Small-area Public Opinion Estimation Using Gaussian Process Regression and Post-stratification

Bryant Moy, Jacob Montgomery, Noah Dasanaike, Santiago Olivella

This draft: April 18, 2022

1 The models

1.1 A grouped binomial model for a single item

Consider a set of n_i respondents to a survey item (e.g. support for a presidential candidate) who share a profile $i \in 1, \ldots, N$ defined by observable demographic characteristics (e.g. age, education, income, etc.).

1.2 A grouped IRT model for multiple items

Now define a group of n_{ij} respondents who again share the same demographic profile $i \in 1, \ldots, N$, and who answer the same survey item $j \in 1, \ldots, J$.[1] Let $0 \leq y_{ij} \leq n_{ij}$ count the number of respondents who answer an item in the affirmative. When considering a battery of items believed to tap into a common unobserved construct of interest (e.g. policy liberalism), we can use observed responses to learn where respondents lie on a space defined by values of the construct. As usual in standard IRT models, we thus define the probability of observing y_{ij} as a function of a latent policy position θ_i and item-specific discrimination and difficulty parameters. More specifically, we define

$$\mu_{ij} = \theta_i \beta_j - \alpha_j$$

and model y_{ij} as a conditionally independent draw from a binomial distribution with probability parameter $\pi_{ij} = \text{logit}^{-1}(\mu_{ij})$, and fixed number of trials

[1] The model allows for differing numbers of respondents for any given item, thus accommodating MAR response patterns conditional on demographic predictors.

n_{ij}. Accordingly, the joint likelihood of aggregated responses $\mathbf{Y}$ given by

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{\beta}}) &\propto \prod_{ij} \frac{\exp(\mu_{ij})^{y_{ij}}}{[1+\exp(\mu_{ij})]^{n_{ij}}} \\ &= \prod_{ij} \exp \kappa_{ij}\mu_{ij}\, \mathbb{E}_{\omega_{ij}}[\exp(-\omega_{ij}\mu_{ij}^2/2)] \end{aligned} \tag{1}$$

where $\kappa_{ij} = y_{ij} - n_{ij}/2$, ω_{ij} is a Pólya-Gamma (PG) random variable distributed PG$(n_{ij}, 0)$, and the equality obtains from the integral identity derived by [6]. In effect, Equation 1 corresponds to the *observed data likelihood*, having marginalized a set of auxiliary variables ω_{ij}. This data augmentation strategy, which incorporates (and then marginalizes over) the ω_{ij}'s, will become useful when we define a computationally efficient EM algorithm for obtaining estimates of ideal points and item parameters.

Our approach's main innovation is to model the expected value of each ideal point θ_i as a flexible function of the demographic characteristics that define demographic profile i. To do so, we let

$$\boldsymbol{\theta} \mid \mathbf{f} \sim N_n(\mathbf{f}, \boldsymbol{\Sigma}_\theta) \tag{2}$$

(where $\boldsymbol{\Sigma}_\theta = \sigma_\theta^2 \mathbf{I}$) and give the vector of ideal point means, $\mathbf{f} = \{f_i\}_{i=1}^N$, a Gaussian Process (GP) prior distribution,

$$\mathbf{f} \mid \mathbf{Z} \sim \text{GP}(\mathbf{0}, \mathbf{K}_\rho) \tag{3}$$

where $\mathbf{K}_\rho = K(\mathbf{Z} \mid \boldsymbol{\rho})$ is an $N \times N$ covariance matrix generated using a kernel computed on a $\mathbf{Z}_{N\times D}$ matrix of demographic predictive features. Effectively, this defines a standard Gaussian Process regression model for the ideal points, as a function of demographic characteristics $\mathbf{Z}$.

Letting ρ_d be the scale-length hyper-parameter the dth feature, we complete the model by defining $K(\cdot)$ as the squared-exponential kernel function,

$$K(\mathbf{z}, \mathbf{z}' \mid \boldsymbol{\rho}) = \exp\left[-\frac{1}{2}\sum_{d=1}^{D} \frac{(z_d - z_d')^2}{\rho_d^2}\right] \tag{4}$$

and by giving (hyper)parameters appropriate (hyper)prior distributions. For each item j, we we thus independent define a bivariate Gaussian priors for the vector of item parameters $\widetilde{\boldsymbol{\beta}}_j = [\beta_j, \alpha_j]^\top$:

$$\widetilde{\boldsymbol{\beta}}_j \sim N_2(\mathbf{0}, \boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}}^{-1});$$

while defining independent hyperpriors for each kernel hyperparameter:

$$\rho_d \sim \text{inverse-gamma}(a, b)$$

for all features in $\mathbf{Z}$.

