Convex Multi-View Clustering Via Robust Low Rank Approximation with Application to Multi-Omic Data

Omar Shetta, Mahesan Niranjan, Srinandan Dasmahapatra

Abstract—Recent advances in high throughput technologies have made large amounts of biomedical omics data accessible to the scientific community. Single omic data clustering has proved its impact in the biomedical and biological research fields. Multi-omic data clustering and multi-omic data integration techniques have shown improved clustering performance and biological insight. Cancer subtype clustering is an important task in the medical field to be able to identify a suitable treatment procedure and prognosis for cancer patients. State of the art multi-view clustering methods are based on non-convex objectives which only guarantee non-global solutions that are high in computational complexity. Only a few convex multi-view methods are present. However, their models do not take into account the intrinsic manifold structure of the data. In this paper, we introduce a convex graph regularized multi-view clustering method that is robust to outliers. We compare our algorithm to state of the art convex and non-convex multi-view and single view clustering methods, and show its superiority in clustering cancer subtypes on publicly available cancer genomic datasets from the TCGA repository. We also show our method's better ability to potentially discover cancer subtypes compared to other state of the art multi-view methods.

Index Terms—Multi-view clustering, outlier robustness, convex optimization, multi-omic data, cancer subtype identification

1 Introduction

Recent advances in high throughput sequencing technologies have made available large amounts of biomedical data consisting of measurements of genomic features across multiple omic scales forming multi-omic datasets when combined. Multi-omic data have been recently used to efficiently visualize and cluster cancer subtypes [4]. Clustering for biomedical data is a useful pattern discovery technique, which is the initial step taken in data exploration. Clustering is especially of great use in the emerging field of precision medicine in discovering cancer subtypes [21]. Separately clustering each omic dataset has the capability of finding patterns in the data. However, using several omics for integrative clustering on the same group of samples has the potential to expose more detailed structures that are not revealed by examining only a single omic measurement. For example, it has been shown that cancer subtypes can be better defined when integrating both DNA methylation and gene expression information [5], [37]. Cancer is a group of diseases caused by DNA alterations that change cell behaviour, which causes malignancy and uncontrolled growth. General treatments for cancer are challenging to develop due to the high genetic heterogeneity of this disease [2]. The field of cancer multi-omics aims to discover potential subtypes and their affiliated molecular biomarkers that can be used for more individualised treatment and prognosis. Cancer multi-omic datasets consist of measuring different molecular parameters, which include multi-omic data such as: RNA expression, microRNA expression, DNA methylation and protein expression, etc. While inference of cellular function or state from any one of these omics is easy to carry out and dominates much of the research reported in the literature, cellular regulation is complex and combined analysis can reveal more information. For example, genes that are transcribed (DNA to mRNA) are not always translated into protein. The mRNA is held (for example, in structures like P-bodies) and is translated only when needed. Similarly, proteins may be synthesized at different rates from the corresponding mRNA by different numbers of ribosomes binding to them. Where disruptions to such regulation is the cause of disease, analysis at only one level can be misleading.

The machine learning community has become interested in the problem of integrating information from different data types to achieve a joint clustering solution, called multi-view clustering. Multi-view clustering acts on multi-view data, where multi-omic datasets are a specific type of this general category of datasets. Multi-view clustering methods found in the literature encompass: Canonical Correlation Analysis (CCA) [6], Co-Training Expectation Maximization (co-EM) [7], multi-view normalized cut [8], co-regularized multi-view spectral clustering [9], multi-view neighbourhood preserving projections [10], CCA regularized with common source graph [31], and Multi-view Non-Negative Matrix Factorization (Multi-NMF) [11]. The CCA method of [6] and multi-view spectral clustering of [9] showed that finding a common latent representation between different views can enhance clustering performance. Moreover, Multi-NMF showed that learning a latent representation for each view, by constraining these representations to be

- *Omar Shetta is a research fellow in the Department of Electronics and Computer Science, Southampton University, Southampton*

  E-mail: os10g13@soton.ac.uk
- *Mahesan Niranjan is with the Department of of Electronics and Computer Science, Southampton University, Southampton .*

Manuscript received April 19, 2005; revised August 26, 2015.

similar to a 'consensus' representation, results in an improved clustering performance. The problem with these multi-view clustering methods is that they either work only with data that have two views [6], [31], or they optimize non-convex objective functions that can only be solved by alternating optimization methods that converge to arbitrary local minima [7], [9], [10], [11].

In contrast, convex methods are found in the literature for the single-view case, where methods for subspace learning make use of convex loss functions [12], [14], [15]. These papers exploit a convex regularizer that reduces rank in place of constraining the dimension of the latent representation with a hard lower bound. Moreover, some authors [16], [17], [18] have approached the problem of multi-view subspace learning by formulating a convex loss function that seeks to find a common latent representation which is then subsequently used for clustering. [18] finds a shared latent representation by minimizing a low rank regularized likelihood of a probabilistic model, which assumes a Gaussian distribution for real valued data. [16] finds a common latent representation by minimizing a regularized l_2 norm squared reconstruction error over the multiple views. Similarly, [17] minimizes a regularized reconstruction loss over the data views. Their reconstruction loss function is generic and can be any convex loss function; however it can only take into account two views. Both [18] and [16] are sentitive to outliers as their loss functions minimize the l_2 norm squared and the Gaussian density function respectively, which are known to be fragile to even one outlier [12], [13]. All the previously mentioned convex multi-view methods do not take into account the local geometric structure of the data; a shortcoming that has been recently addressed by methods involving graph regularizers.

Graph regularizers have recently emerged in both the dimensionality reduction and data clustering areas of applied machine learning that encode the geometric structure of the data in the form of a graph to be exploited by the learning models as an injection of structural knowledge [25], [27], [29], [30], [31], [32], [33], [34], [35], [38], [39]. More specifically, multi-view subspace learning methods of: Graph regularized Multiset Canonical Correlation Analysis (GrMCCA) [35], Graph Multi-view Canonical Correlation Analysis (GMCCA) [34], Integrative Hypergraph regularization Principal Component Analysis (IHPCA) [38], and integrative Graph regularized Matrix Factorization Network Analysis (iGMFNA) [39] (which is a network analysis method that is a generalization of Graph Regularized NMF (GNMF)) are all able to take into account more than two views. They are all formulated as optimizing a non-convex objective function. Both IHPCA and iGMFNA use a matrix factorization model that is capable of finding a shared latent representation for the multiple data views. However, their models minimize an Euclidean distance measure which does not take into account outliers in the data. Moreover, both methods use an alternate iterative method to find the solution to their non-convex optimization problem, which leads to convergence to local optimal solutions. On the other hand, although both GrMCCA and GMCCA have non-convex objective functions they have closed form solutions that are directly computed by eigendecompositions. In the case of GMCCA [34] the graph regularization consists of a common source graph, meaning that it cannot model the graph of each view separately. This is a limitation as it can only be used for applications where common source graphs are available. Moreover, in the case of GrMCCA [35] each view uses a separate graph regularizer. Furthermore, all previously mentioned multi-view subspace learning methods are sensitive to even a small number of outliers. This is mainly because they minimize loss functions that have a quadratic term which will amplify the errors produced by the outliers in the data [13].

In addition to the previously mentioned multi-view methods, there exist multi-omic methods that have been specifically validated on cancer genomic datasets, such as: Similarity Network Fusion (SNF) [26] and iCluster [40]. These are considered as benchmark methods for multi-omic data integration [5]. SNF [26] integrates the different data views by forming networks of samples for each of the views and then fuses the different view-specific networks into one network that will incorporate the complementary structure of the different data views. Moreover, iCluster [40] is a joint latent variable model for integrative clustering that incorporates the different data views to find a shared clustering result.

In this paper we address the above limitations by introducing Convex Graph regularized Robust Multi-view Subspace Learning (CGRMSL) for the problem of multi-view clustering. It is formulated with a convex objective function, that separately takes into account the manifold structure of each view of the data, is robust to outliers, and finds a shared latent representation of the data. We show that our method has superior clustering performance and is able to better visualize the data than other convex and non-convex multi and single view subspace learning methods. We also show the ability of our model to detect potential subtypes with higher confidence than other state of the art multi-view methods. This is shown on genomic cancer datasets from the Cancer Genome Atlas (TCGA) repository [20].

This paper is organized as follows: in Section 2, we will introduce our method CGRMSL and show how it can be solved using a first order optimization method. Moreover, in the same section we will introduce CGMSL, a non-robust version of our method, it will be used as a comparing method to highlight the robustness of CGRMSL. In Section 3 we will show the time complexity of CGRMSL and the hardware settings used in this paper. In Section 4, we will show that our method is robust to outliers, by evaluating it on two multi-view synthetic datasets that have injected outliers. Our method will be compared to the non-robust version, CGMSL, and GrMCCA [35], which is also considered as a benchmark method that is non-robust to outliers. In Section 5, we will describe the multi and single view methods that will be used to compare with CGRMSL, when clustering the TCGA cancer genomic datasets. In Section 6, we will compare the clustering and subtype identification performance of our method to the competing methods on the cancer genomic datasets. Our method will be compared to both multi-view methods and single-view methods applied to each view separately. This section will highlight the better clustering performance on the shared latent space of CGRMSL. In Section 7, we will discuss the results shown in

Section 6. Finally, we present our concluding remarks and potential future work in Section 8.

