Model-Based Survey Weighting Using Hierarchical Gaussian Processes for Structured Populations

Seonghun Lee

1 Introduction

Survey weighting is a fundamental tool for correcting selection bias and achieving population-representative estimates, especially when data suffer from unequal sampling probabilities, nonresponse, or undercoverage. In our previous work, we explored a model-based framework for constructing base survey weights using logistic regression. That approach offered a structured alternative to traditional design-based weights, showing how predictive modeling could align with classical estimators like the Hájek estimator when appropriately normalized. However, logistic regression—while useful—assumes a linear structure on the log-odds scale and does not easily accommodate complex interactions or smooth variation across structured populations.

In this sequel, we introduce a more flexible model-based approach to survey weight construction using Hierarchical Gaussian Processes (HGPs). HGPs offer a principled way to model complex, nonlinear relationships across both continuous and discrete covariates, while borrowing strength across groups through hierarchical structure. This is particularly valuable in survey settings where small subgroup sizes or sparsely observed cells limit the stability of traditional poststratification or regression-based weighting methods.

A key computational innovation in our method is the use of structured covariance matrices that exploit the known grouping of the population. Rather than fitting a full Gaussian process across individual-level covariates—which can be computationally expensive due to the $\mathcal{O}(n^3)$ scaling—we construct a covariance matrix over groups (e.g., poststratification cells from weighting variables). This reduces the dimensionality of the GP to the number of unique groups rather than individuals, yielding substantial computational savings while preserving flexibility in the modeled inclusion probabilities. In practice, this makes HGP-based weighting feasible even for large surveys with thousands of respondents and of poststratification cells.

We develop this approach within a fully Bayesian framework and compare the resulting population estimates to those from alternative models, and the resulting weights to both our previous model-based methods and classical design-based approaches. Our simulations demonstrate that the HGP-based method not only reduces bias in estimated outcomes and variance in estimated weights but also enhances robustness in the presence of data sparsity, providing more stable inference without requiring external raking or trimming. Additionally, while binary outcomes can limit a model's flexibility and stability—especially in sparse cells—our method is designed to accommodate a broad range of outcome types, ensuring wider applicability across survey settings.

This paper contributes a flexible and scalable model-based framework for survey weighting, with a focus on structured populations. Our results support the use of HGPs for smoothing across cells and correcting for complex inclusion mechanisms—expanding the practical toolbox available for researchers seeking principled and efficient weighting strategies.

2 Hierarchical Gaussian Process Regression

Our goal is to estimate the population mean of an outcome variable y across structured subgroups, given a sample that sparsely covers the full poststratification space. Let the population be divided into H hierarchical groups (e.g., demographic strata for which population-level information is available) and K kernel groups (e.g., the full set of poststratification cells spanning all demographic combinations). In typical settings, we observe outcomes for only a small subset of the full set of poststratification cells (e.g., 1,100 observed cells out of 4,480 total), which leads to instability in direct cell-level estimation.

To overcome this sparsity, we propose a hierarchical Gaussian process (HGP) model that enables information sharing across related groups. Specifically, we assume that the latent outcome-generating process is composed of additive components: a structured group-level effect μ_h and a smooth kernel-based effect η_k modeled by a Gaussian process.

Formally, for each observation $i = 1, \ldots, n$ in the sample, we assume the following model:

$$\mathbb{E}[Y \mid X] = g^{-1}(f(X)) + \mathbb{E}[\varepsilon], \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $g(\cdot)$ is a monotonic link function, and the latent function is defined component-wise as

$$f_i = \mu_{h[i]} + \eta_{k[i]}, \quad \text{for } i = 1, \ldots, n.$$

where $\mu_{h[i]}$ is the fixed effect for group $h[i] \in \{1, \ldots, H\}$ and $\eta_{k[i]}$ is the kernel-based effect for group $k[i] \in \{1, \ldots, K\}$. For continuous outcomes, we set $\sigma^2 > 0$. For binary outcomes using a non-identity link (e.g., probit or logit), we set $\sigma^2 = 0$ to reflect the deterministic link from the latent scale to the probability scale.

