MULTIMODAL DATA ANALYSIS
APPLIED TO A MEDICAL SETTING

BY

VAISHNAVI SUBRAMANIAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Professor Minh Do

ABSTRACT

Complex diseases, such as cancer, have traditionally been studied using genetic data, or images alone. To understand the biology of such diseases, joint analysis of multiple data modalities could provide interesting insights. We propose the use of canonical correlation analysis (CCA) as a preliminary discovery tool for identifying connections across modalities, specifically between gene expression and features describing cell and nucleus shape, texture, and stain intensity in histopathological images.

It is also important to capture the interaction between different types of cells, an important indicator of disease status. To that end, it is crucial to quantify and utilize the spatial distribution of various cell types within the examined tissue at different scales. We employ Ripley's K-statistic, a traditional feature employed in geographical information systems, which captures spatial distribution patterns of individual point sets and interactions between multiple point sets. We propose to improve the histopathology image features by incorporating this descriptor to capture the spatial distribution of the cells, and interactions between lymphocytes and epithelial cells.

Applied to 615 breast cancer samples from The Cancer Genome Atlas, CCA revealed significant correlation of several image features with expression of PAM50 genes, known to be linked to outcome. Sparse CCA, an extension of CCA based on sparsity, revealed associations with enrichment of pathways implicated in cancer without leveraging prior biological understanding. The utility of the Ripley's K-statistic on 710 TCGA breast invasive carcinoma (BRCA) patients' histopathology images in the context of imaging-genetics is demonstrated by its superior correlations with gene expressions. These findings affirm the utility of CCA for joint phenotype-genotype analysis of cancer, and the importance of capturing spatial features at multiple scales.

असतो मा सद्गमय ।
तमसो मा ज्योतिर्गमय ।

From ignorance, lead us to truth.
From darkness, lead us to light.

ACKNOWLEDGMENTS

I would like to heartily thank my thesis advisor Prof. Minh Do for his constant encouragement and guidance. Working with him, I have realized the importance of well-directed enthusiasm, in both research and life. His ability to identify the core intuition and broader vision of complex ideas is something I hope to imbibe. His readiness to collaborate with others has enabled me to gain invaluable exposure to important problems in related areas.

I am thankful to Benjamin Chidester for all the advice and support during my first steps into research. His willingness to patiently listen to any idea or problem and provide critical thoughts is an important quality for every researcher. I would like to take this opportunity to thank my collaborators in this project for their valuable contributions: Prof. Jian Ma for advice on genomics; Weizhao Tang for laying the foundations for the work in spatial statistics; Kushagra Tiwary for helping with image processing and detection tasks.

I have gained immense knowledge from the courses in the past two years, and would like to thank Prof. Rayadurgam Srikant, Prof. Alex Schwing, Prof. Pierre Moulin, Prof Zhi-Pei Liang and Prof. Bruce Hajek for the same. My teaching assistant role with Prof. Lav Varshney last semester was rewarding, and I am thankful for the experience.

This thesis would not have been possible without the assistance of the members of my research group. In particular, I would like to thank Mona Zehni and Khoi Nguyen Mac for all the insightful discussions, Trong Nguyen for providing key tips like ‘add more layers’, and Raymond Yeh and Teck Yian Lim for the short and succinct explanations of advanced concepts.

The importance of friends cannot be neglected in any endeavour. I would like to thank Amitha Sandur, Agrima Bansal, Sameer Khan, Abhishek Narwekar, Pratik Deogankar, Ishan Deshpande and Konik Kothari for providing

comfort and support in tough times. Anushri Pampari, Azin Heidarshenas, Krishnakant Saboo, Suraj Jog, Surya Sankagiri, Ashok Makkuva, Amish Goel, Harsh Gupta and Anwesa Choudhuri have helped me become a better researcher and person. Conversations with Shilpa Ramesh and Aditi Udupa on philosophy and principles of life have been critical in molding my outlook in life. I would like to express gratitude towards my friends from IIT Bombay for tolerating my inefficient communications and providing inspiration from miles away.

I am deeply indebted to my brother Vignesh, parents and family for their overwhelming support and care throughout. Finally, I thank God for everything.

CONTENTS

Chapter 1 INTRODUCTION 1
1.1 Multimodal Datasets in Healthcare 1
1.2 Multimodal Data Integration 2
1.3 Spatial Distribution of Cells 3
1.4 Main Contributions 4

Chapter 2 CORRELATION ANALYSIS 6
2.1 Canonical Correlation Analysis 6
2.2 Extensions to CCA 8

Chapter 3 SPATIAL DISTRIBUTION 11
3.1 Ripley's K-function 11
3.2 Examples on Simulated Data 13

Chapter 4 EXPERIMENTS AND RESULTS 16
4.1 Data 16
4.2 CCA with Nuclei Descriptors 17
4.3 CCA with Nuclei Descriptors and Spatial Distribution Features 24

Chapter 5 CONCLUSIONS 29

BIBLIOGRAPHY 30

Chapter 1

INTRODUCTION

1.1 Multimodal Datasets in Healthcare

In 2016, it was reported that the United States healthcare expense was over 3.2 trillion dollars, which is equivalent to the entire world's expenses for the same year on information technology. Providing effective healthcare at reduced cost to the people is a challenge even today. In an NIH study in 2017, it was estimated that the number of cancer cases and deaths due to cancer is expected to increase over 50% worldwide in less than two decades. Thus, it is important to develop a better understanding of diseases such as cancer, so as to provide accurate diagnosis and enable the development of precision medicine.

With an increase in the availability of multimodal datasets, that is, datasets which provide information about the same set of samples/patients from multiple views, there is an opportunity for better understanding and prediction. The Cancer Genome Atlas (TCGA) [1] and The Cancer Imaging Archive (TCIA) [2] are publicly available datasets comprising data of different cancer patients for a range of different cancers. Similar datasets have been recently developed for other important diseases: Alzheimer's Disease Neuroimaging Initiative (ADNI) [3] and Parkinson's Progressive Markers Initiative (PPMI) [4]. These multimodal datasets provide different sets of information, as shown in Figure 1.1. For example, TCGA provides information about genetics, which captures information of the internal state of cells and genes and whole-slide images (WSI) of biopsy tissues which capture information about the immediate micro-environment of the different cells. The accompanying dataset TCIA provides radiology images of MRI and CT scans showing the affected tissue with respect to a broader environment to understand the spread of the tumor.

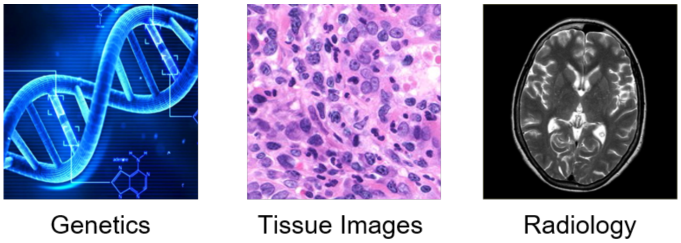


Figure 1.1: Different modalities of patient data.

Given the multimodal datasets which capture different viewpoints of the same internal tissue, there is an inherent shared state of tissue being captured by the different views. In addition, each view will capture important information specific to that view/modality. By leveraging the multimodal datasets it is likely possible to improve the understanding of various diseases by looking at the same affected region from multiple viewpoints and also develop interesting genotype-phenotype associations. It could be possible that we can point out which image properties are affected by which genes (while working with genomics and WSI) or which particular physical regions are affected by which genes (while merging radiology data and genomics).

1.2 Multimodal Data Integration

Simple concatenation of features obtained from each modality, to obtain a long feature vector for every patient, fails to take into account the underlying similarities and the contrasting differences between the viewpoints/modalities or provide an insight on the shared information between the two modalities. To integrate data from different modalities, different techniques have been developed in the past. A comprehensive review on multimodal data fusion for non-neural network techniques is provided by Lahat *et al.* [5], covering data-driven and model-driven methods. A few important methods mentioned for the fusion are independent component analysis (ICA) and its extension, joint ICA, and tensor factorization to separate out r sources which are hidden. More details of these methods can be found in [6].

Another way to work with multimodal data is to identify the common shared information between the different modalities, rather than fusing them. Correlation analyses from statistics can be applied for this task. Standard correlation based analyses which look at correlation between individual features have a limitation that they cannot aggregate information from multiple features. A recent effort to incorporate the interactions between the genes is the work by Cooper *et al.* [7], where each patient's gene expression was represented as a mixture of clustered gene signatures derived from the data. However, this approach still considers individual image features.

