UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2024, Vol. 17, No. 1, pp. 68–80

Bayesian Hierarchical Gaussian Process Mixtures for Regression Analysis

Mezclas de Procesos Gaussianos Jerárquicos Bayesianos para Análisis de Regresión

Juan Sosa[a]
jcsosam@unal.edu.co

Abstract

This paper essentially considers Gaussian process (GP) methods for regression analysis in longitudinal studies. The model is indeed a reasonable choice for description and prediction of phenomena involving repeated measurements in which there is evidence of heterogeneity among batches of measurements. First, we provide all the theoretical and practical details behind our modeling strategy. Then, we study the main properties of the model using simulated data. And finally, we apply analyze an AIDS clinical study developed by the AIDS Clinical Trials Group.

***Keywords*:** Bayesian inference; Gaussian processes; longitudinal data analysis; regression analysis.

Resumen

Este documento considera métodos de proceso gaussiano (GP) para el análisis de regresión en estudios longitudinales. El modelo es una elección razonable para la descripción y predicción de fenómenos que implican mediciones repetidas en las que hay evidencia de heterogeneidad entre lotes de mediciones. Primero, proporcionamos todos los detalles teóricos y prácticos detrás de nuestra estrategia de modelado. Luego, estudiamos las principales propiedades del modelo utilizando datos simulados. Y finalmente, aplicamos y analizamos un estudio clínico sobre el SIDA desarrollado por el Grupo de Ensayos Clínicos sobre el SIDA.

***Palabras clave*:** Inferencia Bayesiana; procesos gaussianos; análisis de datos longitudinales; análisis de regresión.

[a]Departamento de Estadística, Universidad Nacional de Colombia

1. Introduction

A hierarchical Gaussian process (GP) mixture model for regression analysis combines the flexibility and non-parametric nature of Gaussian processes with the ability to model complex hierarchical data structures through a mixture model approach. This framework is particularly useful for capturing heterogeneity and complex dependencies in data, making it a powerful tool for various applications such as longitudinal studies, time series analysis, and spatial data modeling. The following references will provide the reader with a deeper understanding of the theory, implementation, and applications of hierarchical GP mixture models in regression analysis: Neal (1997), Snelson and Ghahramani (2005), Williams and Rasmussen (2006), Titsias and Lawrence (2010), Shi and Choi (2011), and Hensman et al. (2013).

This paper is inspired in findings and methods shown in Shi et al. (2005) and essentially considers GP methods for regression analysis in longitudinal studies. Our purpose here is three-folded. First, we provide all the theoretical and practical details behind our modeling strategy (we also provide brand new R code regarding every empirical finding). Then, we study the main properties of the model using simulated data. And finally, we apply analyze an AIDS clinical study developed by the AIDS Clinical Trials Group (ACTG).

Longitudinal data analysis takes place when several experimental units are observed repeatedly over time measuring a response variable (dependent or output variable) in accordance with one or several covariates (independent or input variables), which may or may not be time-dependent. Such data can often be regarded as consisting of *batches of measurements* (set of values obtained as a result of experimental replication). An experimental unit may have one or several batches attached to it depending on how many times the experiment was carried out with such unit. The main purpose of the analysis is to identify, describe and predict the evolution (mean tendency) of the response variable and to determine how it is affected by the covariates.

Two major challenges arise when a GP regression model is applied to a large dataset with repeated measurements: (1) possible systematic heterogeneity among the different batches and (2) the requirement to invert a covariance matrix with dimension equal to the sample size of the training dataset. For dealing with the above two problems, Shi et al. (2005) propose a hierarchical GP mixture model for regression analysis, in which each batch of observations comes from a batch-specific GP.

This paper is structured as follows. In Section 2, we give an overview of the problem and the main ideas about hierarchical mixture models for regression. In Section 3, we recall some ideas about Bayesian inference in the context of mixtures. In Section 4, we describe the algorithm used in Section 5 and 6 to perform posterior inference using simulated and real data. Finally, in Section 7, we discuss our main findings.