2 Material and Methods

2.1 Convex Graph Regularized Robust Multi-View Subspace Learning

The algorithm we introduce in this paper is called Convex Graph regularized Robust Multi-view Subspace Learning (CGRMSL). It utilizes complementary information from more than one view to find a common latent representation, that will enhance clustering. The dataset to be considered has V views, with each view being represented by a matrix $M_v \in \mathbb{R}^{p_v \times n}$, consisting of n samples arranged in columns with each sample having p_v features, expressed as $M_v = [M_v^1, M_v^2, ..., M_v^n]$. The objective of this method is to decompose each view M_v as the sum of a low rank matrix L_v that gives a low-dimensional representation for the given view and a column sparse matrix C_v that has non-zero columns in the samples that have high reconstruction errors, and are thus treated as outliers. By modelling the reconstruction matrix C_v to be column sparse our method detects (thus is robust to) outlier samples. We make this modelling assumption because in the case of omic data, samples are more likely to be corrupt than a particular genomic feature across all data samples (the latter would have required a row sparse C). The common latent representation is found by constraining the low rank matrices L_v to be similar to a shared matrix between all views, L^*. The graph which has nodes corresponding to samples, is constructed by first finding the K nearest neighbours of each sample, measured in Euclidean distance. Then for each sample we weight the edges to its K neighbours through the Gaussian kernel function $W_v^{ij} = \exp(-\frac{||M_v^i - M_v^j||_2^2}{2\sigma^2})$. All edges to other points that are not in the K nearest neighbours of the sample are weighted as zero. The matrix that incorporates the neighbourhood and similarity information for each view is the affinity matrix $W_v \in \mathbb{R}^{n \times n}$. Then the graph Laplacian matrix $\Phi_v \in \mathbb{R}^{n \times n}$ is defined as $\Phi_v = D_v - W_v$, where D_v is a diagonal matrix where each entry on its diagonal is the row sum of the corresponding row in W_v, $D_v^{ii} = \sum_j W_v^{ij}$. The CGRMSL optimization problem is stated as follows:

$$\min_{L_v, L*, C_v} \sum_{v=1}^{V} \Big(||L_v||_* + \lambda_v ||C_v||_{1,2} + \gamma_v ||L_v - L^*||_F^2 + \alpha \text{tr}(L_v \Phi_v L_v^T) \Big). \quad \text{s.t: } M_v = L_v + C_v. \tag{1}$$

Where λ_v, γ_v, and α are real-valued regularization parameters. The first term in the objective function, $||L_v||_*$, is the nuclear norm of L_v which is the sum of its singular values. It induces low rankness in the matrix L_v. Minimizing the nuclear norm of a matrix is the closest convex surrogate of the intractable and combinatorial rank minimization problem [22], [24]. The second term $||C_v||_{1,2}$ is the sum of the l_2 norms of the columns of C_v. It will induce column sparseness in the matrix C_v. The $l_{1,2}$ norm is the nearest convex surrogate to the number of non-zero columns in a matrix [24]. From the constraint of Problem 1, $M_v = L_v + C_v$, we note that $C_v = M_v - L_v$ is the reconstruction error matrix for the v^{th} view. Therefore, CGRMSL aims to model the outliers by inducing a column sparse structure to the reconstruction error matrix C_v, so that they are filtered out from the low rank matrix L_v. Both the nuclear norm and the $l_{1,2}$ norm have been used in the literature to induce low rankness and column sparseness respectively [12], [24], [25]. Both these norms have been used in our precursor work [33] to induce the structures of the low rankness and column sparseness of the single-view matrix decomposition ($M = L + C$), with an additional graph regularizer (same as the fourth term of Problem 1), to detect outliers and to improve the clustering quality of the recovered subspace. It has been evaluated on single-view data of single cell genomics and cancer genomic datasets. Here, CGRMSL builds on and goes beyond our previous work [33] in being able to model multiple data views to find a shared latent space and is robust to outliers in each view.

The third term of Problem 1 constrains the low rank matrices of each view to be similar to a shared matrix L^*. This term, for a specific view v can be rewritten as $\sum_{i=1}^{n} ||L_v^i - L^{*i}||_2^2$. This constrains each of the column vectors of the low rank matrix of a view, L_v, to be as close as possible in Euclidean distance to each corresponding column vector of L^*. Summing this over all views (as in Problem 1) will integrate the complementary information for all the available views to extract the common latent representation. To extract this we first compute the truncated Singular Value Decomposition (SVD) of L^*, $L^* = U\Sigma V^T$. Then, the common low-dimensional latent representation is the projection of L^* onto its truncated column space U, i.e. $Z = U^T L^*$. The fourth term is a graph regularizer on the low rank matrices. It preserves the intrinsic manifold information of the input data in the form of a graph. To best interpret the function of the graph regularization term for a specific view v, $\text{tr}(L_v \Phi_v L_v^T)$, we can rewrite it in the following way:

$$\text{tr}(L_v \Phi_v L_v^T) = \frac{1}{2} \sum_{i,j=1}^{n} ||L_v^i - L_v^j||_2^2 W_{ij}.$$

The graph regularization term can be better interpreted now as $\frac{1}{2} \sum_{i,j=1}^{n} ||L_v^i - L_v^j||_2^2 W_{ij}$. This function will impose structure in the recovered low rank matrix L_v, in the sense that if two points have high affinity in the original input space the distance of the corresponding columns in L_v needs to be small. Problem 1 is a convex problem which can be solved to find a stable global solution using the Alternating Direction Method of Multipliers (ADMM) optimization method [23].

2.2 CGRMSL Algorithm

Here we use ADMM to optimize the objective function in Problem 1. ADMM has been used in [25], [33] to optimize problems in similar contexts of low rank and sparse matrix decompositions with an additional graph regularizer. The main difference between CGRMSL and our previous work [33] is the summation of the graph regularized decomposition of the input matrix ($M = L + C$) over all the available views, and the third term of Problem 1 that integrates the subspaces recovered from the different views. To solve CGRMSL using ADMM we need to introduce an auxiliary variable so that we can divide the objective function into four separate blocks. We rewrite Problem 1 as follows:

$$\begin{aligned} \min_{L_v,Q_v,L^*,C_v} \sum_{v=1}^{V} \Big(& ||L_v||_* + \lambda_v ||C_v||_{1,2} + \gamma_v ||Q_v - L^*||_F^2 \\ & + \alpha \text{tr}(Q_v \Phi_v Q_v^T) \Big). \\ & \text{s.t: } \ M_v = L_v + C_v \ , \ L_v = Q_v. \end{aligned} \tag{2}$$

Where Q_v with v from 1 to V are the auxiliary variables. Now we can define the augmented Lagrangian function of Problem 2:

$$\begin{aligned} & \mathcal{L}(L_v, L^*, C_v, Q_v, Z_{1,v}, Z_{2,v}) = \sum_{v=1}^{V} \Big(||L_v||_* + \lambda_v ||C_v||_{1,2} \\ & + \gamma_v ||Q_v - L^*||_F^2 + \alpha \text{tr}(Q_v \Phi_v Q_v^T) + \langle Z_{1,v}, M_v - L_v - C_v \rangle \\ & + \frac{p_1}{2} ||M_v - L_v - C_v||_F^2 + \langle Z_{2,v}, Q_v - L_v \rangle \\ & + \frac{p_2}{2} ||Q_v - L_v||_F^2 \Big), \end{aligned} \tag{3}$$

where $\langle X, Y \rangle \ = \text{Tr}(X^T Y)$ is the Frobenius inner product between matrices X and Y. To minimize the augmented Lagrangian with respect to each of the six variables, we use ADMM [23]. The general form of the ADMM algorithm to solve CGRMSL is shown in Algorithm 1, where $Z_{1,v}^k$ and $Z_{2,v}^k$ are Lagrange multiplier matrices and k is the iteration index. Steps 5 to 8 in Algorithm 1 have closed form solutions, derivations of which are shown in Appendix A. The ADMM algorithm has been proven to converge to a global solution for convex objective functions [23].

