Biostatistics (2019) **20**, 4, *pp.* 565–581
doi:10.1093/biostatistics/kxy019
Advance Access publication on May 18, 2018

A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images

QIWEI LI

Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA

XINLEI WANG

Department of Statistics, Southern Methodist University, Dallas, TX, USA

FAMING LIANG

Department of Statistics, Purdue University, West Lafayette, IN, USA

FALIU YI, YANG XIE

Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA

ADI GAZDAR

Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA and Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, TX, USA

GUANGHUA XIAO*

Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA
guanghua.xiao@utsouthwestern.edu

Summary

Digital pathology imaging of tumor tissues, which captures histological details in high resolution, is fast becoming a routine clinical procedure. Recent developments in deep-learning methods have enabled the identification, characterization, and classification of individual cells from pathology images analysis at a large scale. This creates new opportunities to study the spatial patterns of and interactions among different types of cells. Reliable statistical approaches to modeling such spatial patterns and interactions can provide insight into tumor progression and shed light on the biological mechanisms of cancer. In this article, we consider the problem of modeling a pathology image with irregular locations of three different types of cells: lymphocyte, stromal, and tumor cells. We propose a novel Bayesian hierarchical model, which incorporates a hidden Potts model to project the irregularly distributed cells to a square lattice and a Markov random field prior model to identify regions in a heterogeneous pathology image. The model allows us to quantify the interactions between different types of cells, some of which are clinically meaningful. We use Markov chain Monte Carlo sampling techniques, combined with a double Metropolis–Hastings algorithm, in order to simulate samples approximately from a distribution with an intractable normalizing constant. The proposed model was applied to the

*To whom correspondence should be addressed.

pathology images of 205 lung cancer patients from the National Lung Screening trial, and the results show that the interaction strength between tumor and stromal cells predicts patient prognosis ($P = 0.005$). This statistical methodology provides a new perspective for understanding the role of cell–cell interactions in cancer progression.

Keywords: Double Metropolis–Hastings; Hidden Potts model; Lung cancer; Markov random field; Mixture model; Pathology image; Potts model; Spatial point pattern.

1. Introduction

Cancer is a group of diseases characterized by the uncontrolled growth of tumor cells that can occur anywhere in the body. Current guidelines for diagnosing and treating cancer are largely based on pathological examination of hematoxylin and eosin (H&E)-stained formalin-fixed paraffin-embedded tissue section slides. A tumor pathology image harbors a large amount of information, such as growth patterns and interactions between tumor cells and the surrounding micro-environment. Cell growth pattern is associated with the survival outcome of cancer patients (Amin *and others*, 2002; Gleason *and others*, 2002; Borczuk *and others*, 2009; Barletta *and others*, 2010). Furthermore, a recent study shows that the cell growth pattern in tumor tissues predicts treatment response in lung cancer patients (Tsao *and others*, 2015). Different cell types, including lymphocyte (a type of immune cell), stromal, and tumor cells, are commonly seen in tumor tissue images. The interactions among these cells play vital roles in the progression and metastasis of cancer (Mantovani *and others*, 2002; Orimo *and others*, 2005; Merlo *and others*, 2006; Polyak *and others*, 2009; Hanahan and Weinberg, 2011; Gillies *and others*, 2012; Junttila and de Sauvage, 2013). Spatial variations among cell types and their association with patient prognosis have been previously reported in breast cancer (Mattfeldt *and others*, 2009). However, there is a lack of rigorous statistical methods to quantify the cell interactions due to the high complexity and heterogeneity of the disease.

With the advance of imaging technology, H&E-stained pathology imaging is becoming a routine clinical procedure. This process produces massive digital pathology images that capture histological details in a high resolution. Therefore, developing statistical methods for tumor pathology images has become essential to utilize the high-resolution images for patient prognosis and treatment planning. Recent studies have demonstrated the feasibility of using digital pathology image analysis to assist pathologists in clinical diagnosis and prognosis (Beck *and others*, 2011; Yuan *and others*, 2012; Luo *and others*, 2016; Yu *and others*, 2016). Furthermore, the application of computer vision and machine learning techniques allows for the identification and classification of individual cells in digital pathology image analysis (Yuan *and others*, 2012). Recent developments in deep-learning methods have greatly facilitated this process. We have developed a ConvPath pipeline (Figure S1 of supplementary material available at *Biostatistics* online, manuscript under review), which uses a convolutional neural network (CNN) to identify individual cells and predict the cell types (https://qbrc.swmed.edu/projects/cnn/). The CNN was trained using a large cohort of lung cancer pathology images manually labeled by pathologists. This pipeline can process tumor tissue images and determine the cell type and location for each individual cell. It creates new opportunities to study the spatial patterns of and interactions among different types of cells, which may reveal important information about tumor cell growth and its micro-environment. Spatial models, such as the Ising model and Potts model, have been used to extract spatial information for imaging data (Green and Richardson, 2002; Li, 2009; Ayasso and Mohammad-Djafari, 2010). Recently, Li *and others* (2017) proposed a variant of the Potts model to study pathology images, assuming that cell–cell interactions are homogeneous across the whole image. However, tumor cell growth patterns and their surrounding micro-environments are heterogeneous and vary across different spatial locations (see, e.g. Schnipper, 1986; Kirk, 2012; Longo, 2012; Shibata, 2012; Marte, 2013; McGranahan and Swanton, 2017).