We place the following priors on the latent components:

$$\mu \sim \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}, \Sigma_\mu), \qquad \eta \sim \mathcal{GP}(0, K_\eta),$$

where $\boldsymbol{\mu}_{\text{prior}} \in \mathbb{R}^H$ encodes known population-level means, and K_η is the Radial Basis Function (RBF) kernel defined over K structured groups. The squared exponential kernel is computed via:

$$[K_\eta]_{ij} = \alpha^2 \exp\left(-\frac{\|X_i - X_j\|^2}{2\rho^2}\right),$$

where $K_\eta : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$ given that X_i and X_j are P-dimensional covariate vectors for groups i and j, α is the marginal standard deviation, and ρ is the length-scale parameter.

3 Gaussian Processes

3.1 Posterior prediction under gaussian process prior

Notation. For convenience, we denote $g(\mathbb{E}[Y \mid X])$ as $g(y)$ throughout the remainder of this paper.

Theorem 1 (Posterior Prediction in Gaussian Process with Monotonic Link)**.** *Assume a monotonic link function g such that:*

$$g(\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{prior,h}, \boldsymbol{\Sigma}_\mu[h, h'] + \mathbf{K}_\eta[k, k'] + \sigma^2 \mathbf{I}_n),$$

then the posterior prediction of f_* given $g(\mathbf{y})$ is:

$$f_* \mid g(\mathbf{y}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{f_*} + \boldsymbol{\Sigma}_{f_* y}\boldsymbol{\Sigma}_{yy}^{-1}(g(\mathbf{y}) - \boldsymbol{\mu}_y),\ \boldsymbol{\Sigma}_{f_* f_*} - \boldsymbol{\Sigma}_{f_* y}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{y f_*}\right),$$

which we denote as:

$$f_* \mid g(\mathbf{y}) \sim \mathcal{N}(\mathbb{E}[f_* \mid g(\mathbf{y})],\ \mathrm{Cov}(f_* \mid g(\mathbf{y}))),$$

and hence:

$$g(\mathbf{y}_*) \mid g(\mathbf{y}) \sim \mathcal{N}\left(\mathbb{E}[f_* \mid g(\mathbf{y})],\ \mathrm{Cov}(f_* \mid g(\mathbf{y})) + \sigma^2 \mathbf{I}_{n_*}\right).$$

Proof: *The joint distribution of* $\begin{bmatrix} g(\mathbf{y}) \\ f_* \end{bmatrix}$ *is:*

$$\begin{bmatrix} g(\mathbf{y}) \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{prior,h} \\ \boldsymbol{\mu}_{prior,h_*} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\mu[h,h'] + \mathbf{K}_\eta[k,k'] + \sigma^2 \mathbf{I}_n & \boldsymbol{\Sigma}_\mu[h,h_*] + \mathbf{K}_\eta[k,k_*] \\ \boldsymbol{\Sigma}_\mu[h_*,h] + \mathbf{K}_\eta[k_*,k] & \boldsymbol{\Sigma}_\mu[h_*,h'_*] + \mathbf{K}_\eta[k_*,k'_*] \end{bmatrix}\right),$$

which we denote as:

$$\begin{bmatrix} g(\mathbf{y}) \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_{f_*} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yf_*} \\ \boldsymbol{\Sigma}_{f_*y} & \boldsymbol{\Sigma}_{f_*f_*} \end{bmatrix}\right).$$

Applying the standard formula for conditional distributions of multivariate Gaussians, we obtain:

$$f_* \mid g(\mathbf{y}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{f_*} + \boldsymbol{\Sigma}_{f_*y}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{g}(\mathbf{y}) - \boldsymbol{\mu}_y),\ \boldsymbol{\Sigma}_{f_*f_*} - \boldsymbol{\Sigma}_{f_*y}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yf_*}\right).$$

Finally, because $g(\mathbf{y}_*) = f_* + \boldsymbol{\epsilon}_*$ *with* $\boldsymbol{\epsilon}_* \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n_*})$*, we conclude:*

$$g(\mathbf{y}_*) \mid g(\mathbf{y}) \sim \mathcal{N}\left(\mathbb{E}[f_* \mid g(\mathbf{y})],\ \mathrm{Cov}(f_* \mid g(\mathbf{y})) + \sigma^2 \mathbf{I}_{n_*}\right).$$

