Johndrow et. al. Details

Andee Kaplan

12/15/2017

1 M_{th} model with nonparametric prior

Consider a sample of m individuals captured from a population of unknown size N in T lists. x_{it} is a binary representation of each individual's capture history for $i = 1, \dots, m$ and $t = 1, \dots, T$:

$$x_{it} = \begin{cases} 1 & \text{individual } i \text{ is recorded in list } t \\ 0 & \text{otherwise.} \end{cases}$$

These data can be summarized by a contingency table where each cell count is denoted $n(\boldsymbol{x})$ for $\boldsymbol{x} \in \{0,1\}^T$. We will let $\boldsymbol{\zeta}$ denote the zero vector of dimension T, such that $n(\boldsymbol{\zeta})$ is the count of individuals not captured by any list, and is the focus of our inference. We then specify the following model, as in Johndrow, Lum, and Manrique-Vallier (2016):

$$\begin{aligned} x_{it} \mid \theta_i, \beta_t &\overset{ind}{\sim} \text{Bern}(\varphi^{-1}(\theta_i + \beta_t)) \\ \theta_i \mid G^* &\overset{iid}{\sim} G^* \\ G^* &\sim \text{DP}(\alpha_0, N(0, \sigma^2_{G^*})) \\ \beta_t &\overset{iid}{\sim} N(0, \sigma^2_\beta) \\ \alpha_0 &\sim \text{Gamma}(a, b), \end{aligned}$$

where $\varphi^{-1} : \mathbb{R} \to [0,1]$ is a monotone nondecreasing transformation used to parameterize probabilities, such as the logit or probit function.

2 Conditional distribution of $n(\boldsymbol{\zeta})$

The count of individuals not captured by any list, $n(\boldsymbol{\zeta})$, can be thought of as the number of elements not captured in a list *before* m elements are captured by the T lists. In this way, $n(\boldsymbol{\zeta})$ can be thought of as the number of successes (elements not captured in a list) in a sequence of iid Bernoulli trials before a specific (non-random) number of failures (elements captured in a list, m). This leads $n(\boldsymbol{\zeta})$ to be distributed negative binomial random *if* the probability of success (probability of not being captured by any list) is identical across trials (individuals). In general, this is not true, however conditional on the K-length truncation of the stick-breaking process (approximating the DP), it is.

$$
\begin{aligned}
p &= P(\text{an element not being captured by any list} \mid K\text{-length truncation of the stick-breaking process}) \\
&= P(\boldsymbol{x}_i = \boldsymbol{\zeta} | \boldsymbol{\theta}^*_{[1:K]}, \boldsymbol{\beta}_{[1:T]}, \boldsymbol{\nu}_{[1:K]}) \\
&= \int_\Theta P(\boldsymbol{x}_i = \boldsymbol{\zeta} | \theta_i, \boldsymbol{\beta}_{[1:T]}) P(\theta_i | \boldsymbol{\theta}^*_{[1:K]}, \boldsymbol{\nu}_{[1:K]}) d\theta_i \\
&= \int_\Theta \prod_{t=1}^{T} \{1 - \varphi^{-1}(\theta_i + \beta_t)\} \times \sum_{h=1}^{K} \nu_h \boldsymbol{I}(\theta^*_h = \theta_i) d\theta_i \\
&= \int_\Theta \sum_{h=1}^{K} \prod_{t=1}^{T} \{1 - \varphi^{-1}(\theta_i + \beta_t)\} \nu_h \boldsymbol{I}(\theta^*_h = \theta_i) d\theta_i \\
&= \int_\Theta \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \{1 - \varphi^{-1}(\theta^*_h + \beta_t)\} d\theta_i \\
&= \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \{1 - \varphi^{-1}(\theta^*_h + \beta_t)\}
\end{aligned}
$$

When φ is the probit function (as used in Johndrow, Lum, and Manrique-Vallier (2016)), this results in

$$
p = \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \{1 - \Phi(\theta^*_h + \beta_t)\} = \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \{\Phi(-\theta^*_h - \beta_t)\},
$$

where Φ is the standard normal cdf. The result is that $n(\boldsymbol{\zeta})$ is conditionally a negative binomial rabdom variable with the following parameters

$$
n(\boldsymbol{\zeta}) | \boldsymbol{\theta}^*_{[1:K]}, \boldsymbol{\beta}_{[1:T]}, \boldsymbol{\nu}_{[1:K]} \sim \text{Neg-Bin}\left(m, \sum_{h=1}^{K} \nu_h \prod_{t=1}^{T} \{\Phi(-\theta^*_h - \beta_t)\}\right)
$$

when the model assumptions above are made.

References

Johndrow, James E, Kristian Lum, and Daniel Manrique-Vallier. 2016. "Estimating the Observable Population Size from Biased Samples: A New Approach to Population Estimation with Capture Heterogeneity." *arXiv Preprint arXiv:1606.02235*.