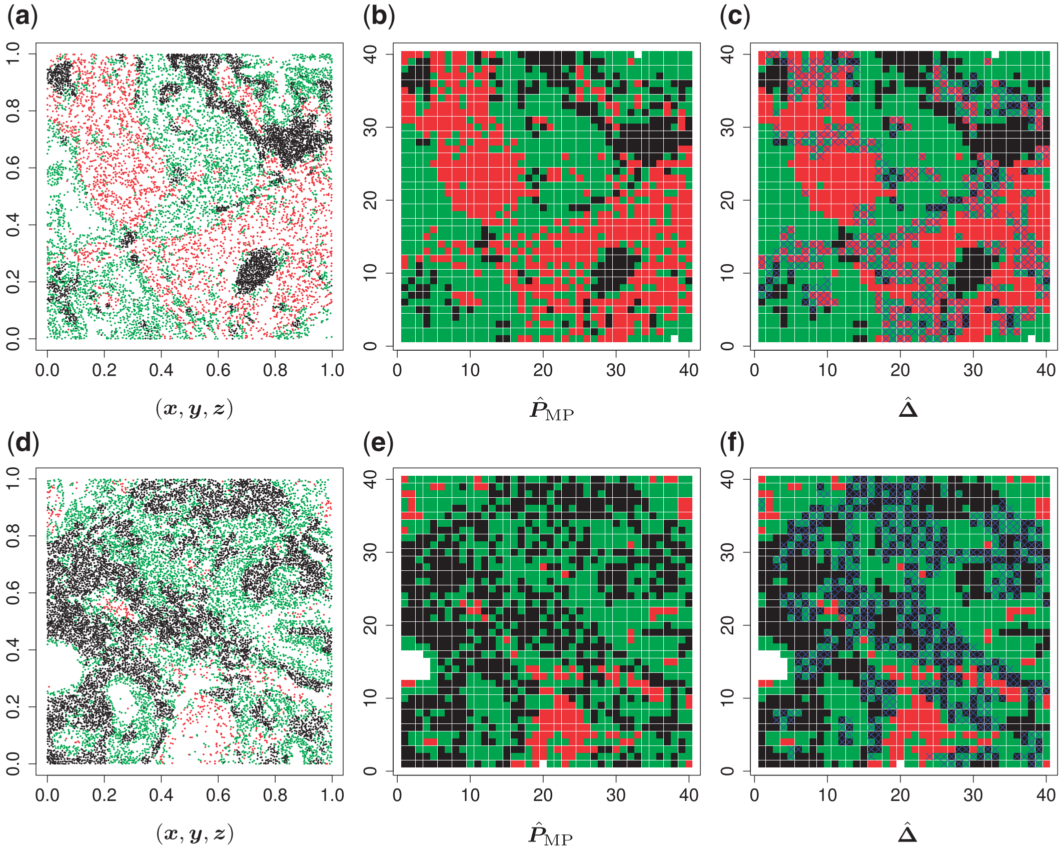

Fig. 1. (a and d) The observed cell distribution maps for two sample images from different patients, where lymphocyte, stromal, and tumor cells are marked in black, red, and green, respectively. (b and e) The estimated hidden spins $\hat{\boldsymbol{P}}_{\mathrm{MP}}$ in the 40-by-40 lattice. (c and f) The estimated AOI indicators $\hat{\boldsymbol{\Delta}}$ by choosing $c = 0.5$ (median model), where the blue indicates $\hat{\delta}_{lw} = 1$. For the first example, of which $\hat{\boldsymbol{\Delta}}$ as shown in (c), $\hat{\theta}_{\mathrm{lym,str}} = -1.036$, $\hat{\theta}_{\mathrm{lym,tum}} = 0.245$, $\hat{\theta}_{\mathrm{str,tum}} = 0.666$, $\hat{\theta}_{0,\mathrm{lym,str}} = -2.039$, $\hat{\theta}_{0,\mathrm{lym,tum}} = -1.039$, $\hat{\theta}_{0,\mathrm{str,tum}} = -1.146$; for the second example, of which $\hat{\boldsymbol{\Delta}}$ as shown in (f), $\hat{\theta}_{\mathrm{lym,str}} = -2.052$, $\hat{\theta}_{\mathrm{lym,tum}} = 0.408$, $\hat{\theta}_{\mathrm{str,tum}} = -0.641$, $\hat{\theta}_{0,\mathrm{lym,str}} = -3.292$, $\hat{\theta}_{0,\mathrm{lym,tum}} = -0.791$, $\hat{\theta}_{0,\mathrm{str,tum}} = -0.762$. Note that in the bottom-left of (d), (e), and (f), the empty region is the alveolus.

In this article, we develop a novel Bayesian hidden Potts mixture model for the cell distribution maps, such as Figure 1(a) and (d), generated by our ConvPath pipeline. The proposed model has several advantages. First, it incorporates a hidden Potts model that projects irregularly distributed cells into a square lattice, which significantly reduces the complexity of the unstructured spatial data. Second, it integrates a Markov random field (MRF) prior model that accounts for the heterogeneity across the image, while partitioning the image into multiple regions with homogeneous cell–cell interactions. The interaction parameters of the Potts model, also called interaction energies, can be used to characterize the strengths of the spatial interactions among different types of cells. The double Metropolis–Hastings (DMH) algorithm (Liang, 2010) is used to sample from the posterior distribution with an intractable normalizing constant in the Potts model. The model performed well in simulated studies.

The proposed model was applied to the 1585 pathology images of 205 lung cancer patients from the National Lung Screening Trial (NLST), and the results show that the interaction strength between tumor and stromal cells is significantly associated with patient prognosis ($P = 0.005$). This statistical methodology provides a new perspective for understanding the role of cell–cell interactions in cancer progression. Last but not least, although this article is motivated by the analysis of tumor pathology images, the proposed model is generally applicable for other types of data from heterogeneous marked point processes. This article, to our best knowledge, is the first attempt to develop a rigorous statistical framework to model the heterogeneous spatial interactions among different types of cells in tumor pathology images.

The remainder of the article is organized as follows. Section 2 introduces the proposed modeling framework and discusses the MRF prior formulation. Section 3 describes the Markov chain Monte Carlo (MCMC) algorithm and discusses the resulting posterior inference. Section 4 assesses performance of the proposed model on simulated data and the results of a lung cancer case study. Section 5 concludes the article with some remarks on future research directions.

2. Models

We first review the Potts model and its interaction energy measurement in Section 2.1, and then we introduce the hidden Potts model in Section 2.2 and the hidden Potts mixture (HPM) model in Section 2.3. The graphical formulation of the HPM model is presented in Figure S2 of supplementary material available at *Biostatistics* online.

2.1. *The Potts model*

Let $G = (V, E)$ denote a graph G composed of a finite set V of vertices and a set E of edges joining pairs of vertices. In statistics, the Potts model can be considered as an undirected graph such that each vertex $v \in V$ is geometrically regular assigned on a lattice (e.g. square, triangular, or honeycomb lattice) and each edge $e \in E$ is at the same distance. Particularly, for an L-by-W square lattice graph in a Cartesian coordinate system, let (l, w), $1 \leq l \leq L$, $1 \leq w \leq W$ denote the coordinate of each vertex. Then, edges are connected between the vertex at location (l, w) and its four neighbors at locations $(l+1, w)$, $(l-1, w)$, $(l, w+1)$, and $(l, w-1)$, if applicable. Every vertex will be assigned a *spin*, which is defined as an assignment of Q $(Q \geq 2)$ different classes. When $Q = 2$, the Potts model is known as the Ising model. Let an L-by-W matrix $\boldsymbol{P}$ denote the collection of all spins, where each element $p_{lw} \in \{1, \ldots, Q\}$. Since the vertices are assigned different spins and react with their neighbors' spins, there will be some measurement of overall energy, named *Hamiltonian*,

$$H(\boldsymbol{P}|\boldsymbol{\theta}) = -\sum_{l=1}^{L}\sum_{w=1}^{W}\sum_{(l',w')\in\text{Nei}_{(l,w)}}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\theta_{qq'}I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q'), \tag{2.1}$$

where $\text{Nei}_{(l,w)}$ denotes the coordinate set of neighbors of vertex (l, w), $\theta_{qq'}$ denotes the interaction energy between adjacent vertices, and I denotes the indicator function. Note that $\theta_{qq'} = \theta_{q'q}$. According to Equation (2.1), only those edges between vertices that have different spins are counted. The negative value of $H(\boldsymbol{P}|\boldsymbol{\theta})$ can be also considered as the weighted average of those edges connecting two different spins among the graph.

The Potts model probability mass function calculates the probability of observing the lattice in a particular state $\boldsymbol{P}$, where a *state* is a choice of spin at each vertex,

$$\begin{aligned}\Pr(\boldsymbol{P}|\boldsymbol{\theta}) &= \frac{\exp(-H(\boldsymbol{P}|\boldsymbol{\theta}))}{\sum_{\boldsymbol{P}'\in\mathcal{P}}\exp(-H(\boldsymbol{P}'|\boldsymbol{\theta}))}\\ &= \frac{1}{C(\boldsymbol{\theta})}\exp\left(\sum_{l=1}^{L}\sum_{w=1}^{W}\sum_{(l',w')\in\text{Nei}_{(l,w)}}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\theta_{qq'}I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q')\right).\end{aligned} \tag{2.2}$$

Here, $\boldsymbol{\theta}$ denotes the collection of interaction energy parameters between different classes, where each element $\theta_{qq'} \in \mathbb{R}, q = 1, \ldots, Q, q' = 1 \ldots, Q, q \neq q'$ and $\mathcal{P}$ denotes the set of all possible states of the lattice. An exact evaluation of the normalizing constant $C(\boldsymbol{\theta})$ needs to sum over the entire space of $\boldsymbol{P}$, which consists of Q^{LW} states. Thus, $C(\boldsymbol{\theta})$ is intractable even for a small size model. Take $Q = 2$, $L = W = 10$ for example, it needs to sum over $2^{100} \approx 1.268 \times 10^{30}$ elements. To address this issue, we will employ the DMH algorithm (Liang, 2010) to estimate $\boldsymbol{\theta}$ without calculating $C(\boldsymbol{\theta})$, which will be illustrated in Section 3. Equation (2.2) serves as the full likelihood of the Potts model, which satisfies the Markov property. Therefore, we can write the probability of observing $p_{lw} = q$ conditional on its neighbor spins, which is

$$\Pr(p_{lw} = q|\boldsymbol{\theta}, \boldsymbol{P}_{-l,-w}) = \frac{\exp\left(\sum_{(l',w')\in \text{Nei}_{(l,w)}} \sum_{q'=1}^{Q} \theta_{qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q')\right)}{\sum_{q'=1}^{Q} \exp\left(\sum_{(l',w')\in \text{Nei}_{(l,w)}} \sum_{q'=1}^{Q} \theta_{qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q')\right)}. \tag{2.3}$$

Here, we use $\boldsymbol{P}_{-l,-w}$ to denote all the spins excluding p_{lw}. According to Equation (2.3), the conditional probability that we observe the vertex (l, w) belonging to class q depends on the interaction energy parameters $\theta_{qq'}, q' = 1, \ldots, Q, q' \neq q$ and the number of edges connecting two different spins. The larger the value of $\theta_{qq'}$, the more likely that p_{lw} is discordant with the majority of its four neighboring spins.