Algorithm 1 ADMM **Convex Graph Regularized Robust Multi-View Subspace Learning (CGRMSL)**

```
input: $M_v \in \mathbb{R}^{p_v \times n}$, $\lambda_v$, $\alpha$, $\gamma_v$, $\Phi_v \ \forall \ v$)
 1) initialize $L_v^0, L^{*,0}, C_v^0, Q_v^0 \ \forall \ v$ to random matrices.
 2) $Z_{1,v}^0 = M_v - L_v^0 - C_v^0$, $Z_{2,v}^0 = Q_v^0 - L_v^0$. $p_1 = 1$ and $p_2 = 1$.
 3) repeat following until convergence
 4)  for $v$=1 to $v$=$V$
 5)  $L_v^{k+1} = \text{argmin}_{L_v} \ \mathcal{L}(L_v, L^{*,k}, C_v^k, Q_v^k, Z_{1,v}^k, Z_{2,v}^k)$
 6)  $C_v^{k+1} = \text{argmin}_{C_v} \ \mathcal{L}(L_v^{k+1}, L^{*,k}, C_v, Q_v^k, Z_{1,v}^k, Z_{2,v}^k)$
 7)  $Q_v^{k+1} = \text{argmin}_{Q_v} \ \mathcal{L}(L_v^{k+1}, L^{*,k}, C_v^{k+1}, Q_v, Z_{1,v}^k, Z_{2,v}^k)$
 8)  $L^{*,k+1} = \text{argmin}_{L^*}\mathcal{L}(L_v^{k+1}, L^*, C_v^{k+1}, Q_v^{k+1}, Z_{1,v}^k, Z_{2,v}^k)$
 9)  $Z_{1,v}^{k+1} = Z_{1,v}^k + p_1(M_v - L_v^{k+1} - C_v^{k+1})$
10)  $Z_{2,v}^{k+1} = Z_{2,v}^k + p_2(Q_v^{k+1} - L_v^{k+1})$
output: $\hat{L}_v = L_v^{k+1}$, $\hat{C}_v = C_v^{k+1}$, $\hat{L}^* = L^{*,k+1}$ when $k$ is last iteration.
```

2.3 Non-Robust version of CGRMSL

Here we introduce a version of CGRMSL which is not robust to outliers to evaluate the contribution of such robustness to the clustering task. We call this multi-view subspace learning algorithm Convex Graph regularized Multi-view Subspace Learning (CGMSL). In CGMSL the $l_{1,2}$ norm for computing the reconstruction errors in Problem 1 is replaced by the standard Frobenius norm squared, $||M_v - L_v||_F^2$. The squared term present in this reconstruction error amplifies the outlier samples giving them much larger weight than non-outlier samples. This in turn skews the low-dimensional subspace towards the outliers making the CGMSL model sensitive to outliers. The optimization problem of CGMSL is as follows:

$$\begin{aligned} \min_{L_v,L^*} \sum_{v=1}^{V} \Big(& ||L_v||_* + \lambda_v ||M_v - L_v||_F^2 + \gamma_v ||L_v - L^*||_F^2 \\ & + \alpha \text{tr}(L_v \Phi_v L_v^T) \Big). \end{aligned} \tag{4}$$

This objective function is also convex; thus a global solution can be found using ADMM. To optimize Problem 4 with ADMM, we need to separate the objective function into three separate blocks by introducing auxiliary variables Q_v (with v from 1 to V):

$$\begin{aligned} \min_{L_v,L^*,Q_v} \sum_{v=1}^{V} \Big(& ||L_v||_* + \lambda_v ||M_v - Q_v||_F^2 + \gamma_v ||Q_v - L^*||_F^2 \\ & + \alpha \text{tr}(Q_v \Phi_v Q_v^T) \Big). \qquad \text{s.t: } \ L_v = Q_v. \end{aligned} \tag{5}$$

Now we can define the augmented Lagrangian function of Problem 5:

$$\begin{aligned} & \mathcal{L}(L_v, L^*, Q_v, Z_{1,v}) = \sum_{v=1}^{V} \Big(||L_v||_* + \lambda_v ||M_v - Q_v||_F^2 \\ & + \gamma_v ||Q_v - L^*||_F^2 + \alpha \text{tr}(Q_v \Phi_v Q_v^T) + \langle Z_{1,v}, Q_v - L_v \rangle \\ & + \frac{p_1}{2} ||Q_v - L_v||_F^2 \Big). \end{aligned} \tag{6}$$

We then minimize the augmented Lagrangian with respect to the four variables separately. The ADMM algorithm for CGMSL is shown in Algorithm 2. Steps 5 to 7 in Algorithm 2 have closed form solutions. Step 7 (Updating L^*) has the same closed form solution as CGRMSL; steps 5 and 6 are different and their derivations are shown in Appendix A.

Algorithm 2 ADMM **Convex Graph Regularized Multi-View Subspace Learning (CGMSL)**

```
input: $M_v \in \mathbb{R}^{p_v \times n}$, $\lambda_v$, $\alpha$ , $\gamma_v$, $\Phi_v \ \forall \ v$)
 1) initialize $L_v^0, L^{*,0}, Q_v^0 \ \forall \ v$ to random matrices.
 2) $Z_{1,v}^0 = Q_v^0 - L_v^0$. $p_1 = 1$.
 3) repeat following until convergence
 4)  for $v$=1 to $v$=$V$
 5)  $L_v^{k+1} = \text{argmin}_{L_v} \ \mathcal{L}(L_v, L^{*,k}, Q_v^k, Z_{1,v}^k)$
 6)  $Q_v^{k+1} = \text{argmin}_{Q_v} \ \mathcal{L}(L_v^{k+1}, L^{*,k}, Q_v, Z_{1,v}^k)$
 7)  $L^{*,k+1} = \text{argmin}_{L^*} \ \mathcal{L}(L_v^{k+1}, L^*, Q_v^{k+1}, Z_{1,v}^k)$
 8)  $Z_{1,v}^{k+1} = Z_{1,v}^k + p_1(Q_v^{k+1} - L_v^{k+1})$
output: $\hat{L}_v = L_v^{k+1}$, $\hat{L}^* = L^{*,k+1}$ when $k$ is last iteration.
```

2.4 Convergence of ADMM

In this section we discuss the convergence of Algorithm 1 and 2 which are multi-block ADMMs.

The general convergence results of ADMM are based on a two-block ADMM structure. This is when the objective function in question is composed of a summation of only two functions. Convergence in the two-block case is guaranteed, for any step-size $p > 0$, when both the functions in the objective are convex [23]. In the three-block and multi-block case of ADMM, convergence is not always guaranteed even when the functions in the objective are convex, and in some cases the algorithm diverges as shown by Chen et al. in [44]. Moreover, Chen et al. in [44] also proved that the presence of a mild condition guarantees the convergence of the extension of ADMM to a multi-block form. The general form of an optimization problem that is optimized by a N-block ADMM is as follows:

$$\begin{aligned} &\min_{\{\boldsymbol{x_i}\}_{i=1}^{\forall i}} \sum_{i=1}^{N} f_i(\boldsymbol{x}_i) \\ &\text{subject to: } \sum_{i=1}^{N} A_i \boldsymbol{x}_i = \boldsymbol{b}. \end{aligned} \tag{7}$$

The condition proved by [44] is the following.
Condition 1 [44]: Convergence of multi-block ADMM (Problem 7) is guaranteed when any two coefficient matrices, A_i, are orthogonal to each other. Therefore, in the case of CGRMSL and CGMSL, if we compare problems 2 and 5 to the general multi-block ADMM form, we find that, for both methods, all the coefficient matrices are equal to the identity matrix, $\{A_i\}_{i=1}^{\forall i} = I$, therefore satisfying the condition for multi-block ADMM to converge for any positive step-size $p > 0$.

3 Computational Complexity

The proposed method CGRMSL (Algorithm 1) is an iterative algorithm that repeats until convergence. The computational complexity of each iteration of Algorithm 1 stems from the following steps: 1) six different updates repeated V times (for each different view present in the data): L_v^{k+1}, C_v^{k+1}, Q_v^{k+1}, $L^{*,k+1}$,$Z_{1,v}^{k+1}$,$Z_{2,v}^{k+1}$. 2) After each iteration the algorithm computes the objective function of Problem 2 to check for convergence. We will break down the time complexity for both steps: 1) The time complexity of each of the six updates is shown as follows:

i $\boldsymbol{L_v^{k+1}}$ **Update**: Computation of one SVD and matrix multiplication of USV^T, with additional matrix additions, subtractions, and multiplication by constants. SVD complexity is $O(p^2n)$, matrix multiplication of USV^T is $O(p^2n + n^2p)$, additional matrix arithmetics is $O(pn)$. The overall complexity after dropping lower order terms is $O(p^2n + n^2p)$.

ii $\boldsymbol{C_v^{k+1}}$ **Update**: Computation of n l_2 norm computations of p-dimensional vectors and additional matrix arithmetics, both with a complexity of $O(pn)$. This gives an overall complexity of $O(pn)$.

iii $\boldsymbol{Q_v^{k+1}}$ **Update**: Computation of matrix arithmetics and inverse of diagonal matrix. The overall complexity is $O(pn + n^2)$.

iv $\boldsymbol{L_v^{*,k+1}}$ **Update**: Computation of matrix addition and multiplication by constant giving a total complexity of $O(pn)$.

v $\boldsymbol{Z_{1,v}^{k+1}}, \boldsymbol{Z_{2,v}^{k+1}}$ **Update**: Computation of matrix arithmetics including addition, subtraction, and multiplication by constant. This gives an overall complexity of $O(pn)$.

2) The time complexity of computing the objective function is broken down as follows:

i $||L_v||_*$ nuclear norm of L_v: Computing the SVD and sum of n singular values (because $p \ll n$). This gives a total complexity of $O(p^2n)$.

ii $||C_v||_{1,2}$ $l_{1,2}$ norm of C_v: Computing the square of every element of a $p \times n$ matrix, sum of n l_2 norms of p-dimensional vectors. This gives an overall complexity of $O(pn)$.

iii $||Q_v - L^*||_F^2$: Computing the subtraction of two $p \times n$ matrices and sum of the squared elements. The total complexity is (pn).

iv $\text{tr}(Q_v \Phi_v Q_v^T)$: multiplication of $p \times n$, $n \times n$, and $n \times p$ matrices with trace of result (sum of p diagonal elements of resulting matrix).The total complexity is $O(pn^2 + np^2)$.

