Biometrics, 2025, 81(4), ujaf149
https://doi.org/10.1093/biomtc/ujaf149
Biometric Methodology

OXFORD

A semiparametric Gaussian Mixture Model with spatial dependence and its application to whole-slide image clustering analysis

Baichen Yu[1], Jin Liu[2,*], Hansheng Wang[1]

[1]Guanghua School of Management, Peking University, Beijing, 100871, China, [2]School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China

[*]Corresponding author: Jin Liu, School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China (liujin@nankai.edu.cn)

ABSTRACT

We develop here a semiparametric Gaussian Mixture Model (SGMM) for unsupervised learning with valuable spatial information taken into consideration. Specifically, we assume for each instance a random location. Then, conditional on this random location, we assume for the feature vector a standard Gaussian Mixture Model (GMM). The proposed SGMM allows the mixing probability to be nonparametrically related to the spatial location. Compared with a classical GMM, SGMM is considerably more flexible and allows the instances from the same class to be spatially clustered. To estimate the SGMM, novel EM algorithms are developed and rigorous asymptotic theories are established. Extensive numerical simulations are conducted to demonstrate our finite sample performance. For a real application, we apply our SGMM method to the CAMELYON16 dataset of whole-slide images for breast cancer detection. The SGMM method demonstrates outstanding clustering performance.

KEYWORDS: EM algorithm; Gaussian Mixture Model; semiparametric Gaussian Mixture Model; spatial dependence; whole-slide image analysis.

1 INTRODUCTION

This work is primarily motivated by the whole-slide image (WSI) analysis for tumor detection (Ghaznavi et al., 2013). WSI is a powerful and widely adopted medical imaging technology, which is used for cancer prognosis prediction (Lee et al., 2022), cancer classification (Breen et al., 2025), and metastasis detection (Bejnordi et al., 2017). More specifically, a WSI is one particular type of color image with ultra-high resolution. For instance, Figure 3(A) demonstrates a WSI example of the sentinel lymph nodes for breast cancer diagnosis. Unfortunately, conventional deep learning-based image classification methods cannot be immediately applied to the WSI data due to its extremely large size. One possible solution is to cut the whole image into many sub-images. By treating the WSI sample as a bag and sub-images as instances contained in the bag, various multiple instance learning (MIL) methods can be applied (Zhou, 2018). Nevertheless, the applicability of those MIL methods hinges on one critical condition. That is the WSI-level labels must be provided for both positive and negative (e.g., tumor or non-tumor) WSI samples. They become inapplicable if only one positive WSI sample is given. One possible remedy is to provide sub-image-level labels for the WSI sample. In this case, essentially any appropriate type of classifier can be trained. Unfortunately, this is seldom the case in real practice. Then, how to learn the sub-image-level labels for the sake of tumor localization in a fully unsupervised way becomes a problem of great interest.

To this end, we develop here a fully unsupervised clustering method. Specifically, we first cut a positive WSI sample into many small-sized sub-images (i.e., instances), which are then represented by feature vectors. In this regard, essentially any well-pretrained self-supervised deep learning models can be used (Kang et al., 2023). Among those models, the UNI model of Chen et al. (2024) demonstrated the state-of-the-art (SOTA) prediction accuracy on a number of important benchmark datasets. Therefore, we are motivated to apply this SOTA model to our WSI instances for feature extraction in this work. Next, a classical Gaussian Mixture Model (GMM) is imposed on the extracted feature vector for clustering analysis (McLachlan et al., 2019). Meanwhile, it has been empirically well documented that the WSI instances are likely to be spatially clustered (Ye et al., 2019). However, to the best of our knowledge, most existing GMM methods, for example the popularly used spectral estimation, do not take the spatial information into consideration (Chen and Zhang, 2024; Löffler et al., 2021). Then, how to modify the classical GMM appropriately so that the valuable spatial information can be effectively incorporated becomes a problem of great interest.

To address this issue, various spatially-informed clustering methods have been developed. These methods can

Received: December 6, 2024; Revised: October 4, 2025; Accepted: October 13, 2025

be classified into three categories. The first category contains those Bayesian methods, which treat each instance as a node in a spatial network. Thereafter, it encourages the spatially adjacent instances in the spatial network to share the same label (Zhao et al., 2021). Unfortunately, those methods often suffer from expensive computational cost, since various MCMC algorithms have to be used for model estimation. The second category contains those methods, which model the mixture component parameters nonparametrically (Lee and Xue, 2018; Zeng et al., 2025). As a consequence, those methods allow the mixture component parameters (e.g., means and covariances) from the same class but different locations with large distances to be very different. In contrast, for most WSI analysis, we should expect the mixture component parameters from the same class to be similar regardless of their locations. The third category contains those methods, which impose a GMM structure not only on the feature vector but also on the spatial locations of the instances (Zhou et al., 2020). The consequence is that the marginal distribution of the instance locations becomes another Gaussian mixture. Unfortunately, the empirical distribution of the instance locations on a WSI seems obviously not a Gaussian mixture.

We are therefore motivated to develop a new GMM method, which allows the mixing probability to be spatially varying in a nonparametric way. Meanwhile, the mixture component parameters (i.e., the means and covariances) are assumed to be constant regardless of the instance locations. For convenience, we refer to the new method as a semiparametric Gaussian Mixture Model (SGMM). Our main contributions are given as follows. Compared with GMM, SGMM allows valuable spatial information to be taken into consideration. Compared with various Bayesian methods, SGMM is computationally much more efficient. Compared with the fully nonparametric methods, SGMM retains the interpretability of the feature vectors and also the parametric convergence rate of the component parameters. Compared with a fully parametric approach, SGMM allows the mixing probability to be fully nonparametric, which provides more flexibility in modeling capability. To estimate an SGMM, efficient estimation methods and computational algorithms are developed. Rigorous asymptotic theories are established. Extensive numerical simulations and a real WSI data analysis are conducted.

The rest of this paper is organized as follows: Section 2 introduces the model setup and the estimation methods. The asymptotic analysis of the theoretical properties is investigated in Section 3. Extensive numerical experiments are presented in Section 4. The article is concluded with a brief discussion in Section 5. All computational details and technical proofs are provided in the Web Appendices.

2 MODEL

2.1 Model and Notations

Let $(\mathbf{X}_i, Y_i, \mathbf{S}_i)$ be the data collected from the ith $(1 \le i \le N)$ sub-image of a given WSI sample. Here, $Y_i \in \{1, 2, \cdots, K\}$ is the class label for the ith sub-image. Note that Y_i is not directly observed. Assume the mixing probability $P(Y_i = k) = \pi_k > 0$ for every $1 \le k \le K$ with $\sum_{k=1}^{K} \pi_k = 1$. Given $Y_i = k$, the feature vector $\mathbf{X}_i = (X_{ij}) \in \mathbb{R}^p$ and the random location $\mathbf{S}_i = (S_{i1}, S_{i2})^\top \in \mathbb{S}$ are generated independently (Zhou et al., 2020), where $\mathbb{S} \subset \mathbb{R}^2$ is a 2-dimensional compact domain. Conditional on $Y_i = k$, we assume for $\mathbf{X}_i$ a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_k = (\mu_{kj}) \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma}_k = (\sigma_{kj_1 j_2}) \in \mathbb{R}^{p\times p}$. We further assume that the random location $\mathbf{S}_i = \mathbf{s}$ given $Y_i = k$ follows an unknown but continuous distribution with a probability density function given by $g_k(\mathbf{s})$ with $g_k(\mathbf{s}) \ge 0$ for any $1 \le k \le K$ and $\mathbf{s} \in \mathbb{S}$.

