A BAYESIAN MIXTURE MODEL FOR ANALYZING LUNG CANCER PATHOLOGY IMAGES

A FINAL PROJECT REPORT FOR STATISTICS 610

Vishal Sarsani
Department of Mathematics and Statistics
University of Massachusetts Amherst

Long Le
Department of Mathematics and Statistics
University of Massachusetts Amherst

August 25, 2021

ABSTRACT

Accurate prediction of cancer patients' survival and clinical outcome is essential for providing effective personalized treatment and improving patients' quality of life. Computational results from high throughput pathology image analysis have great potential to provide prognostic information. Image analysis can provide not only the density information but also spatial heterogeneity information of cancer and immune cells, which provide a great prognostic value for cancer progression. But currently, there are no statistical approaches to model spatial correlations among different kinds of markers generated by protocols such as mIHC(multiplexed immunohistochemistry) restricting the prediction of survival outcomes only to known clinical variables.
In this project, we replicate the methodology section from Xiao et al where they model the spatial correlations among three commonly seen cells (i.e. lymphocyte, stromal, and tumor) observed in tumor pathology images. We use a 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. We use Markov chain Monte Carlo sampling techniques, combined with a double Metropolis-Hastings algorithm, to simulate samples approximately from a distribution with an intractable normalizing constant. The proposed model was applied to the pathology images of 2 lung cancer patients from the National Lung Screening trial

***K*eywords** Digital pathology image · survival outcome · Mixture Model · Ising Model · Double Metropolis

1 Background

Understanding the influence of the tumor microenvironment on cancer development and evolution is of tremendous interest in developing novel treatments. Clinical variables collected during a study provide some prognostic information for cancer patients. In most cases, clinical variables alone will not be able to reduce or explain the total variation in the prognosis. It has been found that strong regional differences and spatial heterogeneity in the tumor microenvironment can greatly impact cancer prognosis. .A recent study showed that traditional histopathological methods to stage colorectal cancer underperformed other prognostic factors that incorporate type, density, and location of immune cells . An array of algorithms currently exists to automatically classify cell and tissue regions in histopathology images. Computational results from such high-throughput pathology image analysis provide more quantified information about the cell abundance, spatial distribution and the spatial context of diverse cell types coexisting within the tumor microenvironment. This enables us to better understand the complexity in spatial data and their relation to prognosis. Previous studies in patients of colorectal and breast cancer showed an association of clinical outcomes with the spatial locations of immune cells [1]. Novel methods need to be explored to obtain more clinically meaningful parameters that can predict the prognosis of patients In this study, will be replicating Bayesian hierarchical mode to study the spatial correlations among three commonly seen cells observed in tumor pathology images from 2 lung cancer patients from the National Lung Screening trial.

2 Model Specification

In this section, we provide a brief summary of the model replicated from the methodology section from Xiao et.al[2] First, we start with an idealized model on regular lattice, known as the Potts Model. Next, we explain how to project our irregular empirical cell distribution onto this lattice via a projection parameter. Then, we improve the model flexibility by allowing different interaction parameters in different regions.

2.1 Potts Model

The Potts Model is simply an extension of the Ising Model. Recall that in an Ising Model, we are given a 2D lattice and each coordinate is assigned a spin up or down. Each configuration of the lattice is also given a probability such that configurations with low Hamiltonian energy is deemed more stable and hence more likely. In a Potts model, we simply allow each vertex to take on possibly more than two classes. In our application, each vertex can either be a lymphocyte, stroma or tumor cell.
Mathematically, the Potts model consists of a $L \times W$ lattice, where each vertex is connected to its four immediate neighbors. Let P be a particular configuration of the lattice. In other words, P specifies p_{lw}, an assignment of vertex (l, w) to one of the three cell types, for each vertex (l, w). Further, let $\theta_{qq'}$ denote the interaction strength between cell type q and q'. Then, the Hamiltonian energy of P is

$$H(P|\theta) = -\sum_{(l,w)} \sum_{(l',w') \in Neigh(l,w)} \theta_{qq'} 1\{p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q'\}.$$

Note that in the equation above, the restriction $p_{lw} \neq p_{l'w'}$ means that we are primarily interested in the interaction between different cell types. The probability of P is then

$$Pr(P|\theta) = \frac{exp(-H(P|\theta))}{\sum_{P'} exp(-H(P'|\theta))} = \frac{exp(-H(P|\theta))}{C(\theta)}.$$

As we will see in section 3, $C(\theta)$ is an intractable constant that we need to deal with in the sampling procedure.