Let I denote the number of iterations needed for CGRMSL to converge. Then the total complexity of the CGRMSL algorithm is the sum of V times the complexity of the updates added with the complexity of computing the objective function. The overall computational complexity expression of CGRMSL is denoted as follows:
$O\Big(I\big(V(p^2n + n^2p)\big)\Big)$.
The hardware settings for this paper are: Intel Core i7-6700, 3.4 GHz (4 cores), 8 GB RAM.

4 Simulation Study

4.1 Data Simulation

In this section we evaluate our model CGRMSL on two synthetic datasets by comparing against GrMCCA [35] and CGMSL. The first synthetic dataset is generated by a mixture of Gaussians (convex shapes). The second synthetic dataset comprises of a mixture of non-convex shapes, namely a mixture of 'moons'. We will show that our model is capable of finding a shared latent space that takes into account all complementary information from the different data views. Furthermore, we will show that our model is robust to outliers by finding a shared latent space that is not affected by their presence.

The first synthetic dataset comprises two 3-dimensional views, M_v $(v = 1, 2)$, with each view containing three different classes generated by a mixture of three Gaussian densities: $p(M_v) = \sum_{i=1}^{3} \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_v^i, \Sigma_v^i)$, where $\boldsymbol{\mu}_v^i$ and Σ_v^i are the mean vector and covariance matrix of the i^{th} Gaussian in the v^{th} view. We generate 500 samples from each Gaussian. For the 1st view the three classes C1, C2, and C3 have the parameters set as follows: $\boldsymbol{\mu}_1^1 = (1\ 2)^T$, $\boldsymbol{\mu}_1^2 = (1\ 4)^T$ and $\boldsymbol{\mu}_1^3 = (6\ 6)^T$. For the 2nd view the three classes C1, C2, and C3 are parametrized by: $\boldsymbol{\mu}_2^1 = (1\ 2)$, $\boldsymbol{\mu}_2^2 = (6\ 6)^T$ and $\boldsymbol{\mu}_2^3 = (1\ 4)^T$. For both views all covariance matrices are set to the identity matrix, and the third dimension is generated by concatenating to the samples from the 2-dimensional Gaussians a standard uniform random variable in the interval (0,0.5). The two views constructed as described will have complementary information to be able to separate the three different classes. Furthermore, for both views we inject two outliers deeper in the third dimension with coordinate vectors: $(2\ 4\ 1.5)^T$ and $(3\ 4\ \text{-}1.5)^T$. Figure 1 shows the input dataset structure of both views generated from a mixture of bivariate Gaussian densities.

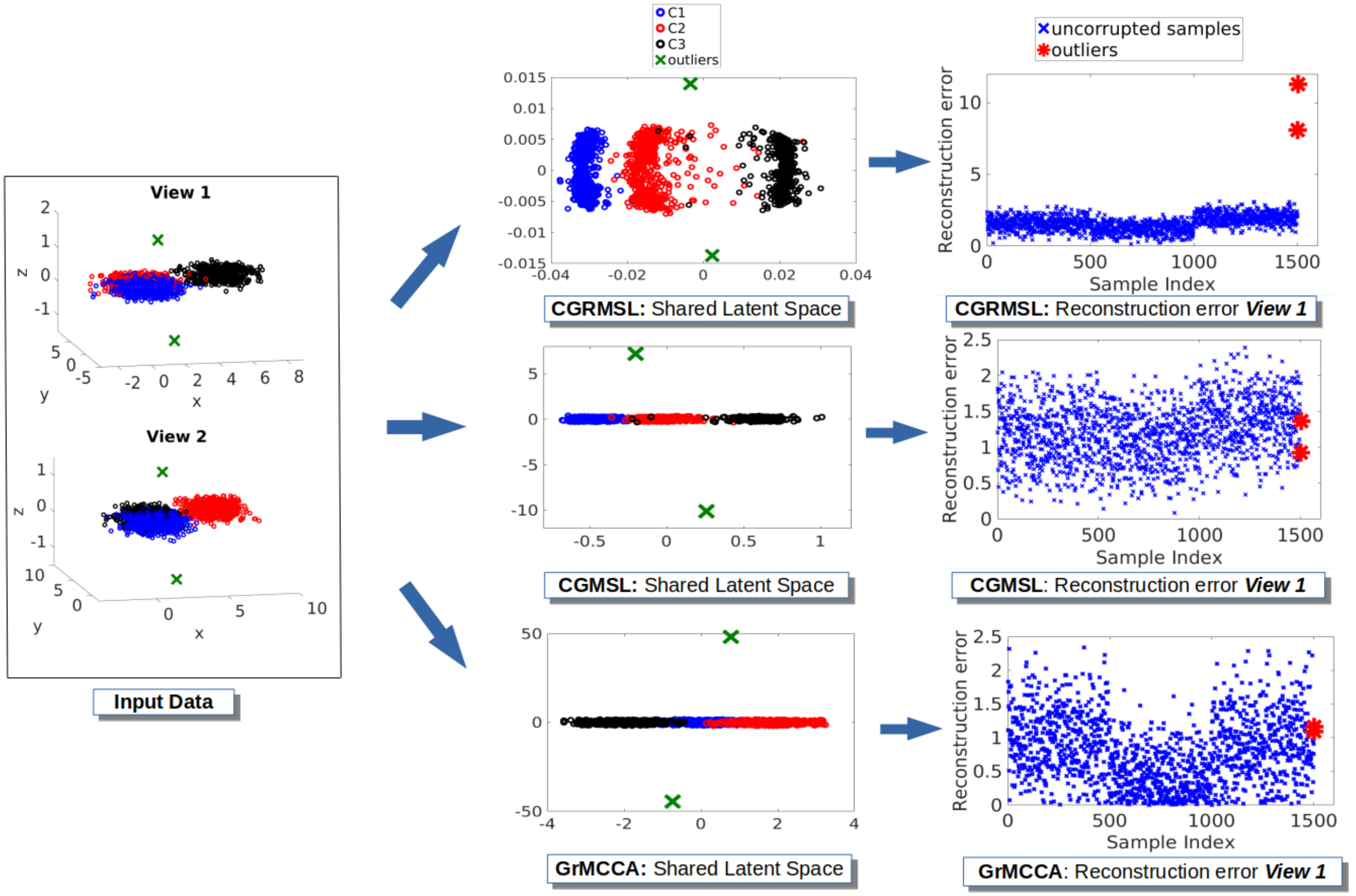


Fig. 1. Synthetic example to compare the three algorithms: CGRMSL, CGMSL and GrMCCA. For each method the shared latent space and the reconstruction error for each sample are shown. We can see that CGRMSL shows robustness to outliers as expected, whereas CGMSL and GrMCCA have resulted in skewed shared latent representations to accommodate the outliers. The robustness of CGRMSL is clearly seen in its reconstruction error plot, it shows that the outliers have much larger reconstruction errors compared to the uncorrupted samples.

The second dataset is also composed of two 3-dimensional views with each view having a mixture of three classes with each class containing 500 samples. Each view is generated as follows. First, three 2-dimensional ‘moons’ are generated. Then, the third dimension is formed by concatenating to the samples from the 2-dimensional ‘moons’ a standard uniform random variable in the interval (0,0.5). The two views are constructed to have complementary information to separate all the three classes as done for the first synthetic dataset. For this dataset, fractions of the entire 1500 samples of the dataset are corrupted to generate the outliers. The fraction of outliers generated are : 0.1 %, 1%, 3 %, 5%, 7%, 10%, 12 %, 15%. The outlier samples are generated by following a ”salt and pepper” corruption model.

4.2 Experimental setting and results

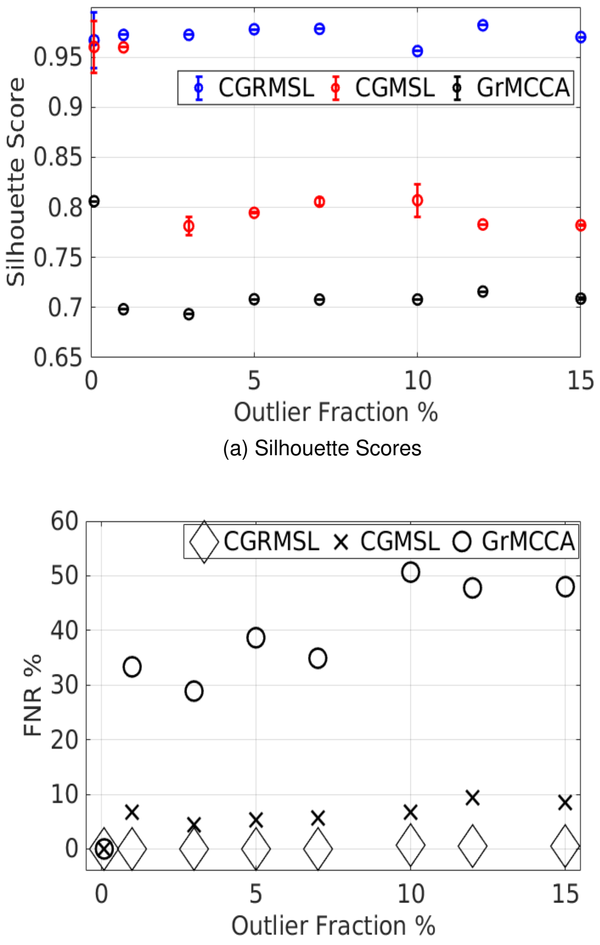
We first construct a K Nearest Neighbour graph for each view. Then, we compute the Gaussian kernel function W_v for each view by setting σ as the squared mean of the Euclidean distances between the K nearest neighbours of all samples. Finally, the graph Laplacian matrices Φ_1 and Φ_2 are constructed from their corresponding W_v as described in section 2.1. The shared latent space for both CGRMSL and CGMSL is found by computing Z as explained in section 2.1. For GrMCCA the shared latent space is computed as described in [35], where Y_{shared} is computed as $Y_{\text{shared}} = Y_1 + Y_2$, and $Y_v = P_v^T M_v$ is the projection of M_v onto the eigenvectors solving the eigendecompositon problem formulated in [35]. The reconstruction error of each sample is computed to show how the outliers affect each method. For CGRMSL and CGMSL the reconstruction error for each sample is computed by the l_2 norm of the error between the i^{th} sample M_v^i of the v^{th} view and its reconstruction from the shared latent space $\hat{L}^{*,i}$: $e_v^i = ||M_v^i - \hat{L}^{*,i}||_2$ for $i = 1, 2..., n$. However, for GrMCCA finding the reconstructions of the shared latent space in the original data space is not feasible. This is because it does not solve directly for a shared latent representation; instead it first solves for the projection vectors P_v of each view, then sums the projections of each view to create a shared latent representation. Hence, finding the reconstruction errors to investigate outliers can only be achieved by investigating reconstruction errors of each projected view. For GrMCCA the reconstruction for the v^{th} view is computed by $R_v = P_v Y_v$ and the reconstruction error for the v^{th} view is expressed by $e_v^i = ||M_v^i - R_v^i||_2$. Both synthetic datasets have been constructed to have three classes with information in both views to be able to separate all three classes. However, each view alone has two out of the three classes with significant overlap and the third class being separate from the first two, as shown in Figure

