Supplement for "Spatial Clustering for Carolina Breast Cancer Study"

Hongqian Niu[1], Melissa Troester[2], and Didong Li[1,†]

[1]Department of Biostatistics, [2]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [†]E-mail: didongli@unc.edu

A. Supplement Overview

In Section B we continue the discussion section of the main paper on the potential limitations of our algorithm and provide some general guidance on usage. In Section C we provide proofs for the two convergence theorems presented in the main paper. Section D contains the full implementation details, and additional figures for the three simulation studies presented in the main paper, as well as an extension to spatially-dependent functions. Section E contains additional details for the real-world application of the main paper as well as additional comparisons to competitor algorithms.

B. Potential Limitations

Here we continue our discussion from the paper on potential limitations of the GPSC model. The first is the question of tuning the appropriate number of clusters. This is a well-known challenge in clustering, which is beyond the scope of our study. However, some clustering algorithms have the capability to automatically determine the number of clusters. In this regard, we found that DBSCAN and GDBSCAN generally performed poorly in our simulation studies, resulting in incorrect and irrelevant clusters. As such, like in our real world application, selecting the appropriate number of clusters for the problem is best handled on a case-by-case basis with input from domain experts or prior background knowledge. The same can be said of the optional tuning parameter λ, which again reinforces contiguous spatial restraints by penalizing assignments to distant clusters. In our application, we were able to compare our results with previous studies on the socioeconomic and environmental cancer risk factors across the state, as well as collaborate with epidemiologists and cancer experts familiar with the datasets.

Next we briefly revisit the modeling assumptions of GPSC. Our main focus is spatial clustering, as motivated by the CBCS application, where different spatial clusters exhibit different functional relations between the response variable y and input features. In this case, according to our theorems presented in the main paper, the performance of GPSC is influenced by several key factors, including D_u, D_l, E_u, and E_l. In simpler terms, if the true underlying functions f_j in different clusters are not clearly distinguishable, or they have unbounded derivatives, it may be challenging to achieve optimal clustering results. However, this is a

common challenge for most clustering algorithms, and overcoming this limitation may require more advanced techniques and designs.

C. Proofs

C.1. *Proof of Theorem 3.1*

We first consider the case of $L = 2$ that is, there are two clusters. Let l_i be the unobserved true cluster label of the x_i and $\widehat{l}_i$ be the cluster label of x_i in the current iteration. Let x_0 be a sample to be clustered with (unobserved) label $l_0 = 1$, that is, x_0 should be assigned to cluster-1. Our goal is to show that GPSC does assign x_0 to cluster-1 under the condition explicitly stated in Theorem 3.1.

Let $(X_1, Y_1) \coloneqq \{(x_i, y_i) : \widehat{l}_i = 1\}$ be the set of samples assigned to cluster-1 with size $n_1 \coloneqq \#\{i : \widehat{l}_i = 1\}$. Similarly, let $(X_2, Y_2) \coloneqq \{(x_i, y_i) : \widehat{l}_i = 2\}$ be the set of samples assigned to cluster-2 with size $n_2 \coloneqq \#\{i : \widehat{l}_i = 2\}$. According to Algorithm 1, we train two GPR models based on (X_1, Y_1) and (X_2, Y_2), to obtain two predictors of y_0 denoted by $\widehat{y}^{(1)}$ and $\widehat{y}^{(2)}$. Under the notation in Definition 3, we have $\widehat{y}^{(1)} = \psi_{X,x_0}(Y_1)$ and $\widehat{y}^{(2)} = \psi_{X,x_0}(Y_2)$, it suffices to show that $e_1 \coloneqq |y_0 - \widehat{y}^{(1)}| < e_2 \coloneqq |y_0 - \widehat{y}^{(2)}|$ as long as

$$\frac{n_{21}}{n_{22}} < \frac{D_l E_l}{D_u E_u} - \frac{\|f\| e^{-c_1 n_1^{\frac{1}{p}}} + \|f\| e^{-c_2 n_2^{\frac{1}{p}}}}{D_u E_u}.$$

To calculate e_1, we introduce the following partially observed dummy variables $\widetilde{Y}_1 \coloneqq f_1(X_1)$ and let $\widetilde{y}^{(1)} \coloneqq \psi_{X_1,x_0}(\widetilde{Y}_1)$. We plug this term in e_1 and apply triangle inequality to obtain the following:

$$e_1 = |y_0 - \widehat{y}^{(1)}| = |y_0 - \widetilde{y}^{(1)} + \widetilde{y}^{(1)} - \widehat{y}^{(1)}| \le \underbrace{|y_0 - \widetilde{y}^{(1)}|}_{①} + \underbrace{|\widetilde{y}^{(1)} - \widehat{y}^{(1)}|}_{②}.$$

