DPGLM: Simulation Study

1 DP-GLM

$$y_i \mid z_i, x_i \sim K(y_i \mid z_i, x_i) = K(y_i \mid z_i), \quad y_i, z_i \in \mathcal{Y} \tag{1}$$

$$z_i \mid x_i = x, \widetilde{\theta}_x, \widetilde{\mu} \sim p_x(z_i) \propto \exp(\widetilde{\theta}_x z_i)\widetilde{\mu}(z_i) \tag{2}$$

$$\widetilde{\theta}_x \mid \theta_x \sim p(\widetilde{\theta}_x \mid \theta_x), \text{ with } b'(\theta_x) = \int_{\mathcal{Y}} z \frac{\exp(\theta_x z)\widetilde{\mu}(z)}{\int_{\mathcal{Y}} \exp(\theta_x u)\widetilde{\mu}(u)du} dz = g^{-1}(x'\beta) \tag{3}$$

$$\widetilde{\mu} \sim \text{ gamma CRM}(\nu), \text{ with } \nu(dw, dm) = \alpha \frac{e^{-w}}{w} dw \cdot G_0(dm) \tag{4}$$

$$\beta \sim \text{ MVN}(\mu_\beta, \Sigma_\beta). \tag{5}$$

1.1 Modeling fractional data

Here $\mathcal{Y} = [0, 1]$.

1.1.1 Hyper-and-tuning parameters

- $K(\cdot \mid z_i) =$ Uniform$(z_i - c_0, z_i + c_0)$. We use $c_0 = 0.025$
- We truncate the CRM at $M = 20$, where the CRM is given by $\widetilde{\mu}(\cdot) = \alpha \sum_{h=1}^{M} w_h \delta_{m_h}(\cdot)$. So, all we need to do is to put priors on w_h and m_h. We take $m_h \sim G_0 =$ Uniform$(0, 1)$ and $w_h \sim$ improper gamma dist with intensity $\rho(dw) = \alpha \frac{e^{-w}}{w} dw$, and the corresponding NRM prior: $\widetilde{\mu}_{nrm}(\cdot) = \alpha \sum_{h=1}^{M} w_h^{normed} \delta_{m_h}(\cdot)$, with $w_h^{normed} \sim$ Beta$(1, \alpha)$. We take $\alpha = 1$.
- $g(\mu) = \ln(\frac{\mu}{1-\mu})$ [logit link]
- We take $\mu_\beta = 0$ and $\Sigma_\beta = \sigma_\beta^2 I_p$. We set $\sigma_\beta^2 = 1$.

1.1.2 How to get pdf and cdf?

The kernel $K(y \mid z) =$ Uniform$(y;\ z - c_0, z + c_0)$, $y \in [0, 1]$. The density of y given x is given by,

$$f(y \mid x) = \int_z K(y \mid z, x) p(z \mid x, \theta_x, \tilde{\mu}) dz = \sum_\ell \frac{1}{2c_0} 1_{\{z_\ell - c_0, z_\ell + c_0\}}(y) \frac{\exp(\theta_x z_\ell) J_\ell}{\sum_{\ell'} \exp(\theta_x z_{\ell'}) J_{\ell'}}$$

. Let's call $\frac{\exp(\theta_x z_\ell) J_\ell}{\sum_{\ell'} \exp(\theta_x z_{\ell'}) J_{\ell'}} = \pi_\ell(\theta_x)$. So,

$$f(y \mid x) = \sum_\ell \pi_\ell(\theta_x) \frac{1}{2c_0} 1_{\{z_\ell - c_0, z_\ell + c_0\}}(y).$$

From here, we get $f_0(y)$ by replacing $\theta_x = 0$. Similarly, the CDF is given by,

$$F(y \mid x) = \int_0^y f(y' \mid x) dy' = \sum_\ell \pi_\ell(\theta_x) \left[\left(\frac{y - z_\ell + c_0}{2c_0} \right) 1_{\{z_\ell - c_0, z_\ell + c_0\}}(y) + \left(\frac{2c_0}{2c_0} \right) 1_{\{z_\ell + c_0 < y\}}(y) + 0 \cdot 1_{y < z_\ell - c_0}(y) \right].$$

From here, we similarly get $F_0(y)$ by replacing $\theta_x = 0$. IMPORTANT!! should we tilt f_0 to have mean μ_0? Then, should we do it for both — truth and estimates, when performing comparisons in simulation study?