2. Hierarchical mixture models for regression

In general, suppose that there are M different batches and N_m observations in the m-th batch. Observations are assumed to be independent among different batches. The response variable y_{mn} in the m-th batch is then modeled by

$$\begin{gathered} y_{mn} = f_m(\boldsymbol{x}_{mn}) + \epsilon_{mn}, \qquad \epsilon_{mn} \overset{\text{iid}}{\sim} \mathsf{N}(0,\sigma^2), \\ m = 1,\ldots,M, \qquad n = 1,\ldots,N_m, \end{gathered} \tag{1}$$

where $\boldsymbol{x}_{mn} = (\boldsymbol{x}_{mn1},\ldots,\boldsymbol{x}_{mnQ})$ is a Q-dimensional vector of covariates and $f_m(\cdot)$ is a nonlinear function of $\boldsymbol{x}_{mn}$. Shi et al. (2005) consider a model in which $f_m(\cdot)$ is assigned a finite mixture of GP prior, i.e.,

$$f_m(\boldsymbol{x}_{mn}) \sim \sum_{k=1}^{K} \pi_k \mathsf{GP}(\boldsymbol{\theta}_k)$$

where K is the given *fixed* number of components in the mixture, and π_k is the weight corresponding to the k-th component. This model can be regarded as a hierarchical model by independently introducing latent indicator variables z_m as follows:

$$\begin{gathered} f_m(\boldsymbol{x}_{mn}) \mid z_m = k \sim \mathsf{GP}(\boldsymbol{\theta}_k), \qquad m = 1,\ldots,M, \qquad n = 1,\ldots,N_m, \\ P(z_m = k) = \pi_k, \qquad k = 1,\ldots,K. \end{gathered}$$

Finally, it is assumed that given z_m, all the components in the mixture have the same structure but with different values of the parameter $\boldsymbol{\theta}_k$. Formally, for each m and each n, $f_m(\boldsymbol{x}_{mn}) \mid z_m = k$ has a multivariate normal distribution with zero mean and covariance function

$$C(\boldsymbol{x}_{mi}, \boldsymbol{x}_{mj} \mid \boldsymbol{\theta}_k) = v_k \, \exp\left\{ -\frac{1}{2} \sum_{q=1}^{Q} w_{kq} (x_{miq} - x_{mjq})^2 \right\} + a_{k0} + a_{k1} \sum_{q=1}^{Q} x_{miq} x_{mjq}, \tag{2}$$

where $\boldsymbol{\theta}_k = (w_{k1},\ldots,w_{kQ}, v_k, a_{k0}, a_{k2}, \sigma_k^2)$. Therefore, given z_m, the m-th output vector $\boldsymbol{y}_m = (y_{m1},\ldots,y_{mN_m})$, has a normal distribution with mean and covariance matrix given by

$$(f_m(\boldsymbol{x}_{m1}),\ldots,f_m(\boldsymbol{x}_{mN_m})) \qquad \text{and} \qquad \boldsymbol{\Psi}(\boldsymbol{\theta}_k) = \mathbf{C}(\boldsymbol{\theta}_k) + \sigma_k^2 \mathbf{I}_{N_m},$$

where $\mathbf{C}(\boldsymbol{\theta}_k)$ is an $N_m \times N_m$ matrix whose ij-th element is given by (2), and $\mathbf{I}_{N_m}$ is the $N_m \times N_m$ identity matrix.

3. Bayesian inference

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$ and $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$, and let $\mathcal{D} = \{(y_m, \boldsymbol{X}_m) : m = 1,\ldots,M\}$ be the collection of training data with $\boldsymbol{X}_m = (\boldsymbol{x}_{m1}^\mathsf{T},\ldots,\boldsymbol{x}_{mN_m}^\mathsf{T})^\mathsf{T}$.Then, the posterior distribution of the unknown parameters given the training data $\mathcal{D}$ is given by

$$p(\boldsymbol{\Theta}, \pi \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\Theta}, \pi)\, p(\boldsymbol{\Theta}, \pi)\,,$$

where

$$p(\mathcal{D} \mid \boldsymbol{\Theta}, \pi) = \prod_{m=1}^{M} \sum_{k=1}^{K} p(\boldsymbol{y}_m \mid \boldsymbol{\theta}_k, \boldsymbol{X}_m).$$

Also, it is assumed that $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ are independent a priori, and the $\boldsymbol{\theta}_k$ are independent and identically distributed, so that

$$p(\boldsymbol{\Theta}, \pi) = p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k).$$

The prior distributions for the rest of parameters are placed as in Rasmussen (1996). Thus, each w_{kq} is assigned an inverse gamma distribution:

$$w_{kq} \sim \mathsf{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2\mu}\right), \qquad k = 1, \ldots, K, \qquad q = 1, \ldots, Q.$$

