Bayesian Estimation Under Informative Sampling

Based on work of T. D. Savitsky, Daniell Toth, Michail Sverchkov

T. D. Savitsky

Office of Survey Methods Research

October 22, 2014

Table of contents

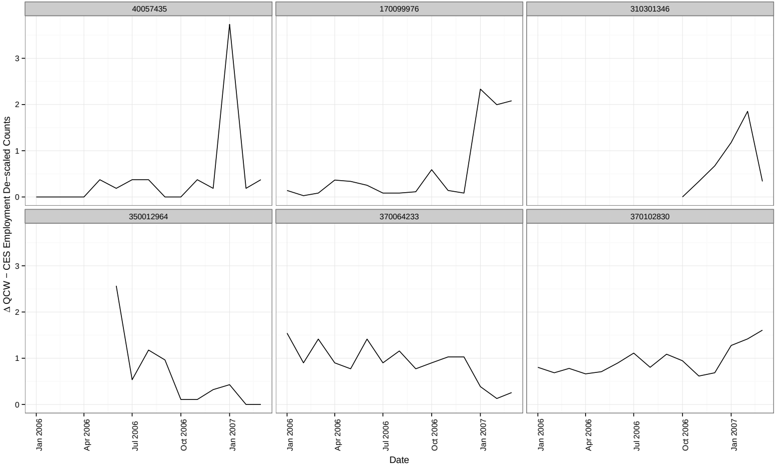
Employment Count Reporting Errors

- Make inference on the population distribution
- When the data are acquired under an informative sampling design
- Where $f\left(\mathbf{y}_o|\boldsymbol{\lambda}\right) \neq f\left(\mathbf{y}_p|\boldsymbol{\lambda}\right)$
- Analyst may not know the sampling design to parameterize it

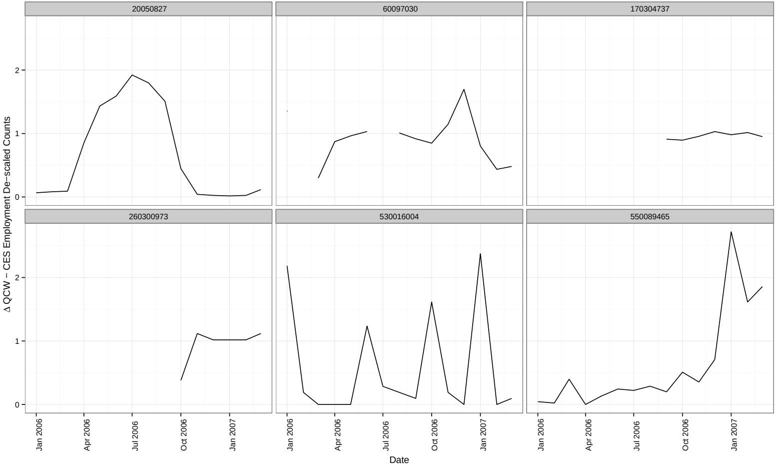
Employment Count Reporting Errors

- Establishments report number of employees
- On both Quarterly Census of Employment and Wages (QCEW)
- And Current Employment Survey (CES)
- Response Analysis Survey (RAS)
- Our data are $y = |y_{\text{QCEW}} - y_{\text{CES}}|$
- Collected on $N = 9777$ establishments, each for $T = 15$ months.
  - June, $2006-$ March, 2007
- Produces N, $T \times 1$ functions, $\{\mathbf{y}_i\}_{i=1,\ldots,N}$
- Functions each standardized to variance 1
- Desired inference about set of latent functional bases

Randomly-selected observed $\mathbf{y}_i$



More Randomly-selected observed $\mathbf{y}_i$



Stratified Sampling Design

- Sampled 2917 establishments - 2013 responded from $N = 9777$
- Stratified design with $H = 6$ strata
- Assigned to strata based on *a priori* “interesting” features
- Focused on selected portions of functions
- e.g. spike at year-end indicates counting checks, rather than employees

Bayesian Estimation Model: DP mixture over GP's

$$
\begin{aligned}
\overset{T\times 1}{\mathbf{y}_i} &\sim \mathcal{N}_T\left(\mathbf{f}_i, \tau_\epsilon^{-1}\mathbb{I}_T\right),\ i = 1,\ldots,N \\
\mathbf{f}_i &\sim \mathcal{N}_T\left(\mathbf{0}, \overset{T\times T}{\mathbf{C}}(\boldsymbol{\theta}_i)\right) \\
\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N|G &\sim G \\
G &\sim \text{DP}(\alpha_0, G_0) \\
&\Updownarrow \\
\mathbf{f}|G &\sim \int \mathcal{N}_T\left(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) G(d\boldsymbol{\theta})
\end{aligned}
$$

