DenVar: Density-based Variation analysis of multiplex imaging data

Souvik Seal, Thao Vu, Tusharkanti Ghosh, Julia Wrobel, and Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health,
University of Colorado CU Anschutz Medical Campus, Aurora, Colorado

Abstract

Multiplex immunohistochemistry (mIHC) and multiplexed ion beam imaging (MIBI) platforms have become increasingly popular for studying complex single-cell biology in cancer patients. In such studies, researchers test for association between functional markers and survival in a two-step process. First, they count the number of positive cells, defined as the number of cells where a functional marker is significantly expressed. Then, they partition the patients into two groups and test for association between the group label with survival. Consequently, the approach suffers from subjectivity and lack of robustness. In this paper, we propose a threshold-free distance metric between patients solely based on their marker probability densities. Using the proposed distance, we have developed two association tests, one based on hierarchical clustering and the other based on linear mixed model. Our method obviates the need

for the arduous step of threshold selection, getting rid of the subjectivity bias. The method also intuitively generalizes to joint analysis of multiple markers. We assessed the performance of our method through extensive simulation studies and also used it to analyze two multiplex imaging datasets.

Keywords: Marker thresholding; Kernel density estimation; Survival analysis; Multiplex tissue imaging

1 Introduction

In recent years, various technologies are being used for probing single-cell spatial biology, for example, multiparameter immunofluorescence (Bataille *and others*, 2006), imaging mass cytometry (IMC) (Giesen *and others*, 2014; Chang *and others*, 2017; Ali *and others*, 2020), multiplex immunohistochemistry (mIHC) (Halse *and others*, 2018; Tan *and others*, 2020) and multiplexed ion beam imaging (MIBI) (Angelo *and others*, 2014; Seal *and others*, 2021*b*). These technologies, often referred to as multiplex tissue imaging, offer the potential for researchers to explore the bases of many different biological mechanisms. Multiplex tissue imaging platforms such as Vectra 3.0 (Akoya Biosciences) (Huang *and others*, 2013), Vectra Polaris (Akoya Biosciences) (Pollan *and others*, 2020), MIBI (Ionpath Inc.) (Keren *and others*, 2019; Ptacek *and others*, 2020) produce images of a similar structure: 2-dimensional images of tissues at the resolution of cells and nuclei, where proteins in the sample have been labeled with antibodies called "markers" that attach to cell membranes and proteins. Typically, mIHC (Vectra platform) images have 6-8 markers, whereas MIBI images can have more than 40 markers. Majority of these markers are surface markers (primarily used for cell type identification), such as CD4, CD3, CD8, CD68 etc. (Jondal *and others*, 1972; Zola *and others*, 2007; Shipkova and Wieland, 2012). There are several functional markers as well, such as HLA-DR

(Jendro *and others*, 1991; Oczenski *and others*, 2003; Saraiva *and others*, 2018), PD-1, PD-L1 and Lag3 (Nguyen and Ohashi, 2015; Han *and others*, 2020; Phillips *and others*, 2021) that dictate or regulate important cell-functions.

With such informative markers, researchers are often interested in finding out if abundance or over-expression of the markers in the tumor micro environment (TME) (Whiteside, 2008; Binnewies *and others*, 2018) have significant impact on clinical parameters, such as patient survival (Sahlberg *and others*, 2014; Koguchi *and others*, 2015; Johnson *and others*, 2020). For this purpose, a commonly used approach, particularly popular in flow and mass cytometry paradigm, is a method based on thresholding of marker values (Bulian *and others*, 2014; Costa *and others*, 2017; Missassi *and others*, 2021). The first step of the method is to identify the cells (or, pixels) positive for a marker defined as the cells which show significant marker expression. Next, researchers proceed in either of the two ways: 1) use the proportion of the positive cells (or, pixels) directly in a univariate linear regression framework to test association with the outcome (Patwa *and others*, 2021), or, 2) consider one or more thresholds to group the patients into two or more clusters based on the proportion of the positive cells and use the vector of cluster labels in a univariate linear regression framework to test association with the outcome (Chen *and others*, 2016; Chang *and others*, 2018; Yang *and others*, 2019; Johnson *and others*, 2021).

The thresholding-based method clearly requires judicious selection of the cut-offs that greatly influence the subsequent steps of the analysis. The result is bound to vary for different choices of the thresholds and a poor choice of thresholds may produce an uninformative and uninterpretable result. There is a plethora of helpful guidelines on choosing these thresholds in different contexts (Barnett *and others*, 1999; Kimball *and others*, 2018; Cossarizza *and others*, 2019; BIO-RAD, 2021). However, there is no universal solution or rule of thumb, and the method remains prone to subjectivity bias and lacks statistical robustness.

In this paper, we propose a threshold-free method for distinguishing the difference between the patients with respect to markers of interest. We treat the expression of every marker as a random variable having realizations in the cells (or, pixels) of a patient. For every marker, we compare its probability density across different patients. Our algorithm is as follows. First, for every patient, the probability density of each marker is estimated using kernel density estimation (KDE) (Silverman, 1981; Sheather and Jones, 1991; Ghosh *and others*, 2006). Then for every marker, a density based distance metric (Basu *and others*, 1998; Jones *and others*, 2001) known as Jensen-Shannon distance (Endres and Schindelin, 2003; Fuglede and Topsoe, 2004) is used to quantify the difference in the estimated density across the patients. The distance matrix obtained for every marker can then be used to classify the patients into meaningful groups using hierarchical clustering (Murtagh, 1985; Murtagh and Legendre, 2014). In a simple linear regression framework, the cluster-labels can then be tested for association with a outcome of interest, such as overall survival of the patients. The distance matrix can also be directly used in a linear mixed model (Hoffman, 2013; Seal *and others*, 2021*a*) or equivalently, a kernel machine regression framework (Liu *and others*, 2008; Hua and Ghosh, 2015; Ge *and others*, 2016; Jensen *and others*, 2019). To study the joint effect of markers, one can also sum up the individual marker-specific distance matrices to create an aggregated distance matrix. Using our proposed method, we have analyzed an mIHC dataset on lung cancer (Johnson *and others*, 2021) from the University of Colorado School of Medicine, finding out that HLA-DR marker probability density impacts 5-year overall survival of the patients. We have also applied the proposed method on a publicly available triple negative breast cancer (TNBC) MIBI data (Keren *and others*, 2018) finding out immunoregulatory protein PD-1 to have significant effect on overall survival. We have performed extensive simulation studies mimicking the real datasets to check the reliability and robustness of our method.

2 Materials and Methods

Suppose there are N patients with the j-th patient having n_j cells and M markers. Let $X_{1ij}, X_{2ij}, \ldots, X_{Mij}$ denote scaled expression of M markers in i-th cell of j-th patient for $i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, N$. Let Y ($N \times 1$ vector) be an outcome of interest and C be an $N \times p$ matrix of p covariates.

2.1 Traditional thresholding based approach for clustering patients

To study if abundance of marker k is associated with a patient's survival or any other outcome of interest (Y), the most popular way is to classify the patients into two or more groups using a thresholding based approach. First, the researcher would consider a threshold, t_1 and check how many of the n_j cells of patient j have marker expression more than that threshold i.e. the number of cells with $X_{kij} > t_1$. Such cells are often referred to as the cells positive for marker k. Lets denote the proportion of the cells positive for a marker k in patient j as, $p_{kj} = \sum_{i=1}^{n_j} I(X_{kij} > t_1)/n_j$, where $I(.)$ denotes the indicator function. Next, the researcher would choose one more threshold, t_2 and check how many patients have more than $t_2\%$ cells positive for a marker k i.e. the patients with $p_{kj} > t_2$. Thus, the patients are clustered into two groups, one group with patients having $p_{kj} \leq t_2$ and the other with $p_{kj} > t_2$. Then, the researcher would test if these two clusters of people have differential rate of survival (or, associated with some outcome of interest). This can easily be extend to allow for more than two groups.

Denote the clustering variable as $Z_{kj} \equiv I(p_{kj} > t_2)$ with Z_{kj} being a binary variable taking values zero and one. When Y is a continuous/categorical outcome, a standard multiple linear regression model with $\mathbf{Z}_k = (Z_{k1}, \ldots, Z_{kN})^T$ as a predictor can be written as

$$Y = C\boldsymbol{\beta} + \mathbf{Z}_k\gamma_k + \epsilon,$$

where $\boldsymbol{\beta}, \gamma$ are fixed effects and ϵ is an $N \times 1$ error vector following multivariate normal distribution (MVN) with mean $\mathbf{0}$ and identity covariance matrix $\sigma^2\mathbb{I}_N$. After estimating the parameters, the null hypothesis, $H_0 : \gamma_k = 0$, can be tested using the Wald test (Gourieroux *and others*, 1982).

