Integrating Multi-omics with Sparse Canonical Correlation Analysis

Michael Komodromos

michael.komodromos19@ic.ac.uk

Abstract

We integrate radiomic and transcriptomic data from patients with ovarian cancer using sparse canonical correlation analysis (sCCA). We demonstrate integration yields prognostic models with greater predictive accuracy in comparison to using radiomics features alone. However, integration does not provide greater predictive accuracy than transcriptomic data alone. Further, we examine network structures providing plausible relational pathways between genes and radiomic features.

Code: our study can be reproduced by running the code available at `https://github.com/mkomod/ovc`. Additionally, we provide an `R` package to perform sCCA found at `https://github.com/mkomod/rcca`.

1 Introduction

Multi-omics is a form of multi-view data, where several distinct feature sets are measured for the same samples; providing complementary information when characterising a biological object (Li et al., 2018). There are several types of "omics" data including: genomics, transcriptomics, protiomics, metabolomics etc.. A recent development in multi-omics is radiomics, which provide a set of quantitative features extracted from an image. Radiomics features include various first and higher order statistics, fractal and shape features characterising disease features not appreciated by the naked eye (Lu et al., 2019). Several studies have explored the integration

of multi-omics data (Pittman et al., 2004; Hasin et al., 2017; Chaddad et al., 2019) and have demonstrated the increased predictive capacity of integration.

Canonical Correlation Analysis (CCA) is a method used to study relations between sets of covariates (Hotelling, 1936). CCA relies on the correlation structure between two groups of data to find a linear combinations of covariates that maximise the correlation. CCA and it's extensions have been successfully used in several disciplines, including: economics, neuroscience and computational biology, to name a few (Zhuang et al., 2020; Yamanishi et al., 2003). Furthermore, there have been several studies using canonical correlation based methods to integrate multi-omics (Witten and Tibshirani, 2009; Hong et al., 2013; Shi et al., 2019; Rodosthenous et al., 2020). For instance, Shi et al. demonstrated that CCA based methods can be used to identify relevant miRNA-mRNA pathways.

In recent years there have been several extensions of CCA (Uurtio et al., 2017; Li et al., 2018). For instance, kernel CCA and deep CCA allow for non-linear relationships between groups of covariates to be discovered through the use (or construction) of a feature map (Andrew et al., 2013). Further, regularised forms of CCA known as sparse CCA (sCCA) have been proposed to tackle high-dimensional datasets, where there are often more covariates than observations (Witten and Tibshirani, 2009; Suo, 2018; Rodosthenous et al., 2020). Finally, CCA has also been framed in a probabilistic perspective allowing for Bayesian interpretations and correspondingly Bayesian equivalents to the mentioned extensions (Bach and Jordan, 2005).

Within our study we consider the integration of radiomic and transcriptomic data through sparse CCA; examining both predictive and exploratory aspects. Regarding the predictive performance we compare our results to current prognostic models as well as models constructed using principal components. Further, we explore the shared structure between these datasets using network based methods.

2 Methods

2.1 Principal Component Analysis

Prior to introducing CCA, we detail principal component analysis (PCA). Principal components (PCs) are linear combinations of random variables with maximal

variance, where linear combinations are found subject to some constraints. Let $w \in \mathbb{R}^p$ and $\boldsymbol{X} \in \mathbb{R}^p$ denote a random vector with covariance matrix Σ, then the *principal directions* are given by solving

$$w^* = \underset{w \in \mathbb{R}^p}{\operatorname{argmax}} \ \operatorname{var}\left(w^\top \boldsymbol{X}\right) \qquad \text{subject to } \|w\|_2 = 1 \tag{1}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. We solve Equation (1) by constructing the Lagrangian,

$$L(\lambda, w) = w^\top \Sigma w - \lambda(\|w\|_2^2 - 1).$$

Taking the derivative of $L(\lambda, w)$ with respect to w and setting to 0 gives

$$\Sigma w - \lambda w = 0 \tag{2}$$

the solution of which is given by the eigenvectors and eigenvalues obtained by the singular value decomposition of Σ,

$$\Sigma = UDV^\top \tag{3}$$

where $U = (u_i)_{i=1}^p$, $V = (v_i)_i^p$ and $\mathrm{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$. Correspondingly, the ith principal direction, $w_i^* = u_i$ and the variance of the ith PC, $w_i^{*\top} \boldsymbol{X}$, is given by λ_i. Letting $W = (w_i^*)_{i=1}^p$, we have, $\operatorname{cov}\left(W^\top \boldsymbol{X}, W^\top \boldsymbol{X}\right) = D$ and $\sum_{i=1}^p \operatorname{var}(\boldsymbol{X}_i) = \operatorname{tr}(D)$ where $\operatorname{tr}(\cdot)$ gives the trace of a matrix.

2.2 Canonical Correlation Analysis (CCA)

CCA is the natural extension of PCA, where CCA is concerned with finding pairs of linear projections of two views that are maximally correlated. Let $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ be random vectors with covariances $(\Sigma_{11}, \Sigma_{22})$ and cross-covariance $\Sigma_{12} = \Sigma_{21}^\top$. Consider a pair of linear projections $(w_1^\top \boldsymbol{X}_1, w_2^\top \boldsymbol{X}_2)$, we are interested in finding the pair of vectors (w_1^*, w_2^*) that maximise the correlation between the linear projections $w_1^\top \boldsymbol{X}_1, w_2^\top \boldsymbol{X}_2$. Formally,

$$(w_1^*, w_2^*) = \underset{w_1, w_2}{\operatorname{argmax}} \operatorname{corr}\left(w_1^\top \boldsymbol{X}_1, w_2^\top \boldsymbol{X}_2\right)$$

which is equivalent to solving

$$\text{maximise } \ w_1^\top \Sigma_{12} w_2, \qquad \text{subject to } \ w_1^\top \Sigma_{11} w_1 = w_2^\top \Sigma_{22} w_2 = 1. \tag{4}$$