1 (Input Data). Therefore, if the method used is capable of integrating the complementary information in both views then the three classes should be all separated from each other in the shared latent space.

For the Gaussian mixture dataset Figure 1 shows the shared lower dimensional latent space of the synthetic data and the reconstruction error of each sample on the 1^{st} view for CGRMSL, CGMSL and GrMCCA (the 2^{nd} view leads to the same conclusion, so only one is shown for simplicity). From Figure 1 we can see that the shared latent space of CGRMSL effectively separates the three different classes present in the two views, whereas the shared latent space from GrMCCA shows less separability. We can also see for CGRMSL that the outliers have considerably higher reconstruction errors compared to all other samples. This indicates that the subspace of the shared latent space is not skewed to accommodate the outliers, thus proving the robustness of our method to outliers. On the other hand, for GrMCCA and CGMSL the reconstruction error of the outliers are in the range of the main samples, showing that the outliers have skewed the shared latent subspace to accommodate them.

For the mixture of 'moons' dataset we evaluate each method's ability to separate clusters in the recovered latent space, and the ability to detect the injected outliers by inspecting the reconstruction errors. This is done for the different outlier fractions mentioned in Section 4.1. The first step to evaluate the ability of a method to separate the three different classes is to compute cluster assignments, by using k-means clustering on the extracted shared latent representation. Then, to evaluate the obtained clusters the silhouette score is computed, which is the mean of the silhouette values of each sample. The silhouette value of each sample is a measure of how similar a sample is to its own cluster compared to the other clusters. For the i^{th} sample, the smallest average distance of the i^{th} sample to all points in any other cluster is denoted as a_i, and the average dissimilarity between the i^{th} sample to all other data points in the same cluster is denoted as b_i. The silhouette value for the i^{th} sample is defined as $s_i = (a_i - b_i)/(\max(a_i, b_i))$. The silhouette ranges from -1 to 1. The silhouette score of a dataset is the average of the silhouette values of all the samples in it. A silhouette score close to 1 indicates that clusters are well separated. Figure 2 (a) shows the errorbar of the silhouette scores of 50 runs of k-means computed on the shared latent spaces extracted from CGRMSL, CGMSL and GrMCCA. It is seen in the figure that CGRMSL has the highest silhouette scores for all fractions of outliers compared to the other two non-robust methods.

The outlier detection performance is assessed by the False Negative Rate (FNR) which computes the number of outliers that have reconstruction errors overlapping with the reconstruction errors of the uncorrupted samples. Therefore, the reconstruction error threshold that is chosen to compute the FNR is the maximum of the reconstruction errors of the uncorrupted samples. Figure 2 (b) shows the FNR for all the three methods for the different outlier fractions. It is seen from Figure 2 (b) that CGRMSL has the best outlier detection performance with FNR starting at zero and remaining close to zero. Moreover, CGMSL and GrMCCA have an increasing overall trend of FNR when a greater fraction of outliers are injected.



(a) Silhouette Scores

(b) False Negative Rate

Fig. 2. Performance of the three different algorithms on the mixture of 'moons' dataset. (a) Silhouette score of clusters computed on the shared latent representation of each method is displayed. (b) Shows the ability of detecting all the injected outliers by inspecting the reconstruction errors.

5 Comparisons

We compare CGRMSL to other convex and non-convex methods, both single and multi-view. The non-convex methods have analytical solutions that are computed by eigendecompositions. Another set of methods that we compare against are the benchmark multi-omic data clustering methods of SNF [26] and iCluster [40].

Single-view Subspace Learning (SSL) (convex) The aim of single view subspace learning is to find a low-dimensional latent representation of the input dataset by taking into account only one view. We will compare against SSL in [14], which finds a sparse low-dimensional latent representation by minimizing a convex objective function.

Single-view Subspace Learning (non-convex) These methods act on a single view and their objective functions are non-convex but have closed form solutions based on eigendecompositions. These are Principal Component

Analysis (PCA) [28] and Graph-Laplacian PCA (GPCA) [27].

Multi-view Subspace Learning (convex) The aim of these methods is to find a common low-dimensional latent representation by using information from multiple views. The methods we compare against are the following convex methods: LRA Cluster from [18], Convex multi-view subspace learning (CMSL) from [17].

Multi-view Subspace Learning (non-convex) These methods are non-convex multi-view methods but have closed form solutions. They are multi-view clustering via canonical correlation analysis (CCA) [6] and GrMCCA [34].

Convex Graph regularized Robust Single-view Subspace Learning (CGRSSL) This method is a single view subspace learning counterpart of our proposed method CGRMSL. It uses the $\hat{L}_v$ found from Algorithm 1 and from there finds the latent representation of the v^{th} view by projecting $\hat{L}_v$ onto its truncated column space: $Z_v = U_v^T \hat{L}_v$, with the SVD of $\hat{L}_v$ being, $\hat{L}_v = U_v \Sigma_v V_v^T$.

Convex Graph regularized Multi-view Subspace Learning (CGMSL) This is the non-robust version of CGRMSL described in section 2.3. It replaces the robust $l_{2,1}$ norm of CGRMSL with the standard Frobenius norm squared for the reconstruction error.

Benchmark Multi-Omic Data Clustering Methods SNF [26] and iCluster [40] are multi-omic data integration and data clustering methods. SNF fuses the different similarity networks of the available data views into a consensus kernel, then spectral clustering is performed on the consensus kernel to find cluster assignments. iCluster is a joint latent variable model that finds a shared clustering result among the data views.

6 Experimental Results Relevant to Cancer

In this section we validate our method against the other state of the art multi-view and single view methods described in section 5. To evaluate our method we conduct experiments on five different TCGA cancer data types [20]: breast cancer (BRCA), esophageal cancer (ESCA), endometrioid cancer (UCEC), kidney renal clear cell carcinoma (KRCCC), and lung squamous cell carcinoma (LSCC). For BRCA, ESCA, and UCEC pre-processed data are gathered from the UCSC Xena browser [36]. For KRCCC and LSCC the pre-processed data are provided by Wang et al. [26].

We first validate the clustering performance by finding a clustering assignment on the projection of the samples on the obtained low-dimensional subspace for each of the benchmark multi-view and single view methods. Subsequently, the clustering assignments are compared to the given subtype labels from the TCGA clinical data for three of the five cancer types: BRCA, ESCA, and UCEC (Section 6.2). For the remaining LSCC and KRCCC cohorts, cancer subtype labels are not present; therefore the objective is to find clusters that can be potential subtypes. Potential subtypes are discovered by performing a survival analysis and comparing how significantly survival times differ between samples in each cluster (Section 6.3). In Section 6.3 we compare only against the benchmark multi-view methods. Table 1 summarizes the different datasets used in this study.

	Patients	features per view	views	subtype labels	subtypes
BRCA	292	250	2	YES	3
ESCA	194	300	2	YES	2
UCEC	112	1000	2	YES	2
KRCCC	122	329	3	NO	to be found
LSCC	106	352	3	NO	to be found

TABLE 1
Summary of the five TCGA cancer datasets used in this paper.