3.2 Computational cost reduction

Theorem 2 (Computational Efficiency via Woodbury Identity)**.** *The computational cost of inverting the* $n \times n$ *covariance matrix* $\Sigma_{\mathbf{yy}}$ *is reduced from* $\mathcal{O}(n^3)$ *to* $\mathcal{O}(n^2(H+K))$*, where* H *and* K *are the dimensions of the group-level covariates and* $H, K \ll n$*.*
Proof: *Write the model as:*

$$f + \boldsymbol{\epsilon} = Z_\mu \boldsymbol{\mu}_h + Z_\eta \boldsymbol{\eta}_k + \boldsymbol{\epsilon}$$

where $Z_\mu \in \mathbb{R}^{n \times H}$*,* $Z_\eta \in \mathbb{R}^{n \times K}$*, and* $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$*. The covariance matrix is:*

$$V = Z_\mu \Sigma_\mu Z_\mu^\top + Z_\eta K_\eta Z_\eta^\top + \sigma^2 I_n$$

Using the Woodbury identity:

$$V^{-1} = \left(\sigma^2 I_n + UCU^\top\right)^{-1} = \sigma^{-2} I_n - \sigma^{-2} U \left(C^{-1} + \sigma^{-2} U^\top U\right)^{-1} U^\top \sigma^{-2}$$

with $U = [Z_\mu\ Z_\eta] \in \mathbb{R}^{n \times (H+K)}$*,* $C = \mathrm{diag}(\Sigma_\mu, K_\eta)$*.*

The major costs are:

- $U^\top U$*:* $\mathcal{O}(n^2(H+K))$
- $(C^{-1} + U^\top U)^{-1}$*:* $\mathcal{O}((H+K)^3)$

Full expression: $\mathcal{O}(2n^2(H+K) + n(H+K)^2 + (H+K)^3)$*, which simplifies to* $\mathcal{O}(n^2(H+K))$*.*

Algorithm for Computational Efficiency Using Cholesky Factorization

Step 1: Construct Covariance Matrices

Compute $\Sigma_{yy}, \Sigma_{yf_*}, \Sigma_{f_*y}, \Sigma_{f_*f_*}$ using the estimated covariance matrices.
[The computational cost reduces from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2(H+K))$ and $\mathcal{O}(n_*^3)$ to $\mathcal{O}(n_*^2(H_*+K_*))$.]
Continuous outcome: $\Sigma_{yy} = \Sigma_{\mu,\text{full}} + K\eta_{\text{full}} + \sigma^2 I_n$
Binary outcome: $\Sigma_{yy} = \Sigma_{\mu,\text{full}} + K\eta_{\text{full}} + \epsilon I_n$, where $\epsilon > 0$ ensures positive definiteness.

Step 2: Cholesky Decomposition & Posterior Mean

Compute $L := \text{Cholesky}(\Sigma_{yy})$ and $g(y)_c := g(y) - \mu_y = f - \mu_y$.
Compute $V_1 := \text{solve}(L^\top, \text{solve}(L, g(y)_c)) = (LL^\top)^{-1} g(y)_c$
Then, $\mathbb{E}[f_* \mid g(y)] = \mu_{f_*} + \Sigma_{f_*y} V_1$

Step 3: Posterior Covariance

Compute $V_2 := \text{solve}(L, \Sigma_{yf^*})$
Then, $\text{Cov}(f_* \mid g(y)) = \Sigma_{f_*f_*} - V_2^\top V_2$

Step 4: Posterior Sampling

Identity link:
$E[Y_* \mid X_*] = \mathbb{E}[f_* \mid y]$
$y_* \sim \mathcal{N}(\mathbb{E}[f_* \mid y], \text{Cov}(f_* \mid y) + \sigma^2 I_{n_*})$
Binary outcome:
$f_* \sim \mathcal{N}(\mathbb{E}[f_* \mid g(y)], \text{Cov}(f_* \mid g(y)))$
$E[Y_* \mid X_*] = g^{-1}(f_*)$
$y_* \sim \text{Bernoulli}(g^{-1}(f_*))$