Next, we consider the case of Y being a survival or recurrence outcome. Let the outcome of the j-th individual be $Y_j = min(T_j, U_j)$, where T_j is the time to event and U_j is the censoring time. Let $\delta_j \equiv I(T_j \leq U_j)$ be the corresponding censoring indicator. Assuming that T_j and U_j are conditionally independent given the covariates for $j = 1, 2, \ldots, n$, the hazard function for the Cox proportional hazards model (Andersen and Gill, 1982; Lin and Wei, 1989; Therneau and Grambsch, 2000) can be written as,

$$\lambda_j(t|C_j, Z_{kj}) = \lambda_0(t)\exp(C_j^T\beta + Z_{kj}\gamma_k), \quad j = 1, 2, \ldots, n \tag{1}$$

where $\lambda_j(t|C_j, Z_{kj})$ is the hazard of the j-th patient at time t, given the vector of covariates C_j and the cluster label Z_{kj} and $\lambda_0(t)$ is an unspecified baseline hazard at time t. To test the null hypothesis: $H_0 : \gamma_k = 0$, a likelihood ratio test (Therneau, 1997) can be considered. The above procedure can be conducted individually for $k = 1, \ldots, M$ and the influential markers can be reported.

As pointed out earlier, the biggest caveat of this approach is finding out appropriate thresholds, t_1 and t_2. In most cases, researchers run the approach with different pairs of (t_1, t_2) and choose the one that gives the most interpretable result. Thus, this step remains entirely subjective and the results may vary largely depending on the selected thresholds.

2.2 Proposed Method: Distance based clustering using individual marker probability density

Keeping the limitation of the thresholding based approach in mind, our aim is to propose a distance metric between the patients based on each marker k that would be devoid of any degree of subjectivity and can easily be tested for association with a outcome of interest.

2.2.1 Formulation of the distance:

For every individual j, we assume that the expression of marker k is a random variable (denoted by X_{kj}) with observed values in n_j cells i.e. $X_{k1j}, X_{k2j}, \ldots, X_{kn_jj}$. Let the probability density function (PDF) of the k-th random variable, X_{kj} be denoted by f_{kj} for $k = 1, \ldots, M$. The distance between two individuals (j, j') w.r.t marker k (denoted by $\text{JSD}_{kjj'}$) can be quantified using density based Jensen-Shannon distance (Endres and Schindelin, 2003; Fuglede and Topsoe, 2004):

$$\text{JSD}_{kjj'} = d(f_{kj}, f_{kj'}) = \sqrt{\int f_{kj}(x) \log \frac{2f_{kj}(x)}{f_{kj}(x) + f_{kj'}(x)} dx + \int f_{kj'}(x) \log \frac{2f_{kj'}(x)}{f_{kj}(x) + f_{kj'}(x)} dx}. \quad (2)$$

A large value of $\text{JSD}_{kjj'}$ will imply that there is a significant difference between the density of k-th marker between patient j and j' and a small value (0 being the smallest) will imply otherwise. The distance matrix between all the individuals (based on k-th marker) can be constructed as, $\text{JSD}_k = [[\text{JSD}_{kjj'}]]$. One can also add up the matrices: JSD_k's ($\text{JSD} = \sum_{k=1}^{M} \text{JSD}_k$) if the goal is to study the distance between individuals w.r.t all the markers jointly.

In real data, the PDF f_{kj} will be unknown. Therefore, we compute corresponding kernel density estimate (KDE) $\hat{f}_{kj}$ (Silverman, 1981; Sheather and Jones, 1991; Ghosh *and others*, 2006) using the observations: X_{kij}'s for $i = 1, \ldots, n_j$. $\hat{f}_{kj}$ typically has the form: $\hat{f}_{kj}(x) = \frac{1}{n_i} \sum_{i=1}^{n_j} w_h (x - X_{kij})$, where w_h is a kernel with bandwidth parameter h. Using the KDEs, $\text{JSD}_{kjj'}$ from Equation (2)

can be estimated as,

$$\text{JSD}_{kjj'} = \sqrt{\sum_{r=1}^{R}\left[\hat{f}_{kj}(x_r)\log\frac{2\hat{f}_{kj}(x_r)}{\hat{f}_{kj}(x_r)+\hat{f}_{kj'}(x_r)} + \hat{f}_{kj'}(x_r)\log\frac{2\hat{f}_{kj'}(x_r)}{\hat{f}_{kj}(x_r)+\hat{f}_{kj'}(x_r)}\right]}, \tag{3}$$

where $x_r, r = 1, \ldots, R$ are grid-points in the interval $(0, 1)$. In our simulations and real data analysis, we keep R at 1024 and choose equidistant grid-points. We have noticed that for $R \geq 512$, the results do not alter. We make sure that the densities integrate to 1 by appropriately scaling them.

2.2.2 Using the distance in association analysis:

Next, we construct suitable association tests for testing the association of the distance matrix with dependent variable, Y.

Test based on hierarchical clustering: The distance matrix (JSD_k) can be subjected to hierarchical clustering (Murtagh, 1985; Murtagh and Legendre, 2014) for clustering the patients into two or multiple groups. Suppose, we obtain a vector of cluster labels: $\mathbf{Z}_k = (Z_{k1}, \ldots, Z_{kN})^T$. Then, exactly the same models, described in Section 2.1 and corresponding tests, can be used to determine if the differential expression of the k-th marker is associated with Y.

Test based on linear mixed model: The distance matrix can be transformed into a similarity matrix (Vert *and others*, 2004) as, $G_k = exp(-\text{JSD}_k)$. When Y is a continuous/categorical outcome, G_k can be incorporated in a linear mixed model framework, particularly popular in the context of heritability estimation (Hoffman, 2013; Seal *and others*, 2021*a*), as,

$$Y = C\boldsymbol{\beta} + g_k + \epsilon,$$

where $\boldsymbol{\beta}$ is the vector of fixed effects, g_k is a vector of random effects following MVN($\mathbf{0}, \sigma^2_{gk} G_k$) and ϵ is an error vector following MVN($\mathbf{0}$, $\sigma^2 \mathbb{I}_N$). After estimating the parameters, the null hypothesis: $H_0 : \sigma^2_{gk} = 0$ can be tested using a likelihood ratio test (Crainiceanu and Ruppert, 2004). Note that, such a linear mixed model setup has been shown to be equivalent to a kernel machine regression framework by Liu *and others* (2008). In a standard kernel machine regression framework, there is one additional width parameter, ρ that has to be estimated.

Next, we consider the case of Y being a survival or recurrence outcome. Using the same definitions and conditional independence assumptions of T_j, U_j and covariates as in Section 2.1, the hazard function for the Cox proportional hazards (PH) model with random effects (Therneau *and others*, 2015) can be written as,

$$\lambda_j(t|C_j, g_{kj}) = \lambda_0(t) \exp(C_j^T \beta + g_{kj}), \quad j = 1, 2, \ldots, n \tag{4}$$

where $\lambda_j(t|C_j, g_{kj})$ is the hazard of the j-th patient at time t, given the vector of covariates C_j and the random effect g_{kj} and $\lambda_0(t)$ is an unspecified baseline hazard at time t. To test the null hypothesis, $H_0 : \sigma^2_{gk} = 0$, a likelihood ratio test based on integrated partial likelihoods (Therneau and Therneau, 2015) can be considered. However, it is to be kept in mind that usually a large sample size is needed to obtain a precise estimate of the random effect variance (Maas and Hox, 2005; Bell *and others*, 2010; Austin and Leckie, 2018). The problem would possibly be exacerbated in the Cox PH model with random effects because the partial likelihood would depend on the number of events (Peduzzi *and others*, 1996; Vittinghoff and McCulloch, 2007; Kocak and Onar-Thomas, 2012; Ogundimu *and others*, 2016). Therefore, we do not recommend using this test unless the sample size is sufficiently large.

