SPARTYN Supplementary material

Nate Osher

2021-11-27

Contents

1 About **5**

2 Introduction **7**

3 Materials and Methods **9**

3.1 Overview 9

3.2 Spatial Point Processes 10

3.3 Intensity Thresholding and Partition 11

3.4 Inference 11

3.5 Infiltration Probability 12

4 Results **13**

4.1 Simulation 13

4.2 Application 14

5 Conclusion **15**

Chapter 1

About

The purpose of this document is to serve as a "living" supplement to the first paper on the SPARTYN pipeline. While the paper itself will be static upon publication, this document will serve as both an ongoing supplement as well as a central location to give updated, detailed information on the pipeline and methods.

If you have a question that is not adequately addressed in this supplement, please email oshern (at) umich.edu.

Chapter 2

Introduction

While the staining and examination of tissue samples has been a ubiquitous medical practice for decades at this point in time, it has only been relatively recently that high definition images of such stainings have begun to be explored through the lens of machine learning and statistical modeling. A considerable amount of work has thus far been put into building and training models capable of remarkably accurate classifications of overall tissue samples as well as subsections of tissue samples [Saltz et al 2018; TODO: other examples]. While valuable in their own right such advances have also created additional opportunities for the application of more traditional statistical techniques. Notably, the ability to quickly and accurately process high definition images of biopsies into cell level classification and location data has allowed for the development and application of more traditional spatial statistical modeling methods.

There is ample prior reason to suspect that such applications may be valuable, particularly in the domain of cancer pathology. The longstanding conventional wisdom among pathologists is that tissue features such as lymphocyte infiltration as well as general tissue heterogeneity are meaningful prognostic indicators [TODO: Citations]. So far, this conventional wisdom has found support in the current quantitative histopathological imaging analysis literature. Li et al. (2018) found that the spatial associations between stromal cells and other types of cells were significantly associated with survival in non-small cell lung cancer using both a Hidden Potts Mixture Model as well as a mark interaction model [Note: these are two different papers, both by Li in 2018- not sure the correct way to acknowledge]. Saltz et al. (2018) specifically examined the presence of lymphocytes across biopsies in several types of cancer, and found that certain summarization metrics of lymphocyte clusters were significantly associated with survival in certain types of cancer.

In this paper, we introduce the SPARTYN (SPatial Analysis of paRtitioned Tumor-lYmphocyte imagiNg) pipeline. SPARTYN is unique in the spatial pathological imaging analysis literature in that it uses statistical models to

assess the association between tumor cells and lymphocytes across an entire partitioned biopsy. This allows for rigorous quantification of uncertainty in the style of Li et al. (2018) while still allowing for the assessment of full images in the style of Saltz et al. (2018). We accomplish this by partitioning the cell-level imaging data of each biopsy into non- overlapping sub-regions, which can then be modeled to capture the local infiltration patterns using standard techniques from point process theory.

Spatial point processes have long been used in the domain of ecology to rigorously investigate spatial relationships between various organisms (H{"o}gmander and S{"a}rkk{"a}, 1999; King et al. 2012). More recently, methods from this paradigm have been successfully applied within the domain of Biostatistics (Kang et al. 2011; Kang et al. 2015). Given the relatively simple and granular nature of cell-level spatial imaging data, marked point processes are a natural way to simultaneously model the randomness in both the quantities and locations of the different cell types along with their relative spatial associations.

Chapter 3

Materials and Methods

3.1 Overview

Raw data was obtained for this project using the image processing tools of Rao and Krishnan. A random forest model was trained on high definition images of cancer biopsies in order to automatically classify cells of different types. Images of Skin Cutaneous Melanoma (SKCM) biopsies from The Cancer Genome Atlas program were processed using this model such that each cell was classified as a tumor cell, a lymphocyte, or other. In addition, the x- and y-coordinates of each cell centroid (relative to the pathology slide) were determined and recorded. The resulting data set for each biopsy consisted of a row for each cell, with a column for the x-coordinate, a column for the y-coordinate, and an indicator of the cell type. Each biopsy was intensity thresholded to define a "fitted" window, i.e. the smallest window that fit all cells in the biopsy within it. This window was then divided into tiles that fully partition it, while containing similar numbers of tumor cells. A bayesian spatial point process model was then fit on each of these tiles separately, yielding a posterior distribution of the local interaction parameter. For each tile, this local distribution was then compared to a tile-specific null distribution, before being combined to summarize the overall level of interaction at the biopsy level.