2.2. *The hidden Potts model*

Potts models have a wide range of applications in many areas since they provide an appealing representation of images and other types of spatial data. However, for images with irregularly distributed dots, it is impossible to apply a Potts model based on a lattice that forms a regular tiling. To overcome this limitation, Li *and others* (2017) proposed a hidden Potts model by introducing an auxiliary lattice to the image and defining a pre-specified projection parameter to control the similarity between the imputed lattice image and the original image. We develop a more flexible hidden Potts model by formulating a prior on the projection parameter. More importantly, our model takes into account the heterogeneity of the imaging data.

We consider a preprocessed pathology image, as shown in Figure 1(a) and (d), with n observed cells, where (x_i, y_i) represents the location and z_i indicates the type of cell i. We denote $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$ as the collection of x_i, y_i, and z_i, $i = 1, \ldots, n$, respectively. Let an L-by-W matrix $\boldsymbol{P}$ denote the hidden spins at the auxiliary lattice, which partitions the image into $(L-1)(W-1)$ squares. The ratio L/W should approximate the ratio of $(x_{\text{upp}} - x_{\text{lwr}})/(y_{\text{upp}} - y_{\text{lwr}})$, where x_{lwr} and x_{upp} denote the lower and upper bounds of the horizontal axis, and y_{lwr} and y_{upp} denote the lower and upper bounds of the vertical axis of the image. The bounds are usually known; if not, they can be estimated from the data itself by: (1) roughly setting $x_{\text{lwr}} = \min\{\boldsymbol{x}\}$, $x_{\text{upp}} = \max\{\boldsymbol{x}\}$, $y_{\text{lwr}} = \min\{\boldsymbol{y}\}$, and $y_{\text{upp}} = \max\{\boldsymbol{y}\}$ or (2) computing the Ripley–Rasson estimator (Ripley and Rasson, 1977) of a rectangle window given $(\boldsymbol{x}, \boldsymbol{y})$. To fit the imaging data into the square-lattice system, we normalize each coordinate (x_i, y_i) by performing a linear transformation $x'_i = 1 + \frac{x_i - x_{\text{lwr}}}{x_{\text{upp}} - x_{\text{lwr}}}(L-1)$ and $y'_i = 1 + \frac{y_i - y_{\text{lwr}}}{y_{\text{upp}} - y_{\text{lwr}}}(W-1)$. Then we assume that the probability of assigning cell i to type q conditional on its adjacent spins at the hidden lattice is

$$\Pr(x'_i, y'_i, z_i = q|\boldsymbol{P}, d) = \frac{\exp\left(d \sum_{\{(l,w): l \le x'_i < l+1, w \le y'_i < w+1\}} I(p_{lw} = q)\right)}{\sum_{q'=1}^{Q} \exp\left(d \sum_{\{(l,w): l \le x'_i < l+1, w \le y'_i < w+1\}} I(p_{lw} = q')\right)}, \tag{2.4}$$

where d is the projection parameter. The larger the value of d, the more likely a cell type is the same as the majority of its adjacent spins. If $d = 0$, then $\Pr(x'_i, y'_i, z_i = q|\boldsymbol{P}, d) = 1/Q$, which means the spatial

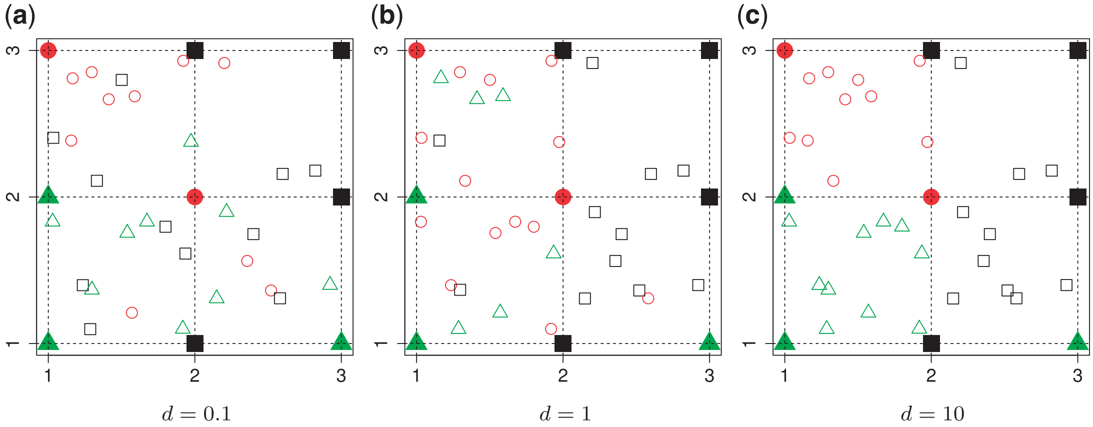

Fig. 2. Illustration of a 3-by-3 auxiliary lattice and the point emission process by different choices of d. The empty points represent the observed cells and the filled points represent the hidden spins in the lattice. The square, circle, and triangle shapes stand for class $q = 1$, 2, and 3, respectively.

distribution of the random cells is independent from the hidden spins. Figure 2 illustrates an example of the point emission process from a given 3-by-3 Potts model under different choices of d. In addition, we also demonstrate how d yields varying spin assignments of the hidden lattice conditional on the observed cells from a larger dataset in our simulation study (see Figure S5 of supplementary material available at *Biostatistics* online). Note that the cell types are independent and identically distributed within each lattice unit, conditional on the four nearby spins. Although this assumption is suitable for most cases, we discuss a location-dependent hidden Potts model in Section S1 of supplementary material available at *Biostatistics* online.

To summarize, the joint likelihood function of the hidden Potts model can be written as

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \boldsymbol{P}, d, \boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \boldsymbol{P}, d) f(\boldsymbol{P} | \boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(x_i', y_i', z_i = q | \boldsymbol{P}, d) \Pr(\boldsymbol{P} | \boldsymbol{\theta}), \tag{2.5}$$

where $\Pr(x_i', y_i', z_i = q | \boldsymbol{P}, d)$ and $\Pr(\boldsymbol{P} | \boldsymbol{\theta})$ are given in Equations (2.4) and (2.2), respectively. The inference of spin p_{lw} depends on: (1) the nearby observed cells within the square area included by coordinates $(l+1, w+1)$, $(l+1, w-1)$, $(l-1, w+1)$, and $(l-1, w-1)$; (2) the nearby spins at locations $(l, w+1)$, $(l, w-1)$, $(l+1, w)$, $(l-1, w)$; and (3) the underlying interaction parameters $\boldsymbol{\theta}$. We specify the prior distribution for d as $d \sim \text{Ga}(a_d, b_d)$. One standard way of setting a weakly informative gamma prior is to choose small values for the two parameters, such as $a_d = b_d = 0.001$ (Gelman, 2006). We conclude this subsection by discussing how to choose the tunable parameters L and W, which determine the size of the auxiliary lattice. Larger values of L and W make the inference computationally expensive, while small values make a rough approximation to the interaction energy. We suggest choosing the values that correspond to $n \approx gLW$, where g can be any integer between 4 and 10. This constraint generally requires observing g cells in each square.

2.3. *A proposal of the hidden Potts mixture model*

Tumor tissues are heterogeneous. Spatial patterns of cell distributions may vary across different spatial regions. Applying a homogenous model described in Section 2.2 may obscure the recovery of the true values of interaction energy by averaging out the real signal from the “areas of interest” with other background areas. In this study, we propose a hidden Potts mixture model in order to take into account the heterogeneity of spatial point patterns observed in the pathology images. We first discuss the general case when the number of mixture components $K \geq 2$ and then focus on the model when $K = 2$. From a biological point of view, these two regions can be referred to respectively: (1) The areas of interests (AOIs) with high interaction strengths among different types of cells. In AOIs, the different types of cells are highly mixed, which may reveal important information about cancer progression and (2) The background areas, in which the same type cells are aggregated into clusters.