Small values of α produce vague priors. The hyperparameter μ is assumed to take the value $\mu_0 Q^{2/\alpha}$, with $\alpha = 1$ and $\mu_0 = 1$. The priors on $\log(\sigma_k^2)$, a_0, and a_1 are taken as $N(0, 3^2)$ and the prior on $\log(v_0)$ is taken as $\mathsf{N}(0, 1)$, which correspond to fairly vague prior distributions. Finally, as in the general mixture model settings, $\boldsymbol{\pi}$ is assigned a Dirichlet distribution, i.e., $\boldsymbol{\pi} \sim \mathsf{Dir}(1, \ldots, 1)$.

4. Algorithm

Instead of generating a sample of $(\boldsymbol{\Theta}, \boldsymbol{\pi})$ from its posterior directly, the implementation is much simpler if the latent variables $\boldsymbol{z} = (z_1, \ldots, z_M)$ are simulated along with the unknown parameters by adopting a Hybrid Monte Carlo algorithm as in Duane et al. (1987). The algorithm consists of a Gibbs algorithm in Step (a) and a Hybrid Monte Carlo algorithm in Step (b). This procedure which referred to as Hybrid MCMC, given in the Appendix of Shi et al. (2005), is defined as follows:

Step (a) Sampling from $p(\boldsymbol{z} \mid \mathcal{D}, \boldsymbol{\Theta})$:

Let c_k be the number of observation for which $z_m = k$, over all $m = 1, \ldots, M$. One sweep of the procedure for sampling $\boldsymbol{z}$ and $\boldsymbol{\pi}$ is as follows:

1. Sample z_m from $p(z_m \mid \boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \propto \pi_k p(\boldsymbol{y}_m \mid \boldsymbol{\theta}_k)$;
2. Sample $(\pi_1, \ldots, \pi_k)$ from $p(\pi_1, \ldots, \pi_k) \propto D(1 + c_1, \ldots, 1 + c_K)$.

Step (b) Sampling from $p(\boldsymbol{\Theta} \mid \mathcal{D}, \boldsymbol{z})$:

Let $p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{z}) \propto \exp\{-\mathcal{E}\}$ where $\mathcal{E}$ is called potential energy. Since, the $\boldsymbol{\theta}$ are independent a priori, then the conditional density function of $\boldsymbol{\Theta}$ is

$$p(\boldsymbol{\Theta} \mid \mathcal{D}, \boldsymbol{z}) = \prod_{k=1}^{K} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{z}),$$

where

$$p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{z}) \propto p(\boldsymbol{\theta}_k) \prod_{\{m:z_m=k\}} p(\boldsymbol{y}_m \mid \boldsymbol{\theta}_k, \boldsymbol{X}_m). \tag{3}$$

Thus, $\boldsymbol{\theta}_k$ are conditionally independent given $\boldsymbol{z}$ and we can deal with each $\boldsymbol{\theta}_k$ separately.

One sweep of a variation of the Hybrid MC algorithm is as follows:

1. Starting from the current state $(\boldsymbol{\theta}, \boldsymbol{\phi})$, calculate the new state $(\boldsymbol{\theta}(\epsilon), \boldsymbol{\phi}(\epsilon))$ by the following *leapfrog steps* with step size ϵ:

$$\begin{aligned} \phi_i(\epsilon/2) &= \phi_i - \frac{\epsilon}{2}\frac{\partial \mathcal{E}}{\partial \theta_i}(\boldsymbol{\theta}), \\ \theta_i(\epsilon) &= \theta_i + \frac{\epsilon}{\lambda}\phi_i(\epsilon/2), \\ \phi_i(\epsilon) &= \phi_i(\epsilon/2) - \frac{\epsilon}{2}\frac{\partial \mathcal{E}}{\partial \theta_i}(\boldsymbol{\theta}(\epsilon)), \end{aligned}$$

   where $\partial\mathcal{E}/\partial\theta_i$ is the first derivative of $\mathcal{E}$ evaluated at $\boldsymbol{\theta}$.