GP covariance function - Rational Quadratic

$$\mathbf{f}_i \sim \mathcal{N}_T\left(\mathbf{0}, \overset{T\times T}{\mathbf{C}}(\boldsymbol{\theta}_i)\right)$$

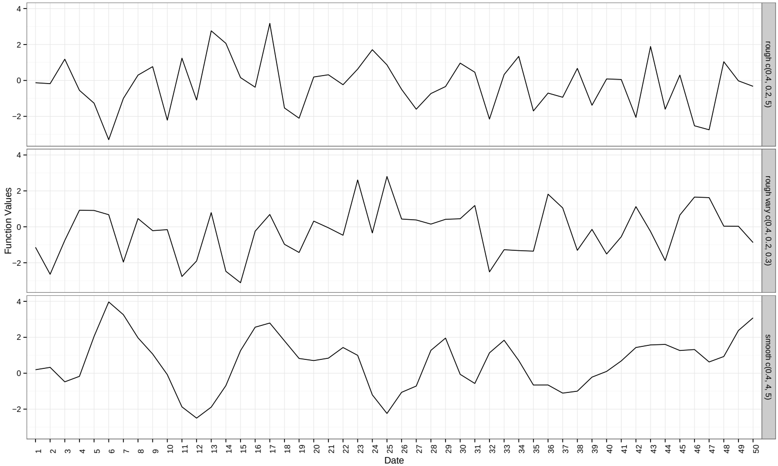
$$\mathbf{C}(\boldsymbol{\theta}_i) \equiv \mathbf{C}_i = \left(C_{f_{ij},f_{i\ell}}\right)_{j,\ell\in(1,\ldots,T)}$$

$$C_{f_{ij},f_{i\ell}} = \frac{1}{\theta_{i,1}}\left(1 + \frac{(t_{ij} - t_{il})^2}{\theta_{i,2}\theta_{i,3}}\right)^{-\theta_{i,3}}$$

- $\theta_{i,1}$ controls the vertical magnitude
- $\theta_{i,2}$ controls the base length-scale $\leftrightarrow$ wavelength
- $\theta_{i,3}$ controls smooth deviations across scales
  - As $\theta_{i,3} \uparrow \infty$, converge to single length-scale, $\theta_{i,2}$

GP Draws

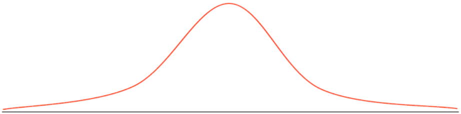
- Top - “Rough”: $(\theta_1, \theta_2, \theta_3) = (0.4, 0.2, 5.0)$
- Middle - “Rough with Variation”: $(\theta_1, \theta_2, \theta_3) = (0.4, 0.2, 0.3)$
- Bottom - “Smooth”: $(\theta_1, \theta_2, \theta_3) = (0.4, 4.0, 5.0)$



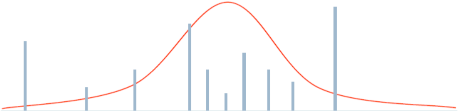
Draws from $G \sim$ DP are discrete, a.s.