Notably, we can find up to $\min(p_1, p_2)$ pairs of vectors denoted as $(w_1^{(i)}, w_2^{(i)})$ for $i = 1, \ldots, \min(p_1, p_2)$. Let $A_1 \in \mathbb{R}^{p_1 \times m}$ and $A_2 \in \mathbb{R}^{p_2 \times m}$ where the columns of A_1 and A_2 correspond to the first m vectors $w_1^{(i)}$ and $w_2^{(i)}$ for $m \leq \min(p_1, p_2)$. Finding the first m vector pairs $(w_1^{(i)}, w_2^{(i)})$ corresponds to solving

$$\underset{A_1, A_2}{\text{maximise}} \ \operatorname{tr}\left(A_1^\top \Sigma_{12} A_2\right) \qquad \text{subject to } \ A_1^\top \Sigma_{11} A_1 = A_2^\top \Sigma_{22} A_2 = I. \tag{5}$$

where $\operatorname{tr}(\cdot)$ is the trace. Let $T \coloneqq \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$, if we suppose Σ_{11} and Σ_{22} are non-singular, then the singular value decomposition of T can be written in the form

$$T = (u_1, \ldots, u_m) D (v_1, \ldots, v_k)^\top \tag{6}$$

where $u_i \in \mathbb{R}^{p_1}$, $v_i \in \mathbb{R}^{p_2}$ and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, then the ith *canonical correlation vectors* are given by

$$w_1^{(i)} = \Sigma_{11}^{-1/2} u_i \tag{7}$$

$$w_2^{(i)} = \Sigma_{22}^{-1/2} v_i, \tag{8}$$

and *canonical correlation coeefficients* $\rho_i = \lambda_i$ for $i = 1, \ldots, m$ (Mardia et al., 1979).

2.3 Sparse Canonical Correlation Analysis (sCCA)

Traditional CCA breaks down in a high dimensional setting. For example, consider (centered) observations $Y \in R^{n \times p}$ where $p > n$, the covariance matrix is given by $\Sigma = Y^\top Y \in \mathbb{R}^{p \times p}$. We notice, $\operatorname{rank}(\Sigma) \leq n - 1 < p$, hence Σ is singular and we are unable to evaluate T. sCCA adapts the objective function presented in Equation (4) so we can perform correlation analysis in a high dimensional setting.

First, we present a generalised framework for regularised CCA, and then present methods by (Witten and Tibshirani, 2009) and (Suo, 2018). Contrary to our description of CCA, we consider matrices of observations, denoted as X. Let $X_1 \in \mathbb{R}^{n \times p_1}$ and $X_2 \in \mathbb{R}^{n \times p_2}$ be matrices of observations where $p \gg n$. Let $w_1 \in \mathbb{R}^{p_1}$ and $w_2 \in \mathbb{R}^{p_2}$ and let $r_1 : w_1 \mapsto \mathbb{R}$ and $r_2 : w_2 \mapsto \mathbb{R}$. Then, the generalised form for the regularised CCA problem is given by

$$\underset{w_1, w_2}{\text{maximise}} \ \operatorname{corr}(X_1 w_1, X_2 w_2) - r_1(w_1) - r_2(w_2). \tag{9}$$

An efficient algorithm to find sparse canonical vectors was proposed by Witten and Tibshirani (2009), under which they set r_1 and r_2 to be the ℓ_1-norm and solve

$$\begin{aligned} &\underset{w_1,w_2}{\text{maximise}} \ \operatorname{corr}(X_1 w_1, X_2 w_2) - \lambda_1 \|w_1\|_1 - \lambda_2 \|w_2\|_1 \\ &\text{subject to } \|w_1\|_2 \leq 1, \|w_2\|_2 \leq 1, \end{aligned} \tag{10}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, $\|\cdot\|_2$ denotes the ℓ_2-norm and $\lambda_1, \lambda_2 \in \mathbb{R}_+$ are regularisation hyperparameters. Under the above formulation the canonical vectors can be inaccurate and non-sparse (Suo, 2018; Rodosthenous et al., 2020); hence Suo relax the constraints in Equation (10) to be

$$\begin{aligned} &\underset{w_1,w_2}{\text{maximise}} \ \operatorname{corr}(X_1 w_1, X_2 w_2) - \lambda_1 \|w_1\|_1 - \lambda_2 \|w_2\|_1. \\ &\text{subject to } \ \operatorname{var}(X_1 w_1) \leq 1, \operatorname{var}(X_2 w_2) \leq 1. \end{aligned} \tag{11}$$

The resulting problem is biconvex, in other words, fixing w_1 makes the problem convex with respect to w_2 and fixing w_2 makes the problem convex with respect to w_1. A description of the algorithm used to solve Equation (11) is presented in (Suo, 2018).

2.3.1 Multiple canonical vector pairs

As with conventional CCA we are able to find multiple pairs of canonical vectors. However, we do so in an iterative manner, by first solving Equation (11) and then solving again where instead of X_1 and X_2 we use

$$\bar{X}_1 = \begin{pmatrix} X_1 \\ W_1^\top X_1^\top X_1 \\ W_2^\top X_2^\top X_1 \end{pmatrix}, \qquad \bar{X}_2 = \begin{pmatrix} X_2 \\ W_2^\top X_2^\top X_2 \\ W_1^\top X_1^\top X_2 \end{pmatrix} \tag{12}$$

where $W_k = \left(w_k^{(i)}\right)_{i=1}^{r-1}$ for $k = 1, 2$ are matrices with the $(r-1)$ previous canonical vectors (Rodosthenous et al., 2020). Solving yields the rth pair of canonical vectors, noting we are able to obtain up to $\min(p_1, p_2)$ pairs.

2.3.2 Tuning the regularisation hyperparameters

There are a number of methods for tuning the regularisation hyperparameters. As our dataset is small we opt to use a permutation based tuning method proposed

by (Witten and Tibshirani, 2009). Witten and Tibshirani's method is outlined in Algorithm 1. Noting, we have adapted the method to optimise Suo's method for sCCA.