2. The new state $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ is such that

$$(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) = \begin{cases} (\boldsymbol{\theta}(\epsilon), \boldsymbol{\phi}(\epsilon)), & \text{with prob. } \min\{1, p(\boldsymbol{\theta}, \boldsymbol{\phi})/p(\boldsymbol{\theta}(\epsilon), \boldsymbol{\phi}(\epsilon))\}; \\ (\boldsymbol{\theta}, -\boldsymbol{\phi}), & \text{otherwise,} \end{cases}$$

   where $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\{-(\mathcal{E} + \mathcal{K})\}$ with $\mathcal{K} = \frac{1}{2}\sum_k \phi_k/\lambda$.

3. Generate υ_i from the standard Gaussian distribution, and update ϕ_i to $\xi\phi_i^* + \sqrt{1-\xi^2}\upsilon_i$.

As suggested in Rasmussen (1996), we set $\epsilon = 0.5N_m^{-1/2}$, $\lambda = 1$, and $\xi = 0.95$.

The algorithm above does not meet the definition of a cycle, since it is a deterministic combination of Gibbs and Metropolis steps that would not themselves converge if applied individually. But each of the components does preserve the stationary distribution of the chain, so, provided the hybrid chain is aperiodic and irreducible, convergence can be obtained (Carlin and Louis, 2000, Sec. 5.4.4).

Shi et al. (2005) argue that in Step (b) the computational burden is much less than that incurred by modelling the data-set by a single GP regression model since the right-hand side of (3) requires the inversion of a covariance matrix of dimension N_m, which is generally much smaller than the total sample size of $N = N_1 + \ldots + N_M$.

5. Simulation

Shi et al. (2005) apply these ideas to study standing-up manoeuvres made by paraplegic patients. The authors consider different settings to implement the model and, in general, they get fairly good results in terms of characterization and prediction. Here, we emulate such methodology by considering the exactly the same data structure in a simulation study.

To perform Bayesian inference under this model, we generate data according to the model (1) with one independent variable ($Q = 1$), three batches ($M = 3$), and $\sigma = m^2/100$, $m = 1, 2, 3$. We consider different variances in order to produce heterogeneity among the batches. We consider input values drawn i.i.d. from a uniform distribution over the interval $(-3, 3)$. For each m, the true regression function is given by

$$f(x) = 0.3 + 0.4x + 0.5\sin(2.7x) + \frac{1.1}{1 + x^2}.$$

See the technical report by Radford Neal, which is available on-line from `http://www.cs.toronto.edu/~radford/mc-gp.abstract.html`.

We generate 200 observations per experimental unit according to model (1). To simulate unbalanced data sets in the batches, which is a main characteristic of the structure in this setting, repeated measures are randomly removed with a rate of $r_m = 0.2$. As in Shi et al. (2005), from the whole data-set, we randomly select about half of the data points from the batches as training data; the rest are used as test data. The sample sizes of the training data are 27, 25 and 29 respectively for the three batches. Figure 1 displays the simulated data. It seems that there is evidence of heterogeneity between different batches. This heterogeneity is not just random variability because we obtain better results by fitting a mixture rather than a single GP.

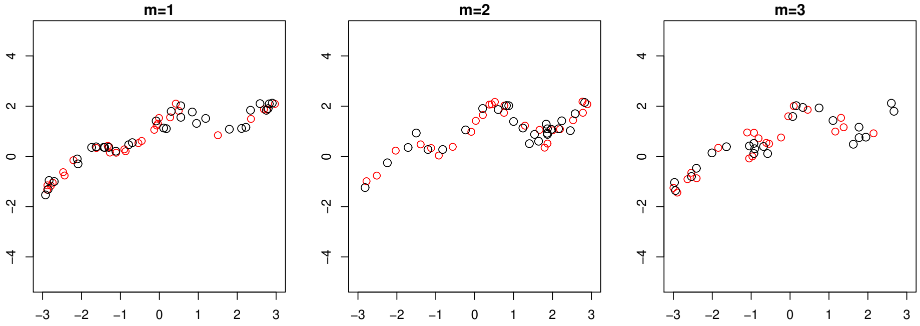

Figure 1: Simulated data according to model 1. Training data in black and test data in red.

We place the prior distributions and pick the hyperparameters as mentioned in section 3. We assume that the number of components in the mixture is $K = 2$ (results with $K = 3$ are very similar). We use the Hybrid MCMC algorithm with

$T = 1200$ iterations to generate samples from the posterior distribution. The chains corresponding to the parameters in the first component achieve convergence really quickly; on the other hand, the chains for the parameters in the second component achieve convergence after approximately iteration 100 approximately. As in Shi et al. (2005), we discard the first 800 iterations. In order to have approximately independent draws, we select one sample every other 5 iterations, and as consequence, a total of 81 samples (for each parameter in the respective component) are selected altogether. Those 81 samples are approximately independent and identically distributed according to the corresponding posterior distribution and form the basis of posterior inference.