Canonical correlation analysis (CCA) [8],[9], which also uses correlations, allows for a linear combination of features of different views/modalities to be correlated, rather than individual features alone. This is an important consideration in the medical setting, because it is more plausible for collections of genes to be related to collections of image features, rather than individually. CCA can be easily extended to incorporate non-linearity in the features by incorporating kernels to obtain kernel CCA [8], [10],[11], non-linearity in weights by using neural networks (Deep CCA) [12], as well as sparsity in its weights [13].

Probabilistic relations between the genes and image features, with genes as drivers for the resulting image features, have been explored by Batmanghelich *et al.* [14] as a joint modelling problem. Techniques based entirely on neural networks have also emerged recently. Multimodal Deep Boltzmann machines (MDBM) [15] models the joint probability distribution of the observed variables of both types of data with a hidden variable. By learning this joint probability distribution, MDBM can sample the hidden representation and recover missing modalities. Another network is the Multimodal Autoencoders [16] by Ngiam *et al.* which uses an autoencoder architecture for capturing the shared information.

1.3 Spatial Distribution of Cells

While advanced deep learning based techniques have been employed to accurately segment out cells, and obtain descriptors, approaches to capture the spatial distribution of cells are still not fully developed. For complex disease such as cancer, an important factor in disease diagnosis is the distribution of cells in the tissue (Figure 1.2). A scenario where the lymphocytes are well mixed with the cancerous epithelial cells (high lymphocyte infiltration) is significantly different from when the two are well-separated in space

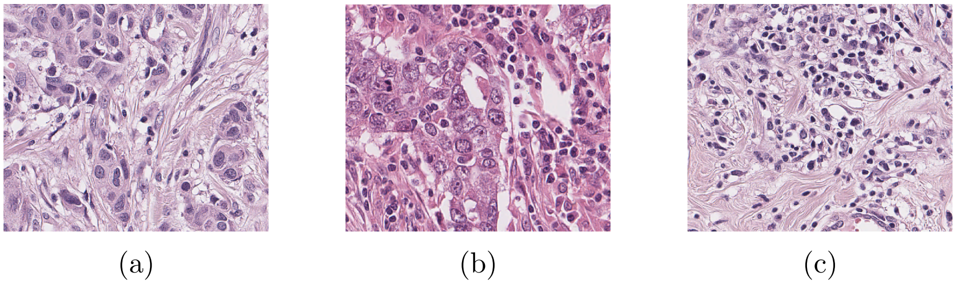


Figure 1.2: Examples of different distribution of cells from TCGA-BRCA.

(low lymphocyte infiltration) and has been shown to be linked to clinical outcome [17].

Traditional methods to capture the distribution of cells in the tissue images in cancer include plane partitioning techniques such as Delaunay triangulation, which partitions based on circumcircles; Voronoi diagrams; and other spatial tessellations [18],[19],[20]. Graph based constructions of cell graphs have also been proposed [21]. These methods, however, only look at the local neighborhood (of a few adjacent nuclei) and do not account for the overall distribution of cells at different scales, or the interactions between different types of cells.

A similar problem arises in the area of geography and ecology, where the task is to quantify the distribution of a population across a region, for example, or the distribution of trees in a forest. Classical tools to identify the level of randomness of spatial point process include the nearest-neighbor statistics, spectral analysis of point processes and location-based functions including F-function, G-function, and Ripley's K-function. These tools can be readily applied to the tissue setting to describe spatial statistics of cells, and even cells of differing types, as proposed recently by Heindl *et al.* [22] and Chang *et al.* [23].

1.4 Main Contributions

As a first step in integrating information from multiple modalities, we look to work with two different viewpoints. In particular, given two sets of features $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ of the same n patients, we would like to identify the amount and type of information shared between $\mathbf{X}$ and $\mathbf{Y}$. It is also important to know if the component of $\mathbf{X}$ and $\mathbf{Y}$ not shared captures any biologically relevant or clinically relevant meaning, or if it is uninformative noise. The goals of the information identification would be (1) to identify important shared biological variables and discover novel relations and (2) to utilize the biologically meaningful shared information in conjunction with the information unique to each modality for better prediction of biological and clinical variables.

This project aims to leverage existing techniques and develop improved methods for these tasks. Our main contribution on this front is the application of CCA to the multimodal medical setting of TCGA cancer patients to understand the underlying shared information and the potential of CCA in this task [24]. To capture information regarding the spatial distribution of cells, we employ Ripley's K-function [25],[26] to capture the second order statistics of the point sets in the context of histopathology images in contrast to first order techniques, which fail to accurately capture the spatial interactions, and higher order statistics, which are computationally expensive.

We observe an increased correlation between the spatial-augmented image feature and gene expressions, indicating a more informative image feature vector. Further, identifying the highly correlated genes and their associated pathways we find an increased association with pathways involved in spatial interactions and cell cycle. Our findings indicate the importance of encoding spatial aspects of the cell distributions in histopathology images and reveal a promising direction for future research.

Chapter 2

CORRELATION ANALYSIS

2.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) [9] is a linear method developed in 1936, by Hotelling, to identify the correlation between two sets of variables $\mathbf{X}$ and $\mathbf{Y}$ containing data about the same n samples, by linearly combining the different features of the samples to obtain meaningful hidden variables. The idea is to combine multiple features linearly to obtain a hidden feature for both $\mathbf{X}$ and $\mathbf{Y}$, such that the resulting feature behaves similarly in both the domains.

2.1.1 Formulation of Canonical Correlation Analysis

Mathematically, given $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ normalized to zero mean and unit variance with $n > \min(p, q)$, CCA looks for $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ based on the optimization problem below, where ρ is an empirical estimate of the Pearson's correlation coefficient, cov(.) is the cross covariance between $\mathbf{X}$ and $\mathbf{Y}$, and σ_X, σ_Y are the standard deviations of $\mathbf{X}$ and $\mathbf{Y}$ respectively:

$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta} \rho(\mathbf{X}\alpha, \mathbf{Y}\beta) \ \text{ such that } \ \|\mathbf{X}\alpha\| = \|\mathbf{Y}\beta\| = 1, \tag{2.1}$$

$$\rho(U, V) = \frac{\text{cov}(U, V)}{\sigma_U \sigma_V}. \tag{2.2}$$

To obtain more than one linear combination (set of weights (α^*, β^*)), the above process can be repeated, imposing orthogonality constraints. It can be shown using the Cauchy-Schwarz inequality that the correlation coefficient ρ always lies in $[-1, 1]$. The vectors α^* and β^* are referred to as the *canonical weights* and $\mathbf{X}\alpha^*$ and $\mathbf{Y}\beta^*$ are the *canonical variates*. The correlations of

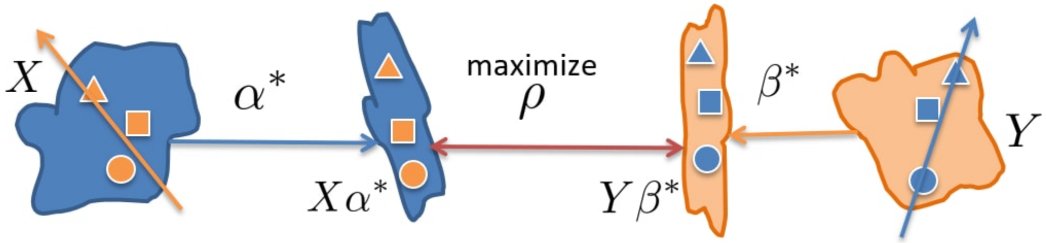


Figure 2.1: An illustration of canonical correlation analysis. CCA identifies the directions α^* and β^* such that, upon transforming ($\mathbf{X}$, $\mathbf{Y}$) to ($\mathbf{X}\alpha^*$, $\mathbf{Y}\beta^*$), the resulting correlation between them is maximized.

each variable of each domain with its corresponding canonical variate are called the *canonical loadings*. For example, for image feature f_1 and the first variate $\mathbf{Y}\beta_1$, both in $\mathbb{R}^p$, the loading $L(f_1, \mathbf{Y}\beta_1) = \rho(f_1, \mathbf{Y}\beta_1)$.

To understand the intuition behind CCA, refer to Figure 2.1. We have two sets of features/variables, $\mathbf{X}$ and $\mathbf{Y}$. Consider three samples (represented by the different shapes) for both $\mathbf{X}$ and $\mathbf{Y}$. CCA identifies the directions α^* and β^* for each of the two sets of variables respectively. These directions are chosen such that, upon transforming ($\mathbf{X}$, $\mathbf{Y}$) to ($\mathbf{X}\alpha^*$, $\mathbf{Y}\beta^*$), the resulting correlation between them is maximized. That is, we look for directions in each of the spaces, such that the features are co-variant, or vary similarly, along these directions.

