Biostatistics, 2024, **00**, 0, *pp.* 1–14
https://doi.org/10.1093/biostatistics/kxae013
Article

A semiparametric Gaussian mixture model for chest CT-based 3D blood vessel reconstruction

Qianhan Zeng[1], Jing Zhou[2,*], Ying Ji[3], Hansheng Wang[1]

[1]Guanghua School of Management, Peking University, Beijing, 100871, China
[2]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China
[3]Department of Thoracic Surgery, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing, 100020, China

*Corresponding author: Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China.
Email: jing.zhou@ruc.edu.cn

SUMMARY

Computed tomography (CT) has been a powerful diagnostic tool since its emergence in the 1970s. Using CT data, 3D structures of human internal organs and tissues, such as blood vessels, can be reconstructed using professional software. This 3D reconstruction is crucial for surgical operations and can serve as a vivid medical teaching example. However, traditional 3D reconstruction heavily relies on manual operations, which are time-consuming, subjective, and require substantial experience. To address this problem, we develop a novel semiparametric Gaussian mixture model tailored for the 3D reconstruction of blood vessels. This model extends the classical Gaussian mixture model by enabling nonparametric variations in the component-wise parameters of interest according to voxel positions. We develop a kernel-based expectation–maximization algorithm for estimating the model parameters, accompanied by a supporting asymptotic theory. Furthermore, we propose a novel regression method for optimal bandwidth selection. Compared to the conventional cross-validation-based (CV) method, the regression method outperforms the CV method in terms of computational and statistical efficiency. In application, this methodology facilitates the fully automated reconstruction of 3D blood vessel structures with remarkable accuracy.

KEYWORDS: 3D reconstruction; blood vessel; computed tomography; Gaussian mixture model; nonparametric kernel smoothing; TensorFlow.

1. INTRODUCTION

Computed tomography (CT) is an advanced 3D imaging technology capable of generating detailed, high-resolution 3D images of internal organs and structural intricacies of the human body. The basic idea of a 3D CT image involves segmenting a lung, as an illustrative example, into several thin pieces. Each piece is then meticulously scanned using a CT machine, yielding detailed, high-resolution grayscale images. Subsequently, these sliced pieces of the lung can be examined individually. Although examining each CT image slice individually offers practical utility, this approach fails to harness the benefits of 3D technology. For example, valuable information embedded within the 3D structure may be overlooked.

In surgical practices, such as lobectomy, wedge resection, or segmentectomy, it is critically important for surgeons to reconstruct blood vessels in three dimensions before surgery. This

Received: July 20, 2023. **Revised:** March 22, 2024. **Accepted:** March 25, 2024

practice enables surgeons to clearly visualize the blood vessels' anatomical structure. This greatly reduces the risk of accidental or missed blood vessel ruptures during surgery (Hagiwara et al. 2014). Furthermore, massive hemorrhages can be avoided (Ji et al. 2021). In addition, the 3D reconstructions of blood vessels offer invaluable educational benefits to novice surgeons by facilitating a comprehensive understanding of the lung anatomy and the shape distribution of blood vessels. However, current standard imaging technology does not yet enable surgeons to perform 3D reconstructions of blood vessels in a fully automated manner. That is, at present, the 3D reconstruction of blood vessels requires substantial expertise. For instance, the widely utilized RadiAnt software (https://www.radiantviewer.com/), recognized for its efficiency and user-friendly interface in viewing CT images, still requires significant manual efforts for such complex reconstructions. On average, it takes approximately 10 min for a very experienced surgeon to reconstruct 3D images of blood vessels from raw CT images. To this end, a sequence of sophisticated but critical, fine-tuned decisions must be made correctly. This could be a challenging task for less experienced surgeons. Consequently, surgical efficiency can be greatly affected. Therefore, developing an automatic method for reconstructing 3D blood vessels from raw CT images is of great interest.

To solve this problem, we propose a novel semiparametric Gaussian mixture model (GMM) for CT-based 3D blood vessel reconstruction. Our method is inspired by the empirical observation that different tissues of different densities manifest different gray intensities in CT images. Even within the same type of tissue, gray intensities may exhibit slight fluctuations according to their different positions within the human body. Consequently, CT values corresponding to different organ tissues are expected to manifest different means and variances, while those related to the same organ tissue are anticipated to have similar means and variances. This seems to be an ideal situation for applying mixture models for unsupervised clustering analysis. In this regard, the classical GMM (McLachlan and Peel 2000) appears to be a natural choice.

In a GMM, we assume that each type of meaningful human organ tissue (e.g. blood vessels) or the image background is represented by a distinct Gaussian mixture component. Each mixture component is assumed to be a parametric Gaussian model with globally constant mean and variance such that the parameters do not flexibly vary by 3D positions in the CT space. To some extent, this assumption imitates the common practice of standard CT imaging software (e.g. RadiAnt), where only two globally constant, fine-tuned parameters, such as window width (WW) and window level (WL), are allowed. Although this simple practice with globally constant fine-tuned parameters is practically useful, it is not fully satisfactory. This is because, even for the same human organ (e.g. blood vessels), the CT density varies according to voxel positions within the 3D CT space. Furthermore, we observe that the globally constant mean (WL) and constant standard deviation (WW) are inadequate to accommodate these variations. Consequently, the accurate reconstruction of blood vessels in a 3D image remains elusive. Thus, extending the classical GMM to consider this variation becomes a key issue.

To solve this problem, we extend the classical GMM by allowing the parameters of interest, including the mean, variance, and class prior probability functions, to vary nonparametrically. This variation is due to the different voxel positions within the 3D CT space. We call this a semiparametric method because it integrates both a parametric component, represented by the GMM, and a nonparametric component, represented by the nonparametric mean, variance, and class prior probability functions. To estimate the parameters of interest for a given voxel position, a kernel-based expectation–maximization (KEM) algorithm is developed. Our extensive numerical experiments suggest that this method is effective. To facilitate computational efficiency, the entire algorithm is presented in a tensor format to efficiently execute the KEM algorithm on a GPU device. This design allows for the full utilization of the GPU's parallel computing capabilities. Once the parameters of interest are estimated accurately, the posterior probability can be computed for the given voxel positions associated with the blood vessel category. This step is crucial for accurately identifying and reconstructing blood vessels within the 3D CT space.

Remarkably, our 3D reconstruction method is very different from the traditional practice commonly used by CT imaging software (e.g. RadiAnt). In commercial software, raw CT densities are used to reconstruct blood vessels three-dimensionally. However, as noted, this method results in 3D blood vessel images that lack accuracy. This is mainly because blood vessels from different voxel positions often have different CT intensity levels, which cannot be fully captured by globally constant fine-tuned parameters. By contrast, our method reconstructs 3D blood vessels using locally computed posterior probabilities. By examining CT intensities locally, we can obtain a much clearer vascular anatomy of the blood vessels. This makes the task of classifying each voxel into the right organ considerably easier. The locally discovered blood vessels are then represented by their posterior probabilities. By employing these posterior probabilities as the image intensities, we find that the expression levels of blood vessels from different voxel positions within the 3D CT space are more comparable. This significantly improves the reconstruction accuracy of the blood vessels.

To summarize, we make two important contributions in this work. Firstly, we contribute to surgical practice by proposing a fully automated method for 3D blood vessel reconstruction with significantly improved accuracy. Second, our study enriches the statistical theory of the classical GMM by extending it from a parametric model to a semiparametric one. The remainder of this article is organized as follows. Section 2 introduces the KEM algorithm and elucidates the primary theoretical properties of the proposed method. Section 3 presents extensive numerical studies on the KEM method. Section 4 concludes this article with a comprehensive discussion of our findings. The technical details are included in the Supplementary Materials.