3.3 Connection to kernel ridge regression and crossed random effects model

Proposition 1 (Equivalence of KRR and GP Posterior Mean under Identity Link)**.** *Consider our hierarchical Gaussian process model with an* ***identity link*** $g(x) = x$*, noise variance* σ^2*, and suppose the group-level prior covariance* $\Sigma_\mu \to 0$ *(so that* $\mu_h \equiv 0$*). Let*

$$\mathbf{K} \;=\; K_\eta[k, k'] \,,$$

the $n \times n$ *kernel matrix over the* n *observed groups, and let* $\mathbf{y} \in \mathbb{R}^n$ *be the observed outcomes. Then the GP posterior mean at a new input* x^* *is*

$$\mathbb{E}\big[f(x^*) \mid \mathbf{y}\big] \;=\; \mathbf{k}(x^*, X)^\top \big(\mathbf{K} + \sigma^2 I_n\big)^{-1} \mathbf{y},$$

which coincides exactly with the Kernel Ridge Regression solution

$$\hat{f}(x^*) \;=\; \sum_{i=1}^{n} \alpha_i \, k(x_i, x^*) \quad \textit{with} \quad \boldsymbol{\alpha} \;=\; \big(\mathbf{K} + \lambda I_n\big)^{-1} \mathbf{y} \quad \textit{and} \quad \lambda = \sigma^2.$$

Proof: Under $\Sigma_\mu \to 0$, our latent function reduces to $f = \eta \sim \mathcal{GP}(0, K_\eta)$. With identity link and Gaussian noise,

$$\mathbf{y} \sim \mathcal{N}\big(0,\; \mathbf{K} + \sigma^2 I_n\big).$$

The standard GP posterior mean at x^* is

$$\mathbb{E}[f(x^*) \mid \mathbf{y}] = \mathbf{k}(x^*, X)^\top \big(\mathbf{K} + \sigma^2 I_n\big)^{-1} \mathbf{y} \,,$$

where $\mathbf{k}(x^*, X)_i = k_\eta(x^*, x_i)$. Let $\mathcal{H}_k$ be a Reproducing Kernel Hilbert Space (RKHS) associated with a positive-definite kernel function $K(\cdot, \cdot)$. Then, KRR solves

$$\hat{f} = \arg\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_k}$$

and admits the dual form

$$\hat{f}(x^*) = \mathbf{k}(x^*, X)^\top \big(\mathbf{K} + \lambda I_n\big)^{-1} \mathbf{y}\,.$$

Thus, setting $\lambda = \sigma^2$ makes the GP posterior mean and the KRR prediction identical.

Proposition 2. *Without loss of generality, let the number of covariates p be 3. Let $X \in \mathbb{R}^{K \times p}$ be the matrix of covariates for the K hierarchical groups, and set*

$$K_\eta \;=\; \tau^2\, X X^\top.$$

Assume the identity link $g(x) = x$, observation noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and a group-intercept prior $\mu \sim \mathcal{N}(\mu_{\text{prior}}, \Sigma_\mu)$. Let

$$b_0 \sim \mathcal{N}(\mu_{\text{prior}}, \Sigma_\mu) \qquad b_1 \sim \mathcal{N}(0,\, \tau^2 I_p)$$

Then the hierarchical GP model

$$y = \mu_h + \eta_k + \varepsilon, \quad \eta \sim \mathcal{N}\big(0,\, K_\eta\big),$$

is equivalent to the crossed random effects model with a random intercept $b_{0,ij}$ *and a* random-slope $b_{1,ijk}$*:*

$$y_{ijk} = b_{0,ij} + X_{ijk}\, b_{1,ijk} + \varepsilon_{ijk}, \textit{ where } i, j, k \textit{ represent the levels of three covariates.}$$

Proof: By construction,

$$\eta \sim \mathcal{N}\big(0,\, K_\eta\big) = \mathcal{N}\big(0,\, \tau^2\, X\, X^\top\big).$$

Introducing $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau^2 I_P)$ and setting $\eta = X\,\mathbf{u}$ yields

$$\mathrm{Cov}(\eta) = X\,\mathrm{Cov}(\mathbf{u})\,X^\top = \tau^2\, X\, X^\top = K_\eta, \quad \mathbb{E}[\eta] = \mathbf{0}.$$

Matching i, j to a group H and i, j, k to a group K gives

$$y_i = \mu_i + \eta_i + \varepsilon_i = b_{0,ij} + X_{ijk}\, b_{1,ijk} + \varepsilon_{ijk},$$

with the specified priors on b_0 and b_1.