It should be noted that CCA is different from obtaining the first dimension of principal component analysis (PCA) for both the sets of variables and correlating them. This difference is due to the fact that the CCA directions need not be equal to the direction of maximum variance, and the PCA directions need not be those of maximum correlation.

By identifying the directions α^*, β^*, we can firstly identify the amount of information shared between the two sets of variables. Further, we have a way to identify which features of $\mathbf{X}$ are well-related with which features of $\mathbf{Y}$, if the correlation is high. In the context of the medical setting in cancer, where we look at gene expressions and image features, we can potentially identify which genes ($\mathbf{X}$ = gene expressions) which are well-related with particular image features ($\mathbf{Y}$ = image features). The hope is that we find biologically meaningful linear combinations, which are informative of clinical variables, such as subtype.

2.1.2 Solving CCA

Given the data matrices $\mathbf{X}$ and $\mathbf{Y}$, an empirical estimate of the covariance matrices between variables in $\mathbf{X}$ and $\mathbf{Y}$ can be obtained:

$$\mathbf{C}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} \mathbf{C_{xx}} & \mathbf{C_{xy}} \\ \mathbf{C_{yx}} & \mathbf{C_{yy}} \end{bmatrix}. \tag{2.3}$$

It can be shown that in the original objective, the choice of scaling of α and β does not affect the solution. Thus, we can enforce unit norm constraints as $\alpha^T \mathbf{C_{xx}} \alpha = 1$ and $\beta^T \mathbf{C_{yy}} \beta = 1$.

Using the method of Lagrange multipliers to enforce the constraints and enforcing stationarity of the Lagrangian $L(\lambda, \alpha, \beta)$ as in [11], we get

$$L(\lambda, \alpha, \beta) = \alpha^T \mathbf{C_{xy}} \beta - \frac{\lambda_x}{2}(\alpha^T \mathbf{C_{xx}} \alpha - 1) - \frac{\lambda_y}{2}(\beta^T \mathbf{C_{yy}} \beta - 1) \tag{2.4}$$

$$\mathbf{C_{xy}} \beta = \lambda_x \mathbf{C_{xx}} \alpha \implies \alpha^T \mathbf{C_{xy}} \beta = \lambda_x \tag{2.5}$$

$$\mathbf{C_{yx}} \alpha = \lambda_y \mathbf{C_{yy}} \beta \implies \beta^T \mathbf{C_{yx}} \alpha = \lambda_y, \tag{2.6}$$

from which we get $\lambda_x = \lambda_y = \rho$, and the generalized eigenvalue problem

$$\begin{bmatrix} 0 & \mathbf{C_{xy}} \\ \mathbf{C_{yx}} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho \begin{bmatrix} \mathbf{C_{xx}} & 0 \\ 0 & \mathbf{C_{yy}} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{2.7}$$

There are multiple ways to solve this eigenvalue problem by simplifying it to a standard eigenproblem. Using Cholesky decomposition, the problem can be simplified as in [8] into a symmetric standard eigenproblem

$$\mathbf{C_{xy}} \mathbf{C_{yy}^{-1}} \mathbf{C_{yx}} \alpha = \lambda^2 \mathbf{C_{xx}} \alpha, \text{ with } \beta = \frac{\mathbf{C_{yy}^{-1}} \mathbf{C_{yx}}}{\lambda} \alpha. \tag{2.8}$$

2.2 Extensions to CCA

The linear CCA is elegant and easy to understand. However, its linearity restricts the extraction of useful descriptors of the data. To overcome this, non-linearities can be introduced in the features through kernels, as in Kernel CCA [11], or through the use of non-linearities in weights through the use of deep neural networks [12].

2.2.1 Kernel CCA

Kernel CCA makes use of a high-dimensional mapping of features through the use of kernels. In reproducing kernel Hilbert spaces, kernels can be used to evaluate the inner product in the feature space, without actually projecting the data into the higher dimensional space. This is commonly referred to as the ‘kernel trick’.

Kernel CCA can be thought of as the linear CCA with the covariance matrix as

$$\begin{bmatrix} K_x^2 & K_xK_y \\ K_yK_x & K_y^2 \end{bmatrix},$$

which can be solved as the generalized eigenvalue problem below, or reduced further to a standard eigenvalue problem.

$$\begin{bmatrix} 0 & K_xK_y \\ K_yK_x & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho \begin{bmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{2.9}$$

As described in [11], the kernel CCA does not provide good estimates of the canonical correlations in general. A regularization is used to overcome the problem, penalizing the norms of f_1 and f_2 to modify the problem to finding α and β through the generalized eigenvalue problem

$$\begin{bmatrix} 0 & K_xK_y \\ K_yK_x & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \rho \begin{bmatrix} (K_x + \kappa I)^2 & 0 \\ 0 & (K_y + \kappa I)^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{2.10}$$

2.2.2 Deep CCA

While kernels enable non-linearities in the features, the weights are still linear. With the advent of neural networks in the past decade, non-linearities can be introduced even on the weights, as shown in Figure 2.2. In this method, weights for non-linearities are learned for both the views, so as to maximize the resulting correlation after transformation.

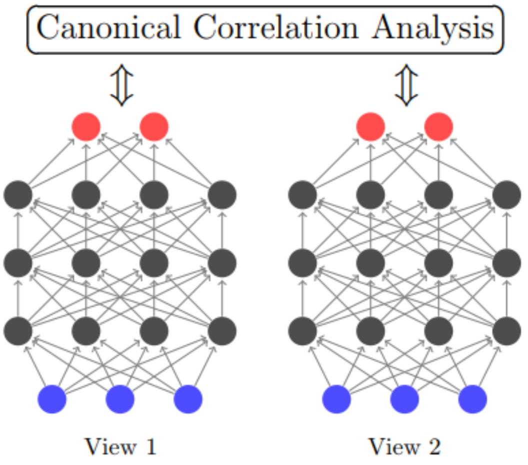


Figure 2.2: Deep canonical correlation analysis: a neural-network based extension to incorporate non-linearity in weights for CCA.

2.2.3 Enforcing Sparsity

A limiting property of CCA is that it is suitable only when $n \geq \max(p, q)$, while most genomic data used today has $n \ll \max(p, q)$. An extension of CCA to enable applying CCA to high-dimensional, low-sample data is through sparsity in the weight α, β.

Many versions of *penalized CCA* have been proposed, which can work for high-dimensional data, while preserving interpretability [13],[27],[28]. One formulation, as proposed by Witten *et al.* [13], optimizes the same objective function subject to the penalty constraints

$$\|\alpha\|^2 \leq 1, \|\beta\|^2 \leq 1, P_x(\alpha) \leq c_x, P_y(\beta) \leq c_y, \tag{2.11}$$

where P_x and P_y are convex penalty functions, often chosen to impose sparsity. For our applications, we will work with the L_1 penalty function. For multiple variates, the algorithm is iterated. An important point to be noted is that, unlike CCA, explicit orthogonality between successive variates is not enforced. This restricts the utility of the method. However, sparse CCA enables us to work with the entire set of features, with the possibility of revealing new connections.

Chapter 3

SPATIAL DISTRIBUTION

3.1 Ripley's K-function

To capture the spatial distribution information of individual cells, and interaction between the two types of cells, Ripley's K-function [25],[26] could prove insightful. This conventional tool in geographic information systems was recently proposed to be applied in the medical domain [23],[29].

3.1.1 Definition of K-function

We present the relevant definitions next.

Spatial Point Processes A spatial point process is a random pattern of points in d-dimensional space, with $d \geq 2$.

Poisson Point Process A Poisson point process (PPP), also known as complete spatial randomness (CSR), is a point process with conditions:

- For any area A and the number of dots within it denoted by $N(A)$, $N(A) \sim \text{Poisson}(\lambda|A|)$.
- For any disjoint areas A, B, $N(A)$ and $N(B)$ are independent.

Then, given condition $N(A) = n$ in an area $A \in S$, the dots within are independently and uniformly distributed in A. This is an important method to generate a PPP model in a rectangle window.

Spatial Functions Let Φ denote the point process and $d(p, S)$ denote the smallest distance between dot p and the dots in set or process S.