2. METHODOLOGY

2.1. Model and notations

Let (Y_i, X_i) be the observation collected from the ith subject, with $Y_i \in \mathbb{R}^1$ being the univariate response of interest and $X_i \in \mathbb{D} = [0, 1]^3$ being the associated voxel position in a 3D Euclidean space. We assume that X_i is uniformly distributed on $\mathbb{D}$. Furthermore, we assume that for each i a latent class label $Z_i \in \{1, \cdots, M\}$, where M is the total number of classes. We then assume that $P(Z_i = m|X_i) = \pi_m(X_i)$ for $1 \le m \le M$. Clearly, we should have $\sum_{m=1}^{M} \pi_m(x) = 1$ for any $x \in \mathbb{D}$. This suggests that we can define $\pi_M(x) = 1 - \sum_{m=1}^{M-1} \pi_m(x)$ for any $x \in \mathbb{D}$ throughout this article. Conditional on $Z_i = m$ and X_i, we assume that Y_i is normally distributed with a mean $\mu_m(X_i)$ and variance $\sigma_m^2(X_i)$. Here, we assume that class prior probability function $\pi_m(x)$, mean function $\mu_m(x)$, and variance function $\sigma_m^2(x)$ are all smooth functions in $x \in \mathbb{D}$ and vary between classes. Furthermore, we assume that there exists an absolute constant $\sigma_{\min} > 0$ such that $\sigma_{\min} \le \sigma_m(x) < \infty$ holds for any $x \in \mathbb{D}$ and $1 \le m \le M$. Next, we consider how to consistently estimate them.

We collect the observed voxel positions and CT values into $\mathbb{X} = \{X_i : 1 \le i \le N\}$ and $\mathbb{Y} = \{Y_i : 1 \le i \le N\}$, respectively. Furthermore, we collect the unobserved latent class labels into $\mathbb{Z} = \{Z_i : 1 \le i \le N\}$. For a given voxel position x, define $\theta(x) = \{\pi^\top(x), \mu^\top(x), \sigma^\top(x)\}^\top \in \mathbb{R}^{3M-1}$, where $\pi(x) = \{\pi_1(x), \cdots, \pi_{M-1}(x)\}^\top \in \mathbb{R}^{M-1}$, $\mu(x) = \{\mu_1(x), \cdots, \mu_M(x)\}^\top \in \mathbb{R}^M$, and $\sigma(x) = \{\sigma_1(x), \cdots, \sigma_M(x)\}^\top \in \mathbb{R}^M$. Define $\Theta = \{\theta(x) : x \in \mathbb{D}\}$. To estimate $\theta(x)$, we develop a novel KEM method. We begin with a highly simplified case with $\pi_m(x) = \pi_m$, $\mu_m(x) = \mu_m$, and $\sigma_m^2(x) = \sigma_m^2$ for some constants $\pi_m > 0$, μ_m, and $0 < \sigma_m < +\infty$. We then have the log-likelihood function for an interior point x in $\mathbb{D}$ as

$$\begin{aligned}\mathcal{L}^*\Big\{\theta(x)\Big\} &= \ln\left\{\prod_{i=1}^{N} f(X_i, Y_i|\Theta)\right\} = \sum_{i=1}^{N} \ln\left\{\sum_{m=1}^{M} f(Y_i|m, X_i, \Theta)P(Z_i = m|X_i, \Theta)f(X_i|\Theta)\right\} \\ &= \sum_{i=1}^{N} \ln\left[\sum_{m=1}^{M} \phi\left\{\frac{Y_i - \mu_m(X_i)}{\sigma_m(X_i)}\right\} \times \left\{\frac{\pi_m(X_i)}{\sigma_m(X_i)}\right\}\right] \\ &= \sum_{i=1}^{N} \ln\left\{\sum_{m=1}^{M} \phi\left(\frac{Y_i - \mu_m}{\sigma_m}\right) \times \left(\frac{\pi_m}{\sigma_m}\right)\right\}, \qquad (2.1)\end{aligned}$$

where $f(x, y|\Theta)$ stands for the joint probability density function of (X_i, Y_i) evaluated at $(X_i, Y_i) = (x, y)$; and $f(y|m, X_i, \Theta)$ is the marginal probability density function of Y_i evaluated at $Y_i = y$ conditional on $Z_i = m$ and X_i. Moreover, $f(x|\Theta)$ is the probability density function of X_i evaluated at $X_i = x$. Since X_i is assumed to be uniformly generated on $\mathbb{D}$, we have $f(x|\Theta) = 1$ for $x \in \mathbb{D}$ and $f(x|\Theta) = 0$ for $x \notin \mathbb{D}$. Moreover, the function $\phi(y) = \exp(-y^2/2)/\sqrt{2\pi}$ is the probability density function of a standard normal random variable.

Nevertheless, class prior probability function $\pi_m(x)$, mean function $\mu_m(x)$, and variance function $\sigma_m^2(x)$ in our case are not constant. Instead, they should vary in response to the value of x (i.e. the voxel position within the 3D CT space) in a fully nonparametric manner. However, for an arbitrarily given x position, we should expect $\pi_m(x) > 0$, $\mu_m(x)$, and $0 < \sigma_m(x) < +\infty$ to be locally approximately constant. This is indeed a reasonable assumption as long as these functions are sufficiently smooth with continuous second-order derivatives. Thus, according to Proposition 3.10 of Lang (2012), we then know that the second-order derivatives of $\pi_m(x)$, $\mu_m(x)$, and $\sigma_m(x)$ are uniformly upper bounded from infinity on the compact set $\mathbb{D}$ for any $1 \leq m \leq M$. Therefore, we draw inspiration from the concept of the local constant estimator (Nadaraya 1965; Watson 1964) and local maximum likelihood estimation (Fan et al. 1998) and introduce the subsequent locally weighted log-likelihood function based on the observed data $(\mathbb{X}, \mathbb{Y})$:

$$\mathcal{L}_x(\theta) = \sum_{i=1}^{N} \ln\left[\sum_{m=1}^{M} \phi\left(\frac{Y_i - \mu_m}{\sigma_m}\right) \times \left(\frac{\pi_m}{\sigma_m}\right)\right] \mathbb{K}\left(\frac{X_i - x}{h}\right), \tag{2.2}$$

where $x = (x_1, x_2, x_3)^\top$ stands for a fixed interior point in $\mathbb{D}$. Note that $\theta = (\pi^\top, \mu^\top, \sigma^\top)^\top \in \mathbb{R}^{3M-1}$ is a set of working parameters with $\pi = (\pi_1, \cdots, \pi_{M-1})^\top \in \mathbb{R}^{M-1}$, $\mu = (\mu_1, \cdots, \mu_M)^\top \in \mathbb{R}^M$, and $\sigma = (\sigma_1, \cdots, \sigma_M)^\top \in \mathbb{R}^M$. For class M, we set $\pi_M = 1 - \sum_{m=1}^{M-1} \pi_m$. Moreover, the 3D kernel function, $\mathbb{K}(\cdot)$, is assumed to be $\mathbb{K}(x) = \mathcal{K}(x_1)\mathcal{K}(x_2)\mathcal{K}(x_3)$, where $\mathcal{K}(t)$ with $t \in \mathbb{R}^1$ is a continuous probability density function symmetric about 0. Here, $h > 0$ is the associated bandwidth. What distinguishes (2.2) from the classical log-likelihood function (2.1) is that information on the peer positions of x is also blended in for local estimations. For convenience, we refer to $\mathcal{L}_x(\theta)$ as a locally weighted log-likelihood function. Then, the local maximum likelihood estimators can be defined as $\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_x(\theta)$. For convenience, we refer to $\hat{\theta}(x)$ as the kernel maximum likelihood estimators (KMLE). We next consider how to optimize $\mathcal{L}_x(\theta)$ so that $\hat{\theta}(x)$ can be computed.

2.2. The KEM algorithm

Recall that the locally weighted log-likelihood function based on the observed data $(\mathbb{X}, \mathbb{Y})$ is already defined in (2.2). Ever since Dempster et al. (1977), a typical way to optimize (2.2) is to develop an EM algorithm based on the so-called complete log-likelihood function. That is the log-likelihood function obtained by assuming that the latent class label Z_i is observed. Thus, if we have access to the complete data $(\mathbb{X}, \mathbb{Y}, \mathbb{Z})$, we can define a complete log-likelihood function for the given voxel position x as follows:

$$\mathcal{Q}_x(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} I(Z_i = m) \ln\left[\phi\left\{\frac{Y_i - \mu_m(x)}{\sigma_m(x)}\right\} \times \left\{\frac{\pi_m(x)}{\sigma_m(x)}\right\}\right] \mathbb{K}\left(\frac{X_i - x}{h}\right), \tag{2.3}$$

where $I(Z_i = m)$ is an indicator function. In theory, any location of interest can be taken as x. In practice, a location of interest is often taken at the recorded voxel positions of CT data. The objective here is to consistently estimate the nonparametric functions $\pi_m(x)$, $\mu_m(x)$, and $\sigma_m(x)$ for $1 \leq m \leq M$. We start with a set of initial estimators as $\hat{\pi}_m^{(0)}(x) = 1/M$ and $\hat{\mu}_m^{(0)}(x) = \hat{\mu}_m$, where $\hat{\mu}_m$

is an initial estimator obtained by (for example) a standard k-means algorithm (MacQueen, et al. 1967). Furthermore, we set $\hat{\sigma}_m^{(0)}(x) = \sigma^{(0)}$, where $\sigma^{(0)}$ can be some prespecified constant. We write $\hat{\pi}_m^{(t)}(x)$, $\hat{\mu}_m^{(t)}(x)$, and $\hat{\sigma}_m^{(t)}(x)$ as the estimators obtained in the tth step.