Dirichlet Process

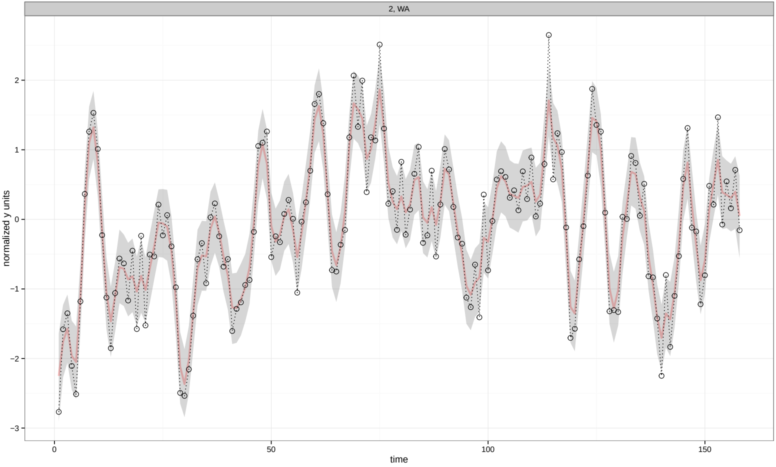
- Consider Gaussian G_0



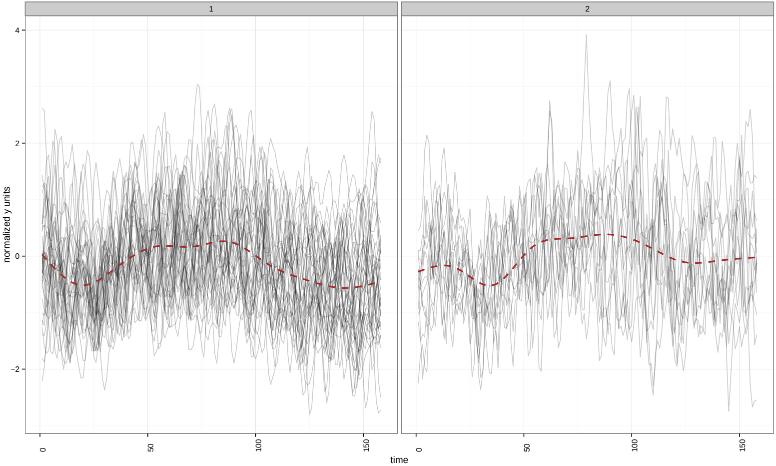
- $G \sim DP(\alpha, G_0)$



f_i: Monthly CPS data



f_*: Monthly CPS data



Estimating Population Density from Sampled Units

- Consider sequence of populations, $\{N_\gamma\}_{\gamma \in \mathbb{N}}$.
- Increasing in size as $\gamma \uparrow$.
- $m_\gamma\left(y\right) := \mathbb{E}\left(I_{\gamma i} | Y_i = y\right)$
- $c_\gamma\left(y_1, y_2\right) :=$ Cov $\left(I_{\gamma i}, I_{\gamma \ell} | Y_i = y_1, Y_\ell = y_2\right)$
- $w_\gamma(y) := \frac{1}{m_\gamma(y)}$
- Assume $m_\gamma\left(y\right) \stackrel{\text{a.s.}}{\rightarrow} m\left(y\right)$
- Assume $\forall y_1, y_2,\ c_\gamma\left(y_1, y_2\right) = o_\gamma(1)$

$$\frac{w_\gamma}{\mathbb{E}_o(w_\gamma(y))} f_o\left(y|\boldsymbol{\lambda}\right) \stackrel{\text{a.s.}}{\rightarrow} f_p\left(y|\boldsymbol{\lambda}\right)$$

- Bonnéry et al. (2013) extends Glivenko-Cantelli

Empirical Generation of Pseudo-populations

The Procedure

- Generate B pseudo-populations, each of size n, by weighted sampling of observed data with replacement using normalized inverse inclusion probabilities
- Estimate $(\boldsymbol{\lambda}_b)_{b=1,\ldots,B}$ under each pseudo-population
- Concatenate estimates, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_B)$

Empirical Generation of Pseudo-populations

The Idea

- Weighted empirical predictive distribution, $\hat{f}\left(\mathbf{y}_p|\mathbf{y}_o\right)$, generates "pseudo-populations" (Pfeffermann et al., 1998, Pfeffermann and Sverchkov, 2003, Sverchkov and Pfeffermann, 2004), applies result of Bonnéry et al. (2013)
- $\mathbf{y}_p$ are treated as parameters

$$\hat{f}\left(\boldsymbol{\lambda}|\mathbf{y}_o\right) = \int f\left(\mathbf{y}_p|\boldsymbol{\lambda}\right)\hat{f}\left(\mathbf{y}_p|\mathbf{y}_o\right)f\left(\boldsymbol{\lambda}\right)d\mathbf{y}_p = \int c(\mathbf{y}_p)f\left(\boldsymbol{\lambda}|\mathbf{y}_p\right)\hat{f}\left(\mathbf{y}_p|\mathbf{y}_o\right)d\mathbf{y}_p.$$

- Full uncertainty accounted for in estimation of $\boldsymbol{\lambda}$
  - population generation, $\hat{f}\left(\mathbf{y}_p|\mathbf{y}_o\right)$
  - $\boldsymbol{\lambda}$ estimation given the finite population, $\hat{f}\left(\boldsymbol{\lambda}|\mathbf{y}_p\right)$
  - conditioned on $\mathbf{y}_o$