Algorithm 1 Permutation validation for (λ_1, λ_2).

for $(\lambda_1, \lambda_2)_j \in \Lambda$ **do**
  Solve Equation (11) for (w_1^*, w_2^*) for X_1 and X_2, taking $(\lambda_1, \lambda_2) = (\lambda_1, \lambda_2)_j$.
  Compute $d_j = \text{corr}(X_1 w_1^*, X_2 w_2^*)$.
  for $i \in \{1, \ldots, B\}$ **do**
    Permute the rows of X_1; constructing X_1^i.
    Solve Equation (11) for (w_1^i, w_2^i) for X_1^i and X_2, taking $(\lambda_1, \lambda_2) = (\lambda_1, \lambda_2)_j$.
    Compute $d_j^i = \text{corr}\left(X_1^i w_1^i, X_2 w_2^i\right)$,
  end for
  Compute $p_j = \frac{1}{B} \sum_{i=1}^{B} 1\left(d_j^i \geq d_j\right)$.
end for
return $(\lambda_1, \lambda_2)_j$ that gives the smallest p_j.

2.4 Semi-supervised sparse CCA

Semi-supervised sparse CCA (SS-CCA) is a method used to select features of X_1 and X_2 related to a response $Y \in \mathbb{R}^n$. The objective function of SS-CCA is given as

$$\begin{aligned} &\underset{w_1, w_2}{\text{maximise}} \ \text{corr}\left(X_1 w_1, X_2 w_2\right) - \lambda_1 \|w_1\|_1 - \lambda_2 \|w_2\|_1 \qquad (13) \\ &\text{subject to } \ \text{var}\left(X_1 w_1\right) \leq 1, \ \text{var}(X_2 w_2) \leq 1, \\ &\qquad w_{1i} = 0 \ \forall i \in Q_1, \ w_{2j} = 0 \ \forall j \in Q_2. \end{aligned}$$

where Q_1 and Q_2 are the set of indices of features of X_1 and X_2 that are least correlated with the response Y (Witten and Tibshirani, 2009). Generalising the criteria for Q_1 and Q_2, we have

$$Q_1 = \{i \ : \ f(X_{1i}, Y) \leq \gamma_1, \ i = 1, \ldots, p_1\}, \qquad (14)$$

$$Q_2 = \{j \ : \ f(X_{2j}, Y) \leq \gamma_2, \ j = 1, \ldots, p_2\} \qquad (15)$$

where γ_1 and γ_2 are thresholds controlling the elements in Q_1 and Q_2, and f is a function to evaluate the relatedness of a feature of X to Y. For example, when considering overall survival, f can return the Cox's statistic. Notably, γ_1 and γ_2 are hyperparameters that need to be tuned.

2.5 Proportional Hazards Model

The proportional hazards model (PHM) is used to model time to failure events. Let T denote a random variable representing time to failure, with density $f(t)$ and survivor function $S(t) = \mathbb{P}(T > t)$. The hazard rate, the rate of failure at time t, is given by

$$\lambda(t) = \frac{f(t)}{S(t)}, \qquad t > 0. \tag{16}$$

The PHM is based on the assumption that the hazard rate is a product of a baseline hazard rate, $\lambda_0(t)$, and a positive functional term $\psi(x;\beta)$, formally,

$$\lambda(t, x) = \lambda_0(t)\psi(x;\beta), \tag{17}$$

for covariates $x \in \mathbb{R}^p$ and unknown parameters $\beta \in \mathbb{R}^p$. A number of forms have been proposed for $\psi(x;\beta)$, the most common being $\psi(x;\beta) = \exp\left(\beta^\top x\right)$. We note, an implication of the form of Equation (17) is that the ratio of any two samples at time t is constant, i.e. $\lambda(t, x_1) \propto \lambda(t, x_2)$ (Cox, 1972). Parameter values β can be estimated using the partial log-likelihood, $\ell(\beta)$ (Cox, 1975).

2.5.1 Evaluating proportional hazard models

The most widely used statistic to evaluate the performance of PHMs is concordance. Concordance is measure of the discriminatory power of a model. Let t_i and $\widehat{t_i}$ be observed and predicted failure times for observation i, then the concordance,

$$c = \mathbb{P}\left(\widehat{t_i} > \widehat{t_j} | t_i > t_j\right), \tag{18}$$

for a pair of observations i, j. A popular estimator for the concordance is Harrell's c-index, given as

$$\widehat{c} = \frac{\sum\sum_{i<j} \mathbb{I}\left(t_i < t_j\right) \mathbb{I}(\widehat{\beta}^\top x_i > \widehat{\beta}^\top x_j)\delta_i + \mathbb{I}\left(t_j < t_i\right) \mathbb{I}(\widehat{\beta}^\top x_j > \widehat{\beta}^\top x_i)\delta_j}{\sum\limits_{i<j}\sum \mathbb{I}\left(t_i < t_j\right)\delta_i + \mathbb{I}\left(t_j < t_i\right)\delta_j}, \tag{19}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\delta_i = 1$ if the observation is uncensored and 0 otherwise. Notably, the c-index is sensitive to the degree of censoring and tends to be higher when the degree of censoring is higher. As a result, an alternate estimator proposed by Gonen and Heller is used for comparison, given as

$$\widehat{k} = \frac{1}{n(n-1)} \sum \sum_{i<j} \left(\frac{\mathbb{I}\left(\widehat{\beta}^\top x_{ji} < 0\right)}{1 + \exp\left(\widehat{\beta}^\top x_{ji}\right)} + \frac{\mathbb{I}\left(\widehat{\beta}^\top x_{ij} < 0\right)}{1 + \exp\left(\widehat{\beta}^\top x_{ij}\right)} \right), \tag{20}$$

where x_{ij} represents pairwise difference $x_i - x_j$. As $\widehat{k}$ is determined through the partial likelihood estimate, $\widehat{\beta}$, and the effect of censoring on the bias of $\widehat{\beta}$ is negligible; we obtain a robust estimator for the concordance probability (Gonen and Heller, 2005). Notably, we are able to obtain an estimate for the standard error of $\widehat{k}$ via a smooth approximation for $\widehat{k}$, details are presented in (Gonen and Heller, 2005).