Accordingly, we can express the joint posterior distribution over ideal points $\boldsymbol{\theta}$, item parameters $\widetilde{\boldsymbol{\beta}}$, and kernel hyper-parameters $\boldsymbol{\rho}$ as

$$\begin{aligned}\pi \equiv p(\boldsymbol{\theta}, \mathbf{f}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\rho} \mid \mathbf{Y}, \mathbf{Z}) \propto & \prod_{ij} \exp(\kappa_{ij}\mu_{ij})\, \mathbb{E}_{\omega_{ij}}[\exp(-\omega_{ij}\mu_{ij}^2/2)] \\ & \times \exp\left[-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}\right] \\ & \times \det(\mathbf{K}_\rho)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{f}^\top \mathbf{K}_\rho^{-1}\mathbf{f}\right] \\ & \times \prod_j \exp\left[-\frac{1}{2}\widetilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\beta}}_j\right] \\ & \times \prod_d \rho_d^{-a-1} \exp\left[-\frac{b}{\rho_d}\right]\end{aligned} \tag{5}$$

Our goal is to learn the parameter and hyper-parameter values that maximize the posterior in Eq. 5, and ultimately obtain predicted ideal points $\theta^\star$ for a new set of demographic profiles $\mathbf{Z}^\star$ that will form the basis of our post-stratification exercise.

2 Fitting the models

2.1 Estimation algorithm

EM Algorithm

We define an expectation-conditional maximization algorithm (ECM) to find the maximum-a-posteriori estimates of ideal points and item parameters in Eq. 5, extending the maximization step to also obtain empirical Bayes estimates of the kernel hyper-parameters, after partially collapsing the target log posterior over the means of ideal points. To do so, we treat the augmented variables $\boldsymbol{\omega}$ as missing data, and proceed by taking the expectation of the complete-data log posterior distribution:

$$\begin{aligned}Q(\boldsymbol{\theta}, \mathbf{f}, \widetilde{\boldsymbol{\beta}}) &\equiv \mathbb{E}_{\boldsymbol{\omega}}\left[\log p(\boldsymbol{\theta}, \mathbf{f}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\omega} \mid \mathbf{Y}, \mathbf{Z}) \mid \boldsymbol{\theta}^{(t-1)}, \widetilde{\boldsymbol{\beta}}^{(t-1)}\right] \\ &= \sum_{ij} \kappa_{ij}\mu_{ij} - \mathbb{E}_{\boldsymbol{\omega}}[\omega_{ij} \mid \boldsymbol{\theta}^{(t-1)}, \widetilde{\boldsymbol{\beta}}^{(t-1)}]\mu_{ij}^2/2 \\ &\quad - \frac{1}{2}(\boldsymbol{\theta} - \mathbf{f})^\top \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \mathbf{f}) - \frac{1}{2}\left(\log\left[\det(\mathbf{K}_\rho)\right] + \mathbf{f}^\top \mathbf{K}_\rho^{-1}\mathbf{f}\right) \\ &\quad - \frac{1}{2}\sum_j \widetilde{\boldsymbol{\beta}}_j^\top \boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\beta}}_j + \text{const.}\end{aligned} \tag{6}$$

At each step, the procedure iterates between computing the expectation in Eq. 6 (the E-step), and maximizing this expression (also known as the "Q function") with respect to $\boldsymbol{\theta}$, $\mathbf{f}$, and $\widetilde{\boldsymbol{\beta}}$ — as well as with respect to hyper-parameters $\boldsymbol{\rho}$ — using a coordinate ascent algorithm (the conditional M-steps).