E step. Based on the current estimate $\hat{\theta}^{(t)}$, we next take expectations on (2.3) conditional on the observed data $(\mathbb{X}, \mathbb{Y})$ and the current estimate $\hat{\theta}^{(t)}$ as follows:

$$E\Big\{\mathcal{Q}_x(\theta)\Big|\mathbb{X}, \mathbb{Y}, \hat{\theta}^{(t)}\Big\} = \sum_{i=1}^{N}\sum_{m=1}^{M} \hat{\pi}_{im}^{(t)}(X_i) \ln\left[\phi\left\{\frac{Y_i - \mu_m(x)}{\sigma_m(x)}\right\} \times \left\{\frac{\pi_m(x)}{\sigma_m(x)}\right\}\right] \mathbb{K}\left(\frac{X_i - x}{h}\right), \tag{2.4}$$

where $\hat{\pi}_{im}^{(t)}(X_i) = P(Z_i = m | X_i, Y_i, \hat{\theta}^{(t)})$. Specifically, $\hat{\pi}_{im}^{(t)}(X_i)$ can be computed as follows:

$$\hat{\pi}_{im}^{(t)}(X_i) = P(Z_i = m | X_i, Y_i, \hat{\theta}^{(t)}) = \frac{P(Y_i | X_i, Z_i = m, \hat{\theta}^{(t)}) P(Z_i = m | X_i, \hat{\theta}^{(t)})}{\sum_{m=1}^{M} P(Y_i | X_i, Z_i = m, \hat{\theta}^{(t)}) P(Z_i = m | X_i, \hat{\theta}^{(t)})}$$

$$= \phi\left\{\frac{X_i - \hat{\mu}_m^{(t)}(X_i)}{\hat{\sigma}_m^{(t)}(X_i)}\right\} \times \left\{\frac{\hat{\pi}_m^{(t)}(X_i)}{\hat{\sigma}_m^{(t)}(X_i)}\right\} \Bigg/ \left[\sum_{m=1}^{M} \phi\left\{\frac{X_i - \hat{\mu}_m^{(t)}(X_i)}{\hat{\sigma}_m^{(t)}(X_i)}\right\} \times \left\{\frac{\hat{\pi}_m^{(t)}(X_i)}{\hat{\sigma}_m^{(t)}(X_i)}\right\}\right]. \tag{2.5}$$

M Step. In this step, we maximize (2.3) and obtain a new estimate $\hat{\theta}^{(t+1)}$ in the $(t+1)$th step. Note that we have $\sum_{m=1}^{M} \pi_m(x) = 1$ for any $x \in \mathbb{D}$. Based on the Lagrange method, we can define the Lagrangian function as $\mathcal{Q} = E\{\mathcal{Q}_x(\theta) | \mathbb{X}, \mathbb{Y}, \hat{\theta}^{(t)}\} + \lambda(\sum_{m=1}^{M} \pi_m - 1)$, where λ is an additional parameter introduced by the Lagrange multiplier method. After maximizing $\mathcal{Q}$, we can update the parameters in the $(t+1)$th step by

$$\hat{\pi}_m^{(t+1)}(x) = \sum_{i=1}^{N} \hat{\pi}_{im}^{(t)}(X_i) \mathbb{K}\left(\frac{X_i - x}{h}\right) \Bigg/ \sum_{i=1}^{N} \mathbb{K}\left(\frac{X_i - x}{h}\right), \tag{2.6}$$

$$\hat{\mu}_m^{(t+1)}(x) = \sum_{i=1}^{N} \hat{\pi}_{im}^{(t)}(X_i) \mathbb{K}\left(\frac{X_i - x}{h}\right) Y_i \Bigg/ \sum_{i=1}^{N} \hat{\pi}_{im}^{(t)}(X_i) \mathbb{K}\left(\frac{X_i - x}{h}\right), \tag{2.7}$$

$$\hat{\sigma}_m^{(t+1)}(x) = \left[\sum_{i=1}^{N} \left\{Y_i - \hat{\mu}_m^{(t+1)}(X_i)\right\}^2 \mathbb{K}\left(\frac{X_i - x}{h}\right) \Bigg/ \sum_{i=1}^{N} \hat{\pi}_{im}^{(t)}(X_i) \mathbb{K}\left(\frac{X_i - x}{h}\right)\right]^{\frac{1}{2}}. \tag{2.8}$$

For convenience, we refer to (2.5)–(2.8) as a KEM algorithm, which should be iteratively executed until convergence. Our extensive numerical experiments suggest that the algorithm works very well.

2.3. Asymptotic properties

Next, we study the asymptotic properties of KMLE $\hat{\theta}(x)$. First, we define some notations. Recall that $\mathcal{L}_x(\theta)$ is the locally weighted log-likelihood function defined in (2.2). We define $\nu_{k,m} = \int t^k \mathcal{K}^m(t) dt$. The following technical conditions are then required:

(C1) (*Kernel Function*) Assume $|\nu_{k,m}| < +\infty$ for $0 \le k \le 2$ and $1 \le m \le 2 + \delta$ with some $\delta > 0$.
(C2) (*Bandwidth and Sample Size*) Assume $h = C_h N^{-1/7}$ for some constant $C_h > 0$.

As mentioned before, we assume that kernel function $\mathbb{K}(t)$ is the product of three univariate kernel density functions symmetric about 0. By Condition (C1), each univariate kernel density function should be a well-behaved probability density function with various finite moments. By Condition (C2), we require the bandwidth h to converge to zero at the speed of $N^{-1/7}$ as the sample size, N, goes to infinity. As a consequence, the locally effective sample size, denoted by Nh^3, diverges toward infinity as $N \to +\infty$. In the meanwhile, the bandwidth h is well constructed so that the resulting estimation bias should not be too large. Both Conditions (C1) and (C2) are fairly standard conditions, which have been widely used in the literature (Fan and Gijbels 1996; Ullah and Pagan 1999; Li and Racine 2007; Silverman 2018).

We define $I\{\theta(x)\} = \int \partial \ln f(x,y|\Theta)/\partial\theta(x) \times \partial \ln f(x,y|\Theta)/\partial\theta(x)^\top \times f(x,y|\Theta)\mathrm{d}y$. Denote the Euclidean norm as $\|z\| = \sqrt{z^\top z}$ for any vector z. Then, we have the following Theorem 2.1, with the technical details provided in Appendices A and B of the Supplementary Materials.

Theorem 2.1 *Assume that both Conditions (C1) and (C2) are satisfied, then (i) there exists a local maximum likelihood estimator* $\hat{\theta}(x)$ *such that* $\|\hat{\theta}(x) - \theta(x)\| = O_p(1/\sqrt{Nh^3})$*; and (ii)* $\sqrt{Nh^3}\left[\hat{\theta}(x) - \theta(x) - \nu_{2,1}h^2 I^{-1}\{\theta(x)\}g(x)/2\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \nu_{0,2}^3 I^{-1}\{\theta(x)\}\right]$*, where* $g(x)$ *is given by* $g(x) = \int\{\partial \ln f(x,y|\Theta)/\partial\theta(x)\}\mathrm{tr}\{\partial^2 f(x,y|\Theta)/\partial x \partial x^\top\}\mathrm{d}y$.