3 Results

We first discuss application of our method on the real datasets. We analyzed two datasets: mIHC lung cancer dataset (Johnson *and others*, 2021) and TNBC MIBI dataset (Keren *and others*, 2018). The first dataset has a single functional marker, HLA-DR and the second dataset has four immunoregulatory proteins, PD-1, PD-L1, Lag3 and IDO. We applied the method proposed in Section 2.2 on both the datasets. In all the analyses, the markers were scaled to have expression value between 0 and 1.

3.1 Application to mIHC Lung Cancer data

In the mIHC lung cancer dataset, there are 153 patients each with 3-5 images (in total, 761 images). The patients have varying number of cells identified (from 3755 to 16949). In total there are 1,590,327 cells from two different tissue regions: tumor and stroma. The cells are classified into either of the six different cell types: CD14+, CD19+, CD4+, CD8+, CK+ and Other based on their expression values of different phenotypic markers, such as CD19, CD3, CK, CD8 and CD14. A functional marker, HLA-DR (also referred to as MHCII), is also measured in each of the cells. Using a threshold-based approach, Johnson *and others* (2021) found out that 5-year overall survival of a patient is associated with the percentage of HLA-DR positive CK+ Tumor cells. Here, we were interested in a more general question: whether 5-year overall survival of a patient is associated with the HLA-DR density in CK+ Tumor cells. We first computed the JSD matrix between the patients as discussed in Section 2.2.1 based on the density of HLADR marker in CK+ tumor cells. Next, we performed a hierarchical clustering based on the computed JSD matrix to classify the patients into two groups and test for association with the survival vector as discussed in Section 2.2.2. Figure 1 shows the Kaplan-Meyer curves (Efron, 1988) of the two groups of patients. We

noticed that Hazard Ratio (HR) is large (> 2) and the p-value is significant indicating that patient survival is associated with the probability density of HLA-DR in CK+ tumor cells. Figure 2 shows individual and mean HLADR probability density of different patients from the two clusters. We noticed that the individual densities from cluster 1 were more right-skewed compared to those from cluster 2 which led to the mean density of cluster 1 having very high mode compared to that of cluster 2. We also used the test based on Cox PH model with random effects from Section 2.2.2 in this case. The estimated variance of the random effect was 0.38. Following Therneau *and others* (2015)'s interpretation of the variance parameter in this context, we concluded that there are multiple patients in the study with quite large relative risks, $exp(\sqrt{0.38}) = 1.855$ fold greater than the average patients. However, the LRT based on integrated partial likelihoods was not significant.

3.2 Application to TNBC MIBI data

The triple-negative breast cancer (TNBC) MIBI dataset has images from 41 patients. Keren *and others* (2018) categorized these patients into three groups: "cold", "compartmentalized" and "mixed" based on the level of immune infiltration in the TIME. We were interested in studying the density of the immunoregulatory protein markers, PD1, PD-L1, and Lag3 which have been shown to have immunological relevance (Keren *and others*, 2018; Patwa *and others*, 2021). PD1 and Lag3 are primarily expressed in immune cells and "cold" patients have very few immune cells expressing them. Thus, we focused our analysis on 33 non-"cold" patients. For PD1 and Lag3, we studied their density only in immune cells of a patient and for PD-L1 we studied its density both in immune and tumor cells of a patient. For every marker, we computed the JSD matrix between the patients and performed a hierarchical clustering to classify the patients into two groups as discussed in 2.2.1. Then, we tested the vector of cluster labels for association with two available outcomes: recurrence and survival. Figure 3 shows the Kaplan-Meyer curves corresponding to the three markers for both

survival (left column) and recurrence (right column). We noticed that the HR of survival was large (HR = 2.824) and significant ($p < 0.0346$) for PD1 marker, indicating that the differences in PD1 marker density is associated with risk of death. For PD1, the HR of recurrence was large as well (HR = 2.065) but was not significant. For other two markers, we did not find any significant results (at level 0.05). However, it is worth pointing out that the HR of both survival and recurrence for PD-L1 were quite large (3.49 and 2.84 respectively), alluding to a possible association of PD-L1 marker density with both risk of death and risk of recurrence. We should also keep in mind that the sample size for this particular analysis was quite low which could limit our power.

3.3 Simulation study application

Next, we assessed the performance of JSD based clustering (from Section 2.2) in different simulation setups. We tried to replicate the characteristics of the real dataset discussed in Section 3.1. In Figure 2, we showcased the mean of HLA-DR probability densities of the patients from the two clusters identified by JSD based clustering method. We found out that these mean densities can be well approximated using Beta distributions (Gupta and Nadarajah, 2004) with different set of parameters (α, β). To find out the set of parameters (α, β) that would adequately replicate the mean densities of the two clusters observed, we considered the following strategy. For example, to replicate the mean density of cluster 1, we first computed its empirical mode (say m_1). We wanted to find out parameters α_1, β_1 so that Beta(α_1, β_1) had the same mode and a density function very similar to the empirical one. Matching the modes imply, $m_1 = \frac{\alpha_1 - 1}{\alpha_1 + \beta_1 - 2}$. For a given value of β_1, α_1 is fixed and can be computed using the last equation. We considered multiple values of β_1 and chose the one for which the simulated density appeared to be closest to the real one. We repeated the above steps for replicating the mean density of cluster 2 as well.

The modes of the mean densities of cluster 1 and 2 were respectively, $m_1 = 0.0039$ and $m_2 =$

0.0176. The mean density of cluster 1 was well approximated by a Beta distribution with $\alpha_1 = 2.17, \beta_1 = 300$ and the mean density of cluster 2 by a Beta distribution with $\alpha_2 = 1.78, \beta_2 = 45$. Refer to Figure 4 and 5 to check how well the real and simulated densities agree. Having figured out the parameters of Beta distribution that best summarized the real data mean densities, we focused on two different simulation studies as follows.

3.3.1 Simulation with densities close to the real mean density of cluster 1:

We assumed that there were two groups with N_1 and N_2 patients ($N = N_1 + N_2$). We considered $N_1 = 60, N_2 = 40$. We assumed that each patient j had same number of cells i.e. $n_j = n$. Two values of n: 200 and 2000 were considered. The marker data for a cell of a patient from group 1 was simulated from Beta(2.17, 300) i.e. the distribution which best summarized the real mean density of cluster 1. The marker data of a patient from group 2 was simulated from Beta(x, 300) where x is such that the mode of this distribution was higher than 0.0039 by a percentage of l i.e. x satisfied

$$0.0039\frac{(100+l)}{100} = \frac{x-1}{x+300-2}.$$

Five different values of l: $10, 20, 100, 150$ and 200 were considered. We wanted to study how well JSD based clustering approach can classify the patients into their respective groups. We used two measures: adjusted rand index (ARI) (Santos and Embrechts, 2009), adjusted mutual information (AMI) (Romano *and others*, 2014) which are popular in semi-supervised learning literature. We compared our method with the thresholding based approach described in Section 2.1. As discussed earlier, the thresholding based approach requires two thresholds t_1 and t_2. Since, we did not know what thresholds would possibly be suitable in this simulation setup, we varied t_1 between 95% and 97.5% quantiles of the full marker data (concatenating marker data of all the individuals) and kept t_2 at 0.01. These two methods were referred to as 95% and 97.5% thresholding respectively.

Table 1 lists the performance of all these methods. We noticed that when the number of cells and difference in modes were both small ($n = 200, l = 10$), all the methods performed poorly in terms of ARI and AMI. However, the performance of JSD based clustering improved hugely when the number of cells increased ($n = 2000$). Even for a moderate difference in modes ($l = 50$), JSD based clustering achieved close to 1 accuracy, whereas thresholding methods kept achieving little to zero accuracy.

3.3.2 Simulation with densities close to the real mean density of cluster 2:

We again considered two groups respectively with N_1 and N_2 patients each of whom had n cells. This time, the marker data for a cell of a patient from group 1 was simulated from Beta(1.78, 45) i.e. the distribution which best summarized the real mean density of cluster 2. The marker data for a cell of a patient from group 2 was simulated from Beta(x, 45) where x is such that the mode of this distribution was higher than 0.0176 by a percentage of l i.e. x satisfied

$$0.0176\frac{(100+l)}{100} = \frac{x-1}{x+45-2}.$$

We again considered $N_1 = 60, N_2 = 40$ (and thus, $N = 100$). Two values of n: 200 and 2000 and five values of l: $10, 20, 100, 150, 200$ were considered. Table 2 lists the performance of all the methods. Once again, JSD based clusetring outperformed the thresholding based approaches in all the cases. One interesting observation is that the thresholding based approaches seemed to be performing worse in this simulation setup compared to the last one. Possibly, a different set of (t_1, t_2) would have been more appropriate in this scenario. It reiterates the point that the subjectivity of the thresholding based approaches can hugely alter or affect the performance.