It is remarkable that we assume the feature vectors and the spatial locations are conditionally independent given the class label Y_i. However, tumor tissues often exhibit heterogeneity to some extent in their expression across different organ locations. Then, it is more appropriate to assume that $\mathbf{X}_i$ and $\mathbf{S}_i$ are conditionally dependent. Nevertheless, a careful analysis reveals that most of the heterogeneity is due to the fact that different tissues from different locations actually belong to different and further refined sub-categories (instead of only two major classes: tumor vs. non-tumor). For the tissues belonging to the same and sufficiently refined sub-category, we find that their expressions are fairly similar regardless of their locations. Therefore, this conditional independence assumption, similar to the one used by Zhou et al. (2020), should hold approximately at least for our intended WSI setting, as long as the number of clusters K is not too small.

Write $\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\big\{-(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\big\}$ as the probability density function of a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Then, the joint probability density function of $(\mathbf{X}_i, \mathbf{S}_i)$ is given by $f(\mathbf{x}, \mathbf{s}) = \sum_{k=1}^{K} \pi_k \phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{x}) g_k(\mathbf{s})$. Write $G(\mathbf{s}) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{s})$. We then have the conditional probability density function of $\mathbf{X}_i$ with $\mathbf{S}_i = \mathbf{s} \in \mathbb{S}$ given by $f_{\mathbf{s}}(\mathbf{x}) = f(\mathbf{x}, \mathbf{s})/G(\mathbf{s}) = \sum_{k=1}^{K} \pi_k(\mathbf{s}) \phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{x})$, where $\pi_k(\mathbf{s}) = \pi_k g_k(\mathbf{s})/G(\mathbf{s}) = P(Y_i = k|\mathbf{S}_i = \mathbf{s})$. It is interesting to note that the conditional distribution of $\mathbf{X}_i$ with $\mathbf{S}_i$ given (but not Y_i) remains a GMM. However, the mixing probability changes from π_k to $\pi_k(\mathbf{s})$, which is a nonparametric function in $\mathbf{s}$. Therefore, we refer to $\pi_k(\mathbf{s})$ as the local mixing probability.

2.2 The Estimation Methods

We next consider how to estimate this SGMM model. For an arbitrary symmetric matrix $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{p\times p}$, define a half-vectorization operator as $\text{vech}(\mathbf{B}) = (b_{ij} : 1 \le i \le j \le p) \in \mathbb{R}^{p(p+1)/2}$. Write $\boldsymbol{\theta}_k = \big(\boldsymbol{\mu}_k^\top, \text{vech}(\boldsymbol{\Sigma}_k)^\top\big)^\top \in \mathbb{R}^{p(p+3)/2}$ for $k = 1, \cdots, K$, and $\boldsymbol{\Theta} = \big(\boldsymbol{\theta}_1^\top, \cdots, \boldsymbol{\theta}_K^\top\big)^\top \in \mathbb{R}^{Kp(p+3)/2}$. Note that the marginal density of $\mathbf{X}_i$ is given by $f(\mathbf{X}_i) = \sum_{k=1}^{K} f(\mathbf{X}_i|Y_i = k)P(Y_i = k) = \sum_{k=1}^{K} \pi_k \phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{X}_i)$. Then, without utilizing the spatial information $\mathbf{S}_i$, the marginal log-likelihood of $\mathbf{X}_i$ can be written as

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k \phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\left(\mathbf{X}_i\right)\right\}. \quad (1)$$

Note that $\sum_{k=1}^{K}\pi_k = 1$. Therefore, we can represent the mixing probability by a $(k-1)$-dimensional vector as $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_{K-1})^\top \in \mathbb{R}^{K-1}$. Then, the classical maximum likelihood estimator (MLE) can be defined as $(\widehat{\boldsymbol{\Theta}}^{mg}, \widehat{\boldsymbol{\pi}}^{mg}) = \arg\max_{\boldsymbol{\Theta},\boldsymbol{\pi}} \mathcal{L}_{\mathbf{x}}(\boldsymbol{\Theta}, \boldsymbol{\pi})$, for which a standard EM algorithm can be used to compute $\widehat{\boldsymbol{\mu}}_k^{mg}$ and $\widehat{\boldsymbol{\Sigma}}_k^{mg}$ (Wu, 1983). Here the subscript "mg" is used to emphasize that $\widehat{\boldsymbol{\Theta}}^{mg}$ is the MLE computed based on the "marginal" distribution of $\mathbf{X}_i$ without taking the location information $\mathbf{S}_i$ into consideration. Then a natural question arises: Can we further improve the estimation accuracy of $\widehat{\boldsymbol{\Theta}}^{mg}$ by utilizing the location information $\mathbf{S}_i$ appropriately?

To address this issue, we develop a novel estimator as follows. Let $\mathbf{s} \in \mathbb{S}$ denote an arbitrary but fixed location. Recall that the conditional density of $\mathbf{X}_i$ given $\mathbf{s}$ is $f_s(\mathbf{X}_i) = \sum_{k=1}^{K}\pi_k(\mathbf{s})\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{X}_i)$. We then construct a locally weighted likelihood function as

$$\mathcal{L}_{\mathbf{s}}\big(\boldsymbol{\Theta}, \boldsymbol{\pi}(\mathbf{s})\big) = \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k(\mathbf{s})\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big(\mathbf{X}_i\big)\right\}\mathbb{K}\left(\frac{\mathbf{S}_i - \mathbf{s}}{h}\right), \tag{2}$$

where $h > 0$ is the bandwidth and $\boldsymbol{\pi}(\mathbf{s}) = \big(\pi_1(\mathbf{s}), \cdots, \pi_{K-1}(\mathbf{s})\big)^\top \in \mathbb{R}^{K-1}$ represents the mixing probability locally around s since $\sum_{k=1}^{K}\pi_k(\mathbf{s}) = 1$. Moreover, $\mathbb{K}(\cdot)$ is a kernel function defined on $\mathbb{R}^2$. Typically, we assume that $\mathbb{K}(\mathbf{s}) = K(s_1)K(s_2)$, where $K(t)$ with $t \in \mathbb{R}$ is a probability density function symmetric about 0. The locally weighted objective function $\mathcal{L}_{\mathbf{s}}\big(\boldsymbol{\Theta}, \boldsymbol{\pi}(\mathbf{s})\big)$ involves two types of unknown parameters. They are, respectively, a fixed dimensional parameter $\boldsymbol{\Theta}$ and a nonparametric parameter $\boldsymbol{\pi}(\mathbf{s}) \in (0,1)^{K-1}$ with infinite dimension. However and fortunately, an initial estimator for $\boldsymbol{\Theta}$ has already been obtained as $\widehat{\boldsymbol{\Theta}}^{mg}$. We can thus replace $\boldsymbol{\Theta}$ in $\mathcal{L}_{\mathbf{s}}\big(\boldsymbol{\Theta}, \boldsymbol{\pi}(\mathbf{s})\big)$ with $\widehat{\boldsymbol{\Theta}}^{mg}$, which leads to a simplified local loss function as $\mathcal{L}_{\mathbf{s}}\big(\widehat{\boldsymbol{\Theta}}^{mg}, \boldsymbol{\pi}(\mathbf{s})\big)$. Here, we adopt the local constant method since the interested parameter $\pi_k(\mathbf{s})$ should be bounded in (0,1). Thereafter, an estimator for $\boldsymbol{\pi}(\mathbf{s})$ can be obtained as $\widehat{\boldsymbol{\pi}}(\mathbf{s}) = \arg\max_{\boldsymbol{\tau}=(\tau_k)\in(0,1)^{K-1}}\mathcal{L}_{\mathbf{s}}(\widehat{\boldsymbol{\Theta}}^{mg}, \boldsymbol{\tau})$.