SKCM is an appealing target for the investigation of lymphocyte infiltration for several reasons. SKCM has been shown to be particularly responsive to Immunotherapy in some cases [CITATION]. It is possible that the ability to quantify lymphocyte infiltration at a large scale may allow for more detailed investigation into the scenarios in which this treatment may be most effectively deployed. What's more, SKCM has an unusually high mutational load amongst the various cancer types [CITATION]. The ability to quantify infiltration may allow for further investigation into not only genomic associations of this occurrence but associations with mutations as well. [TODO: is infiltration

particularly common in SKCM? I feel like it is, but I need to verify this.]

3.2 Spatial Point Processes

Denote the number of biopsies n, and for biopsy i, let c_i denote the number of cells observed and labeled within that biopsy, with corresponding x- and y-coordinates $\mathbf{x}_i = x_{i1}, ..., x_{ic_i}$, $\mathbf{y}_i = y_{i1}, ..., y_{ic_i}$, and marks $\mathbf{m}_i = m_{i1}, ..., m_{ic_i}$. For our application, $m_{ij} \in \{1, 2\}$,with 1 indicating a tumor cell and 2 indicating a lymphocyte. Further, denote the number of tumor cells observed in subject i by T_i, and the number of lymphocytes L_i. Within each patient, this data can be naturally thought of as a marked point process and modeled as such.

We ultimately decided to use a Hierachical Multitype Strauss Model for our data, the density of which is

$$f(\mathbf{p}_1, \mathbf{p}_2) \propto \beta_1^{n_1} \beta_2^{n_2} \gamma_{11}^{S_{R_{11}}(\mathbf{p}_1)} \gamma_{22}^{S_{R_2}(\mathbf{p}_2)} \gamma_{12}^{S_{R_{12}}(\mathbf{p}_1, \mathbf{p}_2)} \quad (1)$$

Where:

- $\mathbf{p}_i$ is a vector of points of type i
- n_i is the number of points of type i
- β_i is the first order intensity of points of type i
- $S_{R_{ij}}(\cdot)$ counts the number of pairs of points of types i and j within R_{ij} one another, where R_{ij} is selected a priori based on subject specific knowledge.
- γ_{ij} captures the tendency of points of type i to be near points of type j

We decided to use this model over the standard Multitype Strauss model for two reasons. Firstly, treating the locations of the lymphocytes as conditional upon the locations of the tumor cells is a priori biologically plausible. Secondly, the hierarchical variant of the Strauss model allows for the proper modeling of positive interaction between points of different types, while the standard multitype model does not. Note that because intra-type interaction is still confined to be negative under the Hierarchical Multitype Strauss Process, we constrained the intra-type interaction parameters γ_{11} and γ_{22} to be 1. Finally, based on prior biological knowledge, we set $R_{12} = 30\ \mu m$. In our context, γ_{12} can be thought of as the degree to which lymphocytes to be close to or far away from tumor cells, conditional upon the locations of the tumor cells. This allows us to distinguish between the
mere relative abundance of different cell types (which is
captured by the β_1 and β_2 parameters) and the actual spatial associations between the different cell types.

3.3 Intensity Thresholding and Partition

Each biopsy was partitioned into non-overlapping sub-regions, each of which was modeled as a point process. The primary motivation for structuring the analysis this came from prior biological knowledge about tumor composition. Tumors are heterogeneous entities in many respects, and lymphocyte infiltration is no exception. Thus, partitioning the biopsy into non- overlapping sub-regions and fitting models on each sub-region is a natural way to capture this heterogeneity. This strategy has additional benefits in that it allows for the parallelization of model fitting across the resulting sub-regions within a single biopsy.