2.4. Bandwidth selection through cross-validation

To ensure the asymptotic properties of the KMLE, we must carefully select the optimal bandwidth h. From Condition (C2), we know that the optimal bandwidth should satisfy $h = C_h N^{-1/7}$. The question is how to make an optimal choice for C_h. To do so, an appropriately defined optimality criterion is required. One natural criterion could be out-of-sample forecasting accuracy. Let (X^*, Y^*) be an independent copy of (X_i, Y_i). Recall that $\hat{\theta}(x) = \{\hat{\pi}_1(x), \cdots, \hat{\pi}_{M-1}(x), \hat{\mu}_1(x), \cdots, \hat{\mu}_M(x), \hat{\sigma}_1(x), \cdots, \hat{\sigma}_M(x)\}^\top$ are the estimators obtained from data $\{(X_i, Y_i) : 1 \le i \le N\}$. Note that $E(Y^*|X^*, \Theta) = \sum_{m=1}^M \pi_m(X^*)\mu_m(X^*)$. Therefore, a natural prediction for Y^* can be constructed as $\hat{Y}^* = \sum_{m=1}^M \hat{\pi}_m(X^*)\hat{\mu}_m(X^*)$. Its square prediction error (SPE) can then be evaluated as $E(Y^* - \hat{Y}^*)^2$. Using a standard Taylor expansion-type argument, we can obtain an analytical formula for SPE using the following theorem:

Theorem 2.2 *Assume that both Conditions (C1) and (C2) are satisfied, we then have*

$$E\left(\hat{Y}^* - Y^*\right)^2 = \left(\sigma_y^2 + \frac{1}{4}C_h^4 N^{-4/7} C_1 + C_h^{-3} N^{-4/7} C_2\right)\{1 + o(1)\},$$

where $\sigma_y^2 = \int_{\mathbb{D}} \sum_{m=1}^M \pi_m(x)\{\sigma_m^2(x) + \mu_m^2(x)\} - \mu^2(x)\mathrm{d}x$, $\mu(x) = \sum_{m=1}^M \pi_m(x)\mu_m(x)$,
$C_1 = \int_{\mathbb{D}}[\nu_{2,1}\dot{\varphi}\{\theta(x)\}^\top I^{-1}\{\theta(x)\}g(x)]^2\mathrm{d}x$*, and*
$C_2 = \int_{\mathbb{D}} \nu_{0,2}^3 \dot{\varphi}\{\theta(x)\}^\top I^{-1}\{\theta(x)\}\dot{\varphi}\{\theta(x)\}\mathrm{d}x$*. Here,*
$\varphi\{\theta(x)\} = \sum_{m=1}^{M-1} \pi_m(x)\{\mu_m(x) - \mu_M(x)\} + \mu_M(x)$ *and* $\dot{\varphi}\{\theta(x)\} = \partial\varphi\{\theta(x)\}/\partial\theta(x)$.

The technical details of Theorem 2.2 are provided in Appendix C in the Supplementary Materials. By optimizing the leading term of $E(\hat{Y}^* - Y^*)^2$ from Theorem 2.2, we know that the optimal bandwidth constant is given by $C_h^* = (3C_2/C_1)^{1/7}$. However, both the critical constants, C_1 and C_2, are extremely difficult to compute without explicit assumptions on the concrete forms of $\pi(x)$, $\mu(x)$, and $\sigma(x)$ with respect to x. To solve this problem, a cross-validation (CV) type method is used to compute the optimal bandwidth. To this end, we need to randomly partition all voxel positions into two parts. The first part contains approximately 80% of the total voxels used for training. The remaining 20% is used for testing. For convenience, the indices of the voxels in the training and testing dataset are collected as $\mathcal{I}_0$ and $\mathcal{I}_1$, respectively. Next, for an arbitrary testing sample, $i \in \mathcal{I}_1$, we have $E(Y_i|X_i) = \sum_{m=1}^{M} \pi_m(X_i)\mu_m(X_i)$. Therefore, we are inspired to predict the Y_i value using $\hat{Y}_i = \sum_{m=1}^{M} \hat{\pi}_m(X_i)\hat{\mu}_m(X_i)$, where the estimates $\hat{\pi}_m(X_i)$ and $\hat{\mu}_m(X_i)$ are computed based on the training data with a given bandwidth constant. Let $\mathbb{C}_{\text{CV}} = \{C_h^{(g)} : 1 \le g \le G_{\text{CV}}\}$ be a set of tentatively selected pilot-bandwidth constants, where G_{CV} is the total number of pilot-bandwidth constants. Therefore, an estimator for SPE can be constructed as $\hat{\mathscr{L}}_{\text{SPE}}(C_h^{(g)}) = |\mathcal{I}_1|^{-1} \sum_{i\in\mathcal{I}_1} (Y_i - \hat{Y}_i)^2$. A CV-based estimator for the optimal bandwidth constant can then be defined as $\hat{C}_h^{\text{CV}} = \operatorname*{argmin}_{C_h^{(g)} \in \mathbb{C}_{\text{CV}}} \hat{\mathscr{L}}_{\text{SPE}}(C_h^{(g)})$; that is, the bandwidth constant in $\mathbb{C}_{\text{CV}}$ with the lowest SPE value is chosen. Consequently, a CV-based estimator for the optimal bandwidth is given by $\hat{h}^{\text{CV}} = \hat{C}_h^{\text{CV}} \times N^{-1/7}$.

2.5. Bandwidth selection through regression

Even though the above CV idea is intuitive, its practical computation is expensive. This is because, for an accurate estimation of the optimal bandwidth constant C_h^*, we typically need to evaluate a large number of pilot-bandwidth constants. To reduce the computation cost, we develop a novel regression method as follows. From Theorem 2.2, we know that $E\{\hat{\mathscr{L}}_{\text{SPE}}(C_h)\} \approx \sigma_y^2 + (C_h^4 C_1/4 + C_h^{-3} C_2)N^{-4/7}$. Note that $E\{\hat{\mathscr{L}}_{\text{SPE}}(C_h)\}$ is approximately a linear function in both C_1 and C_2. The corresponding weights are given by $C_h^4 N^{-4/7}/4$ and $C_h^{-3} N^{-4/7}$, where N is the given total sample size, and C_h is the tentatively selected pilot-bandwidth constant. This immediately suggests an interesting regression-based method for estimating the two critical constants, C_1 and C_2, as follows. We define $\mathbb{C}_{\text{REG}} = \{C_h^{(g)} : 1 \le g \le G_{\text{REG}}\}$ as a set of carefully selected pilot-bandwidth constants. Note that G_{REG} determines the total number of pilot-bandwidth constants to be tested. It must not be too large so that the computation cost can be reduced. For example, we fix $G_{\text{CV}} = 25$ and $G_{\text{REG}} = 5$ in our subsequent numerical studies. We define a pseudo response, $\mathbb{Y} = \{\mathscr{L}_{\text{SPE}}(C_h^{(g)}) - \bar{\mathscr{L}}_{\text{SPE}} : 1 \le g \le G_{\text{REG}}\}^\top \in \mathbb{R}^{G_{\text{REG}}}$, where $\bar{\mathscr{L}}_{\text{SPE}} = G_{\text{REG}}^{-1} \sum_{g=1}^{G_{\text{REG}}} \mathscr{L}_{\text{SPE}}(C_h^{(g)})$. We define $X^{(g)} = (N^{-4/7} C_h^{(g)4}/4, N^{-4/7}/C_h^{(g)3})^\top \in \mathbb{R}^2$. We further define a pseudo-design matrix, $\mathbb{X} \in \mathbb{R}^{G_{\text{REG}}}$, where the gth row is $(X^{(g)} - \bar{X})^\top$ and $\bar{X} = G_{\text{REG}}^{-1} \sum_{g=1}^{G_{\text{REG}}} X^{(g)}$. Then, we have an appropriate regression relationship as $\mathbb{Y} = \mathbb{X}\mathcal{C}$, where $\mathcal{C} = (C_1, C_2)^\top \in \mathbb{R}^2$. Subsequently, an ordinary least squares estimator is obtained for $\mathcal{C}$ as $\hat{\mathcal{C}} = (\mathbb{X}^\top\mathbb{X})^{-1}(\mathbb{X}^\top\mathbb{Y}) = (\hat{C}_1, \hat{C}_2)^\top$. This leads to a regression-based estimator for the optimal bandwidth constant C_h^* as $\hat{C}_h^{\text{REG}} = (3\hat{C}_2/\hat{C}_1)^{1/7}$. For convenience, we abbreviate this method as the REG method.