3.3.3 Simulation mimicking the thresholding based approach

Next, we devised a simulation where the true values of the thresholds: (t_1, t_2) were known. And the marker data generation process was somehow dependent on those. Recall that t_1 controls how we define a cell to be positive for a marker and t_2 controls how we cluster the patients into two groups. The simulation strategy was as follows. We considered two groups with respectively N_1 and N_2 patients, each with n cells. We kept $N_1 = 40$ and $N_2 = 60$ and varied n between 200 and 2000. We wanted the patients in group 1 to have $t_2\%$ positive cells and the patients in group 2 to have more than $t_2\%$ positive cells. For patients in group 1, first we made sure that they had $(100 - t_2)\%$ non-positive cells (by randomly choosing $(100 - t_2)/100$ cells out of the total of n). Let $\mathcal{I}$ denote the set of indices of those non-positive cells for a patient j. Next, the marker data of i-th cell from set $\mathcal{I}$, X_{ij} was simulated from Beta$(2.17, 300)$. To avoid any notational confusion, we highlight that X_{ij} can be thought of as X_{kij} from the methods section. Since we were dealing with a single marker, we dropped the index k for simplicity. Once, all the $X'_{ij}s$ (for $i \in \mathcal{I}$) were generated, the values were scaled to be in the range (0, t_1) using the transformation, $X^*_{ij} = \left(\frac{X_{ij} - min_{i \in \mathcal{I}} X_{ij}}{max_{i \in \mathcal{I}} X_{ij} - min_{i \in \mathcal{I}} X_{ij}}\right) t_1$. Next, the marker data of the positive cells (i.e. X_{ij}'s for $i \in \mathcal{I}^c$) were again simulated from Beta$(2.17, 300)$ and a constant of t_1 was added to them, $X^*_{ij} = max\{X_{ij} + t_1, 1\}$, $i \in \mathcal{I}^c$. Thus, we had generated the whole cell-level data of a patient j, X^*_{ij} for $i \in \{1, \ldots, n\}$ making sure there were only $t_2\%$ cells having marker expression more than t_1. For patients in group 2, we had to make sure that they have more than $t_2\%$ positive cells. So, for a patient j in group 2, we simulated a number, t^*_{2j} from Uniform(t_2, 0.9) and repeated all the steps used in simulating group 1 with t^*_{2j} in place of t_2. Note that for both the groups, we used Beta$(2.17, 300)$ to simulate the initial cell-level data (X_{ij}) and then slightly transformed it (X^*_{ij}) to maintain the threshold criteria. One might as well vary the primary distribution as well between the groups but our goal was to create the hardest possible simulation scenario for our method where there would be no explicit difference in marker

ensities between two groups. We considered two different values of t_1: $0.05, 0.1$ and five different values of $t_2 : 0.005, 0.01, 0.05, 0.1$ and 0.2. Table 3 lists the performance of JSD based clustering for all combinations of the parameters. We found out that our method performed better for higher values of t_2. The value of t_1 and the value of n both did not have any apparent impact on the performance. Keep in mind that using the thresholding approach in this simulation setup with the known values of (t_1, t_2) one would get ARI and AMI accuracy of 1 in all the cases. However, as we have repeatedly pointed out, knowing the true values of (t_1, t_2) will never be possible in real data.

4 Discussion

In multiplexed tissue imaging datasets, it is often of interest to stratify the patients based on the profile of functional markers for the purpose of risk assessment (e.g. risk of recurrence, risk of death etc.). The most common approach of grouping the patients into meaningful clusters is a thresholding based method which requires elaborate tuning of one or more thresholds. In consequence, the method remains remains largely subjective and varies from one researcher to another based on their interpretation of the data. In this paper, we develop a threshold-free method for classifying patients based on the marker density in the tumor micro environment (TME). The method is easy to interpret and free from the subjectivity bias.

In our method, we treat marker expression in a patient as a random variable and compute its kernel density estimate based on the cells of the TME. Once the marker density estimates for all the patients have been computed, a distance metric known as Jensen-Shannon distance is considered to quantify the difference in marker densities between the patients. If the distance between two patients is large, it means that they have very different marker expression profiles. Next, the computed distance matrix can be used in either of the following two ways. It can be subjected

to hierarchical clustering to group the patients into clusters and the cluster-labels can be tested for association with outcomes of interest (e.g. recurrence, survival). Or it can be used directly in a linear mixed model setup (or, equivalently a kernel machine regression model) for testing association with outcomes of interest.

We analyzed two highly complex multiplex tissue imaging datasets, an mIHC lung cancer dataset from University of Colorado School of Medicine and a publicly available triple negative breast cancer MIBI data. In the mIHC dataset, we found out that the difference in HLA-DR marker density between patients was significantly associated with their 5-year overall survival probability. In the MIBI dataset, we found out that the difference in the density of immunoregulatory protein PD-1 was associated with the difference in overall survival. Next, we replicated the characteristics of the mIHC dataset in two simulation scenarios and showcased the robustness of our method in comparison with the thresholding-based method. In the final simulation setup, we aimed to simulate a dataset fully abiding the principles of the thresholding method. We showed that our method performed competently even in that scenario.

One point that we have not covered in great detail in this paper is how to summarize the joint information from different markers effectively. Recall that we have mentioned summing up the distance matrices corresponding to different markers creating a new distance matrix that would be expected to capture some sort of aggregated information. But, it needs to be studied in detail how to make sure that the individual distance matrices do not dominate the aggregated distance matrix. Otherwise, the new matrix will not be very meaningful. We believe that a carefully weighted sum of the individual distance matrices can potentially be one solution and intend to study this in further details as a part of our next work. Also, we would like to apply our method on datasets coming from other imaging platforms, such as CODEX (Goltsev *and others*, 2018) and Visium (Tippani *and others*, 2021).

5 Software

Software in the form of a GitHub *R* package, together with an example dataset and complete documentation is available at this link, https://github.com/sealx017/DenVar.

6 Acknowledgments

We thank the Human Immune Monitoring Shared Resource and support of the University of Colorado Human Immunology and Immunotherapy Initiative for their expert assistance in multiplex IHC and generation of the ovarian and lung datasets. We acknowledge the support of the University of Colorado Cancer Center Support Grant (P30CA046934). S.S. is funded by the Grohne-Stepp Endowment from the University of Colorado Cancer Center.

References

Ali, H Raza, Jackson, Hartland W, Zanotelli, Vito RT, Danenberg, Esther, Fischer, Jana R, Bardwell, Helen, Provenzano, Elena, Rueda, Oscar M, Chin, Suet-Feung, Aparicio, Samuel *and others*. (2020). Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer* **1**(2), 163–175.

Andersen, Per Kragh and Gill, Richard D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Angelo, Michael, Bendall, Sean C, Finck, Rachel, Hale, Matthew B, Hitzman, Chuck, Borowsky, Alexander D, Levenson, Richard M, Lowe, John B, Liu, Scot D,

Zhao, Shuchun *and others*. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature medicine* **20**(4), 436–442.

Austin, Peter C and Leckie, George. (2018). The effect of number of clusters and cluster size on statistical power and type i error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of statistical computation and simulation* **88**(16), 3151–3163.

Barnett, D, Janossy, G, Lubenko, A, Matutes, E, Newland, A and Reilly, JT. (1999). Guideline for the flow cytometric enumeration of cd34+ haematopoietic stem cellsprepared by the cd34+ haematopoietic stem cell working party. *Clinical & Laboratory Haematology* **21**(5), 301–308.

Basu, Ayanendranath, Harris, Ian R, Hjort, Nils L and Jones, MC. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**(3), 549–559.

Bataille, Frauke, Troppmann, Sabine, Klebl, Frank, Rogler, Gerhard, Stoelcker, Benjamin, Hofstadter, Ferdinand, Bosserhoff, Anja-Katrin and Rümmele, Petra. (2006). Multiparameter immunofluorescence on paraffin-embedded tissue sections. *Applied Immunohistochemistry & Molecular Morphology* **14**(2), 225–228.