In the same auxiliary lattice described in Section 2.2, we envision that there are K homogeneous regions with different interaction parameter settings, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. With this assumption, an L-by-W latent matrix $\boldsymbol{\Delta}$ is introduced to indicate the K distinct regions, with $\delta_{lw} = k - 1$ if the spin at location (l, w) belongs to group k. According to Equation (2.2), the probability mass function of the mixture model given the partition $\boldsymbol{\Delta}$ can be written as,

$$\Pr(\boldsymbol{P}|\boldsymbol{\Delta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \\ = \prod_{k=1}^{K} \frac{1}{C_k(\boldsymbol{\theta}_k)} \exp\left(\sum_{\{(l,w):\delta_{lw}=k-1\}} \sum_{(l',w')\in \mathrm{Nei}_{(l,w)}} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{kqq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q') \right), \tag{2.6}$$

where $C_k(\boldsymbol{\theta}_k)$ is the normalizing constants for region k. To encourage two neighbor spins to be more likely in the same region (i.e. have the same δ value), we incorporate the spatial dependency structure into the prior on $\boldsymbol{\Delta}$ via a Potts model that satisfies a spatial Markov property. This prior model is a type of MRF, where the distribution of a set of random variables follows Markov properties that can be presented by an undirected graph. In our model, this graph is defined by the L-by-W auxiliary lattice. The prior can be written as

$$\Pr(\delta_{lw} = k - 1|\boldsymbol{\Delta}_{-l,-w}) = \frac{\exp\left(f \sum_{(l',w')\in \mathrm{Nei}_{(l,w)}} I(\delta_{l',w'} = k - 1)\right)}{\sum_{k'=1}^{K} \exp\left(f \sum_{(l',w')\in \mathrm{Nei}_{(l,w)}} I(\delta_{l',w'} = k' - 1)\right)}, \tag{2.7}$$

where f is a non-negative parameter that controls the spatial interaction and $\boldsymbol{\Delta}_{-l,-w}$ denotes the set of $\delta_{l'w'}$'s excluding $\delta_{l,w}$. A large value of f makes largely clustered configurations of $\boldsymbol{\Delta}$, while a small value corresponds to patterns that do not display any sort of spatial organization. Although the choice of f is very tricky, François *and others* (2006) suggests that the value $f = 1$ can be considered a high level of spatial interaction for $3 \leq K \leq 6$.

When $K = 2$, $\boldsymbol{\Delta}$ becomes a binary latent matrix that indicates the two distinct regions, with $\delta_{lw} = 0$ if the spin at location (l, w) belongs to the background area, and $\delta_{lw} = 1$ if the spin at location (l, w) belongs to the AOI. Let $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$ denote the interaction parameters in the background and AOI regions, respectively. According to Equation (2.6), the probability mass function of the two-component mixture model is written as,

$$\Pr(\boldsymbol{P}|\boldsymbol{\Delta}, \boldsymbol{\theta}_0, \boldsymbol{\theta}) \\ = \frac{1}{C_0(\boldsymbol{\theta}_0)} \exp\left(\sum_{\{(l,w):\delta_{lw}=0\}} \sum_{(l',w')\in \mathrm{Nei}_{(l,w)}} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{0qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q') \right) \\ \times \frac{1}{C(\boldsymbol{\theta})} \exp\left(\sum_{\{(l,w):\delta_{lw}=1\}} \sum_{(l',w')\in \mathrm{Nei}_{(l,w)}} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q') \right), \tag{2.8}$$

where $C_0(\boldsymbol{\theta}_0)$ and $C(\boldsymbol{\theta})$ are the normalizing constants for the two regions. The prior model reduces to an Ising model, characterized by the following probability,

$$\Pr(\delta_{lw}|\boldsymbol{\Delta}_{-l,-w}) = \frac{\exp\left(\delta_{lw}\left(e+f\sum_{(l',w')\in \mathrm{Nei}_{(l,w)}}\delta_{l',w'}\right)\right)}{1+\exp\left(\delta_{lw}\left(e+f\sum_{(l',w')\in \mathrm{Nei}_{(l,w)}}\delta_{l',w'}\right)\right)}, \tag{2.9}$$

where e and f are hyperparameters to be chosen. Compared with Equation (2.7), the extra parameter e controls the number of 1's in $\boldsymbol{\Delta}$ (i.e. the number of spins belonging to the AOI), while f affects the probability of assigning a value according to its neighbor spins. Note that if a vertex does not have any neighbor in the AOI, its prior probability reduces to an independent Bernoulli prior with parameter $\exp(e)/(1+\exp(e))$, which is a logistic transformation of e. Although the parameterization is somewhat arbitrary, some care is needed in deciding the value of f. In particular, a large value of f may lead to a phase transition problem; that is, the expected number of ones in $\boldsymbol{\Delta}$ can increase massively for small increments of f. This problem can happen because Equation (2.9) can only increase as a function of the number of δ_{lw}'s equal to 1. An empirical estimate of the phase transition value can be obtained using the algorithm proposed by Propp and Wilson (1996) and the values of e and f can then be chosen accordingly. In this article, we treat e and f as fixed hyperparameters, following the articles by Li and Zhang (2010) and Stingo *and others* (2013). Table S1 of supplementary material available at *Biostatistics* online lists our recommendation for e and f based on the size of the AOI *a priori*. As for f, any value between 0 and $f_{\max}$, as shown in Table S1 of supplementary material available at *Biostatistics* online, can be considered, with larger values closer to the phase transition point, leading to higher prior probabilities of selection for those nodes whose neighbors are already selected. In Section 4.1, we use simulated data to investigate the sensitivity to the specification of parameters e and f. For $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, we consider normal priors, and we set $\theta_{qq'} \sim \mathrm{N}(\mu, \sigma^2)$ and $\theta_{0qq'} \sim \mathrm{N}(\mu_0, \sigma_0^2)$, where μ can be set to a positive number while μ_0 a negative number. This assumes the AOI has more interaction energy than the background *a priori*.

3. MODEL FITTING

In this section, we describe the MCMC algorithm for posterior inference. Our inferential strategy allows us to simultaneously identify the AOI while quantifying the interaction parameters.

3.1. *MCMC algorithm*

Our primary interest lies in the identification of the AOI, via the matrix $\boldsymbol{\Delta}$, and the inference of the interaction parameters within the AOI and the background area, via the vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. We design a MCMC algorithm based on the DMH (Liang, 2010) and Metropolis search variable selection algorithms (George and McCulloch, 1997; Brown *and others*, 1998) to search the model space that consists of $(\boldsymbol{P}, \boldsymbol{\Delta}, \boldsymbol{\theta}, \boldsymbol{\theta}_0, d)$. We briefly describe why and how DMH is used in the model fitting as follows. The full details of our MCMC algorithm are given in Section S2 of supplementary material available at *Biostatistics* online.

Take the update of $\theta_{qq'}$ as an example. Within each MCMC iteration, we need to sample $\theta_{qq'}$ from its conditional distribution $\pi(\theta_{qq'}|\cdot) \propto \Pr(\boldsymbol{P}|\boldsymbol{\theta})\pi(\theta_{qq'}) = \frac{1}{C(\boldsymbol{\theta})}\exp(-H(\boldsymbol{P}|\boldsymbol{\theta}))N(\theta_{qq'}; \mu, \sigma^2)$. Apparently, the Metropolis–Hastings (MH) algorithm cannot be directly applied to simulate from this distribution as the acceptance probability would involve an unknown normalizing constant ratio $C(\boldsymbol{\theta})/C(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* = \boldsymbol{\theta}$ except the proposed element $\theta^*_{qq'}$ within. To address this issue, Liang (2010) proposed an auxiliary variable MCMC algorithm, which can make the normalizing constant ratio canceled by augmenting appropriate auxiliary variables through a short run of the MH algorithm initialized with the original observation. To do so, an auxiliary variable $\boldsymbol{P}^*$ is simulated starting from $\boldsymbol{P}$ based on the new $\boldsymbol{\theta}^*$. Then, the proposed value $\theta^*_{qq'}$ will be accepted with probability $\min(1, r)$, where the Hastings ratio is computed as

$$r = \frac{\pi(\theta^*_{qq'}|\boldsymbol{P}, \boldsymbol{\theta}_{-q,-q'})}{\pi(\theta_{qq'}|\boldsymbol{P}, \boldsymbol{\theta}_{-q,-q'})} \frac{\Pr(\boldsymbol{P}^*|\boldsymbol{\theta})}{\Pr(\boldsymbol{P}^*|\boldsymbol{\theta}^*)} = \frac{\frac{1}{C(\boldsymbol{\theta}^*)}\exp(-H(\boldsymbol{P}|\boldsymbol{\theta}^*))N(\theta^*_{qq'};\mu,\sigma^2)}{\frac{1}{C(\boldsymbol{\theta})}\exp(-H(\boldsymbol{P}|\boldsymbol{\theta}))N(\theta_{qq'};\mu,\sigma^2)} \frac{\frac{1}{C(\boldsymbol{\theta})}\exp(-H(\boldsymbol{P}^*|\boldsymbol{\theta}))}{\frac{1}{C(\boldsymbol{\theta}^*)}\exp(-H(\boldsymbol{P}^*|\boldsymbol{\theta}^*))}.$$

As we can see, the unknown normalizing constant ratio has been canceled.