For a PPP with intensity λ, the probability that no point lies in a circle with radius r and center u is $\exp(-\lambda\pi r^2)$. The nearest neighbor distance distribution is defined as

$$G(r) = \mathbb{P}\{d(u, \phi\{u\}) \leq r, u \in \Phi\}. \tag{3.1}$$

For a PPP with intensity λ, we can derive that $G_p(r) = 1 - \exp(-\lambda\pi r^2)$ Due to the exponential calculation, it is hard to analyze this function because the errors of estimations are too large. The K function raised by Ripley [25],[26] is more useful.

Ripley's K-function

$$K(r) = \frac{1}{\lambda}\mathbb{E}\{|\phi \cap b(u,r) - \{u\}|, u \in \phi\}. \tag{3.2}$$

K function is the expectation of the number of dots in a ball $b(u,r)$ with center u randomly picked in ϕ and radius r, where u is excluded. We can get

$$K_p(r) = \pi r^2. \tag{3.3}$$

In simpler terms, for a spatial point process $\mathcal{A}$ with point density λ_1, the K-function is defined as

$$K_{\mathcal{A}}(r) = \frac{1}{\lambda_1}\mathbb{E}\{f_{\mathcal{A}}(\mathcal{A}, r)\}. \tag{3.4}$$

For an additional process $\mathcal{B}$ with λ_2, the cross K-function is defined as

$$K_{\mathcal{A},\mathcal{B}}(r) = \frac{1}{2}\Big(\frac{1}{\lambda_1}\mathbb{E}\{f_{\mathcal{A}}(\mathcal{B}, r)\} + \frac{1}{\lambda_2}\mathbb{E}\{f_{\mathcal{B}}(\mathcal{A}, r)\}\Big), \tag{3.5}$$

where $f_{\mathcal{P}_1}(\mathcal{P}_2, r)$ is the number of events for process $\mathcal{P}_2$ within a distance r of a randomly chosen event from $\mathcal{P}_1$, and $\mathbb{E}$ denotes expected value.

The self-K-function yields information regarding the density distribution of a given point process. A larger value of K for a smaller radius implies the presence of more dense regions. The cross-K-function provides information on how far two processes are from each other. If they are far away, K will be 0 for small radii. If the two processes are well mixed with each other, K will be a significant nonzero value. By evaluating the closeness between the

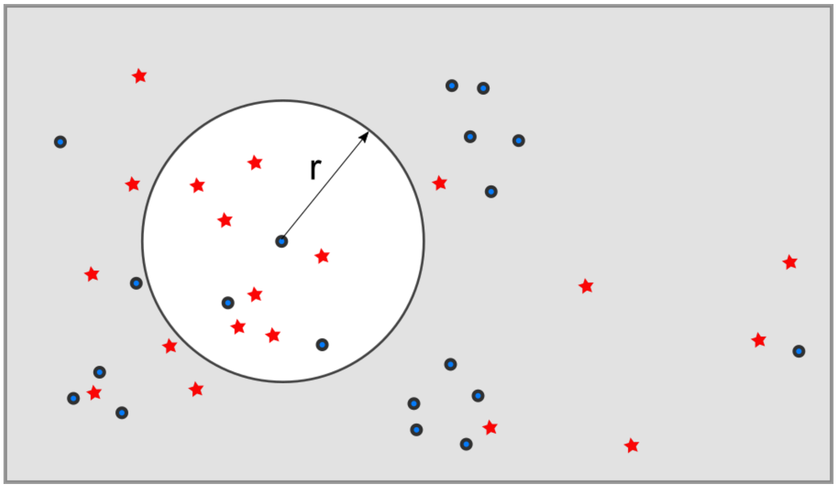


Figure 3.1: Pictorial representation of the K-function evaluated at radius r for the blue process. The circle will be placed at all blue points. For self K-function, count other blue points; for cross K-function, count red stars.

different point sets at different scales, the K-function can capture statistics about the spatial distribution of points at several scales. Therefore, the scale of evaluation of the K-function becomes an important factor.

Note that by way of definition, K-function is an increasing curve with respect to radius r, and the maximum meaningful radius is $\frac{1}{2}\min(h, w)$ where h and w are the height and width respectively of the section of interest. A pictorial representation of the evaluation is shown in Figure 3.1.

In practice, edge effects come into play, with points on the edges not having certain segments of the radius r circle inside the region of interest, and an unbiased estimator of the K-function is used. To estimate the K-function, the average value is computed, in place of the expectation. The resulting spatial feature is then combined with the previously obtained feature to obtain the overall image feature for each patient.

3.2 Examples on Simulated Data

To understand the variation in K-functions, we first present self- and cross-K-functions for four simulated point sets in Figures 3.2 and 3.3. These point sets are generated using CSR in specific intervals. The point sets 1 and 2 are not very well mixed with each other, while point sets 3 and 4 are more uniformly mixed.

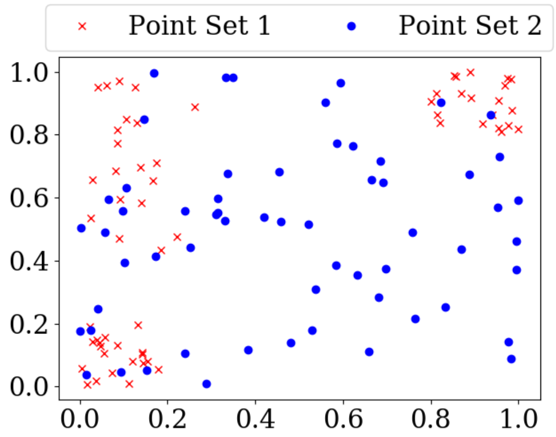


Figure 3.2: Point configuration 1: Point sets 1 and 2 are not well mixed.

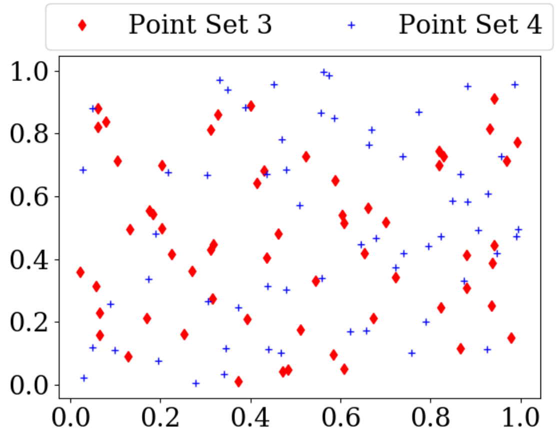


Figure 3.3: Point configuration 2: Point sets 3 and 4 are well mixed.

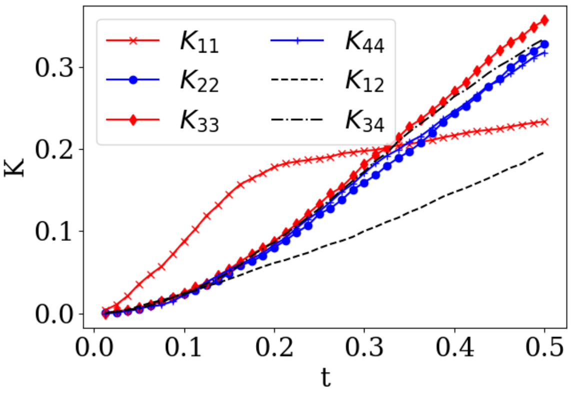


Figure 3.4: K-functions of the two point patterns: their self-K-functions $(K_{ij}, i = j)$ and cross-K-functions $(K_{ij}, i \neq j)$ plotted versus radius (t).

The functions $K_{11}, K_{22}, K_{33}, K_{44}$ represent self-K-functions for the corresponding point processes. It can be seen from Figure 3.4 that due to the clustered nature of point set 1, the resulting self-K-function is significantly different from those for the other point sets: K_{11} is high for small radii, and does not increase as rapidly after the radius $t = 0.25$.

The cross-K-functions K_{12}, K_{34} for the two configurations have large differences for large radii ($t > 0.1$). In particular, K_{12} is well below K_{34}. Looking back at the point sets it can be reasoned that the uniform mixing of the two point sets 3 and 4 result in a higher value of the cross-K-function, in contrast to the point sets 1 and 2, which results in the differing plots of the cross-K-functions.

The differences in the K-functions for the two simulated configurations highlight the ability of the K-function in capturing the spatial statistics of different point configurations. In the cancer setting, the point configurations would represent differing interactions between white blood cells and cancer cells. By accurately capturing the spatial properties of the interactions, it will be possible to better represent the nuclei and cell-based image features.