Bell, Bethany A, Morgan, Grant B, Kromrey, Jeffrey D and Ferron, John M. (2010). The impact of small cluster size on multilevel models: a monte carlo examination of two-level models with binary and continuous predictors. *JSM Proceedings, Survey Research Methods Section* **1**(1), 4057–4067.

Binnewies, Mikhail, Roberts, Edward W, Kersten, Kelly, Chan, Vincent, Fearon, Douglas F, Merad, Miriam, Coussens, Lisa M, Gabrilovich, Dmitry I, Ostrand-

Rosenberg, Suzanne, Hedrick, Catherine C *and others*. (2018). Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine* **24**(5), 541–550.

BIO-RAD. (2021, 09). Flow Cytometry Basics Guide.

Bulian, Pietro, Shanafelt, Tait D, Fegan, Chris, Zucchetto, Antonella, Cro, Lilla, Nückel, Holger, Baldini, Luca, Kurtova, Antonina V, Ferrajoli, Alessandra, Burger, Jan A *and others*. (2014). Cd49d is the strongest flow cytometry–based predictor of overall survival in chronic lymphocytic leukemia. *Journal of Clinical Oncology* **32**(9), 897.

Chang, Boyang, Shen, Lujun, Wang, Kefeng, Jin, Jietian, Huang, Tao, Chen, Qifeng, Li, Wang and Wu, Peihong. (2018). High number of pd-1 positive intratumoural lymphocytes predicts survival benefit of cytokine-induced killer cells for hepatocellular carcinoma patients. *Liver International* **38**(8), 1449–1458.

Chang, Qing, Ornatsky, Olga I, Siddiqui, Iram, Loboda, Alexander, Baranov, Vladimir I and Hedley, David W. (2017). Imaging mass cytometry. *Cytometry part A* **91**(2), 160–169.

Chen, Chang-Long, Pan, Qiu-Zhong, Zhao, Jing-Jing, Wang, Ying, Li, Yong-Qiang, Wang, Qi-Jing, Pan, Ke, Weng, De-Sheng, Jiang, Shan-Shan, Tang, Yan *and others*. (2016). Pd-l1 expression as a predictive biomarker for cytokine-induced killer cell immunotherapy in patients with hepatocellular carcinoma. *Oncoimmunology* **5**(7), e1176653.

Cossarizza, Andrea, Chang, Hyun-Dong, Radbruch, Andreas, Acs, Andreas, Adam, Dieter, Adam-Klages, Sabine, Agace, William W, Aghaeepour, Nima, Akdis, Mübeccel, Allez, Matthieu *and others*. (2019). Guidelines for the use of flow cytometry and cell sorting in immunological studies. *European journal of immunology* **49**(10), 1457–1973.

Costa, AFO, Menezes, DL, Pinheiro, LHS, Sandes, AF, Nunes, MAP, Junior, DP Lyra and Schimieguel, DM. (2017). Role of new immunophenotypic markers on prognostic and overall survival of acute myeloid leukemia: a systematic review and meta-analysis. *Scientific reports* **7**(1), 1–11.

Crainiceanu, Ciprian M and Ruppert, David. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(1), 165–185.

Efron, Bradley. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American statistical Association* **83**(402), 414–425.

Endres, Dominik Maria and Schindelin, Johannes E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory* **49**(7), 1858–1860.

Fuglede, Bent and Topsoe, Flemming. (2004). Jensen-shannon divergence and hilbert space embedding. In: *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*. IEEE. p. 31.

Ge, T, Smoller, JW and Sabuncu, MR. (2016). Kernel machine regression in neuroimaging genetics. *Machine Learning and Medical Imaging*, 31–68.

Ghosh, Anil K, Chaudhuri, Probal and Sengupta, Debasis. (2006). Classification using kernel density estimates: Multiscale analysis and visualization. *Technometrics* **48**(1), 120–132.

Giesen, Charlotte, Wang, Hao AO, Schapiro, Denis, Zivanovic, Nevena, Jacobs, Andrea, Hattendorf, Bodo, Schüffler, Peter J, Grolimund, Daniel, Buhmann, Joachim M, Brandt, Simone *and others*. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11**(4), 417–422.

Goltsev, Yury, Samusik, Nikolay, Kennedy-Darling, Julia, Bhate, Salil, Hale, Matthew, Vazquez, Gustavo, Black, Sarah and Nolan, Garry P. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell* **174**(4), 968–981.

Gourieroux, Christian, Holly, Alberto and Monfort, Alain. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica: journal of the Econometric Society*, 63–80.

Gupta, Arjun K and Nadarajah, Saralees. (2004). *Handbook of beta distribution and its applications*. CRC press.

Halse, H, Colebatch, AJ, Petrone, P, Henderson, MA, Mills, JK, Snow, H, Westwood, JA, Sandhu, S, Raleigh, JM, Behren, Andreas *and others*. (2018). Multiplex immunohistochemistry accurately defines the immune context of metastatic melanoma. *Scientific reports* **8**(1), 1–14.

Han, Yanyan, Liu, Dandan and Li, Lianhong. (2020). Pd-1/pd-l1 pathway: current researches in cancer. *American journal of cancer research* **10**(3), 727.

Hoffman, Gabriel E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one* **8**(10), e75707.

Hua, Wen-Yu and Ghosh, Debashis. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics* **71**(3), 812–820.

Huang, Wei, Hennrick, Kenneth and Drew, Sally. (2013). A colorful future of quantitative pathology: validation of vectra technology using chromogenic multiplexed immunohistochemistry and prostate tissue microarrays. *Human pathology* **44**(1), 29–38.

Jendro, Michael, Goronzy, Jorg J and Weyand, Cornelia M. (1991). Structural and functional characterization of hla-dr molecules circulating in the serum. *Autoimmunity* **8**(4), 289–296.

Jensen, Alexandria M, Tregellas, Jason R, Sutton, Brianne, Xing, Fuyong and Ghosh, Debashis. (2019). Kernel machine tests of association between brain networks and phenotypes. *Plos one* **14**(3), e0199340.

Johnson, Amber M, Boland, Jennifer M, Wrobel, Julia, Klezcko, Emily K, Weiser-Evans, Mary, Hopp, Katharina, Heasley, Lynn, Clambey, Eric T, Jordan, Kimberly, Nemenoff, Raphael A *and others*. (2021). Cancer cell-specific mhcii expression as a determinant of the immune infiltrate organization and function in the non-small cell lung cancer tumor microenvironment. *Journal of Thoracic Oncology*.

Johnson, Amber M, Bullock, Bonnie L, Neuwelt, Alexander J, Poczobutt, Joanna M, Kaspar, Rachael E, Li, Howard Y, Kwak, Jeff W, Hopp, Katharina, Weiser-Evans, Mary CM, Heasley, Lynn E *and others*. (2020). Cancer cell–intrinsic expression of mhc class ii regulates the immune microenvironment and response to anti–pd-1 therapy in lung adenocarcinoma. *The Journal of Immunology* **204**(8), 2295–2307.

Jondal, M, Holm, Gt and Wigzell, H. (1972). Surface markers on human t and b lymphocytes: I. a large population of lymphocytes forming nonimmune rosettes with sheep red blood cells. *The Journal of experimental medicine* **136**(2), 207–215.

Jones, MC, Hjort, Nils Lid, Harris, Ian R and Basu, Ayanendranath. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**(3), 865–873.

Keren, Leeat, Bosse, Marc, Marquez, Diana, Angoshtari, Roshan, Jain, Samir, Varma, Sushama, Yang, Soo-Ryum, Kurian, Allison, Van Valen, David, West, Robert *and others*. (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**(6), 1373–1387.

Keren, Leeat, Bosse, Marc, Thompson, Steve, Risom, Tyler, Vijayaragavan, Kausalia, McCaffrey, Erin, Marquez, Diana, Angoshtari, Roshan, Greenwald, Noah F, Fienberg, Harris *and others*. (2019). Mibi-tof: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances* **5**(10), eaax5851.

Kimball, Abigail K, Oko, Lauren M, Bullock, Bonnie L, Nemenoff, Raphael A, van Dyk, Linda F and Clambey, Eric T. (2018). A beginner's guide to analyzing and visualizing mass cytometry data. *The Journal of Immunology* **200**(1), 3–22.