3.2. *Posterior estimation*

We obtain the posterior inference by post-processing of the MCMC samples after burn-in. Suppose that two sequences of MCMC samples $\theta^{(1)}_{qq'}, \ldots, \theta^{(U)}_{qq'}$ and $\theta^{(1)}_{0qq'}, \ldots, \theta^{(U)}_{0qq'}$ have been collected, where $u, u = 1, \ldots, U$ indexes the iteration after burn-in. An approximate Bayesian estimator of $\hat{\theta}_{qq'}$ and $\hat{\theta}_{0qq'}$ can be simply obtained by averaging over the samples,

$$\hat{\theta}_{qq'} = \frac{1}{U}\sum_{u=1}^{U}\theta^{(u)}_{qq'} \quad \text{and} \quad \hat{\theta}_{0qq'} = \frac{1}{U}\sum_{u=1}^{U}\theta^{(u)}_{0qq'}. \tag{3.1}$$

For $\boldsymbol{\Delta}$, we choose an estimate that relies on the marginal probability of inclusion (PPI) of single spins as the proportion of MCMC iterations in which the corresponding δ_{lw} equal to 1. That is

$$\mathrm{PPI}_\delta(l, w) = \frac{1}{U}\sum_{u=1}^{U}\delta^{(u)}_{lw}. \tag{3.2}$$

A point estimate of $\hat{\boldsymbol{\Delta}}$ is then obtained by identifying those PPI values that exceed a given cut-off c. A simple way is to choose $c = 0.5$ to obtain a median model. An alternative approach is based on a decision theoretic criterion, such as in Newton *and others* (2004), so that an expected rate of false detection (i.e. Bayesian FDR) smaller than a fixed threshold can be guaranteed. For the hidden spins $\boldsymbol{P}$, we construct the estimate by selecting the most likely q for each position (l, w):

$$\hat{\boldsymbol{P}}_{\mathrm{MP}} = \left[\hat{p}_{lw} = q\right]_{L\times W}, \tag{3.3}$$

if $\sum_{u=1}^{U} I\left(p^{(u)}_{lw} = q\right) > \sum_{u=1}^{U} I\left(p^{(u)}_{lw} = q'\right)$ for any $q' \in \{1, \ldots, Q\}, q \neq q'$. We refer to the estimate obtained in this manner as the marginal probability (MP) estimate.

3.3. *Label switching*

In our finite mixture model, the invariance of the likelihood under permutation of the component labels may result in an identifiability problem, leading to symmetric and multimodal posterior distributions with up to $K!$ copies of each “genuine" model. For instance, for the case $K = 2$, we may obtain a model where $\delta_{lw} = 1$ corresponds to the true background area, while $\delta_{lw} = 0$ indicates the true AOI. This will also complicate inference on the parameters. To solve this problem, we simply impose an identifiability constraint on the interaction parameters, $\sum_{q\neq q'}\theta_{qq'} > \sum_{q\neq q'}\theta_{0qq'}$. For a more robust approach, we can also use the relabeling algorithm proposed by Stephens (2000).

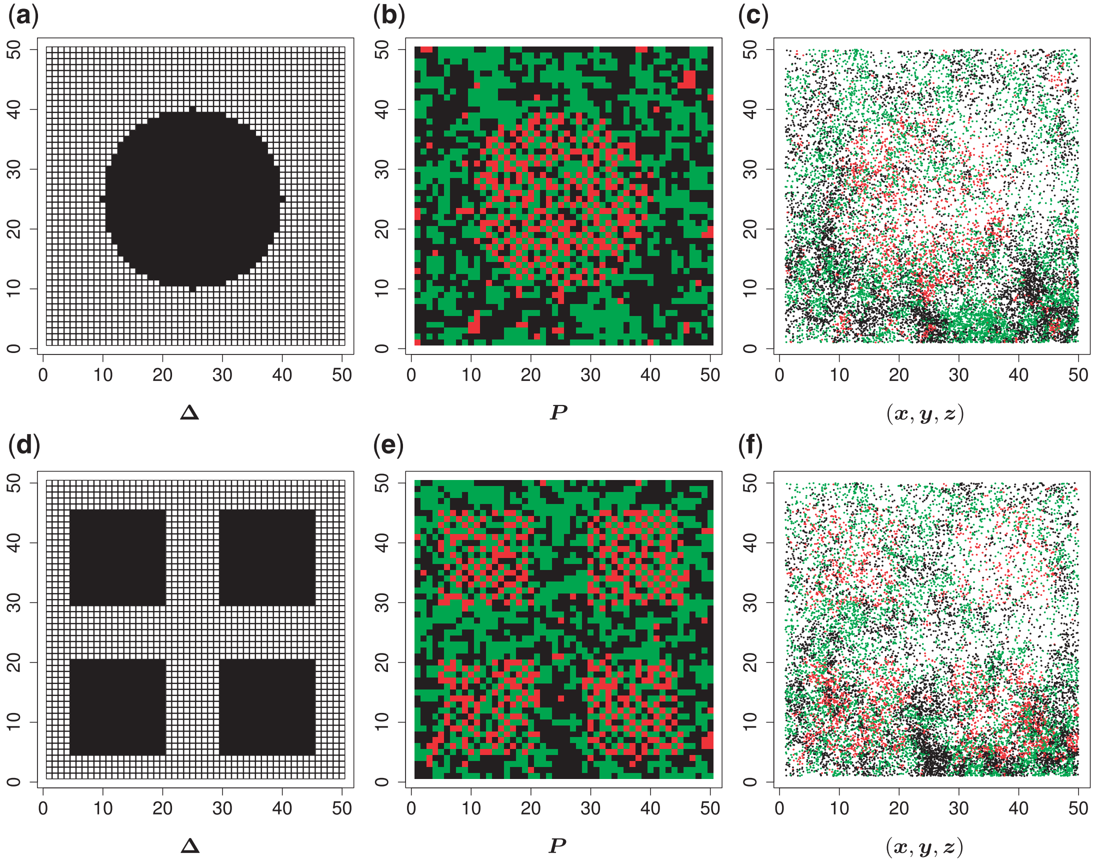

Fig. 3. (a and d) The true maps of the 50-by-50 binary matrix $\mathbf{\Delta}$ for scenarios 1 and 2, respectively. Each vertex in the lattice is represented by either an empty square if $\delta_{lw} = 0$ or a filled square if $\delta_{lw} = 1$. (b and e) The true maps of the 50-by-50 hidden spins $\boldsymbol{P}$ from one dataset generated from scenarios 1 and 2, respectively. The black, red, and green colors stand for class $q = 1$, 2, and 3, respectively. (c and f) The observed cell distribution maps generated from the log Gaussian Cox process conditional on the hidden spins, as shown in (b and e), respectively.

4. Results

We conducted simulation studies to assess the performance of the proposed Bayesian hidden Potts mixture model. The model was then applied to a large cohort of lung cancer pathology images, and it revealed novel potential imaging biomarkers for lung cancer prognosis.