Chapter 4

EXPERIMENTS AND RESULTS

In order to demonstrate the utility of CCA and Ripley's K-function on real data, we apply these techniques on real data. We first employ CCA on the regular image features, through the workflow shown in Figure 4.1. Next, to understand the effect of incorporating spatial features, we follow the workflow in Figure 4.2 and determine the improvements in the correlations and variates learned. Part of this work comes from our recent paper [24].

4.1 Data

We work with 615 breast invasive carcinoma (BRCA) patients from TCGA, for whom the whole slide images (WSI), gene expressions, and clinical information are all available. For the gene features, we use gene expressions retrieved from TCGA using cBioPortal.

BRCAs are tumors that start in the epithelial cells that line organs and tissues throughout the body. Therefore, it is important to correctly segment out the epithelial nuclei present in the WSIs. Following the work by Chidester *et al.* [30], we segment the WSI using a convolutional neural network and obtain features describing the area, shape and texture of the nuclei using the computational tool CellProfiler [31]. Since WSIs can be up to 35000×35000 pixels

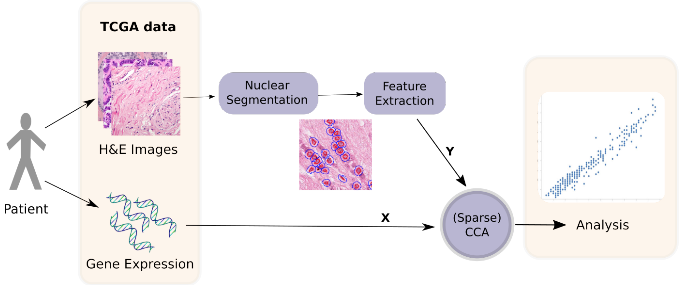


Figure 4.1: Imaging-genomics workflow with nuclei descriptors.

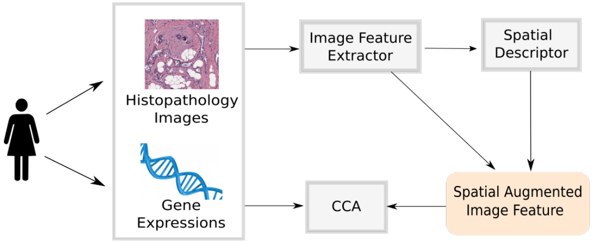


Figure 4.2: Imaging-genomics workflow with nuclei and spatial descriptors.

in dimension, in order to reduce the computational burden of image analysis and to avoid contamination in the analysis by normal cells near the tumor, we manually selected up to 15 representative patches of 1000×1000 pixels from each WSI in the tumor region for segmentation and feature extraction. The extracted nuclei descriptors form the image features.

One of the limitations of the above network is that it positively tags both lymphocytes and epithelial nuclei. However, we turn this limitation to our advantage: to distinguish the lymphocytes from the epithelial cells, a simple thresholding based technique was developed. This yields nuclei of two different types: epithelial - potentially cancerous in nature, and lymphocytes - white blood cells indicating immune activity.

4.2 CCA with Nuclei Descriptors

4.2.1 Features Used

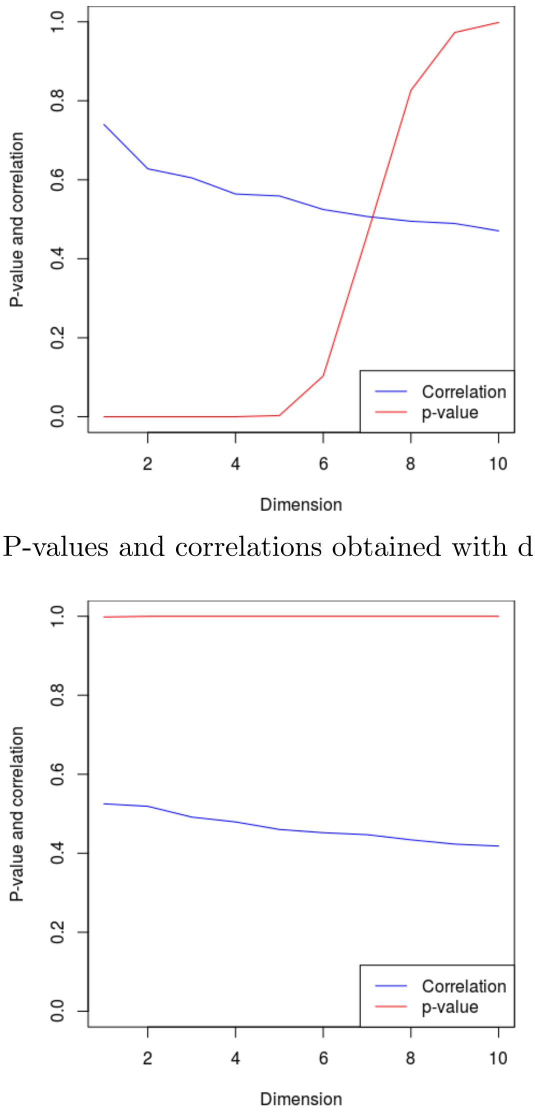
In the setting where we use only nuclei descriptors, we have $\mathbf{X}$ as gene expression data, and $\mathbf{Y}$ as the set of nuclei-based image features. In order to employ CCA, we need to select a subset of the 3000 features available for both images and genes so that $n > \min(p, q)$. For the image features, we used the mean and standard deviation of the shape, texture and color features, which resulted in 84 image features per patient. As a meaningful subset of genes to analyze, we chose the PAM50 set of 50 genes, which has been shown to be discriminative of the general grouping of patients into molecular subtypes [32]. With that, we have $p = 50$, $q = 84$ and $n = 615$.

4.2.2 Results using CCA

Using CCA on these restricted sets of variables, we found four canonical variates of statistical significance (p-value less than 0.05, computed using Wilk's lambda statistic) with strong correlation ($\{0.76, 0.64, 0.61, 0.59\}$ respectively), as shown in Figure 4.3(a). Beyond the first four variates, the significance of the correlation quickly dropped. However, the high values of the correlations inform us of the underlying overlap of information between the two modalities. In order to contrast the results with noise, we simulate random noise with the same dimensions as the data, and apply CCA. The resulting correlations and p-values associated are shown in Figure 4.3(b). It can be seen clearly that while random noise yields high correlations, the resulting p-value is high, and therefore the resulting weights identified are not significant. This highlights the importance of analyzing the correlations obtained using CCA along with the associated p-values.

The standard approach to interpreting the canonical variates is to look at the sign and magnitude of the weights. While the canonical weights define the directions which are meaningful for correlating the two datasets, there could be a potential information sharing between the different features of $\mathbf{X}$ and $\mathbf{Y}$. Therefore, interpreting these can be challenging, e.g., a feature could have a low weight either because it is irrelevant to the covariate, or because it has been shadowed out of the relationship because of a high degree of collinearity with a collection of other features. Therefore, the *canonical loadings* are preferred because they provide more interpretable values. A variable that is highly correlated with a canonical variate is well explained by that canonical variate. Thus, to interpret the learned canonical variates, we obtained the canonical loadings of each image feature and gene with each variate (Figure 4.4).

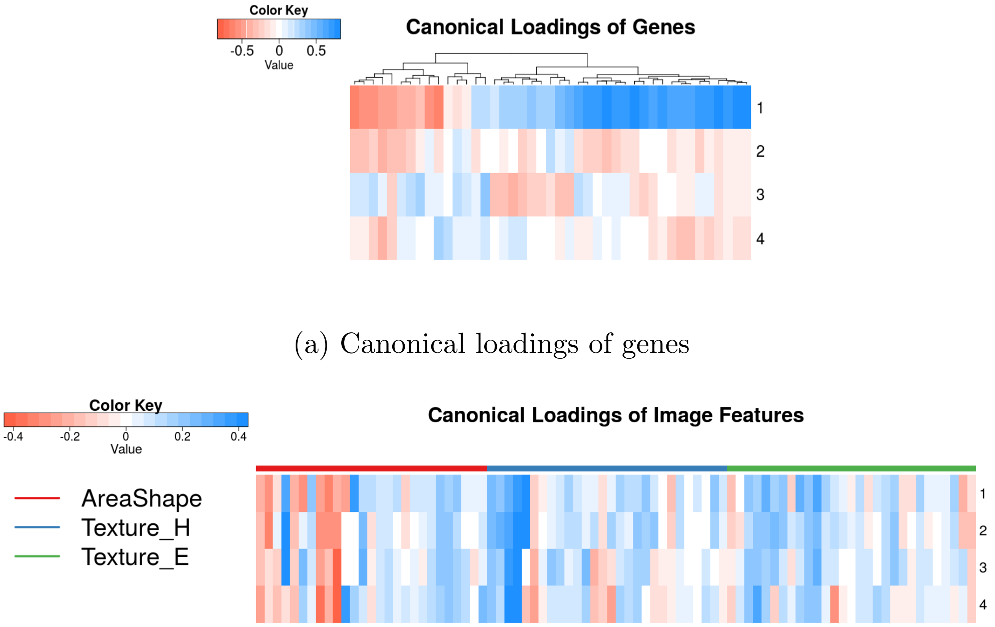
We observe that the first canonical variate is highly correlated with many PAM50 genes, with correlations as high as 0.8, which implies that this variate is highly representative of PAM50 expression. The loadings of the image features in Figure 4.4(b) are grouped by category of the feature represented. The loadings reveal the strongest correlation for most variates is with several texture features of the hematoxylin stain, area, and shape. The first variate shows a strong positive correlation particularly with texture features describing the entropy and variance of the hematoxylin stain within the nucleus and