Substitution "Plug-in" Method

- $f_o\left(\boldsymbol{\lambda}|H\left(\mathbf{Y}_o\right)\right) \equiv$ sample model
- Use $\left(\tilde{w}_i\right)$ to adjust $H\left(\cdot\right)$ from sample to population
- $\widehat{f_p}\left(\boldsymbol{\lambda}|H\left(\mathbf{Y}_o, \tilde{\mathbf{w}}\right)\right) \equiv$ population model
- Statistic, $\widehat{f_p}\left(\boldsymbol{\lambda}|\mathbf{H}\left(\mathbf{y}_o, \tilde{\mathbf{w}}\right)\right) = \displaystyle\prod_{j=1}^{p} \widehat{f_p}\left(\lambda_j|\mathbf{H}\left(\mathbf{y}_o, \tilde{\mathbf{w}}, \boldsymbol{\lambda}_{-j}\right)\right)$

$$\log\widehat{f_p}\left(\theta^{*}_{qm}|H\left(\{\mathbf{y}_i\}_{i\in s_m}, \{\tilde{w}_i\}_{i\in s_m}, \tau_{\epsilon}, \boldsymbol{\theta}^{*}_{m}\right)\right)$$
$$\propto -\frac{1}{2}n_m\log\left(|\mathbf{C}^{\tau}\left(\boldsymbol{\theta}^{*}_{m}\right)|\right) - \frac{1}{2}\sum_{i\in s_m}\mathbf{y}^{'}_{o,i}\mathbf{C}^{\tau}\left(\boldsymbol{\theta}^{*}_{m}\right)\mathbf{y}_{o,i}\tilde{w}_i + \log f(\theta^{*}_{qm})$$

- $\mathbf{C}^{\tau}\left(\boldsymbol{\theta}^{*}_{m}\right) = \mathbf{C}\left(\boldsymbol{\theta}^{*}_{m}\right) + \left(1/\tau_{\epsilon}\right)\mathbb{I}_T$
- s_m collects the units assigned to cluster m
- $n_m = \sum_{i\in s_m}\tilde{w}_i,\ \ \sum_{m=1}^{M} n_m = n$

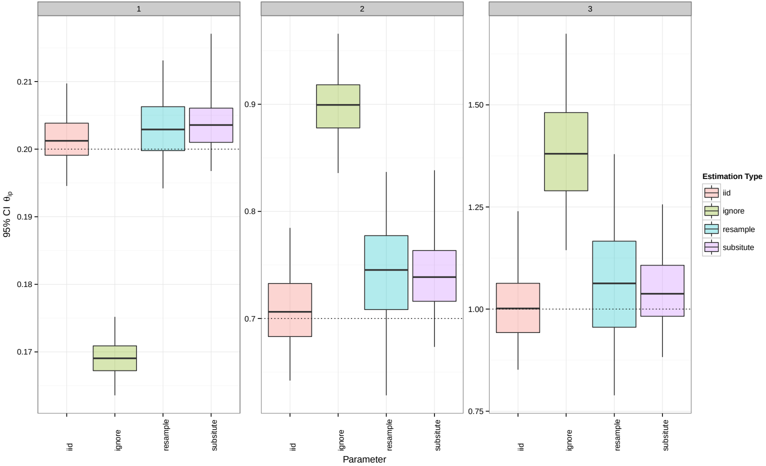
Comments on using Sampling Weights

- Both re-sampling and substitution don't require to re-parameterize
- This estimator employs harmonic averages for full conditionals
  - May break-down under high variation in the weights
- "Gold standard" to jointly model $(\mathbf{y}_i, \pi_i)$

A Warm Up

- Generate $N = 10000$ population units
- $\mathbf{f}_i \sim \mathcal{N}_t\left(\mathbf{0}, \mathbf{C}\left(\boldsymbol{\theta}\right)\right)$
- $\theta = (0.2, 0.7, 1.0)$ are global, not indexed by unit - No clustering
- Generate $y_{ij} = \mathcal{N}\left(f_{ij}, \tau_{\epsilon}^{-1}\right)$
- τ_{ϵ} chosen to achieve noise-to-signal of 0.1
- Employ $H = 4$ strata
- Assign units based on quantiles of variances for N, $T \times 1$ functions, $\mathbf{y}_i$
- Idea is that more interested in high magnitude (error) functions
- (STSI) Sample $n = 750$ from $H = 4$ strata with,
  - $\mathbf{n} = (82, 113, 211, 344)$, $\mathbf{p} = (0.03, 0.05, 0.08, 0.14)$
- Generate B = 25 pseudo-populations for re-weighted estimation