In order to partition a given biopsy into non-overlapping sub-regions, we beganwith the smallest bounding rectangular window that contained cells $1...c_i$, we applied an intensity thresholding algorithm [too much?] in order to find the smallest possible window that still contained all c_i cells. Next, we applied a voronoi tesselation to the T_i tumor cells within the intensity thresholded window, partitioning it into tumor cell
specific sub-windows $1...T_i$ corresponding to tumor cells
$1...T_i$. We then applied a modified version of k-means to
tumor cells $1...T_i$,such that each of the K_i resulting
clusters was constrained to be between a pre-defined range of
cell counts. Finally, each tumor cell specific sub-window within a given cluster was combined into a tile, corresponding to each of the K_i clusters from the k-means clustering. This results in K_i tiles that fully partition the intensity thresholded window. This partition uniquely defines the membership of each of the c_i total cells into one of the resulting K_i total tiles.

It is worth emphasizing that there is nothing particularly unique about this method of partitioning the biopsy. Any other method could be used in its place, so long as the result is a partition of the biopsy into some number of non-overlapping sub-regions on which the subsequent model fitting can proceed.

3.4 Inference

In order to compute posterior distributions of parameters of interest, we used Bayesian techniques in the style of King et al. 2012. This methodology essentially exchanges the likelihood function used in standard Bayesian inference for the pseudolikelihood function (Besag, 1975 [VERIFY]), with the integral approximated via the Berman-Turner device (Baddeley and Turner, 2000) using the spatstat package (Baddeley and Turner, 2005). Finally, each parameter estimated (β_1, β_2, γ_{12}) was assigned a flat prior. [TODO: Technically, normal with mean 0 and variance $1,000,000$ because of limitations of JAGS- should I spell this out?] Posterior distributions were computed using MCMC via the R2jags package (Su, 2015). Using these techniques we were able to compute a

posterior distribution $f_{ik|\mathbf{p}}(y)$ of $\gamma_{ik,12}$ for each for each biopsy $i \in \{1...n\}$ and tile $k \in \{1...K_i\}$.

3.5 Infiltration Probability

In order to compute a localized probability of infiltration, each posterior distribution $f_{ik|\mathbf{p}}(y)$ was compared to a tile specific null distribution, $f_{ik,0}(y)$. Because under null interaction ($\gamma_{ik,12} = 1$)) our model reduces to two independent Poisson processes with intensities β_1, β_2, this was accomplished by treating the observed tumor cells as fixed, simulating s realizations of lymphocytes under a Poisson process with the observed intensity, performing the previously described model fitting procedure on each, and aggregating samples across simulation $1...s$. Under the assumption that the posterior distribution $f_{ik|\mathbf{p}}(x)$ is independent from the null distribution $f_{ik,0}(y)$, we then computed the tile specific Infiltration Probability r_{ik}, given by

$$\text{Infiltration Probability} \quad r_{ik} = \int\int I(x > y) f_{ik|\mathbf{p}}(x) f_{ik,0}(y) dx dy \qquad (2)$$

Note that $r_{ik} \in [0, 1]$. Because r_{ik} is computed as the integral over an indicator variable, this naturally yields an interpretation as a probability. Specifically, r_{ik} can be thought of as measuring the posterior probability that the observed value of γ_{12} (denoted by x in the integral) is larger than the value of γ_{12} we would expect to observe by chance (denoted by y in the integral). We thus refer to r_{ik} as the Infiltration Probability for that particular tile, which can be aggregated across tiles to yield a measure of infiltration on a particular biopsy.

Chapter 4

Results

4.1 Simulation

To compare our model's detection of spatial association between different cell types, we ran a small-scale simulation study in which we tested the ability of this method to accurately classify positive interaction across a range of different simulated cell compositions and spatial associations. For each of four sets of simulations, the number of simulated tumor cells and lymphocytes were set a priori at T_s and L_s respectively. Further, interaction was controlled by a parameter $\phi \in [-1, 1]$, with -1 indicating the most negative possible interaction and 1 indicating the most positive possible interaction.

For a given combination of T_s and L_s, the positive and negative simulations proceeded differently. For the negative simulations ($\phi \in [-1, 0]$), T_s tumor cells and L_s lymphocytes were simulated as independent poisson processes in regions of the window that overlapped to varying degrees. The overlap was controlled by ϕ, such that the processes overlapped on $(100 \cdot (1 + \phi))\%$ of the window in which the simulation occurred. Note that when $\phi = -1$ there was no overlap, and when $\phi = 0$ (complete overlap) the simulation reduced to simulating two independent Poisson Processes within the same window.