(a) P-values and correlations obtained with data

(b) P-values and correlations obtained with random noise

Figure 4.3: Canonical correlation analysis: P-values and correlations obtained with for the real data and simulated random noise.



(a) Canonical loadings of genes

(b) Canonical loadings of image features

Figure 4.4: Canonical loadings of genes and image features for CCA with red: negative correlation, blue: positive correlation, and intensity of color representing the value.

shape features describing the nucleus. Shape features, in particular the form factor of the nucleus, also showed strong positive correlation. Subsequent variates showed much lower loadings, so while still significantly correlated within their imaging counterpart, the interpretation is not as clear.

To further understand the first variate, the 615 patients are mapped into the corresponding variate space. The scatter plot of the mappings ($\mathbf{X}\alpha$, $\mathbf{Y}\beta$ on x and y axes, respectively) is shown in Figure 4.5, with the color representing the true subtype of the cancer patient. Luminal A patients are clustered towards the left, and Basal patients to the right, while HER2 and Luminal B patients are spread out in between. This spread of the subtypes is, interestingly, in accordance with the expected prognosis of the patients. It is also noted that the range of values in the image variate is considerably smaller than those of the genes, suggesting that we should consider a more diverse set of image features. Another possibility is that since the subtypes of cancer are based inherently on the gene and gene expressions of the patients, it is not expected that image features will be able to capture this overlay completely. Therefore, while the overlay plot provides as an interesting way

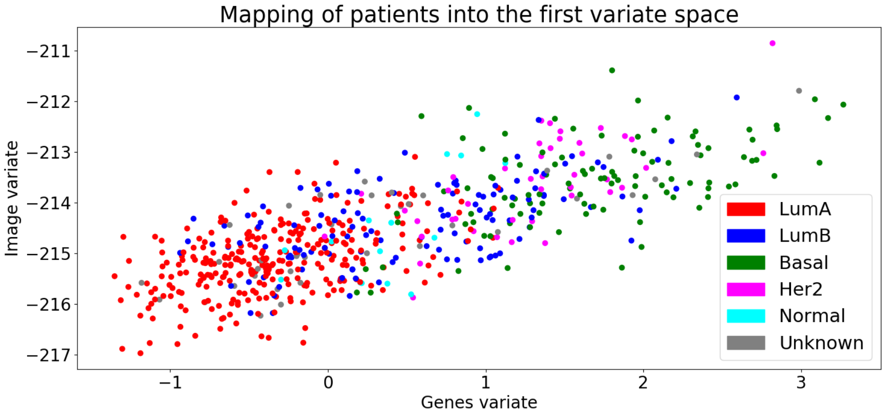


Figure 4.5: Subtype overlaid on the first canonical variate, gene variate along x-axis, image variate along y-axis.

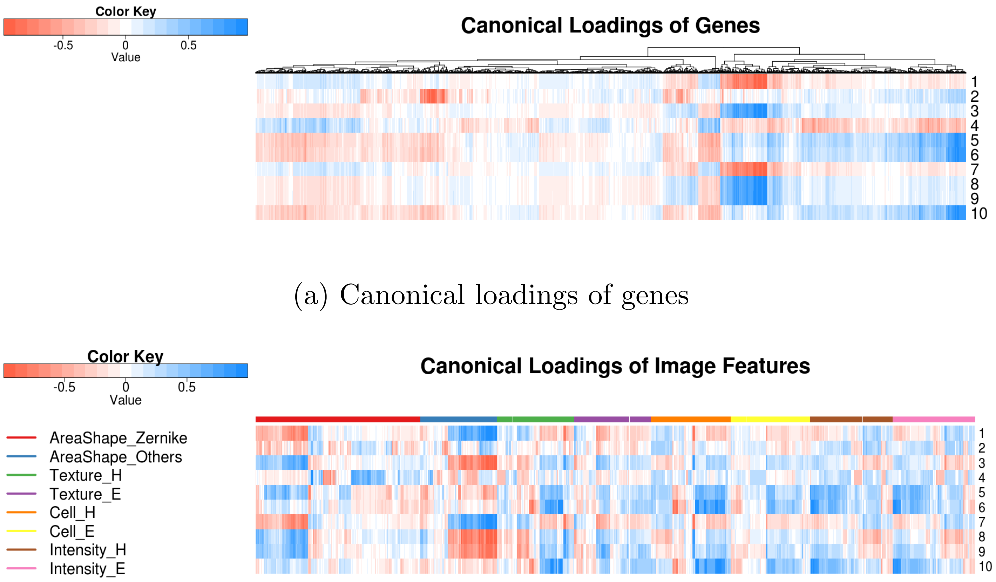
to visualize the separation of the patients, subtypes might not be the best clinical variable to overlay.

4.2.3 Results using Sparse CCA

In contrast to CCA, we were able to analyze all image features and genes using sparse CCA, allowing the algorithm to discover which subset of each is most correlated. We worked with all the 2400 available image features, and the 3000 most variant gene expressions. Thus, for this setting, we have $n = 615, p = 2400$ and $q = 3000$. Using an L1 penalty factor of 0.1 for both image and genomic variables, we obtained sets of 45-60 genes and 30-45 image features with non-zero weights for each of the ten canonical variates, respectively, with correlations in the range 0.35-0.47, with an overall p-value of 0.001. It is noted that the range of correlations obtained for sparse CCA is much lower than those obtained with the regular CCA, possibly due to the algorithm for sparse CCA.

To interpret the learned canonical variates of sparse CCA, we make use of the loadings as before, as shown in Figure 4.6. The category of ‘cell’ indicates that the feature is of the cytoplasmic region surrounding the nucleus, which mostly describes area and shape. All other features are extracted from the nucleus only.

The correlation plot for gene expressions and the canonical variates reveals a grouping of about 500 genes which have a strong correlation with at least



(a) Canonical loadings of genes

(b) Canonical loadings of image features

Figure 4.6: Canonical loadings of genes and image features for sparse CCA with red: negative correlation, blue: positive correlation, and intensity of color representing the value.

one of the variates. We also note the highest weighted genes with the first and second variates have values of correlation of ≈ 0.3, ignoring signs. This could imply that the variate is capturing an aggregate of various genes, rather than individual ones.

Since sparse CCA can consider all genes and image features, it can reveal novel, unbiased phenotype-genotype associations. We selected genes whose expression levels were highly correlated with the canonical variates discovered by sparse CCA and investigated their collective function using the online functional annotation tool DAVID [29], which can test for association of gene sets with KEGG pathways. The KEGG pathways significantly associated are shown in Fig 4.7. The associated pathways for the 1st variate are also presented in Table 4.1.