Observe that ① is the prediction error of standard Gaussian process regression on $(X_1, \widetilde{Y}_1)$, without any misspecified samples. As a result, the upper bound of ① comes from Lemma C.1, the asymptotic theory of Gaussian process regression. That is, $① \le \|f\| e^{-c_1/h_{n_1}}$ for some constant c_1. Assumption (A1) and Dudley's theorem imply that $h_{n_1} = O(n_1^{-\frac{1}{p}})$, so $① \le \|f\| e^{-c_1 n_1^{\frac{1}{p}}}$.

To analyze ②, we first observe that $\widehat{y}^{(1)} = \psi_{X_1,x_0}(Y_1)$ is based on partially correct clusters, while $\widetilde{y}^{(1)} = \psi_{X_1,x_0}(\widetilde{Y}_1)$ is based on true clusters. Then by the differentiability of ψ, we have

$$② = |\psi_{X_1,x_0}(Y_1) - \psi_{X_1,x_0}(\widetilde{Y}_1)| \le \|\nabla \psi_{X_1,x_0}\|_\infty \|Y_1 - \widetilde{Y}_1\| = D_u \|Y_1 - \widetilde{Y}_1\|,$$

where $D_u = \sup_{X_1 \subset X, x_0 \in X} \|\nabla \psi_{X_1,x_0}\|_\infty$. As a result, it suffices to find an upper bound of $\|Y_1 - \widetilde{Y}_1\|$.

Observe that among samples in (X_1, Y_1), some are correctly clustered, denoted by $(X_{11}, Y_{11}) = \{(x_i, y_i) : l_i = 1, \widehat{l}_i = 1\}$ with size n_{11}, while the rest are incorrectly clustered, denoted by $(X_{21}, Y_{21}) = \{(x_i, y_i) : l_i = 2, \widehat{l}_i = 1\}$ with size n_{21}. After reordering the samples, we have $X_1 = \begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix}$ and $Y_1 = \begin{bmatrix} Y_{11} \\ Y_{21} \end{bmatrix}$. By the model assumption, for the correctly clustered samples

$Y_{11} = f_1(X_{11})$, while for the incorrectly clustered samples, $Y_{21} = f_2(X_{21}) \neq f_1(X_{21})$. By the same rule, we can split $\widetilde{Y}_1$ into two components as well, i.e., $\widetilde{Y}_1 = \begin{bmatrix} \widetilde{Y}_{11} \\ \widetilde{Y}_{21} \end{bmatrix}$ with $\widetilde{Y}_{11} = f_1(X_{11}) = Y_{11}$ and $\widetilde{Y}_{21} = f_1(X_{21})$. That is, the difference between Y_1 and $\widetilde{Y}$ only comes from Y_{21} and $\widetilde{Y}_{21}$:

$$\begin{aligned} \|Y_1 - \widetilde{Y}_1\| &= \left\| \begin{bmatrix} Y_{11} \\ Y_{21} \end{bmatrix} - \begin{bmatrix} \widetilde{Y}_{11} \\ \widetilde{Y}_{21} \end{bmatrix} \right\| = \left\| \begin{bmatrix} f_1(X_{11}) \\ f_2(X_{21}) \end{bmatrix} - \begin{bmatrix} f_1(X_{11}) \\ f_1(X_{21}) \end{bmatrix} \right\| \\ &= \|f_2(X_{21}) - f_1(X_{21})\| \leq n_{21} \|f_2 - f_1\|_\infty = n_{21} E_u, \end{aligned}$$

where $E_u = \|f_2 - f_1\|_\infty$. Combining ① and ②, we derive the upper bound of e_1:

$$e_1 \leq C_1 e^{-c_1 n_1^{\frac{1}{p}}} + n_{21} D_u E_u.$$

Then we calculate e_2 by similar idea, but with all inequalities reversed. Again, we introduce the partially unobserved variables $\widetilde{Y}_2 := f_1(X_2)$ and let $\widetilde{y}^{(2)} := \psi_{X_2,x_0}(\widetilde{Y}_2)$. Again, by triangle inequality, we fin the following lower bound of e_2:

$$e_2 = |y_0 - \widehat{y}^{(2)}| = |y_0 - \widetilde{y}^{(2)} + \widetilde{y}^{(2)} - \widehat{y}^{(2)}| \geq \underbrace{|\widetilde{y}^{(2)} - \widehat{y}^{(2)}|}_{③} - \underbrace{|y_0 - \widetilde{y}^{(2)}|}_{④}$$