Features in the case of the five TCGA datasets are: mRNAs for gene expression, DNA methylation sites for DNA methylation, and miRNAs for miRNA expression. BRCA, ESCA, and UCEC have two different views that consist of measurements at two different omic scales: gene expression (transcriptome) and DNA methylation (epigenome). KRCCC and LSCC have three views spanning two different omic scales: gene expression (transcriptome), DNA methylation (epigenome), and miRNA expression (transcriptome). In Table 1 the column 'features per view' describes the number of features retained per view for a specific cancer type. When considering gene expression and DNA methylation the number of sequenced features is considerably large (>20000). Thus it is necessary to reduce the dimensionality of the data to lower the computation time of our proposed methods (CGRMSL and CGMSL), and to get more stable results. Therefore, each view of the datasets present in Table 1 is filtered to retain the most variable genes across samples. This is a commonly used pre-processing procedure for machine learning algorithms applied to genomic datasets [45] to choose the most informative features. The number of features to retain is chosen from a trade-off between the time of computation and the fraction of the total variance explained by the chosen features. Note that choosing a number of features greater than 10000 does not change the conclusion deduced from the results; it only increases the time needed for computation. Moreover, for KRCCC and LSCC cancer types, the miRNA expression view has a small number of sequenced features. Thus, the number of features per view is chosen to be the smallest between the number of features of all views, because our method needs to have the same number of features for all the views.

6.1 Parameter settings

We would like to remind the reader that, from the CGRMSL objective function in Problem 1 in Section 2.1, the graph regularizer on the low rank matrix L_v is constructed based on the original data matrix M_v. Therefore, it is essential that the noise present in the original data is filtered out before constructing the similarity matrices W_v. The parameter K, the number of nearest neighbours used in constructing the kernel matrices $\{W_v\}_{\forall v}$, is important in filtering out the noise in M_v. Choosing a large value of K will fit any noise or outliers present in the data, therefore K should be a moderately small fraction of the size of the data. In such case it can model the local structure of the data without capturing the noise present in M_v. Thus, for our proposed method, CGRMSL, we need to tune K. To find the best value of K, a grid search for all values of K in the range of [1, total number of samples] is conducted. As shown in Figure 3

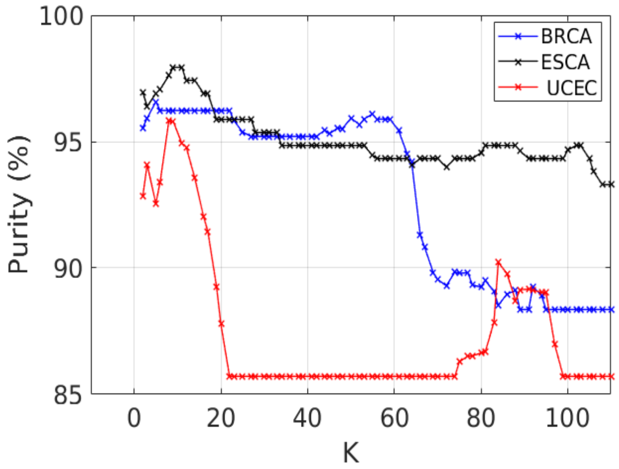
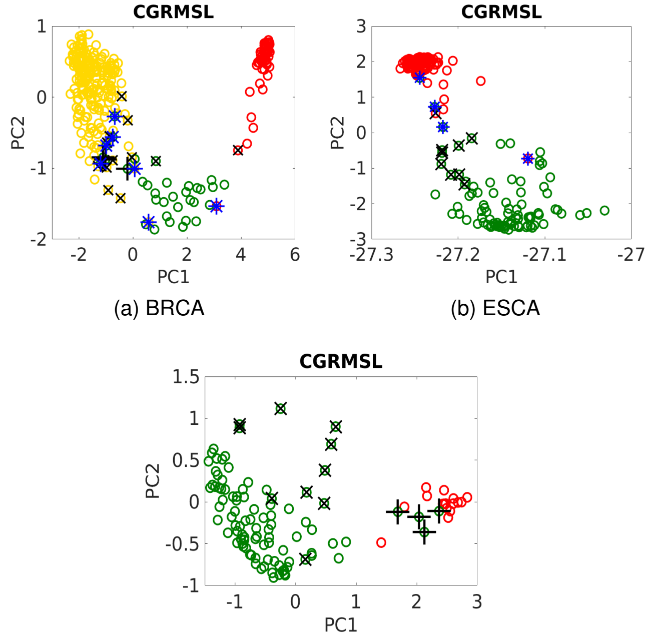


Fig. 3. Change in Purity (%) versus change in K: number of nearest neighbours in kernel $\{W_v\}_{\forall v}$ construction.



(a) BRCA

(b) ESCA

(c) UCEC

Fig. 4. Visualization of CGRMSL for BRCA, ESCA and UCEC. Different subtypes are labelled by: green, red, and yellow 'o'. Misclassified samples by k-means on the CGRMSL subspace are labelled by a **black '+'**. Misclassified samples by k-means on the original space is labelled by a **black 'x'**. Samples that are both misclassified by k-means on the original space and the CGRMSL subspace are labelled by a **blue '$\star$'**.

the mean clustering purity of k-means on the shared latent space Z is recorded for the three cancer types that have given class labels. From Figure 3, it is evident that if K is too large the algorithm fails to capture the local structure of the data and fits redundant noise, thus missing the difference between distinct classes. Therefore, we found empirically that for all the cancer genomic datasets used in our study a K in the range of [2, 20] is sufficient to find an optimal performance of CGRMSL. Moreover, we set σ, the width of the Gaussian function used to construct the Gaussian kernels $\{W_v\}_{\forall v}$, to a fixed value which is the squared mean of the Euclidean distances between the K nearest neighbours of all samples. In addition, the other parameters of CGRMSL found in its objective function (in Problem 1) are also tuned using grid search to obtain the best performance. We found empirically for all of the datasets used in this study that α should be set in the range of [0.1, 100], γ_v in the range of [0.1, 8] for all v, and $\lambda_v \; \forall v = \lambda$ with λ chosen in the range of [0.1, 10]. The effect of these parameters on the clustering performance is found in Appendix C.

We also tuned parameters for all other competing methods by conducting a parameter search. Afterwards, the values with the best performance are recorded in the following sections. For all comparative methods with graph regularizers the value of K is chosen to be the one that gives the best cluster evaluation criterion or p-value, in the range [1, total number of samples]. For GPCA α is chosen in the range [0.01, 2]. η for GrMCCA is chosen in the range [0.5e-4, 1e-2]. α for SSL and CMSL is chosen in the range [1, 100]. For CGMSL the optimal α, γ_v of CGRMSL are used and then λ is tuned again in the range [0.1, 10].

The cancer genomic datasets used in this study all have gold standards, this makes it possible to evaluate clustering performance or p-value when tuning the parameters. In case the gold standard is not available the tuning process is more challenging. To tune the parameters to suitable values we need to measure different metrics, other than accuracies, to evaluate the parameters chosen. For our method, CGRMSL, the two factors that are affected from the regularization parameters are the ranks of L_v and the number of outliers detected in each view. Outliers are defined as samples that have reconstruction errors (l_2 norms of columns of C_v) higher than a specific threshold. Therefore, we can only use both factors to tune the regularization parameters: λ (using $\lambda_v \forall v = \lambda$), α, and γ_v. The tuning process consists of solving CGRMSL for values of λ,α, and γ_v in a specific grid space, and looking for stable regions for the ranks of L_v. We then refine the search space of the parameters to the stable region and record the number of outliers detected in each view. A suitable value of the parameters needs to be chosen in such a way that the number of detected outliers in each view are less than or equal to an expected fraction of outliers. From our studies, we suggest to expect a fraction of outliers that is less than 15 % of the data. This is a suggested procedure to use when tuning parameters in complete absence of gold standards. However, our method's optimal parameters are detected when gold standards are present.

6.2 Clustering

Here we compare the proposed CGRMSL method against the benchmark single and multi-view methods described in Section 5. For the benchmark subspace methods we compute the clustering performance on their learned low-dimensional representations. The clustering performance is evaluated on the three TCGA cancer types with available subtype labels. The problem that will be investigated here is cancer subtype clustering. For BRCA the three most common breast cancer subtypes are: Luminal, Basal, and Her2-enriched. For ESCA the subtypes are: Adenocarcinoma and squamous cell carcinoma. For UCEC the subtypes are: Serous and Endometrioid. For each of the cancer types we first find the common samples between the different views, the three cancer datasets that have labelled subtypes comprise of: n =292, 194, 112 common patients for BRCA, ESCA, and UCEC respectively. After finding the common samples between views, for each view of the cancer types, we only retain the most variable genes across samples. The third column in Table 1 records the number of features retained per view for all the investigated cancer types.

For our method and all of the benchmark methods described above, we evaluate the clustering performance by measuring cluster purity, Normalized Mutual Information (NMI) [43], and Adjusted Rand Index (ARI) [42]. They have been used before in [19], [38], [41] to measure the performance of their multi-view clustering methods for cancer subtype clustering. These evaluation metrics measure the similarity between the predicted cluster labels and the true class labels.

Cluster purity is a measure of how much the clusters contain a single class. It is calculated by counting the number of data points from the most common class in each cluster, and averaging over all clusters:

$$\text{Purity} = \frac{1}{N}\sum_{i=1}^{C}\max_j |w_i \cap t_j|,$$

where w_i is the i^{th} cluster, t_j is the j^{th} class, C is the number of clusters and N is the number of data points.

NMI, is defined as:

$$NMI = 2\frac{I(W,T)}{H(W)+H(T)},$$

where $I(W,T)$ is the mutual information between the predicted cluster labels W and the true class labels T. $H(W)$ and $H(T)$ are the entropy of predicted clusters and true class labels respectively.