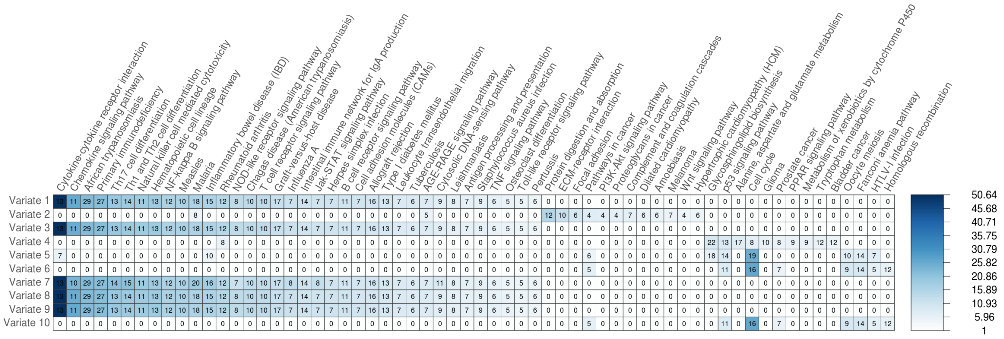


Figure 4.7: Plot of variates vs pathways defined by genes with correlation > 0.35 based on sparse CCA. (Intensity of color represents the -log(p-value), the number is the percentage of pathway genes overlapping with 0 meaning not computed)

Table 4.1: Pathways represented by the top 50 correlated genes based on canonical loadings for the 1st variate, percentage of pathways genes overlapping, and the associated p-value.

Pathway Name	%	p-value
T cell receptor signaling pathway	7.7%	1.81e-09
Th1 and Th2 cell differentiation	7.6%	2.22e-08
Th17 cell differentiation	6.5%	6.36e-08
Primary immunodeficiency	13.9%	1.24e-07
Cytokine-cytokine receptor interaction	3.4%	2.05e-07
Inflammatory bowel disease (IBD)	7.7%	2.53e-06
Chemokine signaling pathway	3.7%	2.83e-06
Natural killer cell mediated cytotoxicity	4.5%	5.57e-06
Measles	4.4%	5.81e-06
Cell adhesion molecules (CAMs)	4.2%	8.10e-06
Hematopoietic cell lineage	4.2%	3.06e-04
Chagas disease (American trypanosomiasis)	3.8%	4.15e-04
HTLV-I infection	1.9%	1.73e-03
Rheumatoid arthritis	3.4%	3.39e-03
Leukocyte transendothelial migration	2.6%	7.09e-03

The first variate and others showed a similar correlation pattern with both image features and gene expressions, which is likely a result of the lack of enforcement of orthogonality by sparse CCA. DAVID revealed that, for the first variate, the highly correlated genes were strongly associated with pathways related to immune response, including primary immunodeficiency, natural killer cell mediated cytotoxicity, and to lymphocytes, including Th1 and Th2 cell differentiation, T-cell and B-cell receptor signaling, and NF-kappa B signaling. Figure 4.6 shows that the expression of these genes has a strong correlation with area and shape features through the latent canonical variates. Given that lymphocytes are easily distinguished by their small size and circular shape, we could hypothesize that these canonical variates are capturing image and genomic descriptions of the presence of lymphocytes within the tumor, which is indeed a biologically relevant association for cancer.

Variates five, six, and ten are not indicative of area or shape, but instead capture texture and cell hematoxylin features, which are indicative of DNA content, and intensity features of both stains. These variates were found to be correlated with gene sets associated with the cell cycle and p53 signaling pathways (related to DNA damage repair and apoptosis), as well as the cell cycle, all of which have important implications for tumor development. The second variate too could have implications for cancer, as it was associated with pathways involved in cell processes such as cell maintenance (ECM-receptor interaction), adhesion (focal adhesion), and proliferation (Wnt signaling and proteoglycans in cancer), and the cycle (PI3K-Akt signaling), though the lack of strong correlation with particular image features necessitates further investigation for a clear interpretation.

4.3 CCA with Nuclei Descriptors and Spatial Distribution Features

To incorporate the spatial features, we use Ripley's K-function presented earlier.

4.3.1 Ripley's K-function on Real Data

To understand the variation in K-functions for real data, we first present self- and cross-K-functions for a couple of point sets (Figure 4.8) obtained after processing the TCGA-BRCA histopathology images.

Figure 4.8c shows different K-functions, with dashed lines for configuration 1, and solid lines for configuration 2. Firstly, all the identified nuclei obtained from the segmentation algorithm are utilized together to obtain the self-K-functions $K_{\text{all},1}$ and $K_{\text{all},2}$ shown in black. Note that the function for configuration 1 lies slightly above that of configuration 2, though the distinction is not prominent.

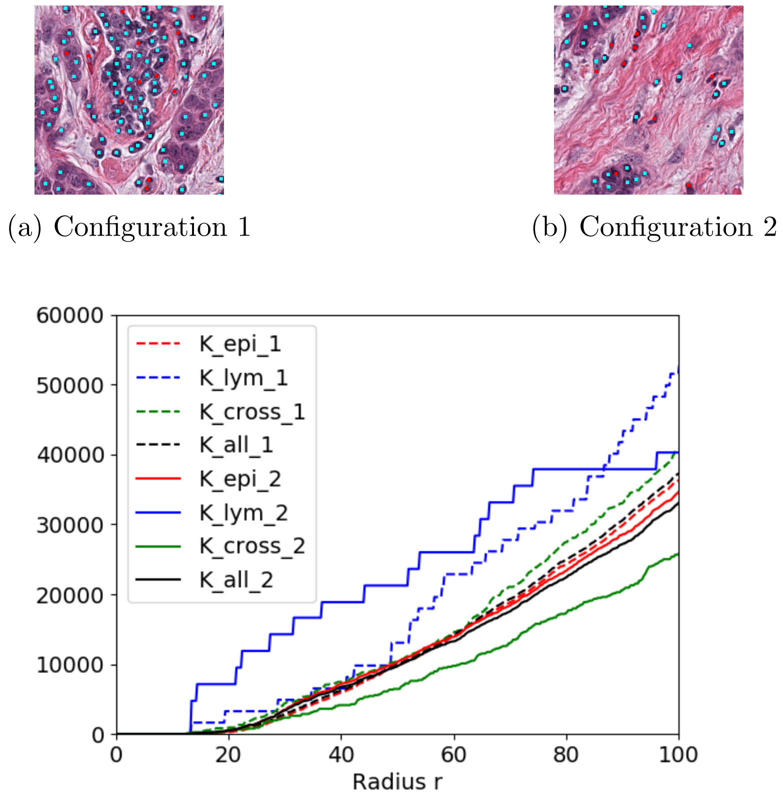
Next, the identified nuclei were classified using a simple thresholding on the area, texture and shape, to obtain two different types of cells: epithelial (in cyan), and lymphocytes (in red). We note that this naive thresholding will provide a noisy classification of the cells. The self K-functions computed for the resulting epithelial cells ($K_{\text{epi},1}$, $K_{\text{epi},2}$) are not very different, while those of the lymphocytes ($K_{\text{lym},1}$, $K_{\text{lym},2}$) show considerable difference, with that of configuration 1 lying below that of configuration 2 for smaller values of radius r, capturing the clustered nature of lymphocytes in configuration 2.

Also, plotting the resulting cross-K-functions ($K_{\text{cross},1}$, $K_{\text{cross},2}$), it is seen that the cross-K-function for configuration 2 lies well below that of configuration 1, indicating the absence of considerable interaction between the two point sets in configuration 2, which can be justified by looking at the definition.

4.3.2 CCA with Spatial-Augmented Features

Having demonstrated the variation in Ripley's K-function with respect to different configurations, we now show the improvement in the information captured by image features upon the inclusion of spatial features by performing correlation analyses using CCA and sparse CCA.

Effect of Spatial Features on CCA correlations To apply CCA to the given setting, subsets of both features need to be chosen. For the nuclei-based image features, those corresponding to the mean and standard deviation of fundamental properties such as the color, texture and shape features are chosen, which yields 84 features as earlier. For the spatial features, the K-function is evaluated at 100 points with different maximum radii (in $\{100, 200, 300, 400, 500\}$). The two are combined to yield an overall 184 element features vector per patient. For the genes, the PAM50 subset of genes



(a) Configuration 1

(b) Configuration 2

(c) Corresponding K-functions

Figure 4.8: The variation in self-K-function and cross-K-function (c) for two configurations (a) and (b) where epithelial cells are shown in cyan, and lymphocytes in red.

Table 4.2: Correlation (ρ) and p-value of spatial features with PAM50 genes using CCA.

Radii	1st variate		2nd variate		3rd variate	
	ρ	P-value	ρ	P-value	ρ	P-value
None	0.740	<1e-15	0.628	4.6e-14	0.605	7.7e-09
$r \leq 100$	0.792	1.9e-15	0.739	1.3e-08	0.712	1.1e-03
$r \leq 200$	0.790	1.9e-11	0.738	4.0e-05	**0.728**	2.2e-02
$r \leq 300$	**0.793**	7.6e-09	**0.748**	1.7e-03	0.714	1.9e-03
$r \leq 400$	0.792	8.8e-08	0.737	6.2e-03	0.710	2.8e-01
$r \leq 500$	0.787	9.4e-09	0.746	1.2e-03	**0.727**	1.5e-03

Table 4.3: Correlation of spatial features with gene expression using sparse CCA for different variates (1st to 5th).