2.6 Networks

Networks can be used to understand the relationships within groups of covariates. To construct networks we use a method proposed by Shi et al. referred to as SmCCNet. SmCCNet uses a similarity matrix constructed by robust canonical vector pairs, details are outlined in Algorithm 2. We note, unlike the original algorithm we use Suo's method to performs sCCA. Following the construction of S, hierarchical clustering is performed and an appropriate level is set to cut the dendogram. Cliques containing features from both X_1 and X_2 are returned and used to construct networks. Full details are presented in (Shi et al., 2019).

Algorithm 2 SmCCNet similarity matrix.

1: **for** $i \in \{1, \ldots, B\}$ **do**
2: Subsample columns of X_1 and X_2, we denote these matrices as X_1' and X_2'.
3: Compute (w_1', w_2') by solving Equation (11) for X_1' and X_2'
4: Compute $A_i = w_1' \otimes w_2'$.
5: **end for**
6: Compute A the element-wise average over the entires in $A_1, \ldots, A_B$
7: Normalise A by $\max(A)$.
8: **return** $S = 1 - A$

3 Results

Data for this study consists of clinical, radiomic and genomic data available from the cancer genome atlas (TCGA). Radiomics data is comprised of features relating to: size, intensity, texture and wavelet decompositions of contrast enhanced CT scans (Lu et al., 2019). The radiomics data for this study were produced by Lu et al., the remaining datasets are summarised in Table 1. Notably, both the radiomics and mRNA datasets have more features than samples.

Data type	Platform	Samples	Features
Clinical		630	19
Radiomics	TexLab 2.0	71	658
mRNA expression	Affymetrix U133	593	12,043

Table 1: Summary of data used.

We merged the datasets to construct our multi-view dataset, upon merging there were $n = 68$ samples, 50 of which had right censored overall survival times (days). Further, prior to our analysis we pre-selected covariates from the Radiomic and mRNA expression datasets by fitting univariate proportional hazards models and selecting covariates with p-values less than $\alpha = 0.05$. Variable selection left $p_1 = 524$ features in our mRNA expression dataset and $p_2 = 32$ features in our radiomics dataset. Notably, applying sCCA to our filtered datasets is equivalent to conducting semi-supervised sparse CCA.

3.1 Principal component analysis

We computed the principal components for the radiomics and mRNA dataset. Figure 1 presents the first two principal components of the mRNA and radiomics datasets. The proportion of variance explained by the leading principal components of our datasets is presented in Table 2. Notably 99.8% of variance is explained by the 20th PC for the radiomics dataset and 99.7% of variance is explained by the 68th principal component for the mRNA dataset. Further, from Table 2 we notice, 75.7% of variance is explained by the first three PCs for the radiomics dataset, whereas

29.2% of variance is explained by the first three PCs for the mRNA dataset.

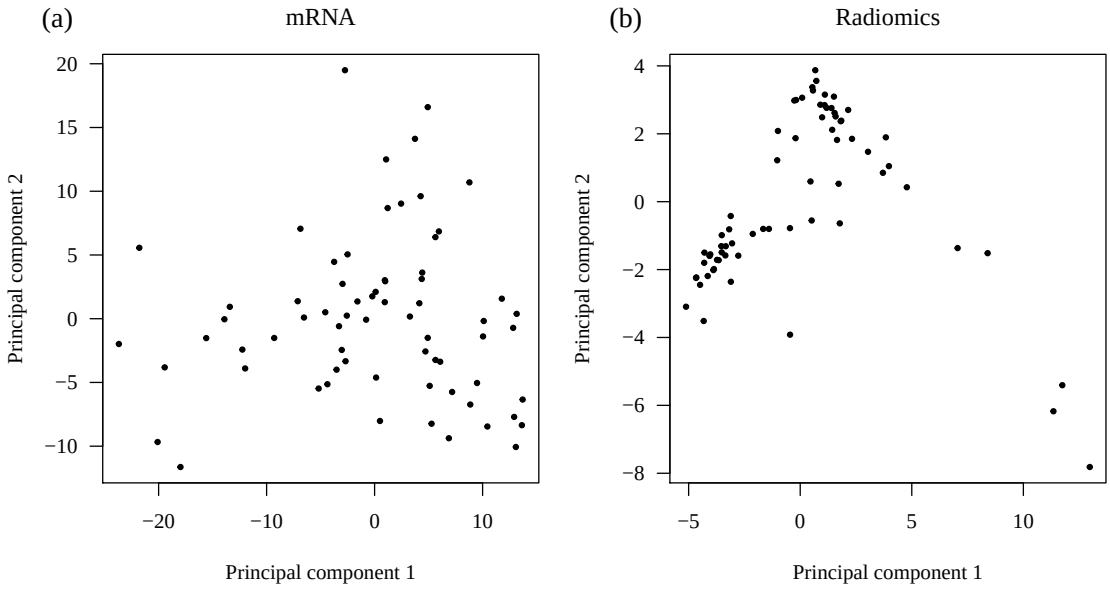

Figure 1: (a) First two principal components of the mRNA expression data. (b) First two principal components of the radiomics data.

	Principal Component							
	1	2	3	4	5	6	7	8
Radiomics	0.488	0.690	0.757	0.797	0.835	0.870	0.897	0.921
mRNA	0.161	0.240	0.292	0.331	0.367	0.401	0.432	0.458

Table 2: Cumulative proportion of variance explained by the ith principal component for the radiomics and mRNA expression datasets.