To solve optimization (2), we develop a kernel-based EM algorithm (i.e., Algorithm 1) with details given in Web Appendix B. By this algorithm, we should compute $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ for every $\mathbf{s} \in \{\mathbf{S}_i : 1 \le i \le N\}$. Consequently, optimization (2) needs to be solved for a total of N times. This leads to heavy computation costs. This computational challenge can be well solved by parallel computation, since the computations of $\widehat{\boldsymbol{\pi}}(\mathbf{S}_i)$s are completely independent of each other. This leads to a significant reduction in computation costs, which is to be shown in the subsequent Section 4. Once $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ is computed for every $\mathbf{s} \in \{\mathbf{S}_i : 1 \le i \le N\}$, we can then consider how to further upgrade $\widehat{\boldsymbol{\Theta}}^{mg}$ for better statistical efficiency.

Recall that the joint probability density of $(\mathbf{X}_i, \mathbf{S}_i)$ is $f(\mathbf{X}_i, \mathbf{S}_i) = G(\mathbf{S}_i)\sum_{k=1}^{K}\pi_k(\mathbf{S}_i)\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\mathbf{X}_i)$. Then, a global log-likelihood function can be constructed as

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\Theta}, \boldsymbol{\Pi}) &= \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big(\mathbf{X}_i\big)g_k\big(\mathbf{S}_i\big)\right\}\\ &= \sum_{i=1}^{N}\log\left[\left\{\sum_{k=1}^{K}\pi_k\big(\mathbf{S}_i\big)\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big(\mathbf{X}_i\big)\right\}G\big(\mathbf{S}_i\big)\right]\\ &= \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k(\mathbf{S}_i)\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big(\mathbf{X}_i\big)\right\} + \sum_{i=1}^{N}\log\left\{G\big(\mathbf{S}_i\big)\right\}.\end{aligned} \tag{3}$$

Next, replace $\boldsymbol{\Pi} = \big(\boldsymbol{\pi}(\mathbf{S}_1)^\top, \cdots, \pi(\mathbf{S}_N)^\top\big)^\top \in (0,1)^{N(K-1)}$ by $\widehat{\boldsymbol{\Pi}} = \big(\widehat{\boldsymbol{\pi}}^\top(\mathbf{S}_1), \cdots, \widehat{\boldsymbol{\pi}}^\top(\mathbf{S}_N)\big)^\top \in (0,1)^{N(K-1)}$, where $\widehat{\boldsymbol{\pi}}(\mathbf{S}_i)$ is the pilot estimator obtained by maximizing (2). This leads to a loss function $\mathcal{L}^*(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Pi}})$ in $\boldsymbol{\Theta}$. We can then optimize $\mathcal{L}^*(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Pi}})$ with respect to $\boldsymbol{\Theta}$. This leads to a joint estimator $\widehat{\boldsymbol{\Theta}}^{jnt} = \arg\max_{\boldsymbol{\Theta}}\mathcal{L}^*(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Pi}})$. The subscript "jnt" is used to emphasize the fact that $\widehat{\boldsymbol{\Theta}}^{jnt}$ utilizes the information from the "joint" distribution of $(\mathbf{X}_i, \mathbf{S}_i)$. Note that the second term in (3) depends on π_k and $g_k(\mathbf{S}_i)$ through $G(\mathbf{S}_i)$. It has nothing to do with $\boldsymbol{\Theta}$. Therefore, we have $\widehat{\boldsymbol{\Theta}}^{jnt} = \arg\max_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Pi}})$, where

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Pi}) = \sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k(\mathbf{S}_i)\phi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\big(\mathbf{X}_i\big)\right\}, \tag{4}$$

for which only the first term in (3) is involved. To optimize $\mathcal{L}(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Pi}})$ with respect to $\boldsymbol{\Theta}$, a standard EM algorithm can be used with a slight modification. Its numerical convergence theory has been well established in the past literature (Balakrishnan et al., 2017; Wu, 1983; Xu and Jordan, 1996). Note that our method is a multi-step estimator that combines parametric and nonparametric estimation methods, which is different from a classical semiparametric efficiency theory. Further note that our estimation process is executed only once and not iteratively. One can also iterate our optimization problems (2) and (4) to estimate $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}^{\text{jnt}}$ iteratively sufficiently, so that a fully iterated estimator $\widehat{\boldsymbol{\Theta}}^{full}$ can be obtained, as long as the time cost is not a serious concern. However, both our subsequent theoretical analysis and numerical simulations in Section 4.1 show that the efficiency gain in this regard is very limited and is ignorable asymptotically.

3 THEORETICAL PROPERTIES

3.1 Theoretical Properties of $\widehat{\boldsymbol{\pi}}_k(\mathbf{s})$

We next study the asymptotic properties of the various estimators. Define $\mathbf{a}_k(\mathbf{s}) = \pi_k^{-1}(\mathbf{s})\mathbf{e}_k \in \mathbb{R}^{K-1}$ for $k = 1, \cdots, K-1$ and $\mathbf{a}_K = -\pi_K^{-1}(\mathbf{s})1_{K-1} \in \mathbb{R}^{K-1}$, where $\mathbf{e}_k \in \mathbb{R}^{K-1}$ denotes the kth column of the $(K-1)$-dimensional identity matrix $\mathbf{I}_{K-1} \in \mathbb{R}^{(K-1)\times(K-1)}$. In addition, $\mathbf{1}_{K-1} \in \mathbb{R}^{K-1}$ refers to the all-one vector with $K-1$ dimensions. For notation convenience, define $\alpha_k(\mathbf{X}_i, \mathbf{s}) = \alpha_k(\mathbf{X}_i, \mathbf{s}|\boldsymbol{\Theta}) = P(Y_i = k|\mathbf{X}_i, \mathbf{S}_i = \mathbf{s})$. Let $\boldsymbol{\delta} = (\boldsymbol{\pi}^\top, \boldsymbol{\Theta}^\top)^\top \in \Delta \subset \mathbb{R}^q$ be the true parameter, where $q = Kp(p+3)/2 + K - 1$, and Δ is assumed to be a

compact parameter space. We first derive the asymptotic results of the marginal MLE $\widehat{\boldsymbol{\Theta}}^{mg}$ as follows.

Theorem 1: *Assume the SGMM model as described in Section 2.1, we have* $\sqrt{N}\left(\widehat{\boldsymbol{\delta}}^{mg} - \boldsymbol{\delta}\right) \to_d N\{0, (\widetilde{\boldsymbol{\Omega}}^{mg})^{-1}\}$ *as* $N \to \infty$, *where* $\widehat{\boldsymbol{\delta}}^{mg} = (\widehat{\boldsymbol{\pi}}^{mg\top}, \widehat{\boldsymbol{\Theta}}^{mg\top})^\top \in \mathbb{R}^q$, *and* $\widetilde{\boldsymbol{\Omega}}^{mg} \in \mathbb{R}^{q\times q}$ *is the asymptotic precision matrix defined in Web Appendix F.*

The proof of Theorem 1 is provided in Web Appendix F. By Theorem 1, we know that $\left\|\widehat{\boldsymbol{\Theta}}^{\text{mg}} - \boldsymbol{\Theta}\right\| = O_p(1/\sqrt{N})$ as $N \to \infty$, where $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top\mathbf{x}}$ stands for the usual ℓ_2-norm for an arbitrary vector $\mathbf{x}$. We next study the asymptotic properties of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$. To this end, a set of commonly used regularity conditions are necessarily needed. They are given in Web Appendix A with detailed discussions. Then, we have the following theorem:

Theorem 2: *Under Conditions (C1)–(C4), we know that for any fixed* $\mathbf{s} \in \mathbb{S} \subset \mathbb{R}^2$, *there exists a local MLE* $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ *such that: (1)* $\left|\widehat{\boldsymbol{\pi}}(\mathbf{s}) - \boldsymbol{\pi}(\mathbf{s})\right| = O_p(1/\sqrt{Nh^2})$; *and (2)*

$$\sqrt{Nh^2}\left\{\widehat{\boldsymbol{\pi}}(\mathbf{s}) - \boldsymbol{\pi}(\mathbf{s})\right\} \to_d N\left\{0, \eta_2^2\mathbf{V}^{-1}(\mathbf{s})\right\}, \tag{5}$$

where $\mathbf{V}(\mathbf{s}) = \int \left\{\sum_{k=1}^K \alpha_k(\mathbf{x}, \mathbf{s})\mathbf{a}_k(\mathbf{s})\right\}\left\{\sum_{k=1}^K \alpha_k(\mathbf{x}, \mathbf{s})\mathbf{a}_k(\mathbf{s})\right\}^\top f_{\mathbf{s}}(\mathbf{x})d\mathbf{x} \in \mathbb{R}^{(K-1)\times(K-1)}$.