Kocak, Mehmet and Onar-Thomas, Arzu. (2012). A simulation-based evaluation of the asymptotic power formulas for cox models in small sample cases. *The American Statistician* **66**(3), 173–179.

Koguchi, Yoshinobu, Hoen, Helena M, Bambina, Shelly A, Rynning, Michael D, Fuerstenberg, Richard K, Curti, Brendan D, Urba, Walter J, Milburn, Christina, Bahjat, Frances Rena, Korman, Alan J *and others*. (2015). Serum immunoregulatory proteins as predictors of overall survival of metastatic melanoma patients treated with ipilimumab. *Cancer research* **75**(23), 5084–5092.

Lin, Danyu Y and Wei, Lee-Jen. (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association* **84**(408), 1074–1078.

Liu, Dawei, Ghosh, Debashis and Lin, Xihong. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics* **9**(1), 1–11.

Maas, Cora JM and Hox, Joop J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* **1**(3), 86–92.

Missassi, G, MRV, Ikoma-Colturato, CM, Bortolucci, JE, Conte-Spilari, AJ, Simioni, MP, Souza and VAR, Colturato. (2021). Immunophenotypic markers associated with minimal residual disease status and outcome in patients with multiple myeloma undergoing autologous stem cell transplantation. *Annals of Hematology & Oncology* **8**, 0–0.

Murtagh, Fionn. (1985). Multidimensional clustering algorithms. *Compstat lectures*.

Murtagh, Fionn and Legendre, Pierre. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification* **31**(3), 274–295.

Nguyen, Linh T and Ohashi, Pamela S. (2015). Clinical blockade of pd1 and lag3—potential mechanisms of action. *Nature Reviews Immunology* **15**(1), 45–56.

Oczenski, Wolfgang, Krenn, Herbert, Jilch, Ruth, Watzka, Herbert, Waldenberger, Ferdinand, Köller, Ursula, Schwarz, Sylvia and Fitzgerald, Robert D. (2003). Hla-dr as a marker for increased risk for systemic inflammation and septic complications after cardiac surgery. *Intensive care medicine* **29**(8), 1253–1257.

Ogundimu, Emmanuel O, Altman, Douglas G and Collins, Gary S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of clinical epidemiology* **76**, 175–182.

Patwa, Aalok N, Yamashita, Rikiya, Long, Jin, Keren, Leeat, Angelo, Michael and Rubin, Daniel. (2021). Multiplexed imaging analysis of the tumor-immune microenvironment reveals predictors of outcome in triple-negative breast cancer. *bioRxiv*.

Peduzzi, Peter, Concato, John, Kemper, Elizabeth, Holford, Theodore R and Feinstein, Alvan R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology* **49**(12), 1373–1379.

Phillips, Darci, Schürch, Christian M, Khodadoust, Michael S, Kim, Youn H, Nolan, Garry P and Jiang, Sizun. (2021). Highly multiplexed phenotyping of immunoregulatory proteins in the tumor microenvironment by codex tissue imaging. *Frontiers in Immunology* **12**.

Pollan, Sara, Hanifi, Arezoo, Nagy, Mate, Stavrou, Nicholas, Parnell, Erinn, Gozo, Maricel, Attanasio, Nickolas, William, Josette and Au, Qingyan. (2020). Profiling exhausted t cells using vectra® polaris™ multiplex immunofluorescence assay in hnscc.

Ptacek, Jason, Locke, Darren, Finck, Rachel, Cvijic, Mary-Ellen, Li, Zhuyin, Tarolli, Jay G, Aksoy, Murat, Sigal, Yari, Zhang, Yi, Newgren, Matt *and others*. (2020). Multiplexed ion beam imaging (mibi) for characterization of the tumor microenvironment across tumor types. *Laboratory Investigation* **100**(8), 1111–1123.

Romano, Simone, Bailey, James, Nguyen, Vinh and Verspoor, Karin. (2014). Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In: *International Conference on Machine Learning*. PMLR. pp. 1143–1151.

Sahlberg, Sara Häggblad, Spiegelberg, Diana, Glimelius, Bengt, Stenerlöw, Bo and Nestor, Marika. (2014). Evaluation of cancer stem cell markers cd133, cd44, cd24: association with akt isoforms and radiation resistance in colon cancer cells. *PloS one* **9**(4), e94621.

Santos, Jorge M and Embrechts, Mark. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In: *International conference on artificial neural networks*. Springer. pp. 175–184.

Saraiva, Diana P, Jacinto, António, Borralho, Paula, Braga, Sofia and Cabral, M Guadalupe. (2018). Hla-dr in cytotoxic t lymphocytes predicts breast cancer patients' response to neoadjuvant chemotherapy. *Frontiers in immunology* **9**, 2605.

Seal, Souvik, Datta, Abhirup and Basu, Saonli. (2021*a*). Rapid estimation of snp heritability using predictive process approximation in large scale cohort studies. *bioRxiv*.

Seal, Souvik, Wrobel, Julia, Johnson, Amber M, Nemenoff, Raphael A, Schenk, Erin L, Bitler, Benjamin G, Jordan, Kimberly R and Ghosh, Debashis. (2021*b*). On clustering for cell phenotyping in multiplex immunohistochemistry (mihc) and multiplexed ion beam imaging (mibi) data.

Sheather, Simon J and Jones, Michael C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3), 683–690.

Shipkova, Maria and Wieland, Eberhard. (2012). Surface markers of lymphocyte activation and markers of cell proliferation. *Clinica chimica acta* **413**(17-18), 1338–1349.

Silverman, Bernard W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)* **43**(1), 97–99.

Tan, Wei Chang Colin, Nerurkar, Sanjna Nilesh, Cai, Hai Yun, Ng, Harry Ho Man, Wu, Duoduo, Wee, Yu Ting Felicia, Lim, Jeffrey Chun Tatt, Yeong, Joe and Lim,

Tony Kiat Hon. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications* **40**(4), 135–153.

Therneau, Terry *and others*. (2015). Mixed effects cox models. *CRAN repository*.

Therneau, Terry M. (1997). Extending the cox model. In: *Proceedings of the First Seattle symposium in biostatistics*. Springer. pp. 51–84.

Therneau, Terry M and Grambsch, Patricia M. (2000). The cox model. In: *Modeling survival data: extending the Cox model*. Springer, pp. 39–77.

Therneau, Terry M and Therneau, Maintainer Terry M. (2015). Package 'coxme'. *R package version* **2**(5).

Tippani, Madhavi, Divecha, Heena Rajesh, Catallini, Joseph L, Weber, Lukas M, Spangler, Abby, Jaffe, Andrew E, Hicks, Stephanie C, Martinowich, Keri, Collado-Torres, Leonardo, Page, Stephanie C *and others*. (2021). Vistoseg: a matlab pipeline to process, analyze and visualize high resolution histology images for visium spatial transcriptomics data. *bioRxiv*.

Vert, Jean-Philippe, Tsuda, Koji and Schölkopf, Bernhard. (2004). A primer on kernel methods. *Kernel methods in computational biology* **47**, 35–70.

Vittinghoff, Eric and McCulloch, Charles E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology* **165**(6), 710–718.

Whiteside, TL. (2008). The tumor microenvironment and its role in promoting tumor growth. *Oncogene* **27**(45), 5904–5912.

Yang, Zhi-Zhang, Kim, Hyo Jin, Villasboas, Jose C, Price-Troska, Tammy, Jalali, Shahrzad, Wu, Hongyan, Luchtel, Rebecca A, Polley, Mei-Yin C, Novak, Anne J and Ansell, Stephen M. (2019). Mass cytometry analysis reveals that specific intratumoral cd4+ t cell subsets correlate with patient survival in follicular lymphoma. *Cell reports* **26**(8), 2178–2193.

Zola, Heddy, Swart, Bernadette, Nicholson, Ian and Voss, Elena. (2007). *Leukocyte and stromal cell molecules: the CD markers*. John Wiley & Sons.