3.2 Sparse canonical correlation analysis

We applied Suo's method for sparse CCA to our Radiomics and mRNA datasets, centering and standardising beforehand. We tuned the regularisation hyperparameters, λ_1 and λ_2, in Equation (13) using permutation validation (Algorithm 1) taking $\Lambda = \{0.4, 0.5, \cdots, 2\} \times \{0.1, 0.2, \cdots, 3\}$ and $B = 1000$. The loss landscape is presented in Figure 2, notably a minimum p-value of 0.018 occurs at $(1.1, 1.6)$.

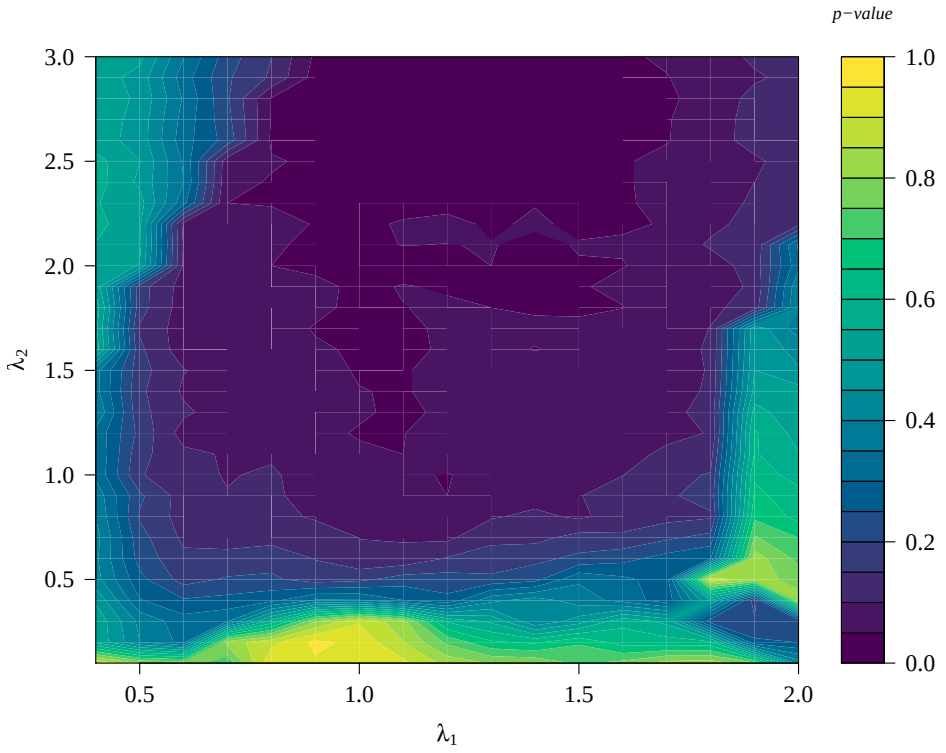

Figure 2: Loss landscape for a grid of values for λ_1 and λ_2. We note, as $\lambda_1 \to 0$ and $\lambda_2 \to 0$ we impose less penalisation on w_1 and w_2, leading to non-sparse results. As such, we avoid choosing regularisation parameters close to the origin.

Taking $\lambda_1 = 1.1$ and $\lambda_2 = 1.6$, we obtain the first three canonical vector pairs by solving Equation (13). We present the canonical vector pairs in Figure 3. Notably, the vector pairs for the mRNA expression data ($w_1^{(i)}$) are unstable, arising from the small sample size of the dataset (Suo, 2018). Further, we notice there are only two non-zero elements in the first canonical vector of the radiomics dataset ($w_2^{(1)}$), and three for the remaining radiomics canonical vectors. Where the non-zero features in the first canonical vector are `NGTDM_Coarse_LHL_25HUgl` and `FD_max_LLH_25HUgl`, and the second and third canonical vectors include `NGTDM_Coarse_LLH_25HUgl`.

We present the projections of the mRNA and radiomics data in Figure 4, noting we refer to these projections as canonical variates. Within Figure 4 we notice the radiomics canonical variate is split into groups, this can be attributed to the `FD_max_LLH_25HUgl` feature. Further, we notice the similarity between the first three canonical variates, suggesting the canonical vectors are not orthogonal, which

is also attributed to the small sample size (Rodosthenous et al., 2020).

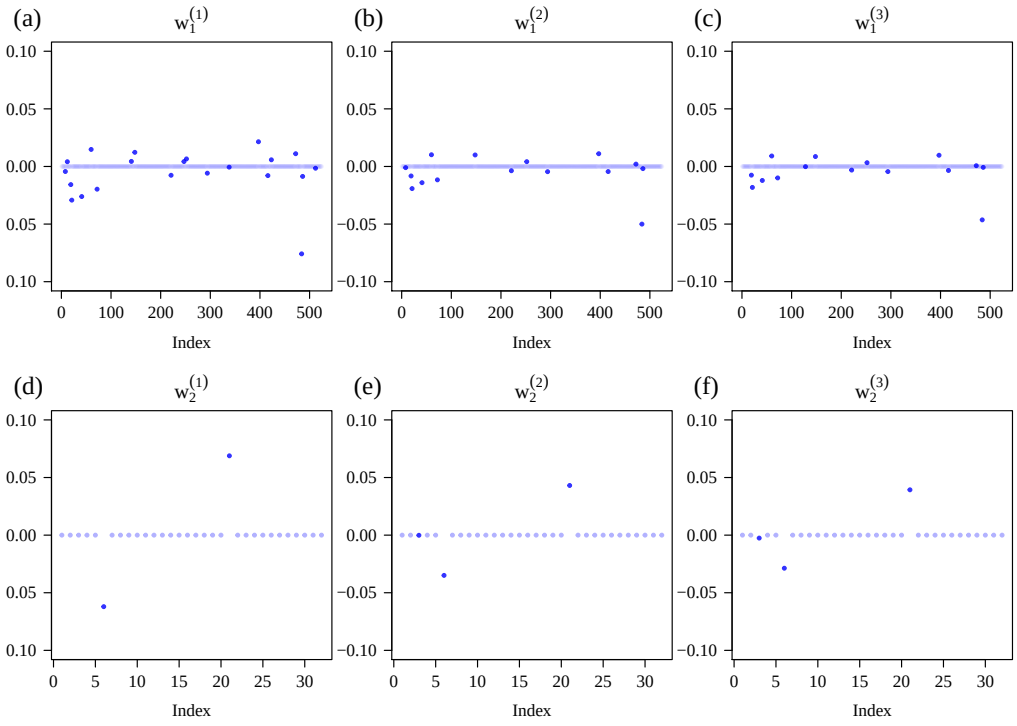

Figure 3: First three sparse canonical vectors for the mRNA expression dataset (a)-(c) and Radiomics dataset (d)-(f). For clarity elements of $w^{(i)}$ with a value of 0 have a lighter shade. We notice a combination of few covariates are needed to maximise the correlation between the mRNA dataset and Radiomics dataset.