3. NUMERICAL STUDIES

3.1. The simulation model

Extensive simulation studies are conducted to validate the finite sample performance of the proposed KEM method. The entire experiment is designed so that the simulation data can mimic real chest CT data as much as possible. Specifically, we first fix a total of $M = 3$ mixture components, which represent the background, bone tissue, and lung tissue, respectively. Next, we need to specify the class prior probability function, $\pi_m(x)$. To make the experiment as realistic as possible, we use the LIDC-IDRI dataset, which is probably the largest publicly available chest CT database (Setio et al. 2017). It consists of 888 human chest CT files from seven medical institutions worldwide. After downloading the dataset, we found that six files were damaged and could not be read into memory. Thus, we use the remaining 882 CT files for the subsequent study.

Let $\mathbb{Y} = (Y_{ijk}) \in \mathbb{R}^{d_x \times d_y \times d_z}$ be an arbitrarily selected chest CT data from the LIDC-IDRI dataset. According to the definition of the database, we should always have $d_x = d_y = 512$ for every CT scan file. However, the number of slices (i.e. d_z) may vary across different CT scan files, but it should fall within the range of $d_z \in [95, 764]$. Next, we consider how to specify the values

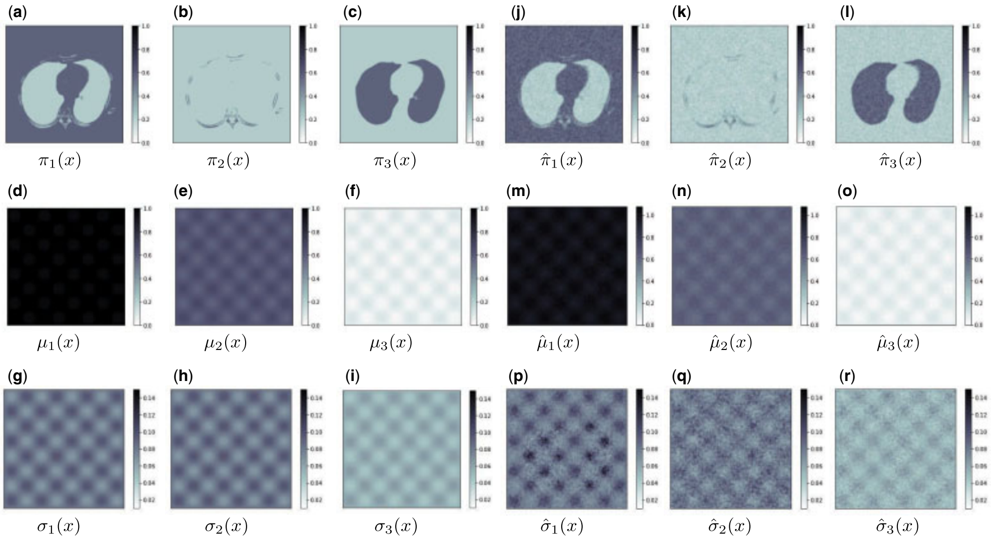


Figure 1. Graphical displays of $\pi_m(x)$, $\mu_m(x)$, $\sigma_m(x)$, $\hat\pi_m(x)$, $\hat\mu_m(x)$, and $\hat\sigma_m(x)$ for $1 \le m \le 3$ based on an arbitrarily selected chest CT slice.

of $\pi_m(X_{ijk})$ for every possible voxel position $X_{ijk} = (i/d_x, j/d_y, k/d_z)^\top \in \mathcal{V}$, where $\mathcal{V} \subset [0,1]^3$ contains all the available voxel positions. Next, we define a class label tensor $\mathbb{Z} = (Z_{ijk}) \in \mathbb{R}^{d_x \times d_y \times d_z}$ with $Z_{ijk} \in \{1,2,3\}$. Specifically, we define $Z_{ijk} = 3$ if its voxel position, X_{ijk}, belongs to the lung-mask data provided by the LIDC-IDRI dataset. Otherwise, we define $Z_{ijk} = 2$ if $Y_{ijk} > 400$. In this case, the (i,j,k)th voxel position should be the bone tissue (Molteni 2013). Finally, we define $Z_{ijk} = 1$ if the previous two criteria are not met.

For the (i,j,k)th voxel position $x = (i/d_x, j/d_y, k/d_z)^\top = (x_1, x_2, x_3)^\top \in \mathcal{V}$, we define x for a local neighborhood as $\mathcal{N}(x) = \big\{(x_1', x_2', x_3')^\top : |(x_1' d_x, x_2' d_y, x_3' d_z)^\top - (x_1 d_x, x_2 d_y, x_3 d_z)^\top|_{\max} < 3\big\}$, where $|x|_{\max} = \max_{1 \le j \le 3} |x_j|$. Then, we define $\pi_m(x) = \{\sum_{x' \in \mathcal{N}(x)} I(Z_{i'j'k'} = m)\}/|\mathcal{N}(x)|$, where $x' = (i'/d_x, j'/d_y, k'/d_z)^\top = (x_1', x_2', x_3')^\top \in \mathcal{V}$. Next, we redefine $\pi_m(x) = \{\pi_m(x) + 0.6\}/2.8$ for $1 \le m \le M$. By doing so, we can guarantee that $0 < \pi_m(x) < 1$ for every $1 \le m \le M$ and $x \in \mathcal{V}$. Next, we define $\mu_m = \{\sum_{x \in \mathcal{V}} I(Z_{ijk} = m) Y_{ijk}\}/\{\sum_{x \in \mathcal{V}} I(Z_{ijk} = m)\}$ and $\sigma_m^2 = \sum_{x \in \mathcal{V}} \{I(Z_{ijk} = m)(Y_{ijk} - \mu_m)^2\}/\{\sum_{x \in \mathcal{V}} I(Z_{ijk} = m)\}$ for $2 \le m \le M$. The $m = 1$ case is set as the background case. To differentiate the background from the other classes, μ_1 is set to 1, and σ_1 is set to σ_2. Then, we set $\mu_m(x) = \mu_m + 0.25 \times \sin(8\pi x_1) \times \sin(8\pi x_2) \times \sin(8\pi x_3)$ and $\sigma_m(x) = |\sigma_m + \sigma_m \times \sin(8\pi x_1) \times \sin(8\pi x_2) \times \sin(8\pi x_3)|$. Thus, we allow both the mean function $\mu_m(x)$ and variance function $\sigma_m^2(x)$ to vary within a reasonable range. For intuitive understanding, $\pi_m(x)$, $\mu_m(x)$, and $\sigma_m(x)$ with $1 \le m \le 3$ generated from an arbitrarily selected CT are graphically displayed in Fig. 1(a)–(i). Throughout this article, the depth index for the displayed 2D slices is consistently set to 100. For any voxel position $x \in \mathcal{V}$, once $\mu_m(x)$ and $\sigma_m(x)$ are given, a normal random variable can be generated. These normal random variables are then combined across different components according to a multinomial distribution with probability $\pi_m(x)$ for class m with $1 \le m \le M$. This leads to a final response Y_{ijk}.

3.2. Implementation details

Although the KEM algorithm developed in (2.5)–(2.8) is theoretically elegant, its computation is not immediately straightforward. If one implements the algorithm faithfully, as (2.5)–(2.8) suggest, the associated computational cost would be unbearable. Consider, for example, the computation of the denominator of equation (2.6), which is $\sum_{i=1}^{N} \mathbb{K}(X_i/h - x/h)$. The resulting computational complexity is $O(N)$, where $N = |\mathcal{V}| = d_x \times d_y \times d_z$ is the total number of voxel positions. Note that this procedure should be performed for each voxel position in $\mathcal{V}$. Then, the overall computational cost becomes $O(N^2)$ order. Consider, for example, a chest CT of size 512 × 512 × 201; then, we have $N = 5.27 \times 10^7$ and $N^2 = 2.78 \times 10^{15}$. Thus, the computational cost is prohibitive, and alleviating the computational burden becomes a critical problem.