In the E-step, we evaluate the expectation in Eq. 6. Conditional on ideal points and item parameters, ω_{ij} can be shown to have a Pólya-Gamma distribution $\text{PG}(n_{ij}, \mu_{ij})$ [6]. Accordingly, the expectation in Eq. 6 has a simple closed-form solution given by

$$\omega_{ij}^{(t)} \equiv \mathbb{E}_{\boldsymbol{\omega}} \left[\omega_{ij} \mid \boldsymbol{\theta}^{(t-1)}, \widetilde{\boldsymbol{\beta}}^{(t-1)} \right] = \frac{n_{ij}}{2\mu_{ij}^{(t-1)}} \tanh\left(\mu_{ij}^{(t-1)}/2 \right) \tag{7}$$

where $\mu_{ij}^{(t-1)}$ is computed with the most current values of all item parameters and ideal points.

In turn, and at the t^{th} iteration, we find the optimal item parameters, ideal points, means of those ideal points, and kernel hyper-parameter values by maximizing Eq. 6 using a (partially collapsed) coordinate ascent approach. As shown in [6] (and illustrated in, for example, [2]), augmenting our data with auxiliary variables $\boldsymbol{\omega}$ allows us to obtain tractable solutions to the conditional maximization steps that form this coordinate ascent process.

First, the update for the item parameters $\widetilde{\beta}_j$ (conditional on data and the most recent values of all ideal points and their means) is given by

$$\widetilde{\boldsymbol{\beta}}_j^{(t)} = \left(\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}} + \mathbf{X}^\top \boldsymbol{\Omega}_j \mathbf{X} \right)^{-1} \mathbf{X}^\top \boldsymbol{\kappa}_j \tag{8}$$

where $\boldsymbol{\Omega}_j = \text{diag}\left(\{\omega_{ij}^{(t)}\}_{i=1}^N \right)$, matrix $\mathbf{X}$ has rows $\mathbf{x}_i = [\theta_i^{(t-1)}, -1]$, and $\boldsymbol{\kappa}_j = [\kappa_{1j}, \ldots, \kappa_{Nj}]^\top$. This can be recognized as the mode of the Gaussian conditional posterior distribution over item parameter vector $\widetilde{\boldsymbol{\beta}}$.

Similarly, the update for each ideal point is the mode of the conditional posterior over θ_i, and is given by

$$\theta_i^{(t)} = \left(\sigma_\theta^{-2} + {\boldsymbol{\beta}^{(t)}}^\top \boldsymbol{\beta}^{(t)} \right)^{-1} \left(f_i^{(t-1)}/\sigma_\theta^2 + {\boldsymbol{\beta}^{(t)}}^\top \widetilde{\mathbf{y}}_i \right) \tag{9}$$

where $\widetilde{\mathbf{y}}_i = [\{\kappa_{ij}/\omega_{ij}^{(t)} + \alpha_j^{(t)}\}_{j=1}^J]^\top$ is a vector of transformed outcomes, and κ_{ij} is defined as before.

Next, we can derive the update for the mean of the ideal point distribution, in closed-form, as the mode of the conditional posterior distribution over $\mathbf{f}$:

$$\mathbf{f}^{(t)} = \mathbf{K}_\rho (\mathbf{K}_\rho + \boldsymbol{\Sigma}_\theta^{-1})^{-1} \boldsymbol{\theta}^{(t)} \tag{10}$$

which is the standard predictive equation of a GP regression model.

The M-step is finalized by finding optimal values of $\boldsymbol{\rho}$ — the kernel hyper-parameters. We do so by maximizing the *marginal* expected log-likelihood — that is, Eq. 6 integrated over the mean vector $\mathbf{f}$ — with respect to the kernel

hyper-parameters. Restricted to terms involving $\mathbf{f}$, and conditional on $\boldsymbol{\theta}$, Eq. 6 can be expressed as the sum of two log-Gaussian densities — one for the set of ideal points and another for the GP-distributed means of these ideal points. This allows us to obtain a closed-form expression for the *marginal* Q function, integrating out ideal point means:

$$\begin{aligned}\bar{Q}_\rho(\boldsymbol{\theta}) &= \int Q(\boldsymbol{\theta}, \mathbf{f}, \widetilde{\boldsymbol{\beta}})\, d\mathbf{f} \\ &= -\frac{1}{2}\Bigg(\boldsymbol{\theta}^\top (\mathbf{K}_\rho + \boldsymbol{\Sigma}_{\boldsymbol{\theta}})^{-1}\boldsymbol{\theta} + \log[\det(\mathbf{K}_\rho + \boldsymbol{\Sigma}_{\boldsymbol{\theta}})] \\ &\qquad + n\log[2\pi] + \text{const.}\Bigg)\end{aligned}$$