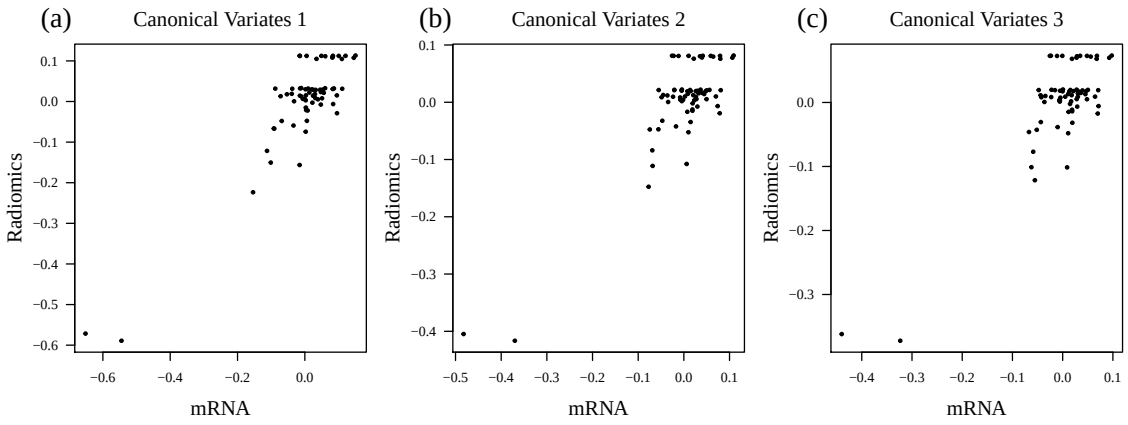

Figure 4: (a) Projections of the mRNA and radiomics data onto the: (a) first, (b) second, (c) third canonical vector pair.

3.3 Models

To evaluate the canonical variates (projections from the canonical vector pairs) we fit multivariate proportional hazards models taking overall survival as our response. We compared the models constructed using the canonical variates against models constructed using principal components of the radiomics and mRNA dataset and the radiomics prognostic vector (RPV) (Lu et al., 2019). We note, within all models we include age (dichotomised at 60 years) and stage as additional predictors.

To evaluate the performance of our models we compute Harrell's c-index, $\widehat{c}$, and Gonen and Heller's c-index, $\widehat{k}$. Noting, we compute $\widehat{k}$ to mitigate the effect of censoring when estimating the c-index. The performance of our models is presented in Table 3. Notably, the first table presents other models, including our baseline model (M0) constructed using the RPV and a model (M1) constructed using age, stage, the first mRNA and second radiomics canonical variates. The second table presents models constructed using mRNA and radiomics canonical variates, noting combinations of mRNA and radiomics vectors were not included due to collinearity between these predictors; resulting in model instability. Our final table presents models constructed using the principal components of our two datasets.

From Table 3 we notice our baseline model gave $\widehat{c} = 0.672$ (0.0978) and $\widehat{k} = 0.707$ (0.0510), which is similar to the results obtained by Lu et al., noting they obtained $\widehat{c}$ ranging from $0.658 - 0.739$ for their discovery dataset, and $\widehat{c}$ ranging from $0.549 - 0.690$ for their TCGA validation dataset. Further, M1 yielded $\widehat{c} = 0.764$ (0.072) and $\widehat{k} = 0.761$ (0.048). Models constructed using radiomics canonical variates had $\widehat{c} = 0.631$ and $\widehat{k}$ ranging from $0.709 - 0.710$. Further, all models constructed using mRNA canonical variates had $\widehat{c}$ ranging from $0.713 - 0.735$ and $\widehat{k}$ ranging from $0.715 - 716$. Finally, models constructed using radiomics PCs yielded $\widehat{c}$ ranging from $0.655 - 0.733$ and $\widehat{k}$ ranging from $0.719 - 0.737$, models constructed using mRNA PCs yielded $\widehat{c}$ ranging from $0.858 - 0.860$ and $\widehat{k}$ ranging from $0.795 - 0.812$. Models with both mRNA and radiomics PCs had $\widehat{c}$ raging from $0.855 - 0.875$ and $\widehat{k}$ ranging from $0.796 - 0.826$.

Other Models			
Label	Predictors	$\widehat{c}$	$\widehat{k}$
M0	Age, Stage, RPV	0.672	0.707
		(0.098)	(0.051)
M1	Age, Stage, RNA_1, $Radiomics_2$	0.764	0.761
		(0.072)	(0.048)

Canonical Variates						
	Radiomics		mRNA		Radiomics, mRNA	
Projections	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$
1	0.631	0.709	0.713	0.716	-	
	(0.100)	(0.055)	(0.099)	(0.050)		
2	0.631	0.710	0.733	0.715	-	
	(0.100)	(0.055)	(0.099)	(0.047)		
3	0.631	0.710	0.735	0.715	-	
	(0.100)	(0.055)	(0.099)	(0.054)		

Principal Components						
	Radiomics		mRNA		Radiomics, mRNA	
Projections	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$	$\widehat{c}$	$\widehat{k}$
1	0.655	0.719	0.860	0.795	0.855	0.796
	(0.102)	(0.049)	(0.046)	(0.039)	(0.044)	(0.039)
1, 2	0.660	0.721	0.860	0.800	0.855	0.814
	(0.102)	(0.055)	(0.046)	(0.038)	(0.050)	(0.039)
1, 2, 3	0.733	0.737	0.858	0.812	0.875	0.826
	(0.065)	(0.054)	(0.049)	(0.041)	(0.049)	(0.039)

Table 3: Model concordance (standard error). Notably, all models include Age and Stage as predictors. M0 denotes our baseline model adapted from (Lu et al., 2019). M1 is a model based on canonical variates, where RNA_1 denotes the first mRNA canonical variate and $Radiomics_2$ the second radiomics canonical variate. We note, canonical variate models (middle table) only include one canonical variate, whereas principal component based models include combinations of mRNA and radiomics PCs.