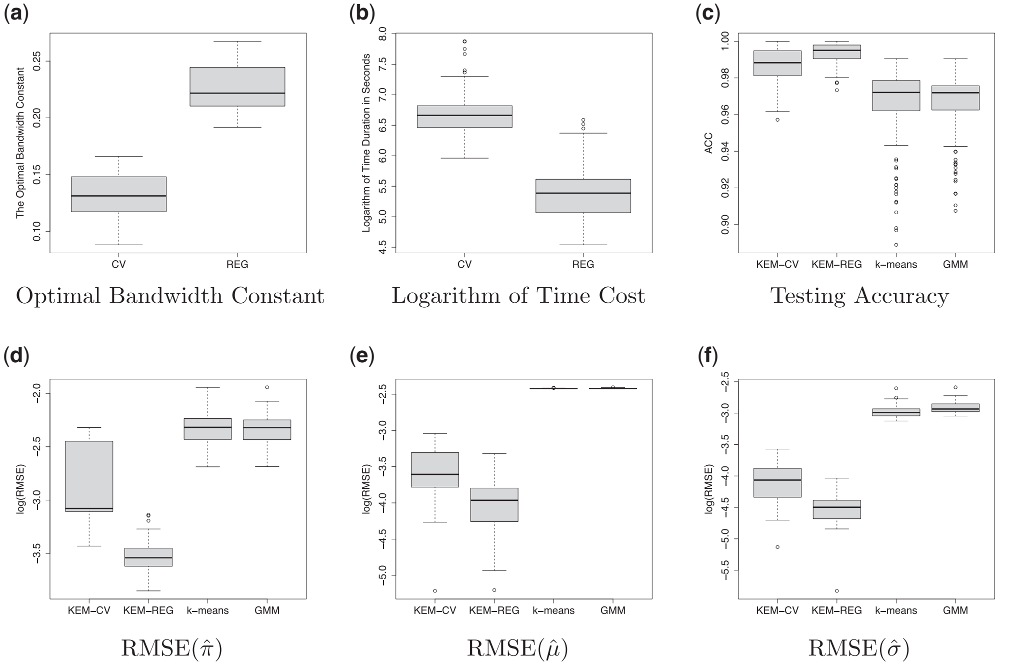
To address this problem, we develop a novel solution. We specify $\mathcal{K}(t) = \exp(-t^2/2)/\sqrt{2\pi}$ as the Gaussian kernel and $\mathbb{K}(x) = \mathcal{K}(x_1)\mathcal{K}(x_2)\mathcal{K}(x_3)$. Accordingly, the weight assigned to its tail decays toward 0 at an extremely fast rate. Therefore, instead of directly using the full kernel weight $\mathbb{K}(x)$, we consider its truncated version as $\mathbb{K}^*(x) = \mathbb{K}(x)I(|(x_1 d_x, x_2 d_y, x_3 d_z)^\top|_{\max} < s)$, where $s > 0$ represents the filter size. We define $\mathcal{N}_s(x) = \{x' = (x'_1, x'_2, x'_3)^\top \in \mathcal{V} : |(x'_1 d_x, x'_2 d_y, x'_3 d_z)^\top - (x_1 d_x, x_2 d_y, x_3 d_z)^\top|_{\max} < s\}$ as a cubic space locally around x with volume $|\mathcal{N}_s(x)| = s^3$. Instead of computing $\sum_{i=1}^{N} \mathbb{K}(X_i/h - x/h)$, we can compute $\sum_{i=1}^{N} \mathbb{K}^*(X_i/h - x/h) = \sum_{1 \le i \le N}^{X_i \in \mathcal{N}_s(x)} \mathbb{K}(X_i/h - x/h)$. Thus, the computational cost for one voxel position can be significantly reduced from $O(N)$ to $O(s^3)$, which is approximately the size of set $\{X_i \in \mathcal{N}_s(x) : 1 \le i \le N\}$. This is typically a much-reduced number compared with N. Furthermore, this operation can be easily implemented on a GPU device using a standard 3D convolution operation (e.g. the Conv3D function in Keras of TensorFlow). This substantially accelerates the computational speed. To this end, the bandwidth must be appropriately specified according to the filter size of the convolution operations (i.e. s). Ideally, the filter should not be too large. Otherwise, the computation cost owing to $|\mathcal{N}_s(x)|$ could be extremely high. However, the filter size, s, cannot be too small compared to bandwidth h. Otherwise, the probability mass of $\mathbb{K}(x)$, which falls into $\mathcal{N}_s(x)$, could be too small.

For the REG method, we consider the filter size as $s^{(g)} = 2g + 1$, where $1 \le g \le G_{\text{REG}}$ and $G_{\text{REG}} = 5$. Accordingly, we set the bandwidth to $h^{(g)} = s^{(g)}/512$. Thus, only odd numbers are considered, so the filter used in the 3D convolution operations can be symmetric about its central voxel position. We wish the filter to be as large as possible. However, as previously discussed, an unnecessarily large filter size would result in prohibitive computational costs. Therefore, we set the largest filter size as 11. For the CV method, following the REG method, we generate $(s^{(g)}, h^{(g)})$ for $1 \le g \le G_{\text{REG}}$. Subsequently, each $h^{(g)}$ is multiplied by 0.3, 0.4, 0.5, 0.6, or 0.7. The bandwidth constant of the two methods is set as $C_h^{(g)} = h^{(g)} \times N^{1/7}$, where N represents the sample size.

3.3. Bandwidth selection and estimation accuracy

With the prespecified filter sizes, both the CV and REG methods introduced in Sections 2.4 and 2.5 can be used to select the optimal bandwidth. Recall that 80% of the voxels are used for training and the remaining 20% are used for testing. Truncated Gaussian kernel and GPU-based convolutions are used to speed up the practical computation. Both the CV and REG methods are used to select the optimal bandwidth constant for each of the 100 randomly selected patients in the LIDC-IDRI dataset. The optimal bandwidth constants selected by the two methods are then boxplotted in Fig. 2(a). We find that the REG method tends to select a larger C_h than the CV method, leading to relatively larger filter sizes for convolution. The time cost of the two methods for selecting the optimal bandwidth for each patient is boxplotted in Fig. 2(b). Evidently, the REG method has a much lower computational cost than the CV method.

To show the superiority of the proposed KEM method, we also compared it with two traditional methods. The two traditional methods in concern are, respectively, k-means (MacQueen, et al. 1967) and GMM (McLachlan and Peel 2000). We then compare the aforementioned methods in terms of testing accuracy and estimation accuracy. Testing accuracy reflects the ability to identify the class labels on the testing set, while estimation accuracy reflects the ability to recover the true parameter values. When comparing the testing accuracy, we train the model on the training set and evaluate the testing accuracy on the testing set. See Fig. 2(c) for the results of the testing accuracy.



Optimal Bandwidth Constant

Logarithm of Time Cost

Testing Accuracy

RMSE($\hat{\pi}$)

RMSE($\hat{\mu}$)

RMSE($\hat{\sigma}$)

Figure 2. (a) is the boxplot of the optimal bandwidth constant C_h selected by the CV and REG methods. (b) is the logarithm of the time cost of the two methods in seconds. (c) is the boxplot of the testing accuracy of the KEM under two bandwidth selection methods, the *k*-means method, and the GMM method. (d)–(f) represent RMSE($\hat{\pi}$), RMSE($\hat{\mu}$), and RMSE($\hat{\sigma}$) of the concerned methods, respectively.