3.4 Uniform Convergence of the Posterior Predictive Mean

Theorem 3. *Consider the HGP regression model with identity link:*

$$y_i = f(x_i) + \varepsilon_i, \quad f_i = \mu_{h[i]} + \eta_{k[i]}, \quad \begin{cases} \mu \sim \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}, \Sigma_\mu), \\ \eta \sim \mathcal{GP}(0, K_\eta), \end{cases} \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $X \in \mathbb{R}^{n \times p}$ and K_η is the squared-exponential kernel over the K unique rows of X. Let

$$\hat{f}_n(x) = \mathbb{E}[\, \mu_{h(x)} + \eta_{k(x)} \mid y_{1:n}, X\,]$$

be the posterior predictive mean, and $f_0(x)$ the true regression function of the same form.

Assume:

1. *f_0 lies in the RKHS of K_η with smoothness $\alpha > 0$.*

2. *The design rows $\{X_{i,:}\}_{i=1}^n$ densely cover the compact covariate domain $\mathcal{X}$.*

3. *The prior covariance Σ_μ has eigenvalues bounded away from zero.*

Then there exists a rate

$$\varepsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^t$$

such that

$$\Pr\Big(\sup_{x\in\mathcal{X}}\big|\hat{f}_n(x) - f_0(x)\big| > M\,\varepsilon_n \;\Big|\; y_{1:n}, X\Big) \xrightarrow{a.s.} 0$$

for every large $M > 0$.

Proof: We combine three key ingredients, adapted to the additive structure $f = \mu + \eta$:

1. Posterior Contraction for (μ, η). Show that the joint posterior concentrates in an ℓ^∞-ball around (μ_0, η_0), at rate ε_n, using standard RKHS-based concentration for η and Gaussian-prior concentration for μ.

2. Finite Dimensional Bernstein–von Mises. We expand the process η using its first J_n eigenfunctions $\phi_j(x)$ and corresponding coefficients θ_j. We also include μ as H additional finite parameters. The joint posterior distribution of μ and the coefficients $\theta_{1:J_n}$ is then approximated by a multivariate Bernstein–von Mises (BvM) theorem. This approximation holds for the parameter vector $(\mu, \theta_{1:J_n})$ with dimension $H + J_n = o(n)$, yielding the following normal approximation:

$$(\mu, \theta_{1:J_n}) \mid y \;\approx\; \mathcal{N}(\widehat{(\mu, \theta)}, n^{-1}I),$$

where $\widehat{(\mu, \theta)}$ represents the posterior mean of the parameters, and I is the identity matrix. This approximation is valid as $n \to \infty$, ensuring that the posterior for these parameters is concentrated around the true values with covariance scaling as n^{-1}.

3. Sup-Norm Control. Write

$$\hat{f}_n(x) - f_0(x) = \underbrace{\sum_{h=1}^{H}(\hat{\mu}_h - \mu_{0,h})\mathbf{1}\{h(x) = h\}}_{O_P(n^{-1/2}H^{1/2})} + \underbrace{\sum_{j=1}^{J_n}(\hat{\theta}_j - \theta_{0j})\phi_j(x)}_{O_P(n^{-1/2}J_n^{1/2})} + \underbrace{\sum_{j>J_n}\theta_{0j}\phi_j(x)}_{O(J_n^{-\alpha})}.$$

Choosing $J_n \asymp n^{1/(2\alpha+1)}$ balances these terms so that $\sup_x|\hat{f}_n - f_0| = O_P(\varepsilon_n)$. A subsequence argument then upgrades to almost-sure convergence.

4 Connection to Survey Weighting

We now present a theorem that connects the regression function $g^{-1}(f)$ to population-calibrated survey weights.