The detailed proof of Theorem 2 is given in Web Appendix G. By the proof in Web Appendix G, we know that the asymptotic distribution of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ remains the same, if the pilot estimator $\widehat{\boldsymbol{\Theta}}^{mg}$ obtained in (1) is replaced by the true value $\boldsymbol{\Theta}$. This result is not surprising, since the marginal estimator $\widehat{\boldsymbol{\Theta}}^{mg}$ is a $\sqrt{N}$-consistent estimator. This is a convergence rate faster than the nonparametric $\sqrt{Nh^2}$ convergence rate of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$. Therefore, the estimation error (i.e., the data re-using effect) induced by $\widehat{\boldsymbol{\Theta}}$ is asymptotically ignorable for $\widehat{\boldsymbol{\pi}}(\mathbf{s})$. Moreover, we are able to extend the pointwise convergence result (5) to a uniform one in the sense that $\sup_{\mathbf{s}\in\mathbb{S}} \left\|\widehat{\boldsymbol{\pi}}(\mathbf{s}) - \boldsymbol{\pi}(\mathbf{s})\right\| = O_p\left(\sqrt{\log N/(Nh^2)}\right)$. The technical details are given in Lemma 1 in Web Appendix H.1.

3.2 Theoretical Properties of $\widehat{\boldsymbol{\Theta}}^{jnt}$

We next study the asymptotic behavior of $\widehat{\boldsymbol{\Theta}}^{jnt}$. We then have the following Theorem 3, whose detailed proof is provided in Web Appendix H:

Theorem 3: *Under Conditions (C1)-(C4), we have* $\sqrt{N}\left(\widehat{\boldsymbol{\Theta}}^{jnt} - \boldsymbol{\Theta}\right) \to_d N\left\{0, \left(\boldsymbol{\Omega}_1^{jnt}\right)^{-1}\boldsymbol{\Omega}^{jnt}\left(\boldsymbol{\Omega}_1^{jnt}\right)^{-1}\right\}$ *as* $N \to \infty$, *where both* $\boldsymbol{\Omega}^{jnt}$ *and* $\boldsymbol{\Omega}_1^{jnt} \in \mathbb{R}^{q'\times q'}$ *are defined in Web Appendix H and* $q' = Kp(p+3)/2$.

By Theorem 3, we know that $\widehat{\boldsymbol{\Theta}}^{jnt}$ remains $\sqrt{N}$-consistent, even if a nonparametric estimator $\widehat{\boldsymbol{\Pi}}$ is involved for computing $\widehat{\boldsymbol{\Theta}}^{jnt}$. To gain some quick insight, we study below some special but also important cases. Specifically, we introduce an ideal estimator $\widehat{\boldsymbol{\Theta}}^{ide}$, which is defined in the same way as $\widehat{\boldsymbol{\Theta}}^{jnt}$ with $\widehat{\boldsymbol{\Pi}}$ replaced by $\boldsymbol{\Pi}$. We then have the following corollary:

Corollary 1: *Under Conditions (C1)-(C4) and assume that the local mixing probability* Π *is given, we then have* $\sqrt{N}\left(\widehat{\boldsymbol{\Theta}}^{ide} - \boldsymbol{\Theta}\right) \to_d N\left\{0, \left(\boldsymbol{\Omega}_1^{jnt}\right)^{-1}\right\}$ *as* $N \to \infty$.

The proof of Corollary 1 is provided in Web Appendix I.1. Moreover, in many real applications, observations from the same class are heavily clustered spatially. Accordingly, the information provided by $\mathbf{S}_i$ should be extremely helpful. To theoretically reflect this interesting phenomenon, we can assume that $E\left[\prod_{k=1}^K\{1 - \alpha_k(\mathbf{X}_i, \mathbf{S}_i)\}\right] \to_p 0$ as $N \to \infty$. We then have the following corollary:

Corollary 2: *Assume Conditions (C1)-(C4) hold and the local mixing probability* Π *is given. Further assume that* $E\left[\prod_{k=1}^K\{1 - \alpha_k(\mathbf{X}_i, \mathbf{S}_i)\}\right] \to_p 0$ *as* $N \to \infty$, *we then have* $\sqrt{N}\left(\widehat{\boldsymbol{\Theta}}^{ide} - \boldsymbol{\Theta}\right) \to_d N\left\{0, \left(\boldsymbol{\Omega}_2^{jnt}\right)^{-1}\right\}$ *as* $N \to \infty$, *where* $\boldsymbol{\Omega}_2^{jnt} \in \mathbb{R}^{q'\times q'}$ *are defined in Web Appendix I.2. It can be verified that* $\boldsymbol{\Omega}_2^{jnt} - \boldsymbol{\Omega}_1^{jnt}$ *is positive semi-definite.*

The proof of Corollary 2 is also provided in Web Appendix I.2. By Corollary 2, we know that, under the situation that the data are heavily clustered spatially, the ideal estimator $\widehat{\boldsymbol{\Theta}}^{ide}$ should be statistically more efficient. In this case, the latent class label Y_is can be recovered by the posterior probability perfectly in the sense that $\alpha_k(\mathbf{X}_i, \mathbf{S}_i) - I(Y_i = k) \to_p 0$. Therefore, the ideal estimator $\widehat{\boldsymbol{\Theta}}^{ide}$ reduces to the simple moment estimator with the best efficiency. We can then reasonably conclude that the joint estimator $\widehat{\boldsymbol{\Theta}}^{jnt}$ should be more efficient than the marginal estimator $\widehat{\boldsymbol{\Theta}}^{mg}$, as long as the spatial information is strong enough.

4 NUMERICAL EXPERIMENTS

4.1 Simulation Studies

To demonstrate the finite sample performance of the different estimating methods, we present here a number of simulation studies.

Study 1. The objective of this study is to compare the statistical efficiency of the initial MLE $\widehat{\boldsymbol{\Theta}}^{mg}$ and the join t estimator $\widehat{\boldsymbol{\Theta}}^{jnt}$. For $\widehat{\boldsymbol{\Theta}}^{mg}$, we apply the K-means algorithm to obtain an initial estimator. For $\widehat{\boldsymbol{\Theta}}^{jnt}$, both the K-means estimator and $\widehat{\boldsymbol{\Theta}}^{mg}$ are considered as the initial estimators. We fix the number of classes $K = 2$, the feature means $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \mathbf{1}_p \in \mathbb{R}^p$ with $p \in \{2, 5, 10\}$, and $\boldsymbol{\Xi}_p = (\rho_{ij})$ with $\rho_{ij} = 0.5^{|i-j|}$. Set $\boldsymbol{\Sigma}_1 = 16\boldsymbol{\Xi}_p$, $\boldsymbol{\Sigma}_2 = 9\boldsymbol{\Xi}_p$, and $\pi_1 = 0.4$. The underlying distribution of the spatial location conditional on the class label is also set to be a Gaussian distribution with mean $\boldsymbol{\mu}_{S1} = -\boldsymbol{\mu}_{S2} = (1, 1)^\top$ and covariance matrix $\boldsymbol{\Sigma}_S = 0.5\boldsymbol{\Xi}_2$ but truncated on $[-5, 5]^2$. The bandwidth h is specified to be $h = 2.5 \times N^{-1/3}$. For a given N, we randomly replicate the experiment for a total of

$R = 500$ times and obtain R mean squared error (MSE) values for every target parameter (i.e., $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$). The MSE values of $\boldsymbol{\mu}_1$ are log-transformed, averaged, and then reported in Table 1. The averaged computational time and the numbers of iterations are also reported. The MSE values of other parameters are qualitatively similar and are presented in Web Appendix D.1.