ARI is the corrected-for-chance version of the Rand index. It is defined by:

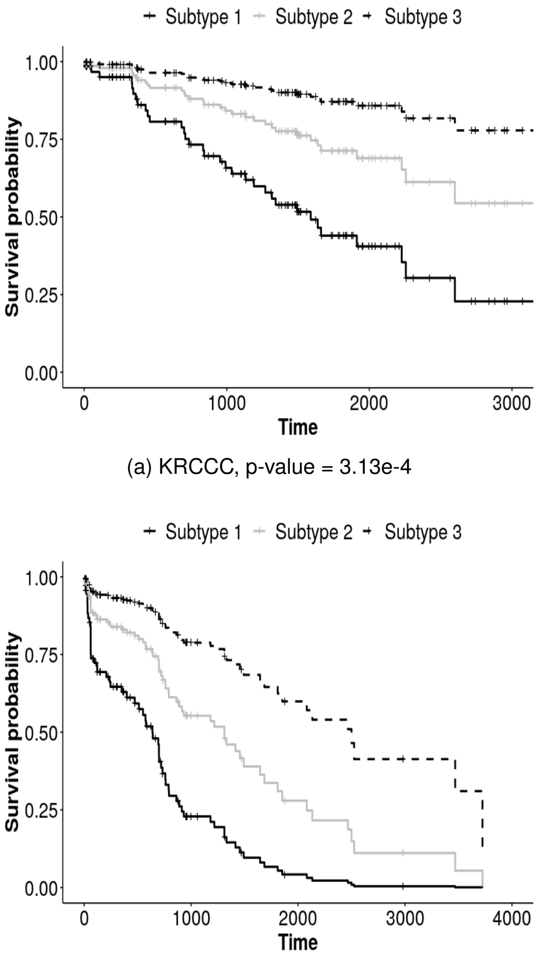
$$ARI = \frac{\sum_{i,j=1}^{C}\binom{n_{i,j}}{2} - \left[\sum_{i=1}^{C}\binom{a_i}{2}\sum_{j=1}^{C}\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{C}\binom{a_i}{2}+\sum_{j=1}^{C}\binom{b_j}{2}\right] - \left[\sum_{i=1}^{C}\binom{a_i}{2}\sum_{j=1}^{C}\binom{b_j}{2}\right]/\binom{n}{2}},$$

where $n_{i,j}$ is the number of times a sample labelled i in the predicted cluster labels W co-occurs with label j in the true class labels T. a_i is the sum of co-occurrences of cluster label i with all the class labels, $a_i = \sum_j n_{i,j}$. b_j is the sum of co-occurrences of class label j with all the cluster labels, $b_j = \sum_i n_{i,j}$. $\binom{n}{2}$ is the number of unordered pairs in a set of n elements.

Clustering is performed by k-means which is run 50 times on the latent representation of each method, the average of each evaluation metric measured on all the 50 runs are reported in Tables 2 and 3. The evaluation metrics of our method (CGRMSL) against all the benchmark single view methods are shown in Table 2. We can see from Table 2 that CGRMSL has superior clustering metrics compared to all the benchmark single view methods applied to each view separately. It is also seen from Table 3 that CGRMSL gives a better clustering performance compared to all other benchmark multi-view methods.

Another result worth highlighting is the capability of our method to visualize the three cancer types. We can see from Figure 4 that CGRMSL tightly places the different subtypes in distinct regions of the two-dimensional latent space. Moreover, Figure 4 also shows the misclassified samples when clustering on the CGRMSL subspace, and misclassified samples by k-means on the original space before dimensionality reduction.

6.3 Subtype Identification and Survival Analysis



(a) KRCCC, p-value = 3.13e-4

(b) LSCC, p-value = 3.27e-5

Fig. 5. Kaplan-Meier survival curves for KRCCC and LSCC. It shows distinct survival times of the subtypes identified by our method (CGRMSL).

Different cancer subtypes are expected to have significantly different survival times [4]. Here we apply our model to identify potential cancer subtypes by performing a survival analysis on the obtained clusters. This is performed on kidney renal clear cell carcinoma (KRCCC) and lung squamous cell carcinoma (LSCC). To measure how significantly the methods have identified different subtypes the Cox survival p-value is used. This is computed using the Cox Wald test to measure whether the subtypes have significantly different survival times. A lower Cox p-value indicates that survival profiles among subtypes are more significantly different. Consequently potential subtypes might be discovered using our multi-view clustering method.

	Criterion	k-means v_1	k-means v_2	PCA v_1	PCA v_2	GPCA v_1	GPCA v_2	SSL v_1	SSL v_2	CGRSSL v_1	CGRSSL v_2	CGRMSL
BRCA	purity	89.14±1.62	84.49 ±5.84	88.36	79.18±0.52	88.03±0.28	73.47 ±0.43	88.60 ±1.22	80.84 ±0.07	96.92	91.79±3.78	**97.26**
	NMI	57.30 ±8.5	38.37 ±8.95	51.44 ±0.25	24.6 ±0.52	60.60 ±0.11	50 ±0.19	62.88 ±0.9	48.77	84.12	82.05	**88.50**
	ARI	42.31 ±13.81	21.34 ±15.18	38.8 ±0.77	19.46 ±2.18	51.53 ±0.30	42 ±0.55	54.27	38.40	89.42	49.27	**91.20**
ESCA	purity	93.68 ±0.22	91.18 ±0.28	93.75 ±0.17	91.25 ±0.26	93.50±0.25	90.72	93.30	90.21	96.90	93.81	**97.94**
	NMI	68.80 ±1.13	61.86 ±1.16	66.51 ±0.85	61.97 ±1.26	67.52 ±0.83	59.32	67.91 ±0.66	62.29	81.12	72.25	**85.88**
	ARI	76.52 ±0.49	67.82 ±0.88	75.47 ±0.86	67.79 ±0.68	75.36 ±0.81	66.16	74.47 ±0.75	67.85	87.94	76.66	**91.88**
UCEC	purity	86.89±1.6	85.71	86.53±1.48	85.71	87.32±2.17	85.71	88.39	85.71	91.96	87.21 ±0.026	**96.43**
	NMI	28.56 ±15.9	25.29 ±2.82	29.69 ±14.69	27.08 ±0.39	25.63 ±17.58	23.09 ±16.20	47.91	28.16 ±6.20	56.63	27.90 ±0.4	**70.52 ±1.75**
	ARI	31.33 ±22.63	16.87 ±5.52	20.91 ±19.40	19.29 ±0.70	17.35 +- 22.45	19.38 ±18.36	53.26	20.90 ±11.39	64.83	20.63 ±0.73	**79.78 ±1.78**

TABLE 2
Cluster purity (average ±std) for single view subspace learning methods, k-means on original space, and CGRMSL. Readings with absent error bars have a std of zero for all 50 k-means runs. v_1 **is the gene expression view** and v_2 **is the DNA methylation view.**

	Criterion	CMSL	LRA Cluster	CCA	GrMCCA	CGMSL	iCluster	SNF	CGRMSL
BRCA	purity	88.36	88.36	88.35	96.92	97.05 ±0.17	84.59	88.36	**97.26**
	NMI	51.26	55.60 ±0.43	50.22 ±0.91	86.25	87.12 ±1.12	51.54	69.41	**88.50**
	ARI	38.57	42.20 ±0.24	33.85 ±1.4	90.15	90.76 ±0.52	60.43	73.24	**91.20**
ESCA	purity	95.36	95.88	94.85	95.36	95.57 ±0.24	96.91	97.42	**97.94**
	NMI	69.5 ±0.85	76.96	73.22	62.10 ±1.22	75.70 ±0.91	81.11	83.42	**85.88**
	ARI	77.69 ±0.92	77.47 ±0.94	80.34	64.09 ±1.63	83.15 ±0.95	87.95	89.90	**91.88**
UCEC	purity	85.94	85.71	90.58±0.44	92.36 ±3.65	95.39 ±0.85	90.18	89.29	**96.43**
	NMI	43.29 ±2.12	40.68 ±0.71	52.92 ±1.12	32.77±28.58	53.23 ±6.01	51.94	49.85	**70.52 ±1.75**
	ARI	46.36 ±3.3	42.31 ±1.13	60.27 ±1.49	48.83 ±33.46	63.50 ±8.38	58.80	55.97	**79.78 ±1.78**

TABLE 3
Cluster purity for the benchmark multi-view subspace learning methods and our method (CGRMSL).

	LRA Cluster	GrMCCA	CGMSL	iCluster	SNF	CGRSSL v_1	CGRSSL v_2	CGRSSL v_3	CGRMSL
KRCCC	1.47e-2	1.2e-3	9.48e-04	1.2e-1	8e-3	4.30e-4	6.30e-2	2.36e-2	**3.13e-4**
LSCC	8.21e-4	2.71e-4	4.46e-5	6.9e-2	1e-3	5.1e-3	4.62e-2	1.1e-3	**3.27e-5**

TABLE 4
Cox-Wald test p-value for all different multi-view methods. Parameters for each method are tuned and the best p-value is reported.