4.1. *Simulation*

We used simulated data to investigate the performance of our strategy for posterior inference on the model parameters. All the generative models were based on a 50-by-50 lattice unless otherwise noted. We considered two scenarios for the true structure of $\mathbf{\Delta}$, as shown in Figure 3(a) and (d). For the first scenario the AOI is a single circle in the center of the image; for the second scenario the AOI is composed of four rectangles. The number of classes was set to $Q = 3$, and therefore, there were three interaction parameters in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. We set $\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{23}) = (0.5, 0.7, 1.0)$ and $\boldsymbol{\theta}_0 = (\theta_{012}, \theta_{013}, \theta_{023}) = (-1.0, -0.5, -1.5)$. For a given $\mathbf{\Delta}$, we first simulated the hidden spins $\boldsymbol{P}$ using the Gibbs sampler, running 100 000 iterations with random starting configurations in both AOI and the background area. Next, we considered generating the points using two different point processes: (1) a homogeneous Poisson point process with a constant intensity $\lambda = 4$ over the space $[1, 50]^2$; (2) a log Gaussian Cox process (LGCP) with an inhomogeneous intensity $\lambda(x, y) = \exp(1 + |x/50 - 0.5| + |y/50 - 0.5| + \mathcal{GP}(x, y)), x \in [1, 50], y \in [1, 50]$, where $\mathcal{GP}$ denotes a zero-mean Gaussian process with variance equal to 0.5 and scale equal to 10. Then, we assigned a class to each point according to its four nearest adjacent spins. Specifically, for point i, its class z_i was drawn from a multinomial distribution $\text{Mn}(\phi_1, \ldots, \phi_Q)$. The parameters $(\phi_1, \ldots, \phi_Q)$ were inferred from a Dirichlet distribution $\text{Dir}(0.1 + \tilde{n}_{i1}, \ldots, 0.1 + \tilde{n}_{iQ})$, where $\tilde{n}_{iq}$ denotes the number of adjacent spins that belong to class q. Mathematically, it can be written as $\tilde{n}_{iq} = \sum_{\{(l,w):l\leq x_i'<l+1,w\leq y_i'<w+1\}} I(p_{lw} = q)$. Note that this mark formulation scheme is different from the model assumption, which is given in Equation (2.4). Figure 3(c) and (f) show examples of the observed data generated from LGCP. We repeated the above steps to generate 30 independent datasets for each setting of $\boldsymbol{\Delta}$ and each point process.

For the priors on $\theta_{qq'}$ and $\theta_{0qq'}$, we used normal distributions $\text{N}(0.5, 1)$ and $\text{N}(-0.5, 1)$, respectively. We set the hyperparameters that control the MRF prior model to $e = -2.94$ and $f = 1.4$, which means that if a spin in the lattice does not have any neighbor in the AOI, its prior probability that it belongs to the AOI is 5% (For the first and second scenarios, the proportions of the AOI over the whole image are 28% and 36%, respectively). As for the gamma prior on the projection parameter d, we set $a_d = b_d = 0.001$, which leads to a vague prior for d with expectation and variance equal to 1 and 1000, respectively. This is one of the most commonly used weak gamma priors (Gelman, 2006). Results we report below were obtained by running the MCMC chain with 50 000 iterations, discarding the first 50% sweeps as burn in. We started the chain from a model by randomly choosing a 5 × 5 window in $\boldsymbol{\Delta}$ to be 1, drawing $\theta_{qq'}$ and $\theta_{0qq'}$ from their prior distributions, and assigning a random mark to each hidden spin p_{lw}. We report the scalability of our methods in Section S3 of supplementary material available at *Biostatistics* online.

Figure S4 of supplementary material available at *Biostatistics* online displays the trace plots of the interaction parameters $\boldsymbol{\theta}$ and the number of spins in the AOI from one simulated dataset generated from LGCP and scenario 2. It clearly shows that each chain converges and stabilizes around its true value in a very short run. Figure S5(a) and (b) of supplementary material available at *Biostatistics* online show the map of marginal posterior probabilities $\Pr(\delta_{lw} = 1|\cdot)$'s and the median model by choosing $c = 0.5$. Figure S5(c) of supplementary material available at *Biostatistics* online shows the map of the MP estimate of the hidden spins. It is evident from the maps that the inspection of the posterior probabilities of $\Pr(\delta_{lw} = 1|\cdot)$ and $\Pr(P_{lw} = q|\cdot)$ allows us to reconstruct the true structure of $\boldsymbol{\Delta}$ and $\boldsymbol{P}$ quite well. Figures S6 and S7 of supplementary material available at *Biostatistics* online show the density plots of MCMC samples of the six interaction parameters, collected from all 30 simulated datasets generated from each point process and scenario. Most of the true values were within the 95% credible intervals. Next we evaluated the overall performance of recovery of the true $\boldsymbol{\Delta}$, in terms of average false positive rate (FPR) and true positive rate (TPR) achieved for different values of threshold c on the posterior probabilities of inclusion $\Pr(\delta_{lw} = 1|\cdot)$'s. Results are reported in Figure 4 by drawing receiver operating characteristic (ROC) curves. The average areas under the ROC curves (AUC) range from 0.903 to 0.950, indicating a satisfactory performance. We also reported the average FPRs, TPRs (i.e. recalls), precisions, and F-1 scores of the median model in Table S2 of supplementary material available at *Biostatistics* online. Lastly, we assessed the overall performance of recovery of the true hidden spins $\boldsymbol{P}$ by plotting the ROC curves for different values of the threshold on the posterior probabilities of inclusion $\Pr(p_{lw} = q|\cdot)$ for each class q (See Figures S8 and S9 of supplementary material available at *Biostatistics* online). We compared our method with simply classifying each spin (w, l) by a k-nearest neighbor (k-NN) algorithm. The training set is the full set of the observed points in the corresponding dataset. Whichever k was chosen, the k-NN algorithm produced a (FPR, TPR) point under the ROC curve that our method generated.

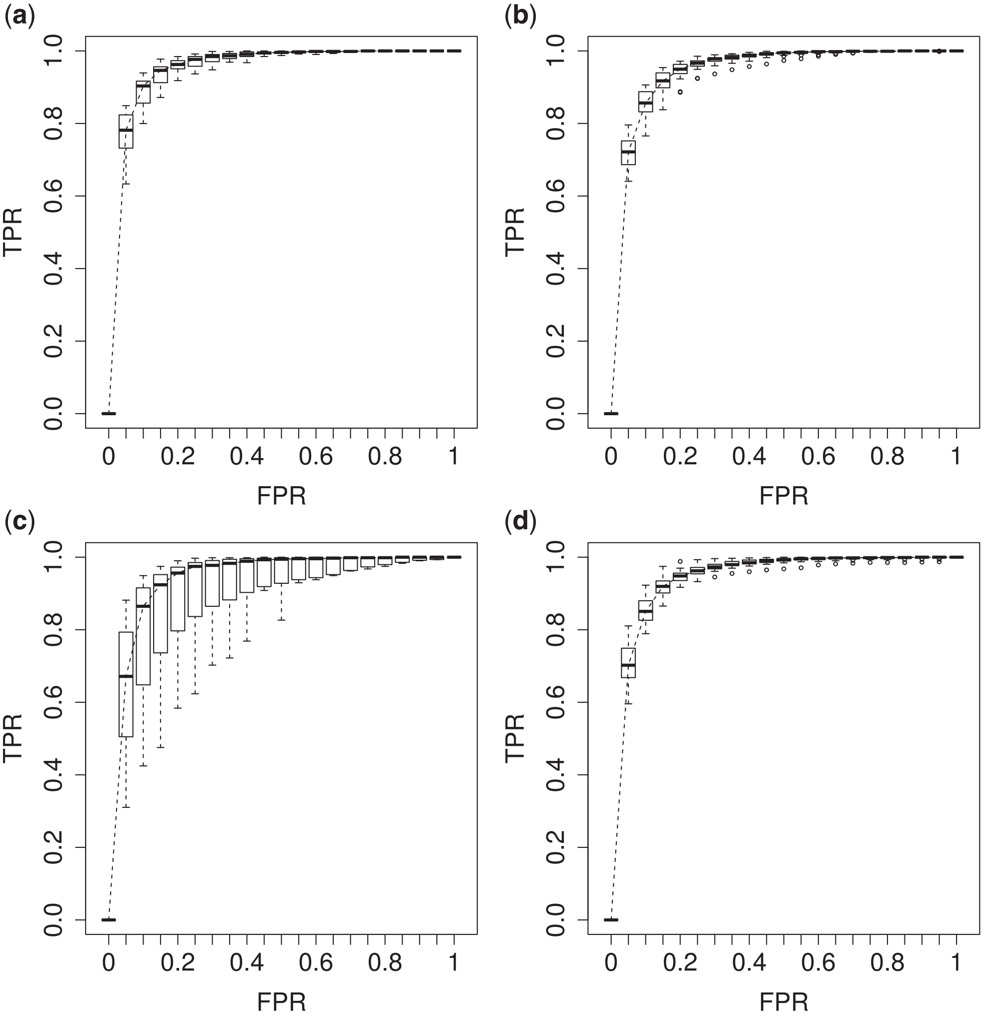

Fig. 4. The ROC curves on the posterior probabilities of inclusion on $\mathbf{\Delta}$, in terms of the boxplots of TPRs under different FPRs, over 30 datasets generated from each point process and each setting of $\mathbf{\Delta}$. (a) Homogeneous Poisson process - scenario 1; (b) Homogeneous Poisson process - scenario 2; (c) Log Gaussian Cox process - scenario 1; (d) Log Gaussian Cox process - scenario 2.