Theorem 4 (Equivalent unit weights). Let w be the equivalent unit weight such that $\sum_i w_i = n$, where w depends only on the weighting variables (independent of the outcome y). Then

$$\boxed{w \;=\; \frac{n}{N}\,(N_k^{\text{pop}})^\top\,\Sigma_{f_* y}\,\Sigma_{yy}^{-1}\,\big(g^{-1}\big)'(f_*)}$$

where N_k^{pop} is the vector of population cell counts, and we use $\mathbb{E}[f_* \mid g(y)]$ for f_*.

Proof. Start by decomposing the weighted mean:

$$\frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{1}{N}\sum_i w_i(y_i - \hat{y}) + \frac{1}{N}\sum_i w_i \hat{y}. \qquad (*)$$

In population mean estimation the left-hand side equals $\frac{\sum_k N_k^{\text{pop}}\,\bar{y}_k}{\sum_k N_k^{\text{pop}}}$. Let $\bar{y}_k = E[Y \mid X]$ and then replace it with $E[Y_* \mid X_*] = g^{-1}(f_*)$ to use Theorem 3. Let $n = \sum_i w_i$ and $N = \sum_k N_k^{\text{pop}}$. For large n, by by Theorem 3, we have

$$\frac{1}{n}\sum_i w_i \hat{y} = \frac{\sum_k N_k^{\text{pop}}\, g^{-1}(f_*)}{\sum_k N_k^{\text{pop}}}$$

Differentiate w.r.t. $\hat{y}$ and solve for w to obtain the stated result.

Special cases.

- **Identity (linear):** $(g^{-1})'(f_*) = 1$, $w = \frac{n}{N}(N_k^{\text{pop}})^\top \Sigma_{f_* y}\Sigma_{yy}^{-1}$.
- **Logit:** $g^{-1}(\eta) = \text{expit}(\eta)$. Let $p^* = g^{-1}(f_*)$. Then,

$$w = \frac{n}{N}(N_k^{\text{pop}})^\top g^{-1}(f_*)(1 - g^{-1}(f_*))\Sigma_{f_* y}\Sigma_{yy}^{-1}\,\frac{1}{p^*(1-p^*)}.$$

- **Probit:** $g^{-1}(\eta) = \Phi(\eta)$,then

$$w = \frac{n}{N}(N_k^{\text{pop}})^\top \phi(f_*)\Sigma_{f_* y}\Sigma_{yy}^{-1}\,\frac{1}{\phi(\Phi^{-1}(f_*))}$$

5 Results

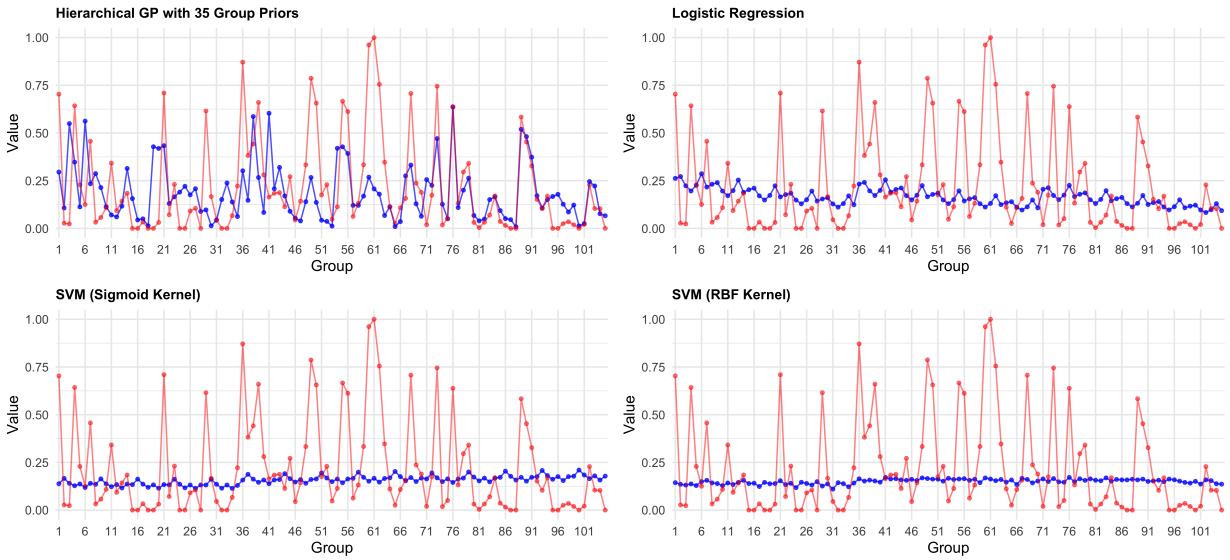
In Table 1, we compare the Mean Squared Error (MSE) of four predictive models—Hierarchical Gaussian Process (HGP), Logistic Regression, Support Vector Machine with a Sigmoid kernel (SVM Sigmoid), and Support Vector Machine with a Radial Basis Function kernel (SVM RBF)—under two prior configurations: one with 7 group priors and another with 35 priors out of the 105 group priors. The MSE values quantify the discrepancy between each model's predicted group rates and the actual population rates, with lower values indicating more accurate predictions.