In this digital pathology slide, we can divide regions in the image into two categories: background region and area of interest (AOI). We expect the interaction between cells in these two regions to differ. Thus, we introduce two sets of parameters θ_0 and θ, which are both $|Q| \times |Q|$ matrices, to account for the interaction strengths in the background and AOI respectively. Then, we also need to provide a $L \times W$ matrix Δ with $\Delta_{lw} \in \{0, 1\}$, specifying whether vertex (l, w) belongs to the background or AOI.

2.2 Hidden Potts Model

As mentioned at the beginning, an image in our data consists of scattered cells. We now need a mechanism to go from the regular lattice model to the data. We are given an empirical list of (x_i, y_i, z_i), where (x_i, y_i) are the physical coordinate of the cell on the 2d plane and z_i is the cell type. We imagine generating (x_i, y_i, z_i) with certain probability given the hidden lattice P.

$$Pr(x_i, y_i, z_i = q|P, d) = \frac{exp(d \sum_{\{(l,w): l \leq x_i \leq l+1, w \leq y_i \leq w+1\}} 1_{p_{lw}=q})}{\sum_{q'} exp(d \sum_{\{(l,w): l \leq x_i \leq l+1, w \leq y_i \leq w+1\}} 1_{p_{lw}=q'})}.$$

Note that d is a hyper-parameter, known as the projection parameter. The above model in a sense encourages the hidden lattice P to assign the same cell type to a hidden vertex as the majority cell type of the vertex's neighboring empirically observed cells. As d increases, the types of the hidden vertex cells and neighboring observed cells are more likely to agree.

2.3 Likelihood of Hidden Potts mixture model

With this assumption, an $L \times W$ latent matrix δ is introduced to indicate the 2 distinct regions:background region(θ_0) and area of interest(θ) with $\delta_{lw} = 0$ if the spin at location (l,w) belongs to group background region and $\delta_{lw} = 1$ if the spin at location (l,w) belongs to group AOI.

The likelihood of the two-component mixture model is written as

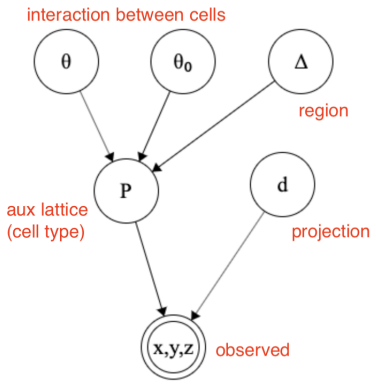


Figure 1: Figure showing the parameters in the likelihood

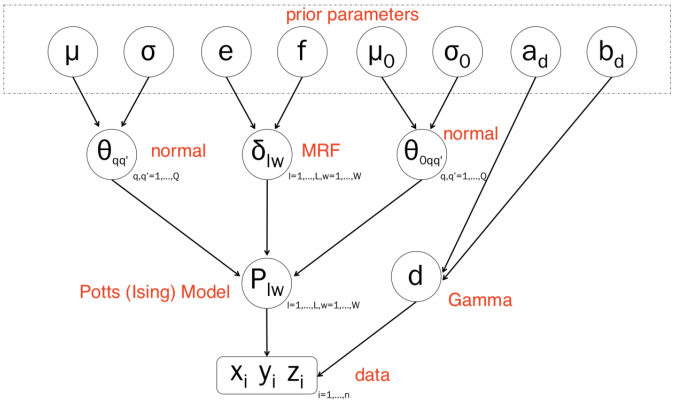


Figure 2: Graphical representation of the Hierarchical Hidden Potts mixture model

