Simulations on "Bayesian hierarchical weighting adjustment and survey inference" by Trangucci et al.

Fall 2018

D.Bonnéry

1 Model description

1.1 Trangucci et al. model0

We describe the model given in (?, Sec. 2).

- Population made of H strata $U_1,\ldots,U_J$.
- Stratum j size: N_j, Total size: $N=\sum_j N_j$
- Sample size in stratum j: n_j
- I vector: sample indicator
- y vector: study variable
- θ_j: average of y in stratum j. $\theta_j=\sum_{i\in U_j} y_i$. In the paper there is an ambiguity: First, θ is called the "population estimand of interest[...] the overall or domain mean". Ambiguity comes from the term "mean". Is $\theta=N^{-1}\sum_i y_i$ or Is $\theta=N^{-1}\sum_k E[y_i]$?
- X : Design variables. In the paper $X^1,\ldots,X^J$ are badly defined. The idea is that X variables are Q categorical variables. The strata correspond to the cells obtained from these categorical variables.

- denote by $1,\dots K_q$ the categories for variable X_q.
- denote by $j[i]$ the stratim of unit i
- Denote by $k[q,j]$ the category for variable X_q in stratum j

Consider the following hierarchical model:

$$y_i \sim \mathcal{N}\left(\theta^{\star}_{j[i]}, \sigma_y^2\right) \tag{1.1}$$

$$\theta^{\star}_j = \alpha_0 + \sum_{\ell=1}^{Q}\left(\sum_{q_1<\dots<q_\ell\in\{1,\dots,Q\}} \alpha_j^{(q_1,\dots,q_\ell)}\right) \tag{1.2}$$

$$\forall \ell \in \{1,\dots,Q\},\ \forall q_1,\dots,q_\ell \in \{1,\dots,Q\},\ \forall j \in \{1,\dots,H\} : \alpha_j^{(q_1,\dots,q_\ell)} \sim \mathcal{N}(0, (\lambda_j^{(q_1,\dots,q_\ell)}\sigma)^2) \tag{1.3}$$

$$\forall \ell \in \{1,\dots,Q\},\ \forall q_1,\dots,q_\ell \in \{1,\dots,Q\},\ \forall j \in \{1,\dots,H\} : \lambda_j^{(q_1,\dots,q_\ell)} = \delta^{(\ell)} \prod_{l=1}^{\ell} \gamma^{(q_l)}_{k[q_l,j]} \tag{1.4}$$

$$\sigma \sim \text{Cauchy}_+(0,1) \tag{1.5}$$

$$\forall q \in \{1,\dots,Q\},\ k \in \{1,\dots,K_q\}\ : \gamma_k^{(q)} \sim \mathcal{N}_+(0,1) \tag{1.6}$$

$$\delta^{(\ell)} \sim \mathcal{N}_+(0,1) \tag{1.7}$$

$$\sigma_y \sim \text{Cauchy}_+(0,5) \tag{1.8}$$

$$\alpha_0 \sim \text{N}(0,10) \tag{1.9}$$

Note I made up the prior on α_0 as I did not find in the paper.

1.2 Competing model

We describe the model given in (**?**, Sec. 2).

Consider the following hierarchical model:

$$y_i \sim \mathcal{N}\left(\theta^{☾}_{j[i]}, \sigma_y^2\right) \tag{1.10}$$

$$(\theta^{☾}_j)_{j=1}^{Q} \sim \mathcal{N}(0, \Sigma) \tag{1.11}$$

$$\Sigma_{j_1,j_2} = \sigma^2 D_\lambda(j_1,j_2) \tag{1.12}$$

$$D_\alpha(j_1,j_2) = \sigma_y^2(\exp(-\sum_{q=1}^{Q} \alpha_q(k[q,j_1] \neq k[q,j_2])) \tag{1.13}$$

