{x}_i^j \tilde{w}_i^j \right) / \sum_{j=1}^J \sum_{i \in \mathbf{S}_{jp}} \tilde{w}_i^j$.

Algorithm A.3: MERGE GLOBAL CLUSTERS

```

1 Compute energy of current state,  $e \leftarrow \sum_{p=1}^K \sum_{j=1}^J \sum_{i \in \{i: s_i^j = p\}} \tilde{w}_i^j \left\| \mathbf{x}_i^j - \boldsymbol{\mu}_p \right\|^2 + \lambda_K K + \lambda_L L$ 
2 for  $p \leftarrow 2$  to  $K$  and  $p' \leftarrow 1$  to  $(p-1)$  do
3   Perform test merge for each pair of global clusters.
4   Set matrix of cluster centers for virtual step,  $\mathbf{M}^* \leftarrow \mathbf{M}$ .
5   for local partitions,  $j \leftarrow 1$  to  $J$  do
6     Let  $\mathbf{S}_{jp'}^* = \{i : s_i^j = p'\}$ 
7     if  $|\mathbf{S}_{jp'}^*| > 0$  then
8       There are some units in  $\mathbf{s}^j$  assigned to global cluster,  $p'$ , Re-assign units linked
9       to  $p'$  to  $p$ .  $\mathbf{S}_{jp}^* = \{i : s_i^j = p \text{ or } s_i^j = p'\}$ 
9   Compute merged cluster centers,  $\boldsymbol{\mu}_p^* = \left( \sum_{j=1}^J \sum_{i \in \mathbf{S}_{jp}^*} \mathbf{x}_i^j \tilde{w}_i^j \right) / \sum_{j=1}^J \sum_{i \in \mathbf{S}_{jp}^*} \tilde{w}_i^j$ 
10  Set  $\mathbf{M}_{p',1:d}^* \leftarrow \mathbf{M}_{p,1:d}^*$ 
11  Compute number of local clusters shed if merge global cluster,  $(p', p)$ .
12  for local partitions,  $j \leftarrow 1$  to  $J$  do
13    Local clusters linked to  $p', p$ :  $\mathbf{c}_{jp'} = \{c : \mathbf{v}^j = p'\}$  and  $\mathbf{c}_{jp} = \{c : \mathbf{v}^j = p\}$ ,
14     $L_{jp'} = |\mathbf{c}_{jp'}|$  and  $L_{jp} = |\mathbf{c}_{jp}|$ 
14  Reduced number of local clusters,  $L^* = L - \sum_{j=1}^J L_{jp}$ , and global clusters,  $K^* \leftarrow K - 1$ .
15  Compute energy under the test merge,
16   $e^* \leftarrow \sum_{p=1}^K \sum_{j=1}^J \sum_{i \in \mathbf{S}_{jp}^*} \tilde{w}_i^j \left\| \mathbf{x}_i^j - \boldsymbol{\mu}_p^* \right\|^2 + \lambda_K K^* + \lambda_L L^*$ 
16  if  $e^* < e$  then
17    Execute merge of global clusters,  $p$  and  $p'$ 
18    for local partitions,  $j \leftarrow 1$  to  $J$  do
19      if  $L_{jp'} > 0$  then
20        Cluster  $p'$  is linked to local partition,  $j$ .
21        if  $L_{jp} = 0$  then  $\mathbf{v}_{c_{jp'}}^j \leftarrow p$ 
22        else
23          Reassign local clusters linked to  $p'$  to  $p$ .
24           $\mathbf{z}_{c_{jp'}}^j = \{i : \mathbf{z}_i^j \in \mathbf{c}_{jp'}\}$  and  $\mathbf{z}_{c_{jp'}}^j \leftarrow c_{jp1}$ .
25          Remove now empty local clusters linked to  $p'$ ,  $\mathbf{v}^j \leftarrow \mathbf{v}_{-c_{jp'}}^j$ .
26          Recode  $\mathbf{z}^j$  local cluster assignments so labels are contiguous by setting
27           $c \leftarrow c - 1, \forall c > c_{jp'}$ .
28          Recode global cluster labels in  $\mathbf{v}^j$  such that  $p \leftarrow p - 1, \forall p > p'$ 
27    Remove global cluster center for  $p'$ .
28     $\mathbf{M}^* \leftarrow \mathbf{M}_{-p',1:d}^*$ 
29    for units,  $i \leftarrow 1$  to  $n_j$  do
30       $s_i^j \leftarrow v_{z_i^j}^j$ .
31     $\mathbf{M} \leftarrow \mathbf{M}^*$ 

```