Accordingly, we can obtain

$$\boldsymbol{\rho}^{(t)} = \arg\max_{\rho} \bar{Q}_\rho(\boldsymbol{\theta}^{(t)}) \tag{11}$$

using a gradient-based optimizer.[2] Effectively, this step turns our ECM algorithm into an instance of ECME [3], which has faster convergence rates than both EM and ECM, and can help regularize the magnitude of the $\boldsymbol{\rho}$ hyper-parameters.[3] We set $\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}} = \text{diag}(0.1)$ and $\sigma^2_\theta = 1.0$ for identification purposes.

In sum, our algorithm iterates between the evaluation of Eq. 7, and the evaluation of Eq.'s 8 through 11.[4]

MCMC Sampler

Rather than finding a posterior mode, a Metropolis-within-Gibbs sampler allows us to obtain samples from the full posterior proportional to Equation 5. Although we can obtain full conditional posterior distributions for most of our model's parameters, there is no closed-form conditional posterior for the hyper-parameters $\boldsymbol{\rho}$. As a result, we embed a simple random-walk Metropolis step at each iteration of our Markov Chain. Accordingly, and at each iteration t of the sampler, we complete the following steps:

1. Sample $\omega_{ij}^{(t)} \sim \text{PG}(n_{ij}, \mu_{ij}^{(t-1)}) \quad \forall i, j$, where $\text{PG}(\cdot)$ is the Pólya-Gamma density function.[5]

[2] The required gradient is given by

$$\frac{\partial}{\partial \rho_d}\bar{Q}_\rho(\boldsymbol{\theta}) = \frac{1}{2}\text{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \mathbf{K}_\rho^{-1})\frac{\partial \mathbf{K}_\rho}{\partial \rho_d}\right)$$

where $\boldsymbol{\alpha} = \mathbf{K}_\rho^{-1}\boldsymbol{\theta}$ [7]. In our implementation, we use the conjugate gradients algorithm for finding the opimal value of $\boldsymbol{\rho}$.

[3] This also explains why the update for $\boldsymbol{\rho}$ *must* come last during the M-step, as other orders can break the monotone convergence property of the algorithm [4].

[4] We implement these steps, as well as those involved in the prediction and post-stratification stage, in a user-friendly, open-source R package, `GrP`.

[5] A sampler for the Pólya-Gamma is available in `R` through the `BayesLogit` package. The run-

2. Sample $\widetilde{\boldsymbol{\beta}}_j^{(t)} \sim \mathrm{N}(m_\beta, V_\beta)$, where $V_\beta = \left(\boldsymbol{\Lambda}_{\widetilde{\boldsymbol{\beta}}} + \mathbf{X}^\top \boldsymbol{\Omega}_j \mathbf{X}\right)^{-1}$, $m_\beta = V_\beta(\mathbf{X}^\top \boldsymbol{\kappa}_j)$, and $\mathbf{X}$, $\boldsymbol{\Omega}_j$, and $\boldsymbol{\kappa}_j$ are all defined as in Equation 8.

3. Sample $\theta_i^{(t)} \sim \mathrm{N}(m_\theta, V_\theta)$, where $V_\theta = \left(\sigma_\theta^{-2} + {\boldsymbol{\beta}^{(t)}}^\top \boldsymbol{\beta}^{(t)}\right)^{-1}$, $m_\theta = V_\theta \left(f_i^{(t-1)}/\sigma_\theta^2 + {\boldsymbol{\beta}^{(t)}}^\top \widetilde{\mathbf{y}}_i\right)$, and $\widetilde{\mathbf{y}}_i$ is defined as in Equation 9.