From Table 1, we find that, with a given feature dimension p, both the marginal estimator $\widehat{\boldsymbol{\Theta}}^{mg}$ and the joint estimator $\widehat{\boldsymbol{\Theta}}^{jnt}$ are statistically consistent, as their log-transformed MSE values steadily decrease as the sample size increases. For example, with $p = 10$, the log(MSE) of $\widehat{\boldsymbol{\Theta}}^{jnt}$ with K-means initialization decreases from -2.066 to -5.014 as N grows from 500 to 5000. Furthermore, we find that the MSE values of the joint estimator are much lower than those of the marginal estimator, which implies that the joint estimator is statistically more efficient. For example, with $p = 10$ and $N = 5000$, the log(MSE) of $\widehat{\boldsymbol{\Theta}}^{jnt}$ with K-means initialization is -5.014, which is much smaller than that of -4.587 of $\widehat{\boldsymbol{\Theta}}^{mg}$. This confirms our theoretical results in Section 3. The results associated with different feature dimensions are qualitatively similar. Moreover, we find that different initial estimators for $\widehat{\boldsymbol{\Theta}}^{jnt}$ (i.e., K-means and $\widehat{\boldsymbol{\Theta}}^{mg}$) lead to very similar estimation results. This suggests that our iterative algorithm is fairly insensitive to the choice of the initial parameter. We next compute the average CPU time consumed by different estimators. From Table 1, we find that the computational time and the average number of iterations can be significantly reduced by using $\widehat{\boldsymbol{\Theta}}^{mg}$ as an initial estimator. This is expected, since $\widehat{\boldsymbol{\Theta}}^{mg}$ is an initial estimator more accurate than K-means. Additionally, the mean integrated square errors (MISEs) of the nonparametric estimator $\widehat{\pi}_k(\mathbf{s})$ on both the in-sample and out-of-sample data points are also reported in Table 1. Specifically, the in-sample MISE is defined to be $\text{MISE}_{\text{IN}} = R^{-1}\sum_{r=1}^{R} N^{-1}\sum_{i=1}^{N}(\widehat{\pi}_1(\mathbf{S}_i) - \pi_1(\mathbf{S}_i))^2$, and the out-of-sample (OOS) MISE is defined to be $\text{MISE}_{\text{OOS}} = R^{-1}\sum_{r=1}^{R} N^{-1}\sum_{i=1}^{N}(\widehat{\pi}_1(\mathbf{S}_i^*) - \pi_1(\mathbf{S}_i^*))^2$, where $\mathbf{S}_i^*$s are an independent copy of $\mathbf{S}_i$s. From Table 1, we find that log(MISE) for both in-sample and out-of-sample data points steadily decreases as sample size N increases. This confirms our theoretical claims in Theorem 2.

Study 2. The objective of this study is to compare the finite sample performance of our method with that of the spatially aided GMM (SAG) of Zhou et al. (2020) in terms of estimation accuracy. The SAG model assumes another independent Gaussian mixture distribution for the spatial location $\mathbf{S}_i$, conditional on the class label Y_i. In this study, we follow the data generating process of Zhou et al. (2020) and set the total number of components to be $L_{\text{true}} = 5$. For the actual parameter estimation, different working numbers of mixture components (denoted by L) are used (e.g., $L = 2$). Denote the resulting estimators of the SAG model with different working numbers of components as $\widehat{\boldsymbol{\Theta}}^{SAG(L)}$. In each replicate, four different estimators are compared. They are, respectively, $\widehat{\boldsymbol{\Theta}}^{mg}$, $\widehat{\boldsymbol{\Theta}}^{SAG(2)}$, $\widehat{\boldsymbol{\Theta}}^{SAG(5)}$, and $\widehat{\boldsymbol{\Theta}}^{jnt}$. Then, we randomly replicate the experiment for a total of $R = 500$ times.Next, the MSE values for estimating $\boldsymbol{\mu}_1$ are log-transformed, averaged, and plotted in Figure 1. From Figure 1, we find that the MSE values of both $\widehat{\boldsymbol{\Theta}}^{jnt}$ and $\widehat{\boldsymbol{\Theta}}^{SAG(5)}$ are uniformly lower than those of $\widehat{\boldsymbol{\Theta}}^{mg}$. Therefore, both the estimators $\widehat{\boldsymbol{\Theta}}^{jnt}$ and the correctly specified SAG estimator $\widehat{\boldsymbol{\Theta}}^{SAG(5)}$ can improve the efficiency of the estimation by incorporating the spatial information. Moreover, we find that our joint estimator $\widehat{\boldsymbol{\Theta}}^{jnt}$ is comparable to the correctly specified fully parametric estimator $\widehat{\boldsymbol{\Theta}}^{SAG(5)}$ as the sample size is large enough. However, when the number of components is mis-specified as $L = 2$, the SAG estimator becomes inconsistent $\widehat{\boldsymbol{\Theta}}^{SAG(2)}$. The MSE values of other parameters are qualitatively similar and are presented in Web Appendix D.1.

Study 3. The objective of this study is to demonstrate the clustering performance of the proposed methods in terms of clustering accuracy. We follow the same setting as the previous Study 2 and compare the clustering performance of the 4 different estimators. They are, respectively, the classical GMM, SAG(2), SAG(5), and the SGMM. For each clustering method and each random replication, various evaluation metrics [(i.e., AUC, Intersection-over-Union (IoU), and Adjusted Rand Index (ARI)] are computed; see Web Appendix C.2 for the technical details. Their average values are then recorded in Table 2. Note that this experiment is based on the setting of $K = 2$. The additional explorations of multiple values of $K > 2$ are provided in Web Appendix D.2.

From Table 2, we find that all the four methods improve as the sample size N increases. The methods of SAG(2), SAG(5), and SGMM all outperform the classical GMM clearly, since the valuable spatial information is not taken into consideration by the GMM. Moreover, recall that the true model is generated from SAG(5). We find that the result of our SGMM method is fairly comparable to that of the SAG(5). The actual performance of the SGMM could be even better if the sample size N is large enough. However, when the model is mis-specified as SAG(2), our SGMM outperforms it significantly in terms of AUC, IoU, and ARI. This further illustrates our model flexibility in terms of clustering analysis.

Study 4. Note that given the initial estimator, $\widehat{\boldsymbol{\Theta}}^{jnt}$ is a one-iteration estimator in terms of subsequently optimizing (2) and (4). Next, we can then take $\widehat{\boldsymbol{\Theta}}^{jnt}$ as the initial estimator to repeat the optimization process (2) and (4) iteratively. We can repeat this process for a sufficient number of times till the algorithm numerically converges. This leads to a fully-iterated estimator $\widehat{\boldsymbol{\Theta}}^{full}$. Then, the log(MSE)s of various estimators (i.e., $\widehat{\boldsymbol{\Theta}}^{mg}$, $\widehat{\boldsymbol{\Theta}}^{jnt}$, and $\widehat{\boldsymbol{\Theta}}^{full}$) are calculated and compared by boxplots. The simulation setup is the same as that of Study 1. The detailed results for estimating $\boldsymbol{\mu}_1$ are then reported in Figure 2. The MSE values of other parameters are qualitatively similar and are presented in Web Appendix D.1.From Figure 2, we find that the finite sample performance of $\widehat{\boldsymbol{\Theta}}^{jnt}$ can be further improved by $\widehat{\boldsymbol{\Theta}}^{full}$. However, the relative improvement margin decreases steadily and disappears eventually as the sample size increases. This is expected, since both $\widehat{\boldsymbol{\Theta}}^{jnt}$ and $\widehat{\boldsymbol{\Theta}}^{full}$ share the same asymptotic efficiency due to the following reasons. Recall that $\widehat{\boldsymbol{\Theta}}^{mg}$ is a global parameter estimator and hence is