Table 1 summarizes the posterior distribution of all the parameters in the mixture. We present the correspondent posterior means and 95 % posterior credible intervals. Note that the posterior estimate of the weights in the mixture π_1 and π_2 are 0.19 and 0.81, respectively, which strongly suggests that most of the inference work is based on the parameters of the second component. This fact also explains why the chains of the parameters in the second component converge faster than in the first one. Notice also that the posterior estimate of σ^2 (variance of the error term in the model) is 2.63 in the fist component and 0.50 in the second one, which clearly establishes some degree of heterogeneity captured by the model.

Par.	Mean	SD	Q2.5 %	Q50 %	Q97.5 %
w_1	5.45	0.78	3.97	5.34	7.09
v_1	5.10	0.92	3.84	4.89	6.91
$a_{0,1}$	1.46	1.04	0.34	1.13	3.49
$a_{1,1}$	-0.67	0.35	-1.19	-0.70	0.09
σ_1^2	2.63	1.45	0.60	2.27	5.11
w_2	0.83	0.04	0.77	0.86	0.89
v_2	1.35	0.02	1.32	1.35	1.38
$a_{0,2}$	2.62	0.04	2.56	2.62	2.67
$a_{1,2}$	2.85	0.04	2.81	2.82	2.91
σ_2^2	0.50	0.02	0.49	0.50	0.53
π_1	0.19	0.15	0.00	0.16	0.56
π_2	0.81	0.15	0.44	0.84	1.00

Table 1: Posterior summaries of all the parameters in the mixture.

To measure the performance of the model and the algorithm, the actual output values of the test data are compared with the predictions. Shi et al. (2005) compute the prediction for a new set of test inputs $\boldsymbol{x}^*$ in the m-th batch, as

$$\hat{y}_m^* = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_m^{*(t)},$$

where $\hat{y}_m^{*(t)} = \psi_m(\boldsymbol{x}^*)^\mathsf{T}\boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})^{-1}\boldsymbol{y}$ with $\psi_m(\boldsymbol{x}^*) = (C(\boldsymbol{x}^*, \boldsymbol{x}_{m1}), \ldots, C(\boldsymbol{x}^*, \boldsymbol{x}_{mN_m}))$. The variance associated with the prediction is calculated similarly, as

$$\hat{\sigma}_m^{*2} = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_m^{*2(t)} + \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_m^{*(t)})^2 - (\hat{y}_m^*)^2,$$

where $\hat{\sigma}_m^{*2(t)} = C(\boldsymbol{x}^*, \boldsymbol{x}^*) - \psi_m(\boldsymbol{x}^*)^\top \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})^{-1}\psi_m(\boldsymbol{x}^*)$. The predictive variance is then $\hat{\sigma}_m^{*2} + \hat{\sigma}^2$.

The results are plotted in Figure 2. We get that the root mean squared error between the prediction and the true test value is 0.26, and that the related correlation coefficient is 0.97. From those summary statistics and from this Figure, the fit seems to be very good.

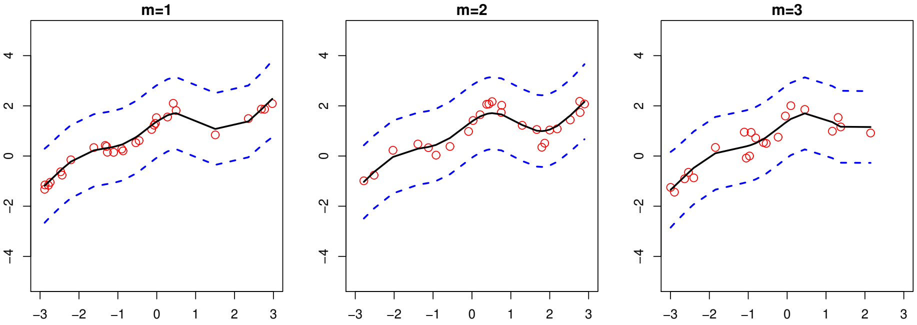

Figure 2: Test data (dots in red), and prediction of the test data (solid line in black) along with the corresponding 95 % confidence intervals (dashed line in blue).