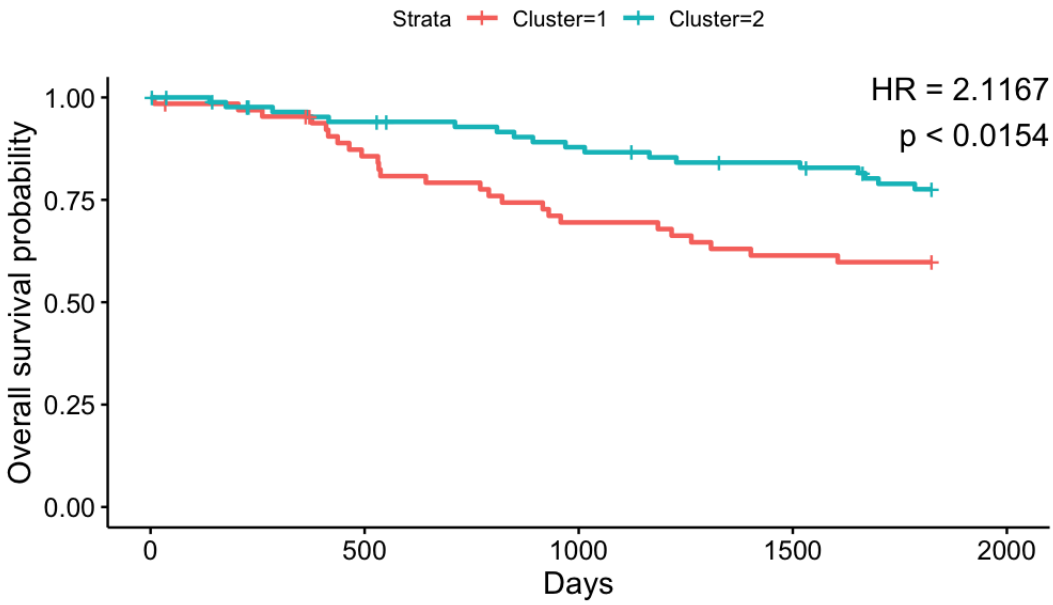

Figure 1: KM curves of 5-year overall Survival of 153 patients color coded by the clusters found using HLA-DR marker density in CK+ Tumor cells. Also, displayed are Hazard ratio (HR) and the p-value corresponding to the test, $H_0 : \gamma = 0$ from Equation 1. Notice that HR is large (> 2) and the p-value is significant as well indicating that the two clusters have significant difference in survival probability.

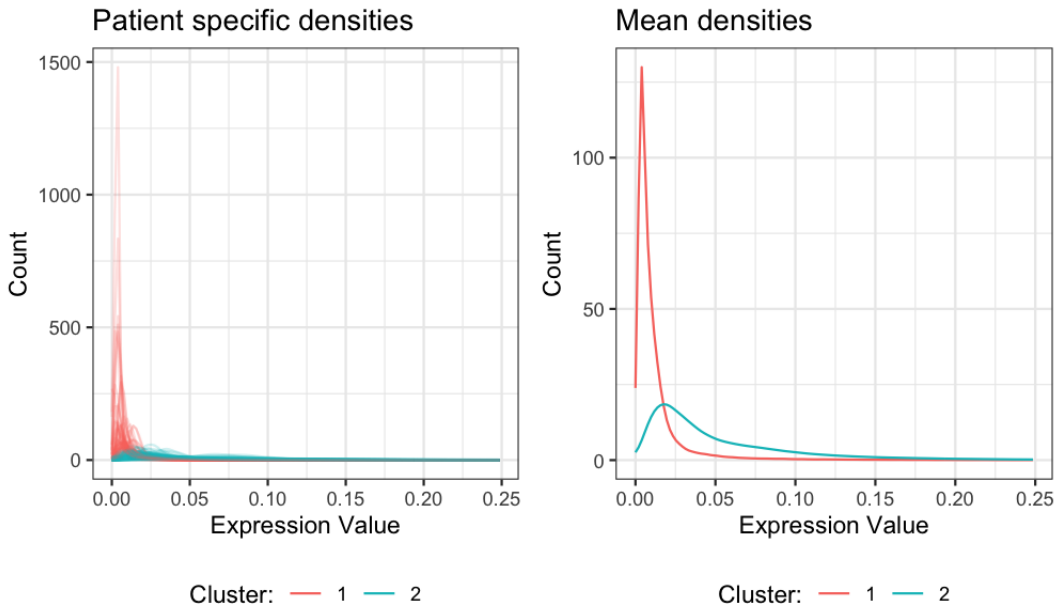

Figure 2: Individual (on the left) and mean (on the right) HLA-DR marker probability density (in CK+ tumor cells) of the patients from the two clusters found using JSD based clustering proposed in Section 2.2. Notice that the individual densities from cluster 1 are more right-skewed than the densities from cluster 2. Consequently, the mean density of cluster 1 is also more right-skewed than that of cluster 2 and has a much higher peak.

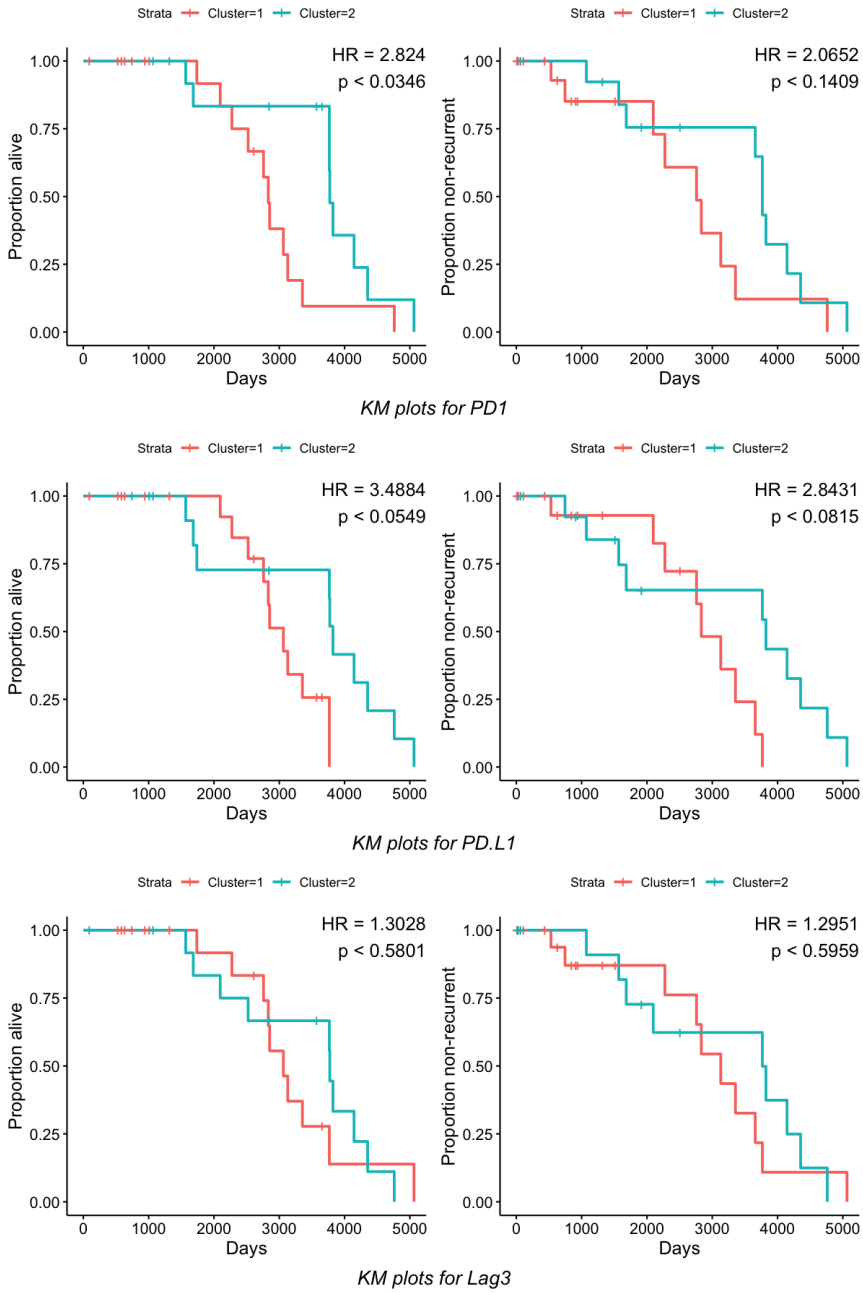

Figure 3: KM Plots of overall survival (left) and recurrence (right) of 33 patients color coded by the clusters found using markers: PD1, PD-L1 and Lag3, using our method. We notice that the difference in PD1 density has significant effect on overall survival.