2 Simulation Studies

We proceed with simulation studies to evaluate the frequentist operating characteristics of the DP-GLM model. Our investigation addresses the following key questions:

(Q1) How does the model perform in terms of predictive accuracy when estimating the baseline density, $f_{\widetilde{\mu}}(y)$, under various scenarios?

(Q2) Do the credible intervals for $f_{\widetilde{\mu}}(y)$ achieve coverage rates close to their nominal levels?

(Q3) In scenarios where the response is independent of predictors, does $\theta_{x;n} := [\theta_x \mid \mathcal{D}_n]$ converge in probability to a constant (in x), or alternatively, do the credible intervals for θ_x attain nominal coverage rates?

(Q4) Do the credible intervals for β_j parameters attain nominal coverage? How is their predictive accuracy?

We consider a data generating mechanism where the response y is sampled from the Speech Intelligibility dataset.

3 Simulation Setting I: Null Case

Let $f_{\widetilde{\mu}}^{(kde)}$ denote the kernel density estimate based on the response data from Speech Intelligibility dataset (ignoring the covariates). We consider $f_{\widetilde{\mu}}^{(kde)}$ as the simulation truth for the baseline density $f_{\widetilde{\mu}}$. Covariates are generated as: $x_{0i} = 1, x_{1i} \sim \text{Normal}(\mu_1, \sigma_1), x_{2i} \sim \text{Normal}(\mu_2, \sigma_2)$, where we take $\mu_1 = 1, \sigma_1 = 0.5, \mu_2 = 2, \sigma_2 = 1$. We sample y independent of x i.e, $y_i \sim f_{\widetilde{\mu}}^{(kde)}$. We use $\mathcal{D}_n$ to refer the observed data $\{x_i, y_i\}_{i=1}^n$. This setting aims to address Q1−Q3.

3.1 Analysis

We get f_0 and its cdf F_0 by replacing $\theta_x = 0$ in the expressions in Section 1.1.2. Similarly we get the estimates.

4 Simulation Setting II: Point masses

Let $f_{\widetilde{\mu}}^{(Beta)}$ denote the Beta(a, b) density estimate based on the response data from Speech Intelligibility dataset (ignoring the covariates). We consider $f_{\widetilde{\mu}}^{(Beta)}$ as the simulation truth for $f_{\widetilde{\mu}}$, with additional point masses p_0 and p_1 respectively at $y = 0$ and $y = 1$. We take $p_0 = 0.1$ and $p_1 = 0.4$. The rest is same as in Setting I. Apart from Q1−Q3, the primary objective here is to assess whether the model accurately estimates the point masses.

4.1 Analysis

We get f_0 and its cdf F_0 as follows — by replacing $\theta_x = 0$ in the expressions in Section 1.1.2 for $y \in (0, 1)$, and let's call it $F_0^\star(y)$. Then, our cdf would be: $F_0(0) = p_0$ and $F_0(y) = p_0 + (1 - p_0 - p_1) \cdot F_0^\star(y),\ y \in (0, 1)$, and $F_0(1) = 1$. Similarly we get the estimates.

5 Simulation Setting III: Regression

We consider the same framework as in Setting I, with one modification: the sampling y is now dependent on $x = (x_1, x_2)$. Specifically, we sample $y_i \sim p(y_i \mid x_i) \propto \exp(\theta_{x_i} y_i) f_{\widetilde{\mu}}^{(kde)}(y_i)$, where $\theta_x \sim \text{Normal}(\widetilde{\theta}_x, \sigma_\theta^2)$, with $\sigma_\theta = 0.001$. Here, $\widetilde{\theta}_x = {b'}^{-1}(g^{-1}(\eta_x))$, with $\eta_x = \beta_0 + x^T \beta$. We set $\beta_0 = -0.7, \beta^T = (0.2, -0.1)$. This setting aims to address Q1, Q2 and Q4.