TABLE 1 Log-transformed mean squared errors for estimating $\boldsymbol{\mu}_1$ using different estimators. The estimators in the quote refer to the initialization methods. The column log(MSE) reports the log-transformed MSE values. The column Time refers to the average CPU time in seconds. The column Iteration is the average numbers of iterations. OOS refers to out-of-sample prediction of the local mixing probability.

	p	2				5				10			
	n	500	1000	2000	5000	500	1000	2000	5000	500	1000	2000	5000
log(MSE)	$\widehat{\boldsymbol{\Theta}}^{\text{mg}}$	−0.996	−1.644	−2.296	−3.389	−1.259	−1.981	−2.877	−4.006	−1.585	−2.601	−3.581	−4.587
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ (K-means)	−1.193	−1.880	−2.860	−3.804	−1.623	−2.521	−3.501	−4.528	−2.066	−3.100	−4.062	−5.014
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ ($\widehat{\boldsymbol{\Theta}}^{\text{mg}}$)	−1.184	−1.880	−2.860	−3.804	−1.626	−2.522	−3.501	−4.527	−2.081	−3.100	−4.062	−5.014
Time (s)	$\widehat{\boldsymbol{\Theta}}^{\text{mg}}$	0.49	0.83	1.44	3.19	0.25	0.46	0.71	1.41	0.20	0.27	0.36	0.93
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ (K-means)	0.06	0.07	0.09	0.23	0.08	0.08	0.11	0.23	0.09	0.09	0.12	0.29
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ ($\widehat{\boldsymbol{\Theta}}^{\text{mg}}$)	0.01	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.04	0.02	0.02	0.03
Iter	$\widehat{\boldsymbol{\Theta}}^{\text{mg}}$	706.5	964.7	1193.5	1231.8	271.9	378.2	375.0	325.0	164.7	170.4	129.6	117.7
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ (K-means)	26.1	22.3	17.6	14.4	36.1	20.2	13.9	11.1	45.2	18.3	11.9	9.9
	$\widehat{\boldsymbol{\Theta}}^{\text{jnt}}$ ($\widehat{\boldsymbol{\Theta}}^{\text{mg}}$)	19.4	17.0	12.9	9.6	32.7	17.0	10.1	7.4	43.6	15.6	8.7	6.8
$\widehat{\pi}_k(\mathbf{s})$	log(MISE)	−2.278	−2.660	−3.262	−3.899	−2.575	−3.374	−4.208	−4.904	−2.928	−4.104	−5.069	−5.668
	Time (s)	1.23	2.36	5.25	19.92	1.15	2.47	5.35	19.26	1.04	2.44	5.18	20.28
	Iter	65.2	34.9	19.9	82.2	56.1	35.0	111.9	76.7	51.0	145.7	98.0	61.6
$\widehat{\pi}_k(\mathbf{s})$(OOS)	log(MISE)	−2.250	−2.674	−3.261	−3.917	−2.555	−3.398	−4.219	−4.941	−2.902	−4.126	−5.112	−5.723
	Time (s)	1.28	2.41	5.22	19.90	1.22	2.50	5.28	19.56	1.09	2.43	5.15	20.63
	Iter	64.8	34.6	19.7	82.2	55.3	34.5	112.5	76.9	51.4	146.0	98.4	61.8

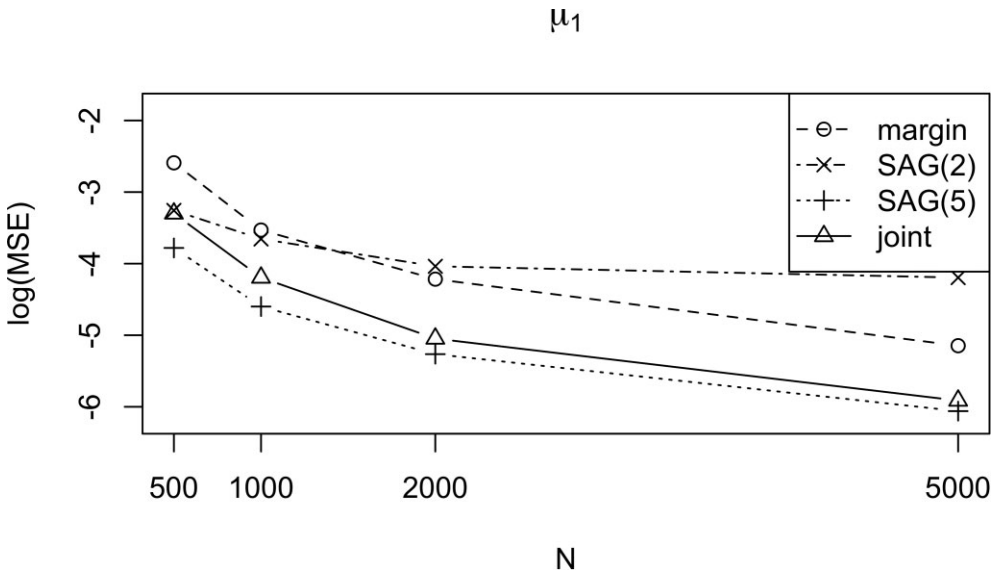

FIGURE 1 Log-transformed mean squared errors for $\boldsymbol{\mu}_1$ using different estimating methods under the SAG model (Zhou et al., 2020) when $L_{\text{true}} = 5$.

TABLE 2 The averaged prediction results using different estimating methods with different evaluation metrics.

Metric	N	GMM	SAG(2)	SAG(5)	SGMM
AUC	500	76.6	92.7	97.9	97.8
	1000	78.0	92.9	98.7	99.1
	2000	78.6	93.4	99.1	99.6
	5000	78.8	93.1	99.3	99.7
IoU	500	49.7	74.7	84.2	85.0
	1000	50.3	73.7	87.1	89.6
	2000	50.5	75.2	88.7	92.2
	5000	50.4	74.4	89.7	93.7
ARI	500	24.7	57.2	74.3	75.8
	1000	25.3	55.5	79.2	83.1
	2000	25.5	57.7	81.8	87.4
	5000	25.4	55.6	83.4	89.8

$\sqrt{N}$-consistent, while $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ is a local estimator and is $\sqrt{Nh^2}$-consistent. Consequently, the estimation error due to $\widehat{\boldsymbol{\Theta}}^{mg} - \boldsymbol{\Theta}$ (i.e., the data reusing effect) is asymptotically negligible for $\widehat{\boldsymbol{\pi}}(\mathbf{s}) - \boldsymbol{\pi}(\mathbf{s})$. Therefore, the asymptotic distribution of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ remains the same as the estimator computed with the true parameter $\boldsymbol{\Theta}$ given. Then, the asymptotic efficiency of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ cannot be further improved by iteratively improving the estimation accuracy of $\boldsymbol{\Theta}$. Since the asymptotic efficiency of $\widehat{\boldsymbol{\pi}}(\mathbf{s})$ cannot be further improved by additional iteration, that of $\widehat{\boldsymbol{\Theta}}^{jnt}$ cannot be further improved asymptotically either.

4.2 Real Data Analysis

To demonstrate the practical usefulness of the proposed methods, we present here a real data example. The dataset used here is the CAMELYON16 dataset, which is an important benchmark dataset about breast carcinoma metastasis detection (Bejnordi et al., 2017). We follow Wang et al. (2022) and select WSIs with medium-size tumors (i.e., lesion ratio greater than 0.1 and smaller than 0.9). This leads to a total of 11 WSIs. For a quick understanding, an arbitrarily selected WSI sample is illustrated in Figure 3(A).