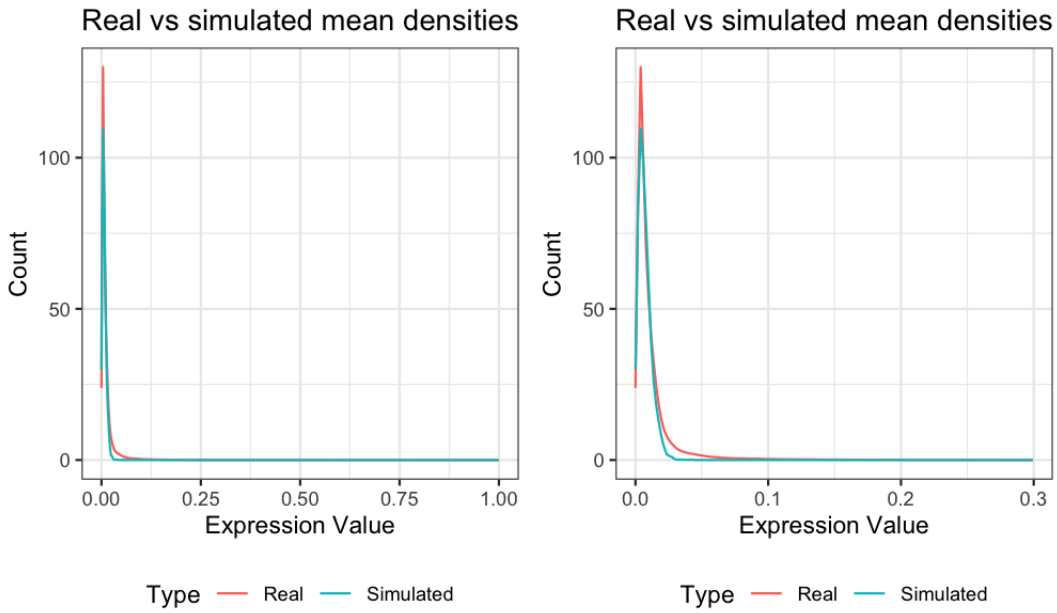

Figure 4: Comparing the probability density of Beta$(2.17, 300)$ to the real mean density of cluster 1. On the left, are shown the densities on the entire range of expression value: (0, 1). On the right, we zoom into the lower expression values and the same densities are shown only between (0, 0.3). Even though there appears to be a difference in the modes of the densities, their overall shapes are quite close.

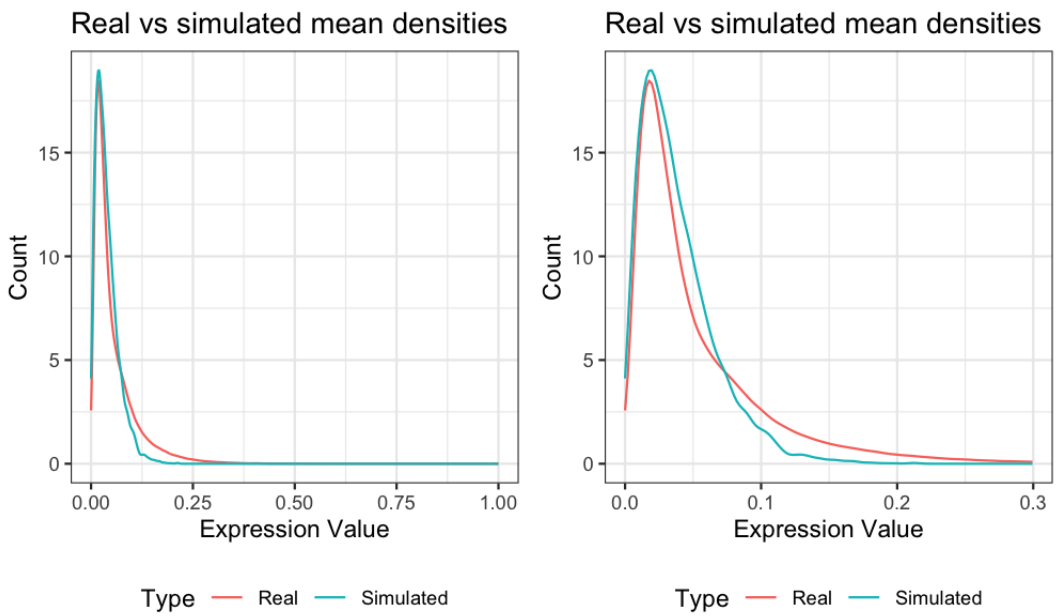

Figure 5: Comparing the probability density of Beta$(1.78, 45)$ to the real mean density of cluster 2. On the left, are shown the densities on the entire range of expression value: (0, 1). On the right, we zoom into the lower expression values and the same densities are shown only between (0, 0.3). The overall shapes of the densities are pretty similar.

Table 1: Performance of different methods in the simulation with densities close to the mean density of cluster 1 as described in Section 3.3.1. JSD based clustering performs sytematically better than the thresholding approaches in all the cases. When the number of cells is large, JSD based clustering performs well even for small differences in modes.

Measure of performance	Number of cells	Percentage difference in modes	JSD based clustering	95% thresholding	97.5% thresholding
ARI	$n = 200$	10	0.0744	0.0014	0.0234
		20	0.3988	0.0029	0.0551
		50	0.9808	0.0236	0.2264
		100	1.0000	0.1979	0.6324
		200	1.0000	0.8628	0.9570
	$n = 2000$	10	0.8029	0.0000	0.0000
		20	0.9530	0.0000	0.0001
		50	1.0000	0.0000	0.0299
		100	1.0000	0.0040	0.8105
		200	1.0000	0.9907	1.0000
AMI	$n = 200$	10	0.0713	0.0026	0.0144
		20	0.3344	0.0041	0.0358
		50	0.9662	0.0283	0.1802
		100	1.0000	0.2001	0.5421
		200	1.0000	0.8088	0.9239
	$n = 2000$	10	0.7233	0.0000	0.0000
		20	0.9976	0.0000	0.0001
		50	1.0000	0.0000	0.0322
		100	1.0000	0.0039	0.7609
		200	1.0000	0.9865	1.0000

Table 2: Performance of different methods in the simulation with densities close to the mean density of cluster 2 as described in Section 3.3.2. JSD based clustering performs systematically better than the thresholding approaches in all the cases. When the number of cells is large, JSD based clustering performs well even for small differences in modes.

Measure of performance	Number of cells	Percentage difference in modes	JSD based clustering	95% thresholding	97.5% thresholding
ARI	$n = 200$	10	0.0345	0.0005	0.0171
		20	0.2003	0.0016	0.0411
		50	0.8656	0.0119	0.1395
		100	0.9996	0.0699	0.4153
		200	1.0000	0.5428	0.8696
	$n = 2000$	10	0.5157	0.0000	0.0000
		20	0.9737	0.0000	0.0001
		50	1.0000	0.0000	0.0045
		100	1.0000	0.0000	0.2627
		200	1.0000	0.4074	0.9984
AMI	$n = 200$	10	0.0363	0.0022	0.0110
		20	0.1727	0.0030	0.0237
		50	0.8035	0.0142	0.1052
		100	0.9993	0.0787	0.3418
		200	1.0000	0.4990	0.7992
	$n = 2000$	10	0.4352	0.0000	0.0000
		20	0.9530	0.0000	0.0001
		50	1.0000	0.0000	0.0052
		100	1.0000	0.0000	0.2558
		200	1.0000	0.3829	0.9971

Table 3: Performance of JSD based clustering in the simulation from Section 3.3.3. The method performs better for larger values of t_2, whereas t_1 does not seem to affect the performance.

Number of cells	Measure of performance	t_2 :	0.005	0.01	0.05	0.1	0.2
$n = 200$	ARI	$t_1 = 0.05$	0.760	0.791	0.801	0.938	1.000
		$t_1 = 0.1$	0.784	0.727	0.815	0.957	0.987
	AMI	$t_1 = 0.05$	0.764	0.772	0.756	0.922	1.000
		$t_1 = 0.1$	0.712	0.719	0.765	0.943	0.987
$n = 2000$	ARI	$t_1 = 0.05$	0.778	0.727	0.800	0.936	1.000
		$t_1 = 0.1$	0.784	0.727	0.808	0.965	1.000
	AMI	$t_1 = 0.05$	0.733	0.678	0.755	0.919	1.000
		$t_1 = 0.1$	0.734	0.680	0.762	0.951	1.000