Finding the upper bound for ④ follows similar logic as for the upper bound for ①. Observe that ④ is the prediction error of standard Gaussian process regression on $(X_2, \widetilde{Y}_2)$, without any misspecified samples. As a result, the upper bound of ④ comes from Lemma C.1 and Assumption (A1). That is, $④ \leq \|f\| e^{-c_2 n_2^{\frac{1}{p}}}$ for some constant c_2.

While, unlike finding upper bound for ②, our goal is to find a lower bound for ③. By mean value theorem,

$$③ = |\psi_{X_2,x_0}(\widetilde{Y}_2) - \psi_{X_2,x_0}(Y_2)| \geq \inf \|\nabla \psi_{X_2,x_0}(Y)\|_\infty \|Y_1 - \widetilde{Y}_1\| = D_l \|\widetilde{Y}_2 - Y_2\|.$$

To find the lower bound of $\|\widetilde{Y}_2 - Y_2\|$, we again split both vectors into two components: $X_2 = \begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix}$ and $Y_2 = \begin{bmatrix} Y_{12} \\ Y_{22} \end{bmatrix}$, where $Y_{12} = f_2(X_{12})$ and $Y_{22} = f_1(X_{22})$. Then,

$$\begin{aligned} \|Y_2 - \widetilde{Y}_2\| &= \left\| \begin{bmatrix} Y_{12} \\ Y_{22} \end{bmatrix} - \begin{bmatrix} \widetilde{Y}_{12} \\ \widetilde{Y}_{22} \end{bmatrix} \right\| = \left\| \begin{bmatrix} f_1(X_{12}) \\ f_2(X_{22}) \end{bmatrix} - \begin{bmatrix} f_1(X_{12}) \\ f_1(X_{22}) \end{bmatrix} \right\| \\ &= \|f_2(X_{22}) - f_1(X_{22})\| \geq n_{22} \inf_{x \in \Omega} |f_2(x) - f_1(x)| \geq n_{22} E_l. \end{aligned}$$

Combining ③ and ④, we find the lower bound of e_2:

$$e_2 \geq n_{22} D_l E_l - C_2 e^{-c_2 n_2^{\frac{1}{p}}}.$$

Finally, we conclude that $e_1 < e_2$ if $C_1 e^{-c_1 n_1^{\frac{1}{p}}} + n_{21} D_u E_u < n_{22} D_l E_l - C_2 e^{-c_2 n_2^{\frac{1}{p}}}$, that is inequality 1.

Lemma C.1 (wendland2004scattered)**.** *When $f \in \mathcal{N}_K(\Omega)$ where K is the RBF kernel in $\mathbb{R}^p$, then let $\widehat{f}_n$ be the approximation to f by GP based on training samples (X, Y) with sample size n and filled distance $h_n := \sup_{x \in \Omega} \min_i \|x - x_i\|$, then*

$$\|f - \widehat{f}_n\|_\infty \leq e^{-c/h_n} \|f\|_K. \tag{3}$$

To prove Theorem 3.1 for arbitrary L and j, the only difference is in the construction of $\widetilde{Y}_j$, which splits into L components. To analyze e_j, $\widetilde{Y}_j = [\widetilde{Y}_{1j}, \cdots, \widetilde{Y}_{Lj}]^\top$ where $\widetilde{Y}_{kj} = f_j(Y_{kj})$ with $\widetilde{Y}_{jj} = Y_{jj}$. As a result, $\|\widetilde{Y}_j - Y_j\| \leq \sum_{k\neq j} n_{kj} D_u E_u$ and

$$e_j \leq \|f\| e^{-c_1 n_j^{\frac{1}{p}}} + \sum_{k\neq j} n_{kj} D_u E_u.$$

Similarly, for e_k with $k \neq j$, $\widetilde{Y}_k = [\widetilde{Y}_{1k}, \cdots, \widetilde{Y}_{Lk}]^\top$ where $\widetilde{Y}_{mk} = f_j(Y_{mk})$ with $\widetilde{Y}_{jk} = Y_{jk}$. As a result, $\|\widetilde{Y}_k - Y_k\| \geq \sum_{m\neq k} n_{mk} D_l E_l$ and