4. Sample $\mathbf{f}^{(t)} \sim \mathrm{N}(m_f, V_f)$, where $V_f = \mathbf{K}_\rho - \mathbf{K}_\rho(\mathbf{K}_\rho + \boldsymbol{\Sigma}_\theta^{-1})^{-1}\mathbf{K}_\rho$ and $m_f = \mathbf{K}_\rho(\mathbf{K}_\rho + \boldsymbol{\Sigma}_\theta^{-1})^{-1}\boldsymbol{\theta}^{(t)}$

5. Propose a vector of log-hyperparameters $l\boldsymbol{\rho}^\star \sim \mathrm{N}(\boldsymbol{\rho}^{(t-1)}), \boldsymbol{\Sigma}_\rho)$, where $\boldsymbol{\Sigma}_\rho$ is the covariance of the proposal (or jumping) density. Then, and given symmetry of the proposal distribution, compute the acceptance ratio as

$$r = \frac{\pi_\rho(\exp(l\boldsymbol{\rho}^\star))}{\pi_\rho(\boldsymbol{\rho}^{(t-1)})}$$

(where π_ρ is the posterior π restricted to terms that involve $\boldsymbol{\rho}$), and sample

$$\boldsymbol{\rho}^{(t)} \sim r^{1(\exp(l\boldsymbol{\rho}^\star))}(1-r)^{1(\boldsymbol{\rho}^{(t-1)})}$$

we iterate these steps until the Markov chain has converged to its stationary distribution, and then run it for M steps to obtain M samples from the posterior π in Equation 5.

2.2 Prediction and post-stratification

To predict average ideal points $\mathbf{f}^\star = [f_1^\star, \ldots, f_M^\star]^\top$ for a given set of M unique demographic profiles $\mathbf{Z}^\star$, we rely on the posterior predictive distribution that conditions on $\boldsymbol{\theta}$. A priori, $\boldsymbol{\theta}$ and $\mathbf{f}^\star$ are jointly multivariate Gaussian distributed:

$$\begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{f}^\star \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} K(\mathbf{Z}, \mathbf{Z} \mid \boldsymbol{\rho}) + \boldsymbol{\Sigma}_\theta & K(\mathbf{Z}, \mathbf{Z}^\star \mid \boldsymbol{\rho}) \\ K(\mathbf{Z}^\star, \mathbf{Z} \mid \boldsymbol{\rho}) & K(\mathbf{Z}^\star, \mathbf{Z}^\star \mid \boldsymbol{\rho}) \end{bmatrix} \right)$$

where $K(\cdot)$ is the exponential kernel function defined in Eq. 4, so that $K(\mathbf{Z}^\star, \mathbf{Z} \mid \boldsymbol{\rho})$ returns the covariance between prediction demographic profiles $\mathbf{Z}^\star$ and observed demographic profiles $\mathbf{Z}$.

Accordingly, the posterior distribution of $\mathbf{f}^\star \mid \boldsymbol{\theta}$ is also Gaussian, and can be easily derived from standard identities [e.g. 1]. Specifically, the mean of this posterior distribution is given by

$$\widehat{\mathbf{f}^\star} = K(\mathbf{Z}^\star, \mathbf{Z} \mid \boldsymbol{\rho})\left[K(\mathbf{Z}, \mathbf{Z} \mid \boldsymbol{\rho}) + \boldsymbol{\Sigma}_\theta\right]^{-1}\boldsymbol{\theta} \tag{12}$$

time for sampling is about linear in the number of counts $n_{i,j}$, which can result in excessive run times. Some performance gains can be made by parallelizing this sampling step across respondent profiles i and items j.

which is, in effect, a linear projection of ideal points associated with demographic profiles in the survey. We of course never truly observe these ideal points, so we replace $\boldsymbol{\theta}$ (and $\boldsymbol{\rho}$) with their maximum-a-posteriori estimates, obtained through the EM algorithm described in the previous section.

The final step of our proposed procedure involves aggregating each of the $\widehat{f}_m^\star$ into S discrete groupings (i.e. strata) that correspond to the target levels at which inferences are desired (e.g. cities in the U.S.), and adjusting these predictions so that they accurately reflect this target. To do so, we first obtain N_m — the total population with demographic profile $\mathbf{Z}_m^\star$ — from a high-quality source (e.g. the Census). The estimated ideal point at the target level is then given by

$$\hat{\theta}_s = \frac{\sum_{m \in s} N_m \widehat{f}_m^\star}{\sum_{m \in s} N_m}$$

as is standard in post-stratification applications [e.g. 5].

References

[1] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] Max Goplerud. A multinomial framework for ideal point estimation. *Political Analysis*, 27(1):69–89, 2019.

[3] Chuanhai Liu and Donald B. Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81(4):633–648, 1994.

[4] Xiao-Li Meng and David Van Dyk. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.

[5] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, pages 375–385, 2004.

[6] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[7] Christopher KI. Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.