We conducted a sensitivity analysis on the choice of the MRF prior hyperparameters, e and f. In particular, we considered 20 settings, by varying e corresponding to the expected proportion of the AOI as 1%, 2%, 5%, 10%, and 15%, and varying f to $0, f_{\max}/4, f_{\max}/2$, and $f_{\max}$. Table S4 of supplementary material available at *Biostatistics* online shows the average AUC for each combination. The model is considerably insensitive to the choice of e if the value of f is equal to its maximum allowed value, as given in Table S1 of supplementary material available at *Biostatistics* online. The result suggests that employing the MRF prior model with a larger f resulted in an increased ability to identify the true AOI, while the independent Bernoulli prior model (i.e. $f = 0$) with no spatial information incorporated performed the worst. For those choices with $f = f_{\max}$, we also calculated the average recalls, precisions, and F1-scores based on the PPI estimates of $\mathbf{\Delta}$ when choosing $c = 0.5$. Again, a close look at Table S3 of supplementary material available at *Biostatistics* online suggests that the proposed model was robust to the choice of hyperparameter e.

As the algorithm consists of two tunable parameters L and W, we also conducted a sensitivity analysis by choosing different values of L and W. We fit one dataset from the homogeneous Poisson process and scenario 2 into models with different lattice sizes 20 × 20, 30 × 30, 40 × 40, and 50 × 50. Figure S10 of supplementary material available at *Biostatistics* online shows the maps of marginal posterior probabilities $\Pr(\delta_{lw} = 1|\cdot)$'s and the median models for each setting. Generally speaking, the model was robust to the choice of L and W.

4.2. *Application*

Lung cancer is the leading cause of death from cancer in both men and women. Non-small-cell lung cancer (NSCLC) accounts for about 85% of deaths from lung cancer. In this case study, we applied the proposed method to the pathology images of 205 NSCLC patients in the NLST project (https://biometry.nci.nih.gov/cdas/nlst/). Each patient had one or more tissue slide(s) scanned at 40× magnification. A lung cancer pathologist first determined and labeled the region of interest (ROI) within tumor region(s) from each tissue slide, and then we randomly chose five square regions per ROI as the sample images. The total number of sample images that we collected was 1585. For each sample image, we used the ConvPath pipeline, as shown in Figure S1 of supplementary material available at *Biostatistics* online, to generate the corresponding cell distribution map as the input of our model. The number of cells in each sample image ranged from 2876 to 26 463.

We applied the proposed model with a 40-by-40 lattice to each preprocessed image. We used the same hyperparameter and algorithm settings as described in Section 4.1. Figure 1(a) and (d) are examples of the observed cell distribution maps from two patients' sample images. Figure 1(b) and (e) visualize the MP estimates of the hidden spins $\boldsymbol{P}$ for the imaging data as shown in Figure 1(a) and (d). We can also consider them as the imputed images by projecting irregularly distributed cells into a 40-by-40 lattice. Figure 1(c) and (f) show the two regions, the AOIs (in blue shadow) and the background areas. The indicator matrix $\boldsymbol{\Delta}$ was estimated by a median model by choosing $c = 0.5$. As we can see, the imputed images are indeed good representations of the original cell distribution maps. Our method appears to separate those regions with intensive cell–cell interaction from a "freeze" background in which the cell–cell interaction energy is relatively low. This observation can be validated through Figure S13 of supplementary material available at *Biostatistics* online, which shows the density plots of $\hat{\boldsymbol{\theta}}$ (red) and $\hat{\boldsymbol{\theta}}_0$ (green) via the hidden Potts mixture model and $\hat{\boldsymbol{\theta}}$ (black) via the hidden Potts model. It is clearly shown that the values of interaction energy of a homogeneous model are the roughly weighted average of the interaction energies of the AOI and the background regions from a mixture model, while the weight is determined by the Δ.

With the estimated interaction parameters in both AOI and the background area in each tissue slide, we conducted a downstream analysis to prove that proposing the hidden Potts mixture model as described in Section 2.3 is needed in practice, compared with its simpler version, which is the hidden Potts model as described in Section 2.2. Specifically, a Cox regression model was first fitted to evaluate the association between those estimated interaction parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$ and patient survival outcomes, after adjusting for other clinical information, such as age, gender, tobacco history, and cancer stage. Multiple sample images from the same patient were modeled as correlated observations in the Cox regression model to compute a robust variance for each coefficient. The overall P-value for the Cox model is 8×10^{-6} (Wald test), and the P-value and coefficient for each individual variable are summarized in Table 1 (P-values smaller than a 5% significance level are in boldface). The results imply that an increased interaction between stromal and tumor cells in the AOI ($\theta_{\text{str, tum}}$) is associated with good prognosis in NSCLC patients ($P = 0.005$). Interestingly, Beck *and others* (2011) also discovered that the morphological features of the stroma in the tumor region are associated with patient survival in a systematic analysis of breast cancer. Besides, the

Table 1. *Survival analysis for NLST lung cancer pathology images. The overall P-value corresponding to a Wald test for the hidden Potts mixture model (heterogeneous) is* 8×10^{-6} *and for the hidden Potts model (homogeneous) is* 0.006

Models	Parameters	Coefficient	exp (Coef.)	SE	*P*-value
Heterogeneous	$\theta_{\text{lym,str}}$	−0.13	0.88	0.062	0.260
	$\theta_{\text{lym,tum}}$	0.07	1.07	0.077	0.584
	$\theta_{\text{str,tum}}$	−0.34	0.71	0.070	**0.005**
	$\theta_{0,\text{lym,str}}$	0.26	1.30	0.077	**0.017**
	$\theta_{0,\text{lym,tum}}$	0.16	1.17	0.072	0.229
	$\theta_{0,\text{str,tum}}$	−0.03	0.97	0.077	0.818
	Number of cells	0.00	1.00	0.000	0.378
	Age	0.03	1.03	0.009	0.248
	Female vs. male	−0.14	0.87	0.092	0.626
	Smoking vs. non-smoking	−0.03	0.98	0.090	0.930
	Cancer stage I vs. II	0.45	1.57	0.147	0.359
	Cancer stage I vs. III	1.43	4.18	0.105	**7 x** $\mathbf{10^{-6}}$
	Cancer stage I vs. IV	1.34	3.81	0.145	**0.006**
Homogeneous	$\theta_{\text{lym,str}}$	0.01	1.01	0.024	0.896
	$\theta_{\text{lym,tum}}$	0.05	1.05	0.065	0.626
	$\theta_{\text{str,tum}}$	−0.13	0.88	0.076	0.295
	Number of cells	0.00	1.00	0.000	0.799
	Age	0.03	1.03	0.009	0.247
	Female vs. male	−0.21	0.81	0.091	0.460
	Smoking vs. non-smoking	−0.04	0.96	0.089	0.887
	Cancer stage I vs. II	0.40	1.49	0.146	0.432
	Cancer stage I vs. III	1.41	4.09	0.105	**1 x** $\mathbf{10^{-5}}$
	Cancer stage I vs. IV	1.25	3.49	0.142	**0.013**

interaction between lymphocyte and stromal cells in the background area ($P = 0.017$) is also a prognostic factor, while the underlying biological mechanism is currently unknown. In comparison, we then fitted a homogeneous hidden Potts model, which is equivalent to our mixture model with all δ_{lw}'s fixed to 1. The estimated interaction parameters $\hat{\boldsymbol{\theta}}_{\text{HP}}$ as well as other clinical variables were used as the predictors of the Cox regression model. The results are summarized in Table 1. As we can see, there is no significant predictor except cancer stage. This indicates that the homogeneous model tends to underestimate the true values of interaction energy between different types of cells. This example demonstrates the advantage of modeling the heterogeneous imaging data via a hidden Potts mixture model, rather than a hidden Potts model.