6. Application

In this section a hierarchical GP model is applied to an AIDS clinical study developed by the *AIDS Clinical Trials Group*[1] (ACTG). With this group, Fischl et al. (2003) evaluated two different 4-drug regimens containing *indinavir* with either *efavirenz* or *nelfinavir* for the treatment of 517 patients with advanced HIV disease (i.e., patients with high HIV-1 RNA levels and low CD4 cell counts).

This study was a randomized, open-label study; it was initially planned to last 72 weeks but later increased to 120 weeks beyond the enrollment of the last subject. The randomization was carried out by using a permuted block design and was stratified according to CD4 cell count and HIV-1 RNA level at screening, as well as previous antiretroviral experience. In addition, clinical assessments, HIV-1 RNA measurements, CD4 cell counts, and routine laboratory tests were performed before study entry, at the time of study entry, at weeks 4 and 8, and every 8 weeks thereafter. More details about design, subjects, treatments and outcome measurements of this study are given in Fischl et al. (2003, p. 626-627).

The hierarchical GP model 1 was used to model the CD4 cell count, which is an essential marker for assessing immunologic response of an antiviral regimen, in arms 1 and 2 of the three treatment arms; notice that in this case treatment arms correspond to the batches in the model. Patients might not exactly follow the de-

[1] Visit the website `https://actgnetwork.org/` for more information about the group.

signed schedule, and missing clinical visits for CD4 cell measurements frequently occurred, which makes this data set[2] (named ACTG 388) a typical longitudinal data set. The main interest of the analysis presented in this section is to characterize the CD4 cell count trajectories over the treatment period by using the proposed estimation method project. Additional analyses of this and other trajectories, as well as more scientific findings of the study, can be found in Fischl et al. (2003, p. 627), Park and Wu (2005, p. 3774), and Wu and Zhang (2006, Chapters 5, 7, 8 and 10), Sosa and Diaz (2012).

Figure 3 shows the CD4 cell counts for some of the 166 patients during the 120 weeks of treatment. This plot indicates that the individual CD4 cell counts are quite noisy over time; it is not easy to see any pattern, and therefore it is not possible to establish if the antiviral treatment was effective (i.e., CD4 cell counts profiles should considerably increase).

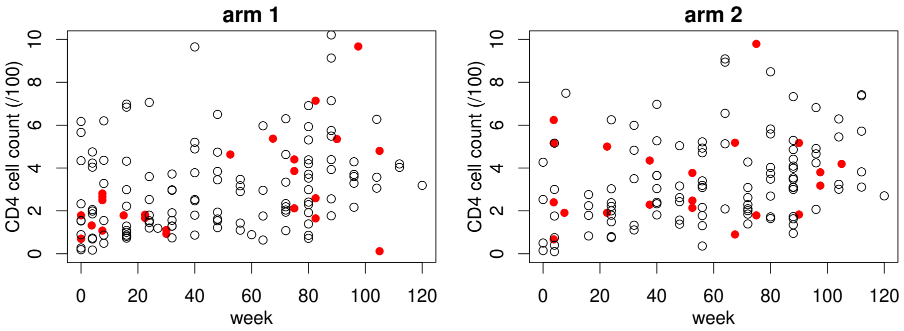

Figure 3: CD4 cell counts considering arms 1 and 2 of the study. Training data in black and test data in red.

In this application, we are just considering a random sample of 133 and 139 observations in treatment arms 1 and 2, respectively. Once again, from the whole data-set, we randomly select data points from the batches as training data; the rest are used as test data. The sample sizes of the training data are 108 and 117 respectively for the two batches.

We place the prior distributions and pick the hyperparameters as mentioned in section 3. We assume that the number of components in the mixture is $K = 2$ since we are dealing with two treatment arms, and use the Hybrid MCMC algorithm with $T = 1200$ iterations to generate samples from the posterior distribution. The chains corresponding to the parameters in the first component achieve convergence really quickly; on the other hand, the chains for the parameters in the second component achieve convergencee after approximately iteration 400 approximately. Again, we discard the first 800 iterations and in order to have approximately independent draws, we select one sample from every other 5 iterations, and as consequence, a total of 81 samples (for each parameter in the respective compo-

[2] Available in `http://www.urmc.rochester.edu/biostat/people/faculty/wusite/datasets/data/ACTG388Data1Arm.cfm`.

nent) are selected altogether. Those 81 samples are approximately independent and identically distributed according to the corresponding posterior distribution and form the basis of posterior inference.