$Pr(P|\Delta, \theta_0, \theta)$
$= \frac{1}{C_0(\theta_0)} \exp\left(\sum_{\{(l,w):\delta_{lw}=0\}} \sum_{(l',w')\in Nei_{(l,w)}} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{0qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q')\right)$
$\times \frac{1}{C(\theta)} \exp\left(\sum_{\{(l,w):\delta_{lw}=1\}} \sum_{(l',w')\in Nei_{(l,w)}} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{qq'} I(p_{lw} \neq p_{l'w'}, p_{lw} = q, p_{l'w'} = q')\right)$
$C_0(\theta_0)$ and $C(\theta)$ are the normalizing constants for the two regions.

2.4 Priors of Hidden Potts mixture model

The prior of the Hidden Potts mixture model model reduces to an Ising model when there are only 2 regions :

$$Pr(\delta_{lw}|\Delta_{-l,-w}) = \frac{\exp\left(\delta_{lw}\left(e+f\sum_{(l',w')\in Nei_{(l,w)}} \delta_{l',w'}\right)\right)}{1+\exp\left(\delta_{lw}\left(e+f\sum_{(l',w')\in Nei_{(l,w)}} \delta_{l',w'}\right)\right)}$$

where e and f are hyperparameters to be chosen.The extra parameter e controls the number number of spins belonging to the AOI while f affects the probability of assigning value according to its neighbor spins.We specify the prior distribution for d as $d \sim Ga(a_d, b_d)$.To use a weakly informative gamma prior,we set $a_d = b_d = 0.001$.For and θ_0 , consider Normal priors $\theta_{qq'} \sim N(\mu, \sigma^2)$) and $\theta_{0qq'} \sim N(\mu_0, \sigma_0^2)$.

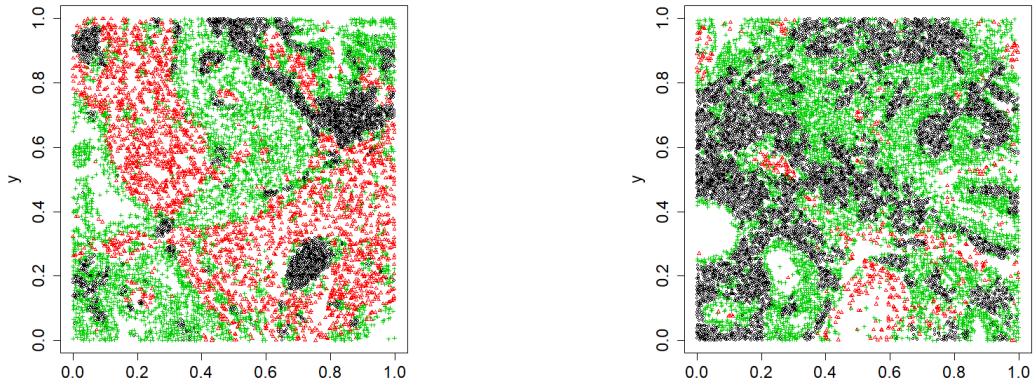


Figure 3: The observed cell distribution maps of lymphocyte, stromal, and tumor cells marked in black, red, and green for two different patients

3 Sampling Algorithm

The inference of the interaction parameters within the Areas of interest via the vector θ. We implemented an MCMC algorithm based on the DMH (Double Metropolis algorithm)[3] and Metropolis search variable selection algorithms[4] to search the model space that consists of $(P, \Delta, \theta, \theta_0, d)$.