This is an image with an ultra-high resolution of $97\,792 \times 221\,184 = 2.2 \times 10^{10}$ pixels. Therefore, appropriate data preprocessing is necessarily needed. In this regard, we follow a widely accepted standard procedure in the literature (Breen et al., 2025; Lee et al., 2022). This preprocessing process contains a total of 3 important steps. They are, respectively, (1) cutting a WSI into a number of sub-images with size 256×256; (2) using the method of Otsu (1979) to discard the background tiles and obtain Figure 3(B); and (3) extracting features for every sub-image by a deep learning method. Here, we apply the UNI model of Chen et al. (2024), which is already pre-trained on a massive 77-TB dataset by self-supervised contrastive learning. This leads to a $d = 512$ dimension feature vector. We then apply the principle component analysis on this feature vector so that its dimension can be further reduced to dimension p. Different values of dimension $p \in \{2, 5, 10\}$ are studied subsequently.

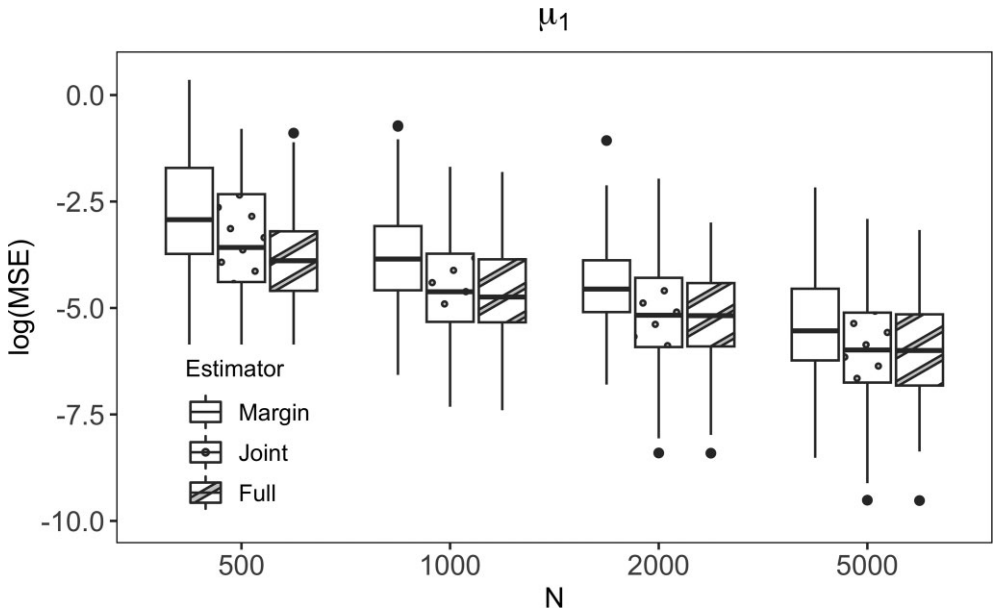

FIGURE 2 Log-transformed mean squared errors for $\boldsymbol{\mu}_1$ obtained by iterating once and fully iterating.

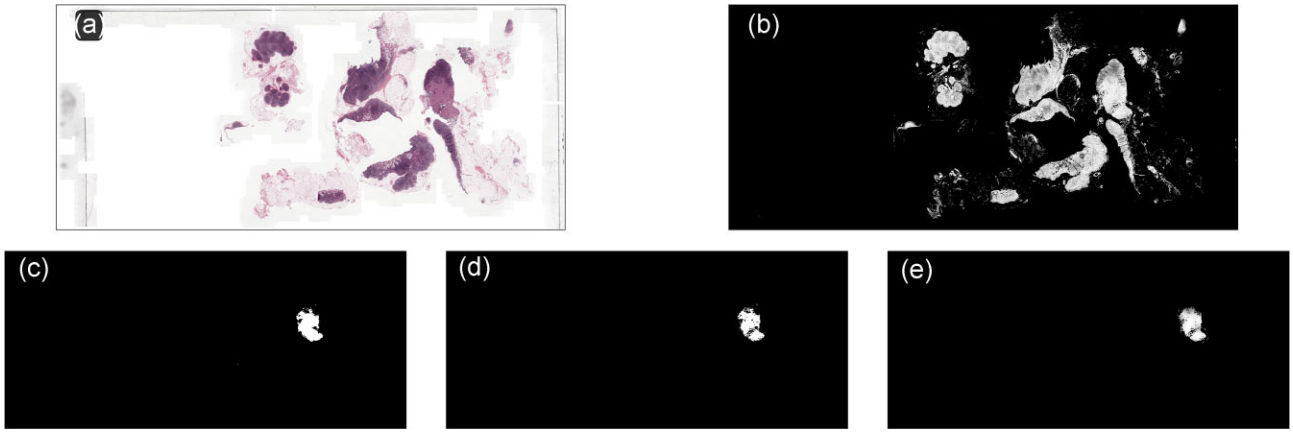

FIGURE 3 The prediction result of an arbitrarily selected WSI sample. (A) The original WSI. (B) The tissue region of the WSI. The bright area displays the informative region obtained by the method of Otsu (1979). (C) The human-annotated ground truth of the tumor region. (D) The posterior probability of being a tumor by our SGMM. E, the estimated local mixing probability by our SGMM.

We next apply the proposed SGMM method to those instances. To this end, the number of clusters needs to be specified. Recall that the goal here is to distinguish between tumor and non-tumor regions. Therefore, it is natural to set the number of clusters to $K = 2$. However, as we mentioned before in Section 2, the tumor cells and normal cells can be further classified into more refined sub-categories, since different refined sub-categories vary in morphology and hence statistical distribution. If we directly set $K = 2$, the tumor related sub-categories can possibly be confused with those non-tumor related ones. This often leads to sub-optimal classification accuracy. Therefore, we are motivated to consider a relatively larger range of choices for the number of clusters. This leads to more refined sub-category discovery. More specifically, we have tried various $K \in \{2, \cdots, 8\}$. Accordingly, the joint estimator $\widehat{\boldsymbol{\Theta}}^{jnt}$ can be computed and the posterior probabilities $\widehat{\alpha}_k(\mathbf{X}_i, \mathbf{S}_i)$ can be evaluated. For a robust evaluation, we compared their average performance on different dimension ps and different WSI samples as follows.

Note that for this particular dataset, the human-annotated ground truth $Y_i^{\text{tumor}} \in \{0, 1\}$ is only provided for two major classes (i.e., $Y_i^{\text{tumor}} = 1$ for tumor and $Y_i^{\text{tumor}} = 0$ otherwise) as shown in Figure 3(C). To compare our clustering results with this human-annotated label, we have to convert the refined K sub-categories into two major classes (i.e., tumor vs. normal). The technical details for this converting process are given in Web Appendix C.1. By this converting process, we obtain for each instance an integrated probability of being a tumor. Matching the predicted probability with the human-annotated ground truth Y_i^{tumor}, we are able to compute the AUC (Ling et al., 2003) and IoU (Rezatofighi et al., 2019). We also compute the ARI (Hubert and Arabie, 1985). The technical details of those evaluation metrics are also introduced in Web Appendix C.2.

TABLE 3 The averaged prediction results for different K values by using four competing methods. The top performance is highlighted in boldface for each row. The average time cost is reported in parentheses in seconds.