3.4 Networks

To examine relationships between radiomic and mRNA expression features we constructed networks using SmCCNet. We present our results in Figure 5 and Table 4. Notbaly, networks can be used to examine potential radiomic-gene pathways.

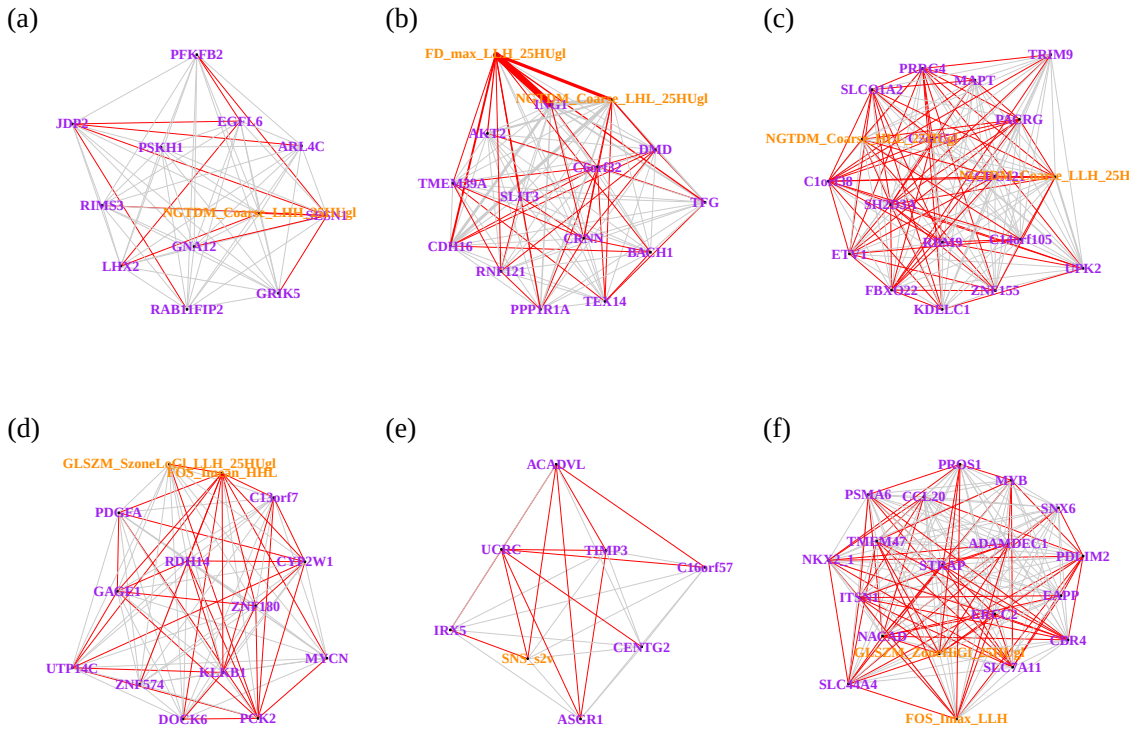

Figure 5: Networks constructed using SmCCNet. Radiomics features are coloured orange and genes are coloured purple. Connective lines indicate the correlation between the features with red lines indicating negative correlation and grey positive.

Network	Features
1	JDP2, LHX2, PFKFB2, RAB11FIP2, GRIK5, PSKH1, ARL4C, GNA12, RIMS3, SESN1, EGFL6
	NGTDM_Coarse_LHH_25HUgl
2	TFG, TEX14, CDH16, PPP1R1A, C6orf32, SLIT3, DMD, BACH1, TMEM39A, AKT2, RNF121, CRNN, ING1
	NGTDM_Coarse_LHL_25HUgl, FD_max_LLH_25HUgl
3	UPK2, PRRG4, MAPT, SH2D3A, PACRG, SLCO1A2, ZNF155, TRIM9, KDELC1, ETV1, RBM9, CCDC121, C2orf54, C1orf38, FBXO22, C14orf105
	NGTDM_Coarse_LLH_25HUgl, NGTDM_Coarse_HLL_25HUgl
4	MYCN, PDGFA, PCK2, DOCK6, UTP14C, RDH14, ZNF180, C13orf7, GAGE1, ZNF574, CYP2W1, KLKB1
	FOS_Imean_HHL, GLSZM_SzoneLoGl_LLH_25HUgl
5	C16orf57, TIMP3, IRX5, ACADVL, CENTG2, ASGR1, UCRC
	SNS_s2v
6	SLC44A4, PDLIM2, CCL20, ERCC2, NACAD, SLC7A11, NKX2_1, MYB, SNX6, STRAP, CBR4, EAPP, ITSN1, PSMA6, TMEM47, PROS1, ADAMDEC1
	GLSZM_ZoneHiGl_25HUgl, FOS_Imax_LLH
7	HLA_DOB, PDE3A, ADAM17, ORAI3, MSC, SPOCK1
	NGTDM_Coarse_LLL_25HUgl
8	RECK, SLPI
	GLRLM_SRHGLE_25HUgl
9	EEF1G
	FOS_Imedian_LHH
10	SCGB2A1
	NGTDM_Coarse_HHL_25HUgl
11	KLHL2
	NGTDM_Coarse_HLH_25HUgl, GLRLM_SRLGLE_LHH_25HUgl

Table 4: Networks discovered using SmCCNet. Each network row is split into two cells, the top cell refers to mRNA features and the bottom to radiomics features.