$\theta = (0.2, 0.7, 1.0)$ and $\mathbf{n} = (82, 113, 211, 344)$



2− stage Sampling Design

- Form $n_I = 10$ homogenous blocks by variance quantiles (of $\mathbf{y}_i$)
- 1^{st} stage: Sample 7 blocks proportional to variance
  - $p = (0.04, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.14, 0.19)$
- 2^{nd} stage: $H = 4$ strata based on variance quantiles within blocks

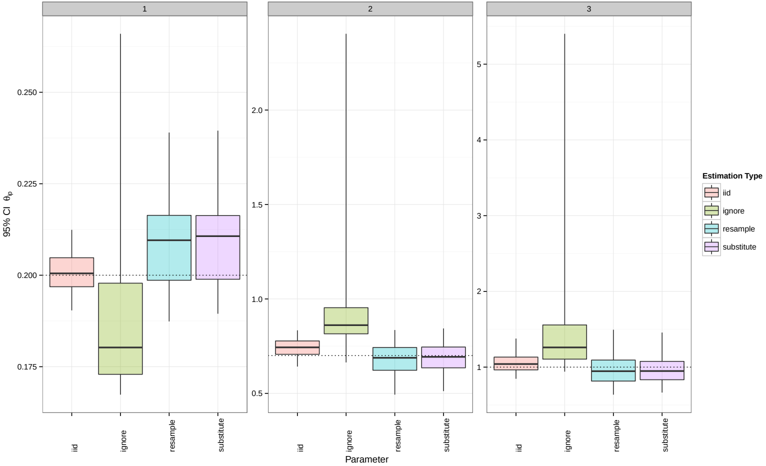
$p_I(\mathbf{s}_I) * p_i(\mathbf{s}_{Ii})' =$

block/stratum	H=1	2	3	4
1	0.001	0.002	0.004	0.006
2	0.002	0.003	0.006	0.009
3	0.002	0.004	0.007	0.010
4	0.003	0.004	0.008	0.012
5	0.003	0.005	0.009	0.013
6	0.004	0.005	0.010	0.015
7	0.004	0.006	0.011	0.017
8	0.004	0.006	0.012	0.019
9	0.005	0.007	0.014	0.022
10	0.007	0.010	0.019	0.029

- $n = 750$ units from $N = 10000$ population

$2-$ stage Posterior Distributions for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$

- Monte Carlo draws of 10 samples, each selecting 7 of 10 blocks
- "Most" population elements have non-zero selection probabilities
- Conditional inclusion dependence is low



Generate Functions Under $M = 3$ Clusters

- Generate $N = 10000$ population units
- Here, $\{ \overset{P\times 1}{\boldsymbol{\theta}_i} \}_{i=1,\ldots,N}$
- 3 clusters with location values, $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_M^*)$:

$\boldsymbol{\Theta}^* =$

parameter/cluster	$m=1$	2	3
$p=1$	0.30	0.30	0.30
2	0.30	2.50	0.70
3	0.80	1.50	2.00
	rough-varied	smooth	rough

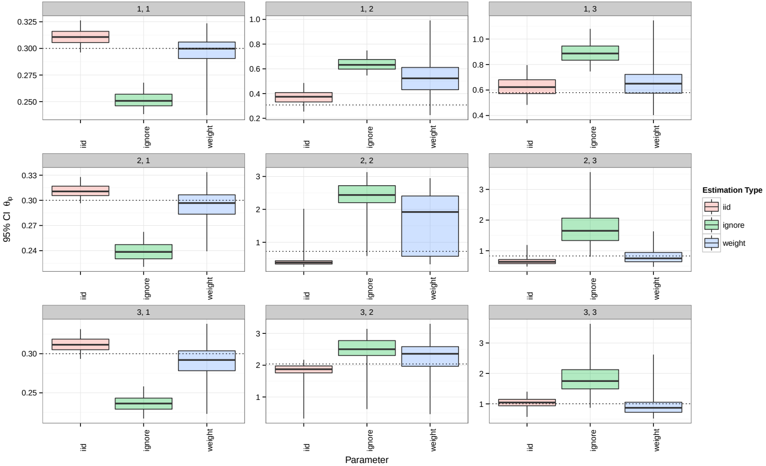
- $\mathbf{s} = s_1, \ldots, s_N$ indexes randomly-assigned cluster memberships
- $s_i \in (1, \ldots, M = 3)\,,\; i = 1, \ldots, N$
- $\boldsymbol{\theta}_i \equiv \boldsymbol{\theta}^*_{s_i}$