Metric	K	GMM	SAG	BayesSpace	SGMM
AUC	2	93.0 (2.1)	93.6 (25.6)	82.7 (947.1)	**94.5** (2.5)
	3	89.8 (3.2)	**93.8** (41.3)	86.7 (939.8)	91.1 (3.7)
	4	88.8 (5.8)	**92.6** (66.1)	90.6 (950.8)	92.2 (6.5)
	5	90.2 (8.2)	**93.2** (82.4)	92.9 (947.4)	93.0 (9.2)
	6	90.8 (10.1)	93.2 (109.0)	93.2 (946.8)	**93.7** (11.3)
	7	91.4 (16.3)	92.3 (120.6)	93.6 (948.3)	**94.1** (27.7)
	8	91.0 (30.3)	92.8 (158.6)	93.7 (948.3)	**93.9** (32.1)
IOU	2	53.2	51.3	50.5	**53.7**
	3	55.7	**60.3**	58.0	56.2
	4	60.3	61.2	59.9	**61.3**
	5	61.0	**63.2**	**63.2**	62.4
	6	62.5	63.2	62.6	**63.8**
	7	62.3	63.6	**64.3**	63.6
	8	62.0	63.6	**64.3**	63.1
ARI	2	60.2	56.3	48.5	**63.4**
	3	41.0	42.0	34.9	**42.7**
	4	29.8	30.5	28.5	**31.1**
	5	24.4	**27.2**	24.2	27.0
	6	20.8	22.6	20.2	**22.9**
	7	18.1	19.1	17.1	**20.1**
	8	16.1	17.4	14.9	**18.5**

Our SGMM method is then compared with a total of three competing methods in terms of out-of-sample prediction accuracy. The first method is the classical GMM, which makes no use of spatial information. The other two competitors are the SAG method of Zhou et al. (2020) and the BayesSpace method of Zhao et al. (2021). Both methods are spatially informed. The detailed results are given in Table 3.

From Table 3, we find that GMM performs worst to a large extent for all three performance metrics. This is expected, since no spatial information is used by GMM. Among the other three methods, our SGMM method performs best in most cases. Specifically, we obtain the top performance on four cases in AUC, three cases in IoU, and six cases in ARI. Moreover, we find that usually better IoU results can be obtained by a relatively larger number of clusters K, since there may exist some sub-categories in the WSI data. To summarize, we find that our SGMM method seems to be very competitive in terms of various prediction measures as compared with its competitors. More importantly, it offers a key advantage in computational efficiency. Specifically, the averaged computational time of SAG and BayesSpace is about 86 and 947 seconds, respectively. On the other side, our SGMM only takes 12 seconds on average. In contrast, human annotation by a well-trained expert takes about 15 minutes for one single WSI on average (Bejnordi et al., 2017). The computational advantage of our method is mainly because the kernel-based EM algorithm can be executed in a fully parallel way by tensor-based GPU computing systems; see Web Appendix B for implementation details. For a graphical illustration, we take the WSI of Figure 3 as an example. We provide in Figure 3(D) and Figure 3(E) not only the estimated mixing probability but also the estimated posterior probability by SGMM. We find that our SGMM method can recover the tumor region more smoothly and continuously. Practically, pathologists can overlay the heatmaps produced by SGMM on the original WSI as an interactive layer. Consequently, the pathologists can view histological images and our prediction results simultaneously. Therefore, the pathologists can be more focused on the suspicious area for more accurate and refined tumor diagnosis.

5 CONCLUDING REMARKS

To conclude this paper, we discuss here some interesting topics for future study. The SGMM is a semiparametric model, which contains a parametric component (i.e., a Gaussian component). Then, how to relax this parametric assumption for better flexibility is a problem of great interest. Second, our method includes nonparametric smoothing estimation. Therefore, it suffers from both the boundary issue and the curse of dimensionality. Many useful boundary bias reduction techniques, including pseudo data augmentation (Chen and Hong, 2012) and boundary-modified kernels (Su and Wang, 2017) might be helpful to some extent. Although WSIs are on two-dimensional planes, theoretically, it remains interesting to study the curse of dimensionality. One possible solution is to slice a high dimensional WSI sample into multiple two-dimensional slices, analyse them separately, and integrate them together using appropriate ensemble methods (Tong et al., 2019). Another possible solution is to leverage the strength of deep neural networks, which have been shown to have excellent capability in counteracting the curse of dimensionality (Jiao et al., 2023). Third, our current theoretical results only support feature vectors of a fixed dimension. How to allow the features with high dimension is another future topic. In this regard, appropriate model structures, such

as the sparsity assumption (Cai et al., 2019) and the factor structure (Zhao et al., 2018), might be helpful.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices A–I, and data and code referenced in Section 4 are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

J. L.'s research is supported by the National Natural Science Foundation of China (No.12201316). H. W.'s research is supported by the National Natural Science Foundation of China (No.12271012 and 72495123).

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The whole-slide image (WSI) data that support the findings in this paper are available at https://camelyon16.grand-challenge.org.

REFERENCES

Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45, 77–120.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G. et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318, 2199–2210.

Breen, J., Allen, K., Zucker, K., Godson, L., Orsi, N. M. and Ravikumar, N. (2025). A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. *NPJ Precision Oncology*, 9, 33.

Cai, T. T., Ma, J. and Zhang, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47, 1234–1267.

Chen, B. and Hong, Y. (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica*, 80, 1157–1183.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H. et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30, 850–862.

Chen, X. and Zhang, A. Y. (2024). Achieving optimal clustering in Gaussian Mixture Models with anisotropic covariance structures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Vancouver: Curran Associates.

Ghaznavi, F., Evans, A., Madabhushi, A. and Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8, 331–359.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Jiao, Y., Lai, Y., Lu, X., Wang, F., Yang, J. Z. and Yang, Y. (2023). Deep neural networks with ReLu-Sine-Exponential activations break curse of dimensionality in approximation on Hölder class. *SIAM Journal on Mathematical Analysis*, 55, 3635–3649.

Kang, M., Song, H., Park, S., Yoo, D. and Pereira, S. (2023). Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3344–3354. Vancouver: IEEE.

Lee, K. H. and Xue, L. (2018). Nonparametric finite mixture of Gaussian graphical models. *Technometrics*, 60, 511–521.

Lee, Y., Park, J. H., Oh, S., Shin, K., Sun, J., Jung, M. et al. (2022). Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, 6, 1–15.

Ling, C. X., Huang, J. and Zhang, H. (2003). AUC: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 3, 519–524. Acapulco: Morgan Kaufmann Publishers.

Löffler, M., Zhang, A. Y. and Zhou, H. H. (2021). Optimality of spectral clustering in the Gaussian Mixture Model. *The Annals of Statistics*, 49, 2506–2530.

McLachlan, G. J., Lee, S. X. and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 6, 355–378.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9, 62–66.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666. Long Beach: IEEE.

Su, L. and Wang, X. (2017). On time-varying factor models: estimation and testing. *Journal of Econometrics*, 198, 84–101.

Tong, L., Sha, Y. and Wang, M. D. (2019). Improving classification of breast cancer by utilizing the image pyramids of whole-slide imaging and multi-scale convolutional neural networks. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1, 696–703. Milwaukee: IEEE.

Wang, Z., Saoud, C., Wangsiricharoen, S., James, A. W., Popel, A. S. and Sulam, J. (2022). Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. *IEEE Transactions on Medical Imaging*, 41, 3952–3968.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11, 95–103.

Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8, 129–151.

Ye, J., Luo, Y., Zhu, C., Liu, F. and Zhang, Y. (2019). Breast cancer image classification on WSI with spatial correlations. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1219–1223. Brighton: IEEE.

Zeng, Q., Zhou, J., Ji, Y. and Wang, H. (2025). A semiparametric Gaussian Mixture Model for chest CT-based 3D blood vessel reconstruction. *Biostatistics (Oxford, England)*, 26, kxae013.

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T. et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39, 1375–1384.

Zhao, Y., Shrivastava, A. K. and Tsui, K. L. (2018). Regularized Gaussian Mixture Model for high-dimensional clustering. *IEEE Transactions on Cybernetics*, 49, 3677–3688.

Zhou, S., Bondell, H., Tordesillas, A., Rubinstein, B. I. and Bailey, J. (2020). Early identification of an impending rockslide location via a spatially-aided Gaussian Mixture Model. *The Annals of Applied Statistics*, 14, 977–992.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5, 44–53.

Received: December 6, 2024; Revised: October 4, 2025; Accepted: October 13, 2025