Bayesian Sparse Propensity Score Estimation for Unit Nonresponse

Hejian Sang, Iowa State University
Gyuhyeong Goh, Kansas State University
Jae Kwang Kim, Iowa State University

Nonresponse weighting adjustment using propensity score (PS) is a popular tool for handling unit nonresponse. However, including all the auxiliary variables into the propensity model can lead to inefficient estimation and the consistency is not guaranteed if the dimension of the covariates is large. For the PS setup, let Y is a scalar response and X is a p-dimensional vector of covariates. Y is subject to missingness and X is fully observed. δ is the indicator function of observing Y. Suppose we are interested in estimating parameter $\theta \in \Theta$, which is the unique solution to the population estimating equation $E\{U(\theta; X, Y)\} = 0$. Furthermore, we define the propensity score for the i-th observation as $\Pr(\delta_i = 1|x_i) = \pi(\phi; x_i) = G\left(x_i^T \phi\right)$, where $G : \mathbb{R} \to [0, 1]$ is a known distribution function and $\phi = (\phi_1, \phi_2, \ldots, \phi_p)^T$ is a p-dimensional unknown parameter. However, when ϕ is sparse, that is, ϕ contains many zero values, the MLE often involves large variance and fails to be consistent

To formulate our proposal, we first introduce a latent variable z, which indicates nonzero elements of ϕ. To account for the sparsity of the response model, we assign the Spike-and-Slab Gaussian mixture prior for ϕ, denoting as $p(\phi \mid z)$. Assign independent Bernoulli prior for z, denoting as $p(z)$. Let $L_1(\phi|x, \delta)$ be the likelihood of ϕ obtained under the assumption of independence. Then, our proposed Bayesian sparse propensity score (BSPS) method can be described as following two steps:

Step 1: Generate z^* from the marginal posterior distribution of z given x and δ.

Step 2: Generate θ^* from an approximate posterior distribution of θ given the z^* generated from **Step 1**.

To generate z^* in **Step 1** efficiently, the data augmentation algorithm can be applied. That is, the marginal posterior distribution of z given x and δ can be obtained by iterating the following two steps until convergence:

I-step: Given ϕ^*, generate z^* from

$$\begin{aligned} z^* \sim p(z|x, \delta, \phi^*) &= \frac{L_1(\phi^*|x, \delta)p(\phi^*|z)p(z)}{\int L_1(\phi^*|x, \delta)p(\phi^*|z)p(z)dz} \\ &= \frac{p(\phi^*|z)p(z)}{\int p(\phi^*|z)p(z)dz} = p(z|\phi^*). \end{aligned}$$

P-step: Given z^*, generate ϕ^* from

$$\phi^* \sim p(\phi|x, \delta, z^*) = \frac{L_1(\phi|x, \delta)p(\phi|z^*)}{\int L_1(\phi|x, \delta)p(\phi|z^*)d\phi}.$$

I-step can be explicitly expressed as a Bernoulli distribution. However, the normalizing constant in **P-step** (1) is not tractable. Thus, we propose to use the Approximate Bayesian Computation method to replace the original likelihood function of ϕ. And, in **Step 2**, we apply the similar technique due to no explicit likelihood function of θ.

Model consistency and asymptotic normality are established. The finite-sample performance of the proposed method is investigated in limited simulation studies, including a partially simulated real data example from the Korean Labor and Income Panel Survey.