The medians of the testing accuracy results for the KEM method with the bandwidth selected by the CV method (KEM-CV), the KEM method with the bandwidth selected by the REG method (KEM-REG), the *k*-means method, and the GMM method are, respectively, 0.9883, 0.9950, 0.9721, and 0.9719. We can see that both the traditional methods suffer from worse testing accuracy as compared with the proposed KEM methods. When comparing the estimation accuracies, all voxel positions of the CT data are used. We can compute $\hat{\pi}_m(x)$, $\hat{\mu}_m(x)$, and $\hat{\sigma}_m(x)$ using the KEM-CV, KEM-REG, *k*-means, and GMM methods for every $1 \leq m \leq M$ and every $x \in \mathcal{V}$. Because this is a simulation study with the true parameter values specified in Section 3.1, which are known to us, we can evaluate the accuracy of the resulting estimates using the following RMSE criteria:

$$\text{RMSE}(\hat{\pi}) = \left[|\mathcal{V}|^{-1} m^{-1} \sum_{x \in \mathcal{V}} \sum_{m=1}^{M} \left\{ \hat{\pi}_m(x) - \pi_m(x) \right\}^2 \right]^{1/2},$$

$$\text{RMSE}(\hat{\mu}) = \left[|\mathcal{V}|^{-1} m^{-1} \sum_{x \in \mathcal{V}} \sum_{m=1}^{M} \left\{ \hat{\mu}_m(x) - \mu_m(x) \right\}^2 \right]^{1/2},$$

$$\text{RMSE}(\hat{\sigma}) = \left[|\mathcal{V}|^{-1} m^{-1} \sum_{x \in \mathcal{V}} \sum_{m=1}^{M} \left\{ \hat{\sigma}_m(x) - \sigma_m(x) \right\}^2 \right]^{1/2}.$$

The RMSE results of the aforementioned four methods are boxplotted in Fig. 2(d)–(f). We find that the KEM method outperforms the *k*-means and GMM methods substantially. This is not surprising, since the two traditional methods cannot estimate the parameters of interest nonparametrically.

Table 1. Mean RMSE values under different sampling ratios.

r	RMSE($\hat{\pi}$)	RMSE($\hat{\mu}$)	RMSE($\hat{\sigma}$)
0.01	0.07398	0.04150	0.02299
0.10	0.04552	0.02267	0.01368
0.50	0.03031	0.01664	0.01138
1.00	0.02440	0.01457	0.01002

Comparatively speaking, the REG method performs slightly better than the CV method. Thus, the median value of the bandwidth constants (i.e. $C_h = 0.2217$) chosen by the REG method is then fixed for the subsequent simulation studies.

With the fixed bandwidth constant $C_h = 0.2217$, we conduct more comprehensive simulation studies to evaluate the finite sample performance of the estimators. These are, respectively, $\hat{\pi}_m(x)$, $\hat{\mu}_m(x)$, and $\hat{\sigma}_m(x)$ for $1 \leq m \leq M$. For every chest CT scan in the LIDC-IDRI dataset, only $r \times 100\%$ of the voxel positions are used for parameter estimation. Different sample sizes can be examined by varying the sampling ratio r in $(0, 1]$. For an intuitive understanding, we arbitrarily select a replicated experiment. In this experiment, visual depictions of 2D slices for $\hat{\pi}_m(x)$, $\hat{\mu}_m(x)$, and $\hat{\sigma}_m(x)$ with $r = 1$ are presented in Fig. 1(j)–(r). We find that those estimated functions resemble the true functions very well. For a fixed r value, the experiment is randomly replicated 1000 times on a randomly selected chest CT scan in the LIDC-IDRI dataset. The sample means are summarized in Table 1. We find that as the value of r increases, the reported mean RMSE values steadily decrease toward 0 for every estimate. This suggests that they should be consistent estimates of the target parameters. Remarkably, the convergence rate shown in Table 1 seems to be slow. This is expected because we are dealing with a nonparametric kernel-smoothing procedure in a 3D space. The optimal convergence rate is indeed slow.

3.4. Case illustration

Recall that the ultimate goal of our method is to reconstruct fine-grained 3D images of blood vessels. We aim to accomplish this critical task by using chest CT data in a fully automatic manner. To this end, we randomly select three CT files from the LIDC-IDRI dataset. The PatientIDs of the CTs are, respectively, LIDC-IDRI-0405, LIDC-IDRI-0811, and LIDC-IDRI-0875. To avoid distractions from undesirable human tissue, we follow Liao et al. (2019) and generate a lung mask for each CT slice. Then, we multiply the lung mask by each slice. This leads to more concentrated chest CT slices, as shown in Fig. 3(a)–(c). Consider Fig. 3(c) as an example: The background is pure black. There is a lung-shaped area floating in the middle of the image, the color of which is slightly lighter than the background. Blood vessels, which appear as light spots, are dispersed inside the lung-shaped area. Because image signals from irrelevant organ tissue are excluded, detecting blood vessels in the focused chest CT image is much easier than in the raw chest CT image.

Next, we apply the developed KEM method to the concentrated chest CT data. In this case, the surgeon's primary objective is to accurately identify the blood vessels. Therefore, those voxels associated with blood vessels form one latent class. Second, since chest CT is mainly for diagnosing early-stage lung cancer, the lung tissue is also of great interest. Thus, the voxels associated with lung tissue form another important category. Lastly, in order to extract the lung tissue from the entire CT image, one needs to separate the background from the lung tissue. Therefore, the voxels associated with the background constitute the third latent class. It seems that all of these three classes can cover almost all the voxel positions in this real example. Therefore, we set $M = 3$ in this experiment.

The CT values are then rescaled to between 0 and 1. We initialize $\hat{\mu}_m^{(0)}(x)$ for any $x \in \mathcal{V}$ by using a standard k-means algorithm. We set $\hat{\pi}_{im}^{(0)} = \hat{\pi}_m^{(0)}(x) = 1/M$ and $\hat{\sigma}_m^{(0)}(x) = 0.1$ for $1 \leq i \leq N$, $1 \leq m \leq M$, and $x \in \mathcal{V}$. The filter size is set to 3 for fast computational speed. Subsequently, the KEM algorithm is used to estimate the parameters of interest. In our case, the estimated

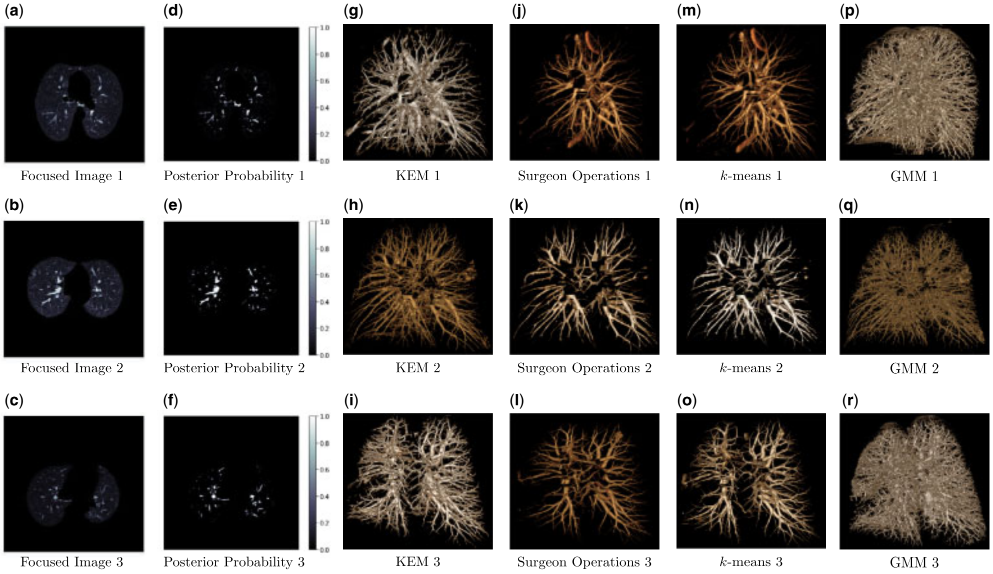


Figure 3. Graphical displays of blood vessels. (a)–(c) show the concentrated 2D slices of the randomly selected CTs. (d)–(f) show 2D slices of the posterior probabilities of the randomly selected CTs. (g)–(i) display the 3D reconstructions of the blood vessels by the KEM algorithm. (j)–(l) display the 3D reconstructions of the blood vessels by a surgeon. (m)–(o) display the 3D reconstructions of the *k*-means method. (p)–(r) display the 3D reconstructions of the GMM method.

posterior probabilities with respect to the blood vessels are of primary interest. See Fig. 3(d)–(f) for a visual representation of the 2D slices. The original lung-shaped area in Fig. 3(a)–(c) no longer obstructs the dance of the blood vessels in Fig. 3(d)–(f). Moreover, the blood vessels are successfully highlighted, whereas all other irrelevant human tissue is discarded. Subsequently, the estimated posterior probabilities with respect to the blood vessels are rescaled back to the original data range and saved as a Dicom file for the subsequent 3D reconstruction. We then utilize RadiAnt software to load the Dicom file and then reconstruct the blood vessels; see Fig. 3(g)–(i) for example. As shown, the blood vessels in the lung are preserved completely, while all the other human tissue is beautifully erased.