Table 2 summarizes the posterior distribution of all the parameters in the mixture. We present the correspondent posterior means and 95 % posterior credible intervals. Note that the posterior estimate of the weights in the mixture π_1 and π_2 are 0.34 and 0.66, respectively. Notice also that the posterior estimate of σ^2 (variance of the error term in the model) is 1.87 in the fist component and 2.54 in the second one, which clearly establishes some degree of heterogeneity captured by the model.

Par.	Mean	SD	Q2.5 %	Q50 %	Q97.5 %
w_1	1.31	0.76	0.24	1.54	2.42
v_1	1.98	0.58	1.14	1.99	2.81
$a_{0,1}$	1.33	0.34	0.70	1.36	1.95
$a_{1,1}$	-0.85	0.88	-2.12	-0.87	0.55
σ_1^2	1.87	0.84	0.29	2.06	2.99
w_2	4.71	0.29	4.12	4.77	5.11
v_2	1.65	0.39	0.58	1.79	2.07
$a_{0,2}$	2.58	0.22	2.20	2.54	3.07
$a_{1,2}$	-0.37	0.26	-1.05	-0.34	0.20
σ_2^2	2.54	0.09	2.37	2.54	2.70
π_1	0.34	0.28	0.01	0.24	0.95
π_2	0.66	0.28	0.05	0.76	0.99

Table 2: Posterior summaries of all the parameters in the mixture.

To measure the performance of the model and the algorithm, the actual output values of the test data are compared with the predictions as before. Figure 4 displays the estimated curve. Note that the estimated population mean function is smooth. Here we can see that it increased sharply during the first 40 weeks, and continued to increase at a slower rate until about week 100. This shows that with this anti-viral treatment, the overall CD4 counts increased dramatically during the first 40 weeks, but the effect of the drug therapy faded over time.

7. Discussion

In what follows we point out the main characteristics and findings of the methodology described in this project:

- This model is indeed a reasonable choice (even better than network models according to Shi et al. 2005) for description and prediction of phenomena involving repeated measurements in which there is evidence of heterogeneity among batches.

- The number of mixture components K is assumed as fixed, and it is determined empirically in practice.

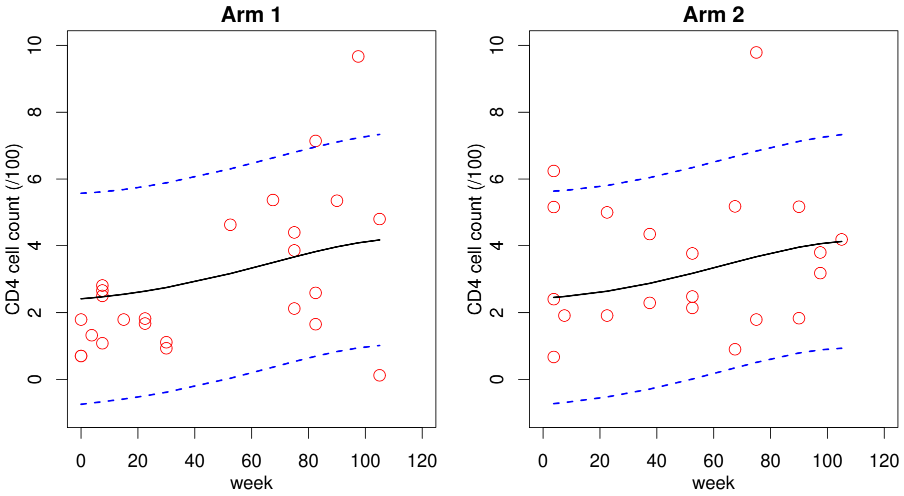

Figure 4: Test data (dots in red), and prediction of the test data (solid line in black) along with the corresponding 95 % confidence intervals (dashed line in blue).

- The version of the Hybrid MCMC algorithm is more efficient than the standard algorithm used in a regular GP for regression.
- The approach is robust in the sense that when different values of the hyperparameters are chosen, the final results are almost the same; the sample size is generally quite large, so the data dominate the prior.