Figures 1 and 2 visualize the predicted (blue) versus actual population (red) group rates under the 7-prior and 35-prior settings, respectively. These visualizations allow us to assess how closely each model's predictions track the true group-level patterns. The analysis is conducted under intentionally sparse sampling conditions: the sample includes only 300 units from a total population of 10,000, and 70 out of the 105 possible population-defined groups are observed in the sample. This setup evaluates model performance in small-sample, partially observed group settings. Due

to computational constraints, we chose to work with 105 total groups rather than the hundreds or thousands that may be encountered in large-scale population modeling. This choice balances realism with tractability while still providing a meaningful test of each model's predictive accuracy.

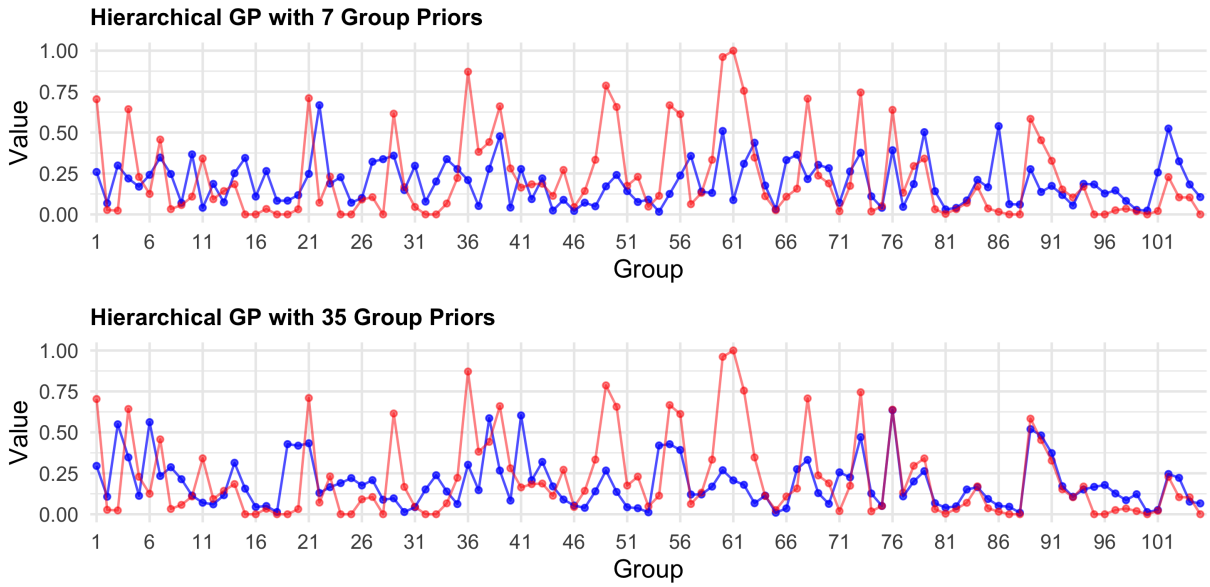
Priors	**HGP**	**Logistic**	**SVM Sigmoid**	**SVM RBF**
7 priors	0.062	0.066	0.068	0.065
35 priors	0.053	0.066	0.068	0.065

Table 1: MSE for Different Methods



Blue = Predicted Rates, Red = Population Rates.