The use of DMH is mainly due to the presence of the parameter-dependent intractable constant $C(\theta)$ needed to sample the full conditional $Pr(\theta_{qq'}|.)$. Within each standard MCMC iteration, we need to sample $\theta_{qq'}$ from its conditional distribution.

$$\pi(\theta_{qq'}|\cdot) \propto Pr(P|\theta)\pi(\theta_{qq'}) = \frac{1}{C(\theta)} \exp(-H(P|\theta))N(\theta_{qq'}; \mu, \sigma^2)$$

If we want to sample from the full conditional, we need to calculate the acceptance ratio.

$$a = \frac{\pi(\theta^*_{qq'}|\cdot)}{\pi(\theta_{qq'}|\cdot)} = \frac{C(\theta)}{C(\theta^*)} \frac{\exp(-H(P|\theta^*))N(\theta^*_{qq'}; \mu, \sigma^2)}{\exp(-H(P|\theta))N(\theta_{qq'}; \mu, \sigma^2)}$$

DMH eliminates the need to calculate $C(\theta)$ by simulating an auxiliary variable P^* based on the proposed parameter θ^*. Specifically, sample $P^* \sim \pi(P|\theta^*_{qq'}, .)$ using Metropolis. Now, the acceptance ratio becomes

$$a = \frac{\pi(\theta^*_{qq'}|P, \theta_{-q,-q'})}{\pi(\theta_{qq'}|P, \theta_{-q,-q'})} \frac{Pr(P^*|\theta)}{Pr(P^*|\theta^*)}.$$

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

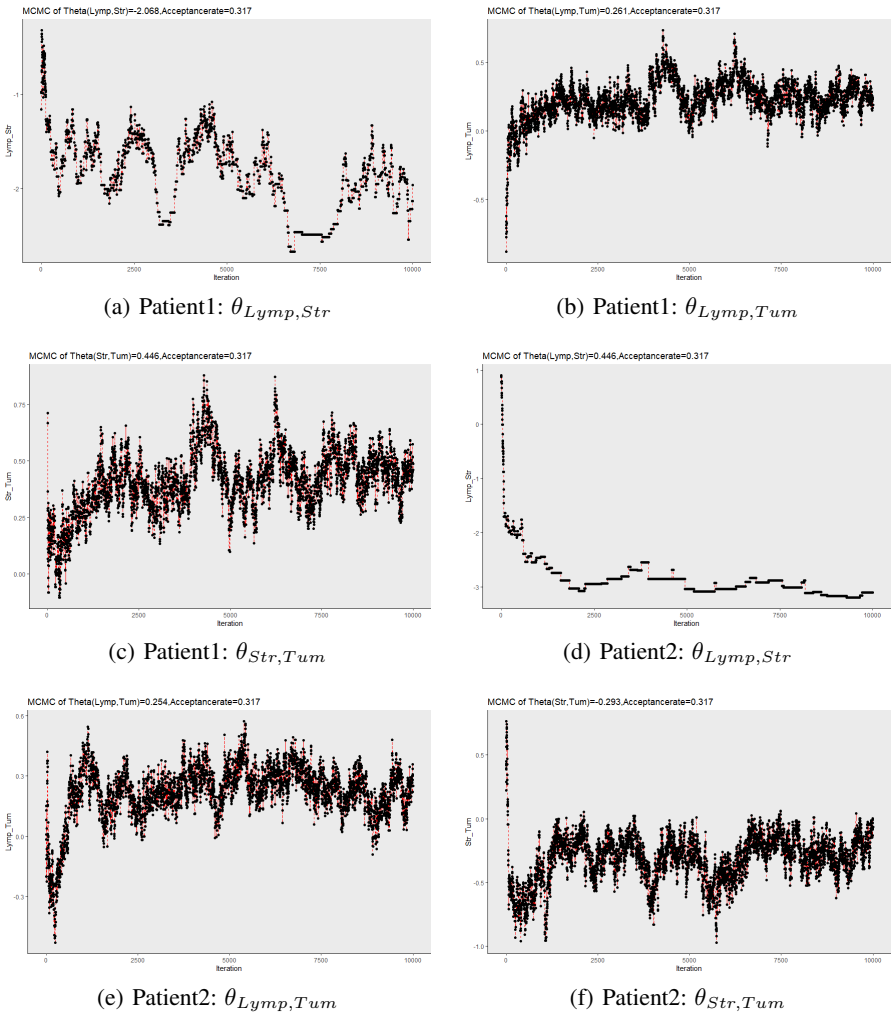
4 Result and Future Work

4.1 Data Preprocessing

The model was applied to the pathology images of 2 NSCLC patients in the NLST project[5].Each patient had a tissue slide scanned at 40 magnification.The annotation of region of interest (ROI) within tumor region(s) was manually done by a pathologists.We used a Convpath pipeline[6] to generate the corresponding cell distribution map as the input of model.