Radii	L1-penalty	1st	2nd	3rd	4th	5th
None	0.05	0.490	0.404	0.321	0.424	0.399
$r \leq 100$	0.05	0.489	0.403	0.466	0.424	0.382
$r \leq 300$	0.52	0.470	0.345	0.457	0.478	**0.460**
$r \leq 500$	0.05	0.489	0.403	**0.466**	**0.535**	0.424

which have been shown to be informative in breast cancer subtyping is chosen.

The results are presented in Table 4.2, which records the correlations and p-values (computed using Wilk's lambda statistic). It can be observed that the augmentation of spatial features significantly improves the correlation for the first 3 variates. The correlation achieved by the first variate increases by a factor of 5%, while both second and third variates show an improvement in correlation by a factor of 10%. Beyond the 3 variates, the combined spatial image features did not yield statistically significant results.

4.3.3 Sparse CCA with Spatial-Augmented Features

In order to get a deeper understanding of the spatial features, we run Sparse CCA to obtain 5 variates on the set of 3400 most variant genes, and 3400 image features comprising the 2400 dimensional nuclei features augmented with the K-function evaluated at 1000 different radii, with different maximum radii (in $\{100, 300, 500\}$). The penalty factor to be chosen was determined by the algorithm to obtain the result with the highest statistical significance. The results of the correlations obtained are shown in Table 4.3. We observe that inclusion of the spatial features increases the correlation obtained for variates 3 through 5, while having little effect on the first two variates.

Further, for the scenario leading to highest correlations (max radius = 500), we identify the image features and genes which are highly correlated with the variates. For the image features, we observe that the 4th variate is dominated by spatial features, while being uncorrelated with the nuclei-based features (Figure 4.9). The presence of such a variate highlights the importance of spatial features in correlations with genes, implying that the features capture important properties of gene expression variation. Upon

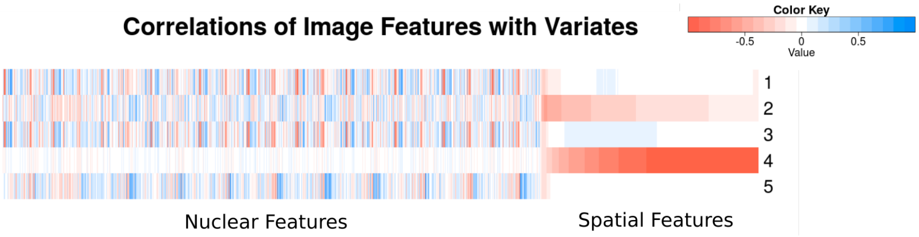


Figure 4.9: Canonical Loadings: Correlation of spatial-augmented image features with variates (K-function's maximum radius $r = 500$).

investigating the KEGG pathways of the correspondingly correlated genes, we obtain the primary pathways (1) cell cycle, (2) oocyte meiosis, (3) p53 signaling pathway, (4) cytokine-cytokine receptor interaction and (5) cell adhesion molecules. Of these, (1)-(3) belong to the category 'cell growth and death', while (4),(5) fall under 'signaling molecules and interaction', all of which is important information carried by the spatial distribution of cells in the histopathology images.

Chapter 5

CONCLUSIONS

We presented the application of a simple tool, canonical correlation analysis, to multimodal data analysis on the TCGA breast cancer data set. We observed a considerable amount of shared information between the image features and gene expressions. We were also able to identify a subset of image features closely related to the 50 chosen genes. While the resulting linear combinations contained information relevant to subtyping in cancer, we saw that we cannot predict subtype correctly as yet. However, these experiments took us one step closer to answering questions about the shared information between image features and genes.

We demonstrated the utility of CCA and sparse CCA in discovering connections between cellular features and gene expressions for breast cancer. The learned canonical variates represent latent spaces that link the two modalities and provide insight into their joint variation. Their biological relevance was shown through their association with diverse pathways with implications for cancer, and could benefit from a more diverse range of image features. For sparse CCA, imposing orthogonality in the variates and understanding the sensitivity of the penalty factor for sparse CCA would be important for use in a clinical setting. We envision that such a correlation analysis could be a preliminary step in studies of phenotype and genomic traits, with follow-up affirmation by biologists, toward new insights into genetic diseases.

We also demonstrated the use of Ripley's K-function in the histopathology setting to encode spatial information between epithelial cells and lymphocytes. We showed that incorporating spatial features increases correlation with PAM50 gene expression by up to 10%. Further, employing sparse CCA, we confirm that the spatial feature is able to capture important aspects of the spatial interaction between the cells of different types. We thus highlighted the importance of spatial information and revealed a promising direction for future research.

BIBLIOGRAPHY

[1] Cancer Genome Atlas Network *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, p. 61, 2012.

[2] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle et al., “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository,” *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[3] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward et al., “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods,” *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[4] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury et al., “The Parkinson Progression Marker Initiative (PPMI),” *Progress in Neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[5] D. Lahat, T. Adali, and C. Jutten, “Multimodal data fusion: an overview of methods, challenges, and prospects,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[6] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, “Multimodal data fusion using source separation: Two effective models based on ICA and IVA and their properties,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1478–1493, 2015.

[7] L. A. Cooper, J. Kong, D. A. Gutman, W. D. Dunn, M. Nalisnik, and D. J. Brat, “Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images,” *Laboratory Investigation*, vol. 95, no. 4, p. 366, 2015.

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[9] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[10] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[11] F. R. Bach and M. I. Jordan, “Kernel independent component analysis,” *Journal of Machine Learning Research*, vol. 3, no. Jul, pp. 1–48, 2002.

[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[13] D. M. Witten, R. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[14] N. K. Batmanghelich, A. V. Dalca, M. R. Sabuncu, and P. Golland, “Joint modeling of imaging and genetics,” in *International Conference on Information Processing in Medical Imaging.* Springer, 2013, pp. 766–777.

[15] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2222–2230.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.

[17] S. E. Stanton and M. L. Disis, “Clinical significance of tumor-infiltrating lymphocytes in breast cancer,” *Journal for Immunotherapy of Cancer*, vol. 4, no. 1, p. 59, 2016.

[18] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, “Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2010.

[19] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, “Histopathological image analysis: A review,” *IEEE Reviews in Biomedical engineering*, vol. 2, pp. 147–171, 2009.

[20] C. Demir and B. Yener, “Automated cancer diagnosis based on histopathological images: a systematic survey,” *Rensselaer Polytechnic Institute, Technical Report*, 2005.

[21] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell graphs of cancer," *Bioinformatics*, vol. 20, no. suppl_1, pp. i145–i151, 2004.

[22] A. Heindl, S. Nawaz, and Y. Yuan, "Mapping spatial heterogeneity in the tumor microenvironment: A new era for digital pathology," *Laboratory Investigation*, 2015.

[23] Y. H. Chang, G. Thibault, V. Azimi, B. Johnson, D. Jorgens, J. Link, A. Margolin, and J. W. Gray, "Quantitative analysis of histological tissue image based on cytological profiles and spatial statistics," in *Engineering in Medicine and Biology Society (EMBC).* IEEE, 2016, pp. 1175–1178.

[24] V. Subramanian, B. Chidester, J. Ma, and M. N. Do, "Correlating cellular features with gene expression using CCA," *ArXiv e-prints (to appear in International Symposium on Biomedical Imaging)*, Feb. 2018.

[25] B. D. Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 172–212, 1977.

[26] P. M. Dixon, "Ripley's K function," *Encyclopedia of Environmetrics*, 2002.

[27] S. Waaijenborg and A. H. Zwinderman, "Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks," *BMC Bioinformatics*, vol. 10, no. 1, p. 315, 2009.

[28] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–34, 2009.

[29] D. Huang, B. Sherman, and R. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2008.

[30] B. Chidester, M. Do, and J. Ma, "Discriminative bag-of-cells for imaging-genomics," in *Pacific Symposium on Biocomputing*, 2018.

[31] A. Carpenter, T. Jones, M. Lamprecht, C. Clarke, I. Kang, O. Friman, D. Guertin, J. Chang, R. Lindquist, J. Moffat, P. Golland, and D. Sabatini, "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." *Genome Biology*, vol. 7, no. 10, p. R100, 2006.

[32] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, p. 1160, 2009.