4 Discussion

Our results highlight the predictive and exploratory utility of sCCA, in particular we have shown the subspaces found by sCCA are as good or better than our baseline comparison. Furthermore, our method of discovering linear combinations of covariates is semi-supervised, i.e. covariates are preselected based on some criteria and canonical covariates are found without knowledge of the response. In comparison, in our baseline model the RPV is constructed through a supervised approach, where the RPV is a linear combination of features found using an ℓ_1 regularised proportional hazards model. Hence, not only does sCCA perform better than current models, subspaces are found without knowledge of the response. However, if prediction is our ultimate goal then a PHM comprised of age, stage and the first mRNA principle component yields a model with $\widehat{c} = 0.86$ (0.046) and $\widehat{k} = 0.795$ (0.039). Outperforming all canonical variate based models and our baseline model in Table 3. The difference in model performance suggests either the shared structure between the radiomics and the mRNA data is insubstantial (from a predictive perspective), or the signal to noise ratio is inhibiting the discovery of low dimensional structure. Furthermore, the lack of validation data has limited the scope of our study. Given we have not evaluated our models performance on unseen, we are unable to comment on whether our models are overfit.

Further, the interpretation of SmCCNet results requires caution. The networks formed are based on the outer product of the absolute canonical vector pairs; hence groups constructed are based on features with large absolute values. Furthermore, network pathways (connective lines) are constructed from the correlation matrix between the mRNA and radiomics data. As estimation of the correlation matrix suffers in a high-dimensional setting, erroneous conclusions may be drawn from these pathways. As a results, care is needed when interpreting the results and networks should be used as a primer to direct further research rather than provide conclusive relations between mRNA and radiomics features.

From a methodological standpoint care has been taken to limit external input. However as shown by Figure 3, the permutation validation scheme for λ_1 and λ_2

yields results where the mRNA canonical vectors are unstable. We can correct for instability by increasing the value of λ_1. Suggesting, different validation schemes may need to be explored and their output examined; a comparison of these schemes may prove to be an insightful avenue for future research. In addition, when selecting the features to include in our study we arbitrarily set a p-value threshold of 0.05, examining robust variables selection schemes may yield better results. Further, as seen Figure 4, the multiple pairs of vectors discovered using sCCA are not orthogonal; suggesting methods can be developed further to produce multiple pairs of orthogonal vectors. Furthermore, extending sCCA to support integration with a censored response, such as overall survival, may yield vector pairs that better explain the response, and consequently canonical variates that are better predictors.

5 Conclusions

We have demonstrated by integrating radiomics and mRNA expression data we can increase the predictive accuracy of prognostic models. In particular, we have shown a combination of projections obtained from semi-supervised sparse canonical correlation analysis results in a models that are better than radiomics based prognostic models. However, we have also shown that we can construct highly predictive models using mRNA based principal components, suggesting greater explanatory structure within the datasets than between. Further, we have uncovered pathways between radiomic and transcriptomics features that may prove insightful under further examination.

References

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. *30th Int. Conf. Mach. Learn. ICML 2013*, 28(3):2284–2292, 2013.

F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. *Dept. Stat. Univ. California, Berkeley, CA, Tech. Rep*, pages 1–11, 2005.

A. Chaddad, P. Daniel, S. Sabri, C. Desrosiers, and B. Abdulkarim. Integration of radiomic and multi-omic analyses predicts survival of newly diagnosed IDH1 wild-type glioblastoma. *Cancers (Basel).*, 11(8):1–16, 2019.

D. R. Cox. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B*, 34(2): 187–220, Feb 1972.

D. R. Cox. Partial Likelihood. *Biometrika*, 62(2):269–276, 1975.

M. Gonen and G. Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biol.*, 18(1):1–15, 2017.

S. Hong, X. Chen, L. Jin, and M. Xiong. Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.*, 41(8):1–15, 2013.

H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321, 1936.

Y. Li, F. X. Wu, and A. Ngom. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, 19(2):325–340, 2018.

H. Lu, M. Arshad, A. Thornton, G. Avesani, P. Cunnea, E. Curry, F. Kanavati, J. Liang, K. Nixon, S. T. Williams, M. A. Hassan, D. D. Bowtell, H. Gabra, C. Fotopoulou, A. Rockall, and E. O. Aboagye. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic-

and molecular-phenotypes of epithelial ovarian cancer. *Nat. Commun.*, 10(1): 1–11, 2019.

K. V. Mardia, J. Kent, and J. Bibby. *Multivariate analysis*. Academic Press, 1979. ISBN 9780124712522.

J. Pittman, E. Huang, H. Dressman, C. F. Horng, S. H. Cheng, M. H. Tsou, C. M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins, and M. West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci. U. S. A.*, 101(22):8431–8436, 2004.

T. Rodosthenous, V. Shahrezaei, and M. Evangelou. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics*, 36(17):4616–4625, 2020.

W. J. Shi, Y. Zhuang, P. H. Russell, B. D. Hobbs, M. M. Parker, P. J. Castaldi, P. Rudra, B. Vestal, C. P. Hersh, L. M. Saba, and K. Kechris. Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, 35(21):4336–4343, 2019.

X. Suo. *Topics In High-Dimensional Statistical Learning*. PhD thesis, Stanford, 2018.

V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu. A tutorial on canonical correlation methods. *arXiv*, 50(6), 2017.

D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, 8(1): 1–27, 2009.

Y. Yamanishi, J. P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19 Suppl 1:323–330, 2003.

X. Zhuang, Z. Yang, and D. Cordes. A technical review of canonical correlation analysis for neuroscience applications. *Hum. Brain Mapp.*, 41(13):3807–3833, 2020.