Figure 1: Predicted Group Rates vs. Population Rates for 35 out of 105 priors.



Blue = Predicted Rates, Red = Population Rates.

Figure 2: HGP 7 Group Priors vs. 35 Group Priors.

We now present the results of the survey weights after applying Theorem 4. Table 2 compares the original population counts across cities with the corresponding weighted estimates derived from the adjusted survey weights. Table 3 summarizes the effective sample size after weighting, along with descriptive statistics of the normalized weights, highlighting their distribution and variability.

Table 2: Population Counts and Weighted Estimates by City

City	Pop. Count	Weighted Est.
1	1108	922.24
2	1221	985.26
3	1619	1959.19
4	1678	2540.39
5	1611	1537.32
6	1342	918.94
7	1421	1136.67

Table 3: Summary of Normalized Weights and Effective Sample Size

Statistic	Value
ESS	165 out of 300
Minimum	0.01125
1st Quartile	17.21502
Median	29.09463
Mean	34.36426
3rd Quartile	41.72785
Maximum	217.29819

6 Discussion

The data-generating process incorporated demographic covariates—education (3 levels), age group (5 levels), and city of residence (7 levels)—sampled to reflect empirical marginal distributions. Since we are dealing with weighting variables, all predictors are categorical. Ordinal variables, such as education, were treated numerically, while nominal variables, like city of residence, were encoded using one-hot encoding to appropriately capture their categorical nature. We simulated a binary outcome using both structured main effects and a smooth, nonparametric latent component to cap-

ture complex group heterogeneity. Structured effects were generated by applying linear coefficients to these covariates, while latent variation was introduced through a smooth, nonparametric function defined over 105 group cells formed by the interaction of city, education, and age. This latent function was drawn from a Gaussian process with a radial basis function (RBF) kernel, enabling flexible modeling of unobserved structure across cells. The final outcome probability was computed by applying a sigmoid transformation to the sum of the structured and latent components, resulting in probabilities that reflect both observed covariates and underlying heterogeneity.

Figure 1 and Table 1 highlight the clear advantage of the Hierarchical Gaussian Process (HGP) model in capturing latent structure across demographic subgroups, each defined by a unique combination of education, age group, and city of residence (yielding 105 total groups). Unlike logistic regression and support vector machines (SVMs), which only account for structured covariate effects and do not incorporate priors, the HGP flexibly adapts to variation across subgroups by leveraging prior information. This flexibility becomes increasingly effective as the amount of prior information grows: the HGP with 35 priors achieves the lowest mean squared error (MSE = 0.053), outperforming all other models, whose MSEs remain comparatively high and flat (0.065–0.068) regardless of the number of priors. In Figure 1, the HGP demonstrates a stronger capacity to track the wide range of subgroup-level empirical outcome rates, particularly in areas where unobserved heterogeneity plays a dominant role. In contrast, the alternative models struggle to account for this broad distribution of subgroup signals, resulting in underfitting in regions with complex structure. These findings underscore the value of incorporating latent interaction patterns and prior structure, especially when observable covariates alone are insufficient to explain outcome variation. By modeling at the group level, our approach also reduces the dimensionality of the latent function and enables effective sharing of information across similar subgroups.

The target survey population consists of live births occurring in in large U.S. cities with population over 200,000 between 1998 and 2000. This focus is motivated by the process of constructing base weights of the Future of Families and Child Wellbeing Study (FFCWS). The FFCWS sample is a stratified and multistage design with 4,898 children, oversampling births to unmarried mothers at a ratio of 3 to 1 with the inclusion of a large number of Black, Hispanic, and low-income families. Follow-up interviews were conducted across seven waves, when children were approximately ages 1, 3, 5, 9, 15, and 22. In constructing the FFCWS weights, four demographic variables were used for poststratification, and geographical information was used to estimate population birth counts using the Centers for Disease Control and Prevention (CDC) annual natality data. These variables are mother's marital status, race/ethnicity, age, education, and city of birth, resulting in 4,480 group cells. The size and complexity of the FFCWS sample and its cell structure impose substantial computational and storage demands. As such, we plan to apply our method to this target dataset in the near future.