5. Conclusion

In this article, we focus on modeling cell distribution maps that arise in a lung cancer pathology image study. A hierarchical Bayesian framework was proposed in order to achieve three goals: (1) to reduce the complexity of the imaging data with thousands of irregularly distributed cells; (2) to quantify the interaction among different types of cells; and (3) to identify regions in the image where the interaction patterns significantly differ from each other. The proposed model utilizes the spatial information of thousands of irregularly distributed cells in the image. The introduction of auxiliary lattice helps to reduce the complexity of imaging data and defines a concise and explicit neighborhood for each spin in the Potts model. Our model is able to not only quantify the interaction energy between different types of cells, but also to distinguish clinically meaningful patterns from the background area via a Markov random field model.

For the lung cancer pathological imaging data, our study shows the interaction strength between stromal and tumor cells in the AOI is significantly associated with patient prognosis. This parameter can be easily measured using the proposed method and used as a potential biomarker for patient prognosis. This biomarker can be translated into real clinical tools at low cost because it is based only on tumor pathological slides, which are available in standard clinical care. In addition, this statistical methodology provides a new perspective to understand the biological mechanisms of cancer.

Several extensions of our model are worth investigating. First, the proposed model can be extended to more flexible finite mixture models by imposing a prior distribution on the number of components K. Second, the correlation among interaction parameters could be taken into account by modeling them as a multivariate normal distribution. Third, we can learn about fixed hyperparameters, such as e and f in the MRF prior by formulating its hyperpriors (see, e.g. Liang, 2010; Stingo and Vannucci, 2011). Fourth, although the use of MRF prior models encourages neighboring spins to clump together, there is no guarantee that the AOI is spatially contiguous, even if choosing $f = f_{\max}$. How to generate a clinically useful and smooth AOI based on the MP matrix of inclusion PPI_δ could be another future research direction. Last but not least, the proposed model provides a good opportunity to investigate the performance of other approximate Bayesian computation methods, such as variational Bayes, or even exact algorithms for sampling from distributions with intractable normalizing constants, such as Liang *and others* (2016). These could be future research directions.

6. Software

Software in the form of `R/C++` code is available on GitHub https://github.com/liqiwei2000/BayesHiddenPottsMixture. All the simulated datasets analyzed in Section 4.1 and two real datasets corresponding to the two sample images shown in Figure 1(a) and (d) of the manuscript are available on figshare https://figshare.com/projects/Bayesian_hidden_Potts_mixture_models/29659.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank Jessie Norris for helping us in proofreading the manuscript. *Conflict of Interest*: None declared.

Funding

This work was partially supported by the National Institutes of Health [R01CA172211, P50CA70907, P30CA142543, R01GM115473, R01GM117597, R15GM113157, and R01CA152301], and the Cancer Prevention and Research Institute of Texas [RP120732].

References

Amin, M. B., Tamboli, P., Merchant, S. H., Ordóñez, N. G., Ro, J., Ayala, A. G. and Ro, J. Y. (2002). Micropapillary component in lung adenocarcinoma: a distinctive histologic feature with possible prognostic significance. *The American Journal of Surgical Pathology* **26**, 358–364.

Ayasso, H. and Mohammad-Djafari, A. (2010). Joint NDT image restoration and segmentation using Gauss–Markov–Potts prior models and variational Bayesian computation. *IEEE Transactions on Image Processing* **19**, 2265–2277.

BARLETTA, J. A., YEAP, B. Y. AND CHIRIEAC, L. R. (2010). Prognostic significance of grading in lung adenocarcinoma. *Cancer* **116**, 659–669.

BECK, A. H., SANGOI, A. R., LEUNG, S., MARINELLI, R. J., NIELSEN, T. O., VAN DE VIJVER, M. J., WEST, R. B., VAN DE RIJN, M. AND KOLLER, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine* **3**, 108ra113.

BORCZUK, A. C., QIAN, F., KAZEROS, A., ELEAZAR, J., ASSAAD, A., SONETT, J. R., GINSBURG, M., GORENSTEIN, L. AND POWELL, C. A. (2009). Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. *The American Journal of Surgical Pathology* **33**, 462.

BROWN, P. J., VANNUCCI, M. AND FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 627–641.

FRANÇOIS, O., ANCELET, S. AND GUILLOT, G. (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**, 805–816.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–534.

GEORGE, E. I. AND MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

GILLIES, R. J., VERDUZCO, D. AND GATENBY, R. A. (2012). Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nature Reviews Cancer* **12**, 487–493.

GLEASON, D. F., MELLINGER, G. T.; THE VERETANS ADMINISTRATION COOPERATIVE UROLOGICAL RESEARCH GROUP (2002). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of Urology* **167**, 953–958.

GREEN, P. J. AND RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association* **97**, 1055–1070.

HANAHAN, D. AND WEINBERG, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.

JUNTTILA, M. R. AND DE SAUVAGE, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501**, 346–354.

KIRK, R. (2012). Genetics: personalized medicine and tumour heterogeneity. *Nature Reviews Clinical Oncology* **9**, 250–250.

LI, F. AND ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* **105**, 1202–1214.

LI, Q., YI, F., WANG, T., XIAO, G. AND LIANG, F. (2017). Lung cancer pathological image analysis using a hidden Potts model. *Cancer Informatics* **16**, 1176935117711910.

LI, S. Z. (2009). *Markov Random Field Modeling in Image Analysis*. New York: Springer Science & Business Media.

LIANG, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation* **80**, 1007–1022.

LIANG, F., JIN, I. H., SONG, Q. AND LIU, J. S. (2016). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *Journal of the American Statistical Association* **111**, 377–393.

LONGO, D. L. (2012). Tumor heterogeneity and personalized medicine. *New England Journal of Medicine* **366**, 956–957.

LUO, X., ZANG, X., YANG, L., HUANG, J., LIANG, F., CANALES, J. R., WISTUBA, I. I., GAZDAR, A., XIE, Y. AND XIAO, G. (2016). Comprehensive computational pathological image analysis predicts lung cancer prognosis. *Journal of Thoracic Oncology* **12**, 501–509.

Mantovani, A., Sozzani, S., Locati, M., Allavena, P. and Sica, A. (2002). Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. *Trends in Immunology* **23**, 549–555.

Marte, B. (2013). Tumour heterogeneity. *Nature* **501**, 327–327.

Mattfeldt, T., Eckel, S., Fleischer, F. and Schmidt, V. (2009). Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *Journal of Microscopy* **235**, 106–118.

McGranahan, N. and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628.

Merlo, L. M. F., Pepper, J. W., Reid, B. J. and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924–935.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

Orimo, A., Gupta, P. B., Sgroi, D. C., Arenzana-Seisdedos, F., Delaunay, T., Naeem, R., Carey, V. J., Richardson, A. L. and Weinberg, R. A. (2005). Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* **121**, 335–348.

Polyak, K., Haviv, I. and Campbell, I. G. (2009). Co-evolution of tumor cells and their microenvironment. *Trends in Genetics* **25**, 30–38.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Ripley, B. D. and Rasson, J. P. (1977). Finding the edge of a poisson forest. *Journal of Applied Probability* **14**, 483–491.

Schnipper, L. E. (1986). Clinical implications of tumor-cell heterogeneity. *New England Journal of Medicine* **314**, 1423–1431.

Shibata, D. (2012). Heterogeneity and tumor history. *Science* **336**, 304–305.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 795–809.

Stingo, F. C., Guindani, M., Vannucci, M. and Calhoun, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association* **108**, 876–891.

Stingo, F. C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27**, 495–501.

Tsao, M.-S., Marguet, S., Le Teuff, G., Lantuejoul, S., Shepherd, F. A., Seymour, L., Kratzke, R., Graziano, S. L., Popper, H. H., Rosell, R. *and others*. (2015). Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *Journal of Clinical Oncology* **33**, 3439–3446.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L. and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* **7**, 12474.

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H. *and others*. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science Translational Medicine* **4**, 157ra143.

[*Received September 25, 2017; revised March 9, 2018; accepted for publication March 18, 2018*]