Future work to improve this methodology should be done in many directions, but we consider that the most important is to tackle the problem of assessing the value of K and parameter estimation at the same time.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this art

Recibido: 20 de marzo de 2024
Aceptado: 20 de junio de 2024

Referencias

B. Carlin and T. Louis. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor &

Francis, 2000. ISBN 9781420057669. URL http://books.google.com/books?id=484r1P5o0hQC.

S. Duane, A. Kennedy, and D. Roweth. Hybrid monte carlo. *Physics Letters*, B (195):216–222, 1987.

M. A. Fischl, H. J. Ribaudo, A. C. Collier, A. Erice, M. Giuliano, M. Dehlinger, J. J. Eron, Jr., M. S. Saag, S. M. Hammer, S. Vella, G. D. Morse, and J. E. Feinberg. A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. *The Journal of Infectious Diseases*, 188:625–634, 2003.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

R. M. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.

J. G. Park and H. Wu. Backfitting and local likelihood lethods for nonparametric mixed-effects models with longitudinal data. *Journal of Statistical Planning and Inference*, 136(2006):3760–3782, 2005.

C. Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression*. PhD thesis, University of Toronto, 1996.

J. Q. Shi and T. Choi. *Gaussian process regression analysis for functional data*. CRC press, 2011.

J. Q. Shi, R. Murray-Smith, and D. Titterington. Hierarchical gaussian process mixtures for regression. *Statistics and Computing*, 15:31–41, 2005.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

J. Sosa and L. G. Diaz. Random time-varying coefficient model estimation through radial basis functions. *Revista Colombiana de Estadistica*, 35(1):167–184, 2012.

M. Titsias and N. D. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 844–851. JMLR Workshop and Conference Proceedings, 2010.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

H. Wu and J. T. Zhang. *Nonparametric Regression Methods for Longitudinal Data Analysis*. John Wiley and Sons, 2006.

A. Notation

The cardinality of a set A is denoted by $|A|$. If P is a logical proposition, then $\mathbf{1}\{\text{P}\} = 1$ if P is true, and $\mathbf{1}\{\text{P}\} = 0$ if P is false. $\lfloor x \rfloor$ denotes the floor of x, whereas $[n]$ denotes the set of all integers from 1 to n, i.e., $\{1, \dots, n\}$. The Gamma function is given by $\Gamma(x) = \int_0^\infty u^{x-1}\, e^{-u}\, \mathrm{d}u$. Matrices and vectors with entries consisting of subscripted variables are denoted by a boldfaced version of the letter for that variable. For example, $\boldsymbol{x} = (x_1, \dots, x_n)$ denotes an $n \times 1$ column vector with entries $x_1, \dots, x_n$. We use $\mathbf{0}$ and $\mathbf{1}$ to denote the column vector with all entries equal to 0 and 1, respectively, and $\mathbf{I}$ to denote the identity matrix. A subindex in this context refers to the corresponding dimension; for instance, $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The transpose of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\mathsf{T}$; analogously for matrices. Moreover, if $\mathbf{X}$ is a square matrix, we use $\operatorname{tr}(\mathbf{X})$ to denote its trace and $\mathbf{X}^{-1}$ to denote its inverse. The norm of $\boldsymbol{x}$, given by $\sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$, is denoted by $\|\,\boldsymbol{x}\|$.

Now, we present the form of some standard probability distributions used in this article:

- Multivariate normal:

  A $d \times 1$ random vector $\boldsymbol{X} = (X_1 \dots, X_d)$ has a multivariate Normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathsf{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density function is

  $$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}\, |\boldsymbol{\Sigma}|^{-1/2}\, \exp\left\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}.$$

- Inverse Gamma:

  A random variable X has an Inverse Gamma distribution with parameters $\alpha, \beta > 0$, denoted by $X \mid \alpha, \beta \sim \mathsf{IG}(\alpha, \beta)$, if its density function is

  $$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{-(\alpha+1)}\, \exp\{-\beta/x\}, \quad x > 0\,.$$

- Dirichlet:

  A $K \times 1$ random vector $\boldsymbol{X} = (X_1, \dots, X_K)$ has a dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$, where each $\alpha_k > 0$, denoted by $\boldsymbol{X} \mid \boldsymbol{\alpha} \sim \mathsf{Dir}(\boldsymbol{\alpha})$, if its density function is

  $$p(x \mid \boldsymbol{\alpha}) = \begin{cases} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1}, & \text{if } \sum_{k=1}^K x_k = 1; \\ 0, & \text{otherwise.} \end{cases}$$