Additionally, we ask the third author of this work (an experienced surgeon working in one of the major hospitals in Beijing, P.R. China) to manually reconstruct the blood vessels. To ensure a fair basis for comparisons, the surgeon initiated the reconstruction by using the same concentrated CT as those employed in the KEM method. The reconstruction results are visualized in Fig. 3(j)–(l). Moreover, we used both the *k*-means and GMM methods for blood vessel reconstruction. The respective outcomes are visually presented in Fig. 3(m)–(r). First, the GMM method seems over-detailed. Its 3D reconstruction results are distinct from those of the surgeon's manual operations. A simulation with ground truth blood vessels generated from the result of the KEM method is given in Appendix D of the Supplementary Materials. Although the ground truth blood vessels are not over-detailed, the result of the GMM method is still over-detailed. This confirms that the excess details of the GMM method are largely due to estimation errors. Second, the 3D reconstruction results of the KEM method, the *k*-means method, and the surgeon's operations are similar to each other. However, the connectivity quality of blood vessels achieved through the KEM method appears superior to that of the *k*-means method and the surgeon's operations. This is expected because manual operations and the *k*-means method cannot adapt to subtle spatial variations within the 3D CT space. Third, the KEM result is free of both the disconnected problem and the over-detailed problem; see Fig. 3(g)–(i) for example. This is because KEM is a more accurate estimation method

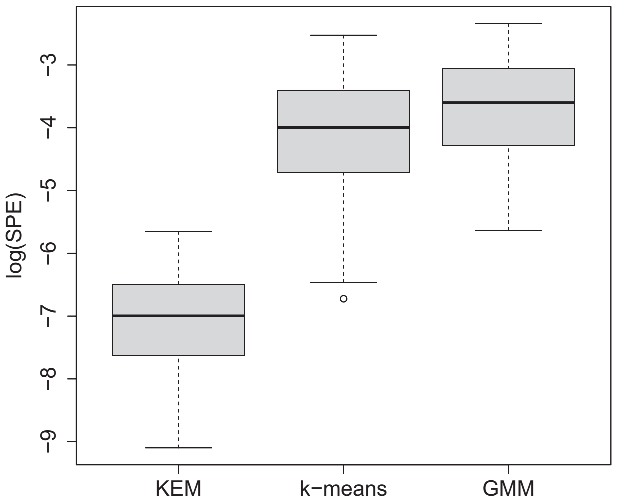


Figure 4. Logarithm of the SPE results of the KEM, *k*-means, and GMM methods.

due to its adaptability to spatially varying parameters in the CT space. This makes KEM a more reliable method.

Furthermore, we also conduct a comparative analysis for prediction in terms of the SPE for the 882 CT files in the LIDC-IDRI dataset, using the KEM, *k*-means, and GMM methods on the testing set. The SPE metric is computed as $\text{SPE} = |\mathcal{I}_1|^{-1}\sum_{i\in\mathcal{I}_1}(Y_i - \hat{Y}_i)^2$, where $\mathcal{I}_1$ denotes the testing set, Y_i denotes the CT value, and $\hat{Y}_i$ denotes the predicted CT value by one particular method (i.e. the KEM, *k*-means, and GMM methods). For the KEM method, the predicted CT value is given by $\hat{Y}_i = \sum_{m=1}^{M}\hat{\pi}_m(X_i)\hat{\mu}_m(X_i)$, for which both $\hat{\pi}_m(X_i)$ and $\hat{\mu}_m(X_i)$ are bandwidth-dependent estimates. In contrast, for the other two competing methods (i.e. the *k*-means and GMM methods), the predicted CT value is computed by $\hat{Y}_i = \sum_{m=1}^{M}\hat{\pi}_m\hat{\mu}_m$, where $\hat{\pi}_m$ and $\hat{\mu}_m$ are the parameter estimates computed by the *k*-means method or the GMM method, which are bandwidth-independent. Here, we use the SPE metric rather than the RMSE metric because one should know in advance the values of the true parameters to compute RMSE. In simulation, we can compute the RMSE for each method because the true parameters are known to us. Unfortunately, in real data analysis, the ground truth values of the parameters are unknown. Therefore, we are unable to calculate the RMSE for the real data analysis. In contrast, SPE remains to be computable without the true parameters. Figure 4 illustrates boxplots of the logarithms of the SPE results. It is evident that the KEM method outperforms both the *k*-means and GMM methods. In Appendix D of the Supplementary Materials, we provide example slices of the reconstructed responses of the three methods. We find that the reconstructed responses of the KEM method outperform the other alternatives.

4. CONCLUDING REMARKS

In summary, we developed a semiparametric GMM tailored for an innovative application of 3D blood vessel reconstruction. This work contributes significantly to both the literature on statistical theory and medical imaging applications. From a theoretical standpoint, we introduced a semiparametric extension to the classical GMM. We also developed a KEM algorithm for model estimation. To underpin this methodology, we established a rigorous asymptotic theory. Additionally, we devised a regression-based approach for optimal bandwidth selection, which surpasses the traditional cross-validation method in both computational and statistical efficiency. On the application front, we addressed a critical challenge in medical imaging—3D blood vessel reconstruction—by defining it within a statistical framework of semiparametric Gaussian mixtures. In conclusion, we would like to discuss a topic for future research. The proposed method allows only a fixed number of mixture components. However, the human body is a highly sophisticated system with many different organs and tissues. Therefore, exploring the extension of the current method to accommodate a diverging number of mixture components is a topic worthy of future investigation.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

J.Z.'s research is supported in part by the National Natural Science Foundation of China (No. 72171226) and the National Statistical Science Research Project (No. 2023LD008). Y.J.'s research is supported in part by the National Natural Science Foundation of China (No. 62306189). H.W.'s research is partially supported by National Natural Science Foundation of China (No. 12271012).

Conflict of interest statement. None declared.

DATA AVAILABILITY

Software in the form of R and Python codes, together with sample input data and complete documentation are available online at `https://github.com/Helenology/Paper_KEM`.

REFERENCES

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977:39(1):1–22.

Fan J, Gijbels I. Local polynomial modelling and its applications. New York (NY): Springer; 1996.

Fan J, Farmen M, Gijbels I. Local maximum likelihood estimation and inference. J R Stat Soc B. 1998:60(3): 591–608.

Hagiwara M, Shimada Y, Kato Y, Nawa K, Makino Y, Furumoto H, Akata S, Kakihana M, Kajiwara N, Ohira T, et al. High-quality 3D image simulation for pulmonary lobectomy and segmentectomy: results of preoperative assessment of pulmonary vessels and short-term surgical outcomes in consecutive patients undergoing video-assisted thoracic surgery. Eur J Cardio-Thoracic Surg. 2014:46(6):e120–e126.

Ji Y, Zhang T, Yang L, Wang X, Qi L, Tan F, Daemen JHT, de Loos ER, Qiu B, Gao S. The effectiveness of three-dimensional reconstruction in the localization of multiple nodules in lung specimens: a prospective cohort study. Transl Lung Cancer Res. 2021:10(3):1474–1483.

Lang S. Real and functional analysis, volume 142. Springer, New York: Springer Science & Business Media; 2012.

Li Q, Racine JS. Nonparametric econometrics: theory and practice. Princeton, New Jersey: Princeton University Press; 2007.

Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE Trans Neural Netw Learn Syst. 2019:30(11):3484–3495.

MacQueen J. et al. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. Oakland, CA, USA, p. 281–297.

McLachlan GJ, Peel D. Finite mixture models. New York: Wiley; 2000.

Molteni R. Prospects and challenges of rendering tissue density in hounsfield units for cone beam computed tomography. Oral Surg Oral Med Oral Pathol Oral Radiol. 2013:116(1):105–119.

Nadaraya E. On non-parametric estimates of density functions and regression curves. Theory Probab Appl. 1965:10:186–190.

Setio AAA, Traverso A, de Bel T, Berens MS, van den Bogaard C, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Med Image Anal. 2017:42:1–13.

Silverman BW. Density estimation for statistics and data analysis. New York: Chapman & Hall; 2018.

Ullah A, Pagan A. Nonparametric econometrics. New York: Cambridge University Press; 1999.

Watson GS. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A. 1964:359–372.