For the positive simulations ($\phi \in (0, 1]$), T_s tumor cells were simulated under a Poisson process. After their locations were determined, L_s lymphocytes were simulated. For each lymphocyte l_i, a Bernoulli random variable $C_i \sim Bern(\phi)$ was drawn. If $C_i = 1$, l_i was simulated within 30 μm of a
randomly selected tumor cell t_j. Otherwise, l_i was simulated from a Poisson process. Thus, the level of interaction was again controlled by p, with $p = 0$ now corresponding to two independent Poisson Processes and $p = 1$ corresponding to a situation in which all lymphocytes are within 30 μm of at least one tumor cell.

Across the different simulation settings, accurate classification was possible using IP, with the minimum AUC across simulations being 0.84. For detailed descriptions of simulation settings, simulated data, and results, see Figure [N].

4.2 Application

We applied our method to a data set consisting of 335 images of skin cutaneous melanoma taken from The Cancer Genome Atlas. Images were processed as described in section 2.1. Our methods were applied to the resulting data sets. [GIVE SUMMARY STATISTICS ABOUT TILES?]

4.2.1 Survival Analysis

In order to assess association between PPPI and survival, we fit a Cox Proportional Hazards model. In addition to adjusting for average logit-PPPI, we adjusted for patient level factors such as cancer stage, age, and sex. In addition, we adjusted for readily calculable tumor level features, such as number of tumor cells (as a proxy for size) and logit lymphocyte proportion (the number of TILs divided by the number of tumor cells and TILs). We found that after adjusting for these other factors, an increase in logit-PPPI was significantly associated with increased risk of death ($p < 0.05$). The same model was fitted with average logit-PPPI exchanged for the average normalized value of the estimated Mark Correlation Function evaluated at $r = 30$ (the same as the radius of interaction used in our model fitting). See Table 1 for coefficients and standard errors in each model.

4.2.2 Genomic Associations

In addition to associations with survival, we investigated the association between our measurement and gene expression. Gene expression [RNA-seq] data was acquired for all 335 patients in our sample using TCGA Assembler [citation here]. Additionally, 42 significantly mutated genes of interest were identified using previous work investigating the genomic differences in SCM [TCGA Network, 2015]. Of the 335 patients in our sample, 240 had gene expression data for all genes of interest, while 95 were missing data for all genes of interest. We examined the marginal association between the normalized gene expression data for the 240 patients with complete data and average logit PPPI values. After correcting for multiple testing using the Benjamini-Hochberg procedure, we found that the expression of three genes were significantly associated with average logit PPPI: LRRC37A3, B2M, and TP53. See supplementary Table [N] for full details on significance of associations with each gene.

Chapter 5

Conclusion

As algorithms for cell-level image classification improve, the opportunities for more and more granular quantitative analysis of histopathological imaging data will become both more numerous and more fruitful. Moreover, as spatial gene expression data becomes more and more common, so too will opportunities for synthesizing data on the relative spatial locations of different cell types along with local gene expression data through complex modeling.

The SPARTYN pipeline represents a valuable contribution in and of itself to the histopathological imaging analysis literature through its ability to model and quantify lymphocyte infiltration across entire biopsies in a way that captures meaningful variation across patients. However, it also creates numerous opportunities for future work. Partitioning each biopsy into non-overlapping sub-regions such that each is assigned a value invites the usage of other tools from the spatial statistical canon. Specifically, tools from areal data analysis may be readily applied without any modifications or further theoretical development. Moreover, recall that our usage of Bayesian methods allows for not only the computation of scalar summary statistics for each sub-region but also posterior distributions. Future work may involve adapting more traditional methods to model correlated density functions, or developing new methods as necessary.

It also bears mentioning that while we limited ourselves to investigating two cell types throughout this paper, there is nothing in our pipeline or model that depends on this limitation in such a way that would render future investigation of more cell types impossible. It is highly likely that future iterations of our pipeline and methodology will seek to investigate the spatial associations between three or more cell types.

Finally, our work could very easily take advantage of the burgeoning field of spatial transcriptomics. It would be quite trivial (particularly in a Bayesian framework) to model a function of one or all of the parameters of interest as a linear combination of the local gene expression data for genes that are known

to be relevant to immune response. Alternatively, future methods in this area could be developed at the intersection of spatial data and high dimensional data in order to identify such genes.