4.2 Application

We applied the model with a 50-by-50 lattice to each pathology image of the patient.We used normal distributions priors of N(0.5,1) and N(-0.5,1) for $\theta_{qq'}$ and $\theta_{0qq'}$.Hyperparameters that control the MRF prior model were set to



(a) Patient1: $\theta_{Lymp,Str}$

(b) Patient1: $\theta_{Lymp,Tum}$

(c) Patient1: $\theta_{Str,Tum}$

(d) Patient2: $\theta_{Lymp,Str}$

(e) Patient2: $\theta_{Lymp,Tum}$

(f) Patient2: $\theta_{Str,Tum}$

Figure 4: MCMC posterior estimates of interaction energy of lymphocyte, stromal, and tumor cells for two patients

e=2.94 and f=1.4, which means that if a spin in the lattice does not have any neighbor in the AOI.The gamma prior on the projection parameter were set to $a_d = b_d = 0.001$.We obtain the posterior inference by post-processing of the MCMC samples(iterations=10000) after burn-in(5000 iterations).The approximate Bayesian estimator of $\theta_{qq'}$ can be simply obtained by averaging over the iterations

$$\hat{\theta}_{qq'} = \quad \frac{1}{U}\sum_{u=1}^{U}\theta_{qq'}^{(u)}$$

where $u, u = 1, \ldots, U$ represent the indexes the iteration after burn-in

4.3 Future Work

We want to speed up the current implementation of doubly MH so that we can run 100k iterations as the convergence is not reached in some of the parameters for 10k iterations. We want to perform parameter estimation on all 205 patients of the NLST trial and with the estimated interaction parameters in the AOI, we plan to conduct a downstream survival analysis. A Cox regression model will be first fitted to evaluate the association between estimated interaction parameters and patient survival outcomes, after adjusting for other clinical information, such as age, gender, tobacco history, and cancer stage.

5 Conclusion

In this project, we model the cell distribution maps from 2 patients of a lung cancer pathology image study. We used a hierarchical Bayesian framework proposed in et al to quantify the interaction among different types of cells. We implemented a double Metropolis–Hastings algorithm, to simulate samples approximately from a distribution with an intractable normalizing constant.Although the convergence of psoterior estimates is not reached for n=10000 iterations for some parameters, our future work includes speeding the current implemntation. We also want to investigate whether the interaction strength between stromal , lymphocytw and tumor cells in the AOI is significantly associated with patient prognosis. This parameter can be used as a potential biomarker for patient prognosis and gives ius a new perspective to understand the biological mechanisms of cancer.

References

[1] Nawaz, S. and Heindl, A. and Koelble, K. and Yuan, Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer,Mod. Pathol,2015-28

[2] Li, Q. and Wang, X. and Liang, F. and Yi, F. and Xie, Y. and Gazdar, A. and Xiao, G. A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images,Biostatistics,2019-20

[3] I. Murray, Z. Ghahramani, and D. J. C. MacKay "MCMC for doublyintractable distributions," in Uncertainty in Artificial Intelligence (UAI), pp. 359–366, AUAI Press, 2006

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

[5] NLST Project https://biometry.nci.nih.gov/cdas/nlst/.

[6] Wang, S. and Wang, T. and Yang, L. and Yang, D. M. and Fujimoto, J. and Yi, F. and Luo, X. and Yang, Y. and Yao, B. and Lin, S. and Moran, C. and Kalhor, N. and Weissferdt, A. and Minna, J. and Xie, Y. and Wistuba, I. I. and Mao, Y. and Xiao, G. ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network, Journal="EBioMedicine",2019-50