$$\sigma_y \sim \text{Cauchy}_+(0,5) \tag{1.14}$$

$$\alpha_q \sim \text{Cauchy}_+(0,1) \tag{1.15}$$

2 Data Generation

Step 1

Create procedures that can

1. generate a population of size N and with Q random categorical variables $X_1, \ldots, X_Q$. X_q is generated by drawing with replacement in $\{1, \ldots, K_q\}$ where $K_q > 1$, K_q can be generated by drawing from $\mathcal{U}_{1,\ldots,p}$.
2. compute the corresponding stratum $j = 1, \ldots, J$ and the correspondances $(k[q,j])_{j=1,\ldots,J, q=1,\ldots,Q}$.
3. generate, for such a population, the hyper parameters σ_y, α_0, $(\delta^\ell)_{\ell=1,\ldots,Q}$,$\lambda_k^l$,$\sigma$.
4. compute the $\lambda_{k_1,\ldots,k_\ell}^{(q_1,\ldots,q_\ell)}$, for possible values of $k_1, \ldots, k_\ell{}^{(q_1,\ldots,q_\ell)}$.
5. generate the $\alpha_{j,(k_1,\ldots,k_\ell)}^{(q_1,\ldots,q_\ell)}$.
6. compute $\theta^\star$.
7. generate a number r of realisations of y for such hyperparameters and such strata.
8. Set seed.
9. Create a population with $N = 10000$,$Q = 2$, $p = 5$. Display J, N_j, $\theta^\star$.

For 1-8, see R code below.
9.
9. This are tables we obtained with different seeds.

Table 2.1:

	X_1	X_2	Stratum (j)	N_j	$\theta^\star$
S1	1	1	S1	113	-5.151
S2	1	2	S2	111	-3.938
S3	1	3	S3	119	-2.698
S4	2	1	S4	101	-0.163
S5	2	2	S5	106	-4.607
S6	2	3	S6	105	-1.853
S7	3	1	S7	140	0.434
S8	3	2	S8	104	-1.733
S9	3	3	S9	101	-6.088

Table 2.2:

	X_1	X_2	Stratum (j)	N_j	$\theta^\star$
S1	1	1	S1	91	3.215
S2	1	2	S2	73	0.055
S3	1	3	S3	83	2.687
S4	1	4	S4	86	2.681
S5	2	1	S5	91	-5.146
S6	2	2	S6	82	-10.338
S7	2	3	S7	80	3.576
S8	2	4	S8	78	0.260
S9	3	1	S9	95	6.074
S10	3	2	S10	80	0.900
S11	3	3	S11	91	2.567
S12	3	4	S12	70	-4.289

```
library(SimuTrangucci)
set.seed(11)
GG<-Generate_all(N=1000,Q=2,p=5)
Strata1<-data.frame(GG$XX$Strata,
                    N_j=GG$XX$N_j,
                    thetastar=GG$thetastar)
set.seed(7)
GG<-Generate_all(N=1000,Q=2,p=5)
Strata2<-data.frame(GG$XX$Strata,
                    N_j=GG$XX$N_j,
                    thetastar=GG$thetastar)
```

3 Bayesian computations

Step 1

Using Jags,

1. Draw posterior distribution of $\theta_j^\star$ for the largest value of $\theta_j^\star$, obtained from the observation of $y^{(1)}$ only. Add real value of θ_1 and prediction to the plot.
2. Draw distribution of predictions of $\theta_j^\star$. Add real value of θ_j to the plot.
3. Draw real values of all $\theta_j^\star$ vs $r = 30$ predictions.

Some functions to generate the jags model file:
library(SimuTrangucci)

Step 2

same thing with Stan

Step 3

1. Design a sampling scheme that favors some cells
2. Compute Trangucci estimator for θ for the $r = 30$ realisations.

Step 4

1. Use another model to generate the population, a model that does not fit the current one, to fail the product structure.
2. Look at predictions for pop total.

1. Assume the following model:
2.

4 Bayesian computations