$$e_k \geq \sum_{m\neq k} n_{kj} D_l E_l - \|f\| e^{-c_k n_k^{\frac{1}{p}}}.$$

C.2. *Proof of Theorem 3.2*

For simplicity, we show the case of $L = 2$ only, the extension to general case is similar to the proof of Theorem 3.1. Following the proof in Section C.1, it sufficies to analyze ①. Recall that $y_0 = f_1(x_0) + \epsilon$, we first define $y_{0*} := f_1(x_0)$, then $|y_0 - y_{0*}| \leq |\epsilon| \leq 3\tau$ with probability 99.7% since $\epsilon \sim N(0, \tau^2)$. Since y_{0*} is the clean observation without any noise, the previous analysis carries to $|y_{0*} - \widetilde{y}^{(1)}|$ naturally, that is,

$$\begin{aligned} ① &= |y_0 - \widetilde{y}^{(1)}| = |y_0 - y_{0*} - y_{0*} + \widetilde{y}^{(1)}| \\ &\leq |y_0 - y_{0*}| + |y_{0*} - \widetilde{y}^{(1)}| \leq |\epsilon_0| + \|f\| e^{-c_1 n_1^{\frac{1}{p}}} \end{aligned}$$

where $\epsilon_0 \sim (0, \tau^2)$. To analyze ②, by the same argument, it suffices to bound $\|Y_1 - \widetilde{Y}_1\|$.

$$\begin{aligned} \|Y_1 - \widetilde{Y}_1\| &= \left\| \begin{bmatrix} Y_{11} \\ Y_{21} \end{bmatrix} - \begin{bmatrix} \widetilde{Y}_{11} \\ \widetilde{Y}_{21} \end{bmatrix} \right\| = \left\| \begin{bmatrix} f_1(X_{11}) \\ f_2(X_{21}) \end{bmatrix} + \Delta - \begin{bmatrix} f_1(X_{11}) \\ f_1(X_{21}) \end{bmatrix} \right\| \\ &= \|f_2(X_{21}) - f_1(X_{21})\| + \|\Delta_1\| \leq n_{21} \|f_2 - f_1\|_\infty = n_{21} E_u + \|\Delta_1\|, \end{aligned}$$

where Δ_1 is the vector of noise ϵ's so $\|\Delta_1\| \sim \chi(n_1)$. As a result,

$$e_1 \leq C_1 e^{-n_1^{\frac{1}{p}}} + n_{21} D_u E_u + |\epsilon_0| + \|\Delta_1\|.$$

In the same logic, we have $③ \geq n_{22} D_l E_l - \|\Delta_2\|$, $④ \leq |\epsilon| + \|f\| e^{-c_2 n_2^{-\frac{1}{p}}}$, and

$$e_2 \geq n_{22} D_l E_l - C_2 e^{-n_2^{\frac{1}{p}}} - |\epsilon_0| - \|\Delta_2\|.$$

Finally, we conclude that $e_1 < e_2$ if

$$C_1 e^{-n_1^{\frac{1}{p}}} + n_{21} D_u E_u + |\epsilon_0| + \|\Delta_1\| < n_{22} D_l E_l - C_2 e^{-n_2^{\frac{1}{p}}} - |\epsilon_0| - \|\Delta_2\|,$$

that is,

$$n_{21} D_u E_u < n_{22} D_l E_l - C_1 e^{-c_1 n_1^{\frac{1}{p}}} - C_2 e^{-c_2 n_2^{\frac{1}{p}}} - 2|\epsilon| - \|\Delta_1\| - \|\Delta_2\|.$$

Note that $2|\epsilon| = 2\tau\chi(1)$, $\|\Delta_1\| = \tau\chi(n_1)$ and $\|\Delta_2\| = \tau\chi(n_2)$. Then Theorem 3.2 follows by setting $\xi = 2|\epsilon| + \|\Delta_1\| + \|\Delta_2\|$, the sum of independent χ-distributions with degrees of freedom $1, n_1$ and n_2 rescaled by 2τ, τ and τ respectively.

The limiting case holds since $\chi(n)/n \xrightarrow{n\to\infty}$, that is, the χ random variable grows sub-linearly with the degree of freedom.