Generate Functions Under $M = 3$ Clusters

- $\mathbf{f}_i \sim \mathcal{N}_t\left(\mathbf{0}, \mathbf{C}\left(\boldsymbol{\theta}^*_{s_i}\right)\right)$
- Generate $y_{ij} = \mathcal{N}\left(f_{ij}, \tau_\epsilon^{-1}\right)$
- τ_ϵ chosen to achieve noise-to-signal of 0.1
- Employ $H = 4$ strata
- Assign units based on quantiles of variances for N, $T \times 1$ functions, $\mathbf{y}_i$
- (STSI) Sample $n = 800$ from $H = 4$ strata with unequal probabilities, $\mathbf{p} = (0.03, 0.06, 0.07, 0.16)$, $\mathbf{n} = (75, 155, 178, 392)$
- Generate B = 25 pseudo-populations for re-weighted estimation

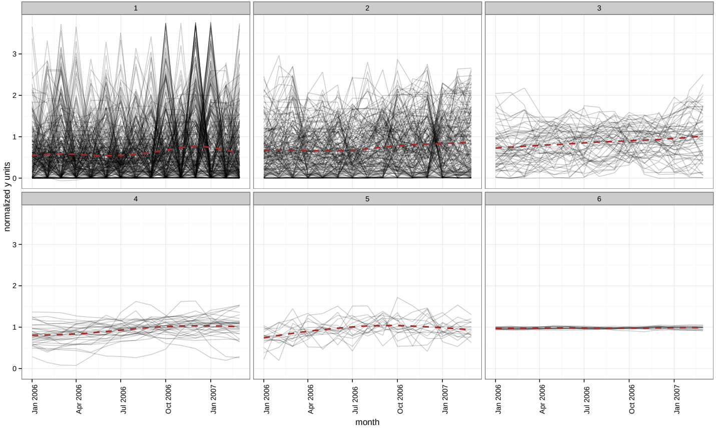
Posterior Distributions for $\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2}, \theta_{i,3})$

- One unit randomly chosen within each cluster
- Non-informative performs worst on members of smooth cluster
- Does best on rough cluster - over-samples ↑ variance functions



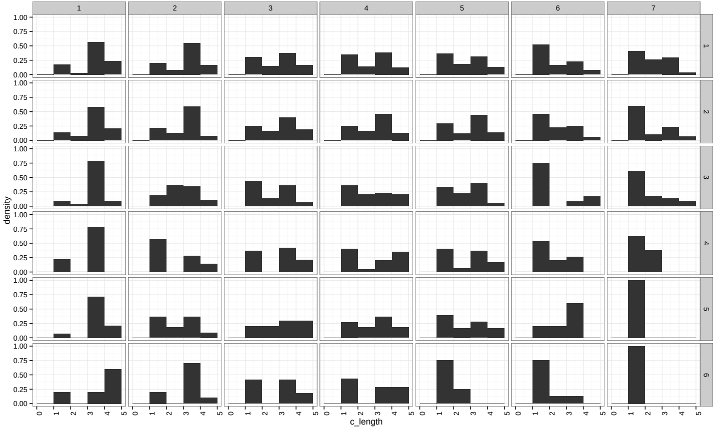
Clustering of $\{f_i\}$ for RAS Data

- 6 clusters contain $(950, 648, 204, 106, 57, 48)$ establishments
- Based on relative magnitude of errors committed than on shape



Closing Period by Size × Cluster

- 7 employee size categories, $(1-9,\ 10-19,\ 20-49,\ 50-99,\ 100-249,\ 250-499,\ \ >499)$
- BLS defines four time-indexed closing periods
- larger sized establishments in lower error clusters report sooner



References

Bonnéry, D., Breidt, F. J., and Coquet, F. (2013). Uniform convergence of the empirical cumulative distribution under informative selection from a finite population. Technical report, Submitted to Bernoulli.

Pfeffermann, D., Krieger, A., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. Statistica Sinica 8, 1087-1114 (1998).

Pfeffermann, D. and Sverchkov, M. (2003). *Fitting Generalized Linear Models under Informative Sampling*, pages 175–195. John Wiley & Sons, Ltd.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30(1):79–82.