After projecting the samples onto the subspace given by CGRMSL we perform k-means clustering 50 times and report the lowest Cox Wald test p-value. The lowest p-value over the parameters of each method is reported. Here we partition into three clusters as it gives the best clustering result when compared to partitioning into two, four, and five clusters. The way in which we determine the optimal number of clusters for both LSCC and KRCCC cancer datasets is by inspecting the silhouette score of the samples in the low dimensional representation found by CGRMSL. Figure B.1 in Appendix B shows the silhouette scores computed on the shared latent space extracted from CGRMSL by using different values of K nearest neighbours; in the predefined range of [2,20]. This is repeated for: two, three, four, and five clusters. From the Figure we can see that the highest mean silhouette score is achieved for three clusters.

We compare our method to other state of the art multi-view methods that can take into account more than two views; these results are shown in Table 4. It it seen from Table 4 that our method, CGRMSL, scores a more significant p-value compared to the other multi-view methods and the single-view version of our algorithm CGRSSL (for each view). Moreover, the table shows that the outlier sensitive version of our algorithm, CGMSL, performs better than the other multi-view clustering methods. In addition, to show the distinct survival curves between identified subtypes, we show in Figure 5 the Kaplan-Meier survival curves for both cancer types using the subtypes identified by our method. From Figure 5 (a) and (b) it is evident that for both cancer types the three identified subtypes have significantly different survival profiles, a property that was not labelled in the datasets.

7 Discussion

The results comparing CGRMSL to the single-view methods in Table 2 show that our method is successful in integrating different views by taking into account the complementary information present in each view. It is also worth noting from CGRSSL (the single view counterpart of our method) that the 1st view has a higher clustering effect than the 2nd view, for each of the three datasets. This implies that the gene expression view has more subtype separability information than the DNA methylation view, for the three cancer genomic datasets: BRCA, ESCA, and UCEC. The benchmark SNF method in [26] analysed the same TCGA breast cancer genomic dataset, BRCA. The authors of the paper show in their study a similar conclusion, that for the BRCA cohort the gene expression view by itself has a lower p-value (more significant clustering results) compared to the DNA methylation and miRNA expression views.

The results in Table 3 and 4 show that our method, CGRMSL, is better able to integrate the complementary information present in the different omic views than the other benchmark multi-view clustering methods. CGRMSL uses a variety of different algorithmic features, which gives it the capability of performing better than the benchmark multi-view clustering methods used in this study. CGRMSL performs better because it: 1) uses and integrates the sample similarity information of each view (in the form of graph regularizers, similar to SNF and GrMCCA); 2) it is a subspace learning method where the resulting representation is a shared low-dimensional representation of the different views (similar to GrMCCA and CMSL); 3) it does not assume a simple Gaussian distribution noise like iCluster, GrMCCA, and CGMSL, which makes CGRMSL more robust to outliers; 4) it is not an early integration method like LRACluster, where the different views will be treated as one, which by doing this the structural diversities between different views are ignored; 5) it is formulated as a convex optimization problem that has all the previous features. In summary, CGRMSL blends in its framework the different algorithmic features of the benchmark methods, while including a model robust to outliers. This makes it a better multi-view clustering algorithm than the benchmark methods.

8 Conclusion

In this paper we propose an efficient convex multi-view clustering method that learns a common latent representation which takes into account the complementary information found in the separate views of the data. Moreover, it is robust to outliers in the data and takes into account its intrinsic manifold structure in each view. We have shown that our method CGRMSL is superior to state of the art convex and non-convex multi-view and single view methods found in the literature. Furthermore, we have shown that CGRMSL takes advantage of learning a shared latent representation, through the matrix L^*, as compared to the single view version of our method. We have shown better clustering performance on the important biomedical problem of cancer subtype clustering. Finally, we have shown the ability of our method to potentially discover new subtypes.

One limitation to take into account for our method is that the different views of the input data need to be filtered to have the same number of features before applying CGRMSL. In the case of genomic data this can be done with a feature filtering pre-processing step, which is quite common in the literature as many genes that are sequenced are not involved in the specific biological function that is being investigated.

Novel multi-omic single-cell datasets have started to emerge in the life-sciences community, they have a much higher resolution of biological information compared to traditional bulk sequencing techniques. Therefore, a promising future work direction is to investigate these new multi-omic datasets with CGRMSL, for subtype identification and even rare cell type identification. Our method will extend nicely to single-cell datasets as in our model we extract a low-rank approximation of the data, which is known to be beneficial with single-cell data as it can act as an imputation for the dropped-out measurements present in such data. Another direction of future work will be to extend CGRMSL by injecting biological prior knowledge extracted from interaction networks in the form of a graph between genes. A graph regularizer based on these biological interaction networks can be added to the CGRMSL model. In this case, not only can we include in our method information in different genomic views but also incorporate prior gene regulatory knowledge. This could potentially aid in drug target discovery for complex diseases.

Acknowledgments

OS was funded via a PhD studentship from the University of Southampton via the Engineering and Physical Sciences research Council (EPSRC). MN's contribution to the work was funded through EPSRC grant EP/S005463/1. We also thank Professor Akiko Takeda, University of Tokyo, for helpful comments on the convergence of the estimator.

Code and Reproducibility

All data used in this paper are available online from cited sources in the main text. Code is available on GitHub. https://github.com/omarshetta/IEEE_manuscript

References

[1] R. Siegel, K. Miller and A. Jemal, "Cancer statistics, 2015", *CA: A Cancer Journal for Clinicians*, vol. 65, no. 1, pp. 5-29, 2015.

[2] P. Bedard, A. Hansen, M. Ratain and L. Siu, "Tumour heterogeneity in the clinic", *Nature*, vol. 501, no. 7467, pp. 355-364, 2013.

[3] M. McCafferty, N. Healy and M. Kerin, "Breast cancer subtypes and molecular biomarkers", *Diagnostic Histopathology*, vol. 15, no. 10, pp. 485-489, 2009.

[4] M. Cai and L. Li, "Subtype Identification from Heterogeneous TCGA Datasets on a Genomic Scale by Multi-View Clustering with Enhanced Consensus", *BMC Medical Genomics*, vol. 10, no. 4, pp. 65-79, 2017.

[5] N. Rappoport and R. Shamir, "Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark", *Nucleic Acids Research*, vol. 47, no. 2, pp. 1044-1044, 2018.

[6] K. Chaudhuri, S. M. Kakade, K. Livescu and K. Sridharan, "Multi-View Clustering via Canonical Correlation Analysis", in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 129-136.

[7] S. Bickel and T. Scheffer, "Multi-View Clustering", in *Proceedings of IEEE international Conference on Data Mining*, vol. 4, pp. 19-26, 2004.

[8] D. Zhou and C. JC. Burges, "Spectral Clustering and Transductive Learning with Multiple Views", in *Proceedings of the 24th International Conference on Machine learning*, 2007, pp. 1159-1166.

[9] A. Kumar, P. Rai and H. Daume, "Co-regularized Multi-View Spectral Clustering", in *Proceedings of Advances in Neural Information Processing Systems*, 2011, pp. 1413-1421.

[10] N. Quadrianto and C. Lampert, "Learning Multi-View Neighborhood Preserving Projections", in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 425-432.

[11] J. Liu, C. Wang, J. Gao and J. Han, "Multi-View Clustering via Joint Non-negative Matrix Factorization", in *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013, pp. 252-260.

[12] E. J. Candès, X. Li, Y. Ma and J. Wright, "Robust Principal Component Analysis?", *Journal of the ACM*, vol. 58, no. 3, pp. 1-37, 2011.

[13] P.J. Huber, "Robust statistics", In International Encyclopedia of Statistical Science, pp. 1248-1251. Springer, Berlin, Heidelberg, 2011.

[14] X. Zhang, Y. Yu, M. White, R. Huang and D. Schuurmans, "Convex Sparse Coding, Subspace Learning, and Semi-Supervised Extensions", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011.

[15] F. Bach, J. Mairal and J. Ponce, "Convex Sparse Matrix Factorizations", arXiv preprint arXiv:0812.1869, 2008.

[16] Y. Guo, ”Convex Subspace Representation Learning from Multi-View Data”, in *Proceeding of the AAAI Conference on Artificial Intelligence*, vol. 27, no.1, 2013.

[17] M. White, X. Zhang, D. Schuurmans and Y. Yu, ”Convex Multi-view Subspace Learning”, in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1673-1681.

[18] D. Wu, D. Wang, M. Zhang and J. Gu, ”Fast Dimension Reduction and Integrative Clustering of Multi-Omics Data using Low-Rank Approximation: Application to Cancer Molecular Classification”, *BMC Genomics*, vol. 16, no. 1, 2015. Available: 10.1186/s12864-015-2223-8.

[19] P. Chalise and B. Fridley, ”Integrative clustering of multi-level *omic* data based on non-negative matrix factorization algorithm”, *PLOS ONE*, vol. 12, no. 5, 2017. Available: 10.1371/journal.pone.0176278.

[20] ”The Cancer Genome Atlas”, National Cancer Institute, 2019. [Online]. Available: http://cancergenome.nih.gov/. [Accessed: 01-Jul- 2019].

[21] V. Prasad, T. Fojo and M. Brada, ”Precision oncology: origins, optimism, and potential”, *The Lancet Oncology*, vol. 17, no. 2, 2016. Available: 10.1016/s1470-2045(15)00620-8.

[22] J. Cai, E. Candès and Z. Shen, ”A Singular Value Thresholding Algorithm for Matrix Completion”, *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956-1982, 2010.

[23] S. Boyd, ”Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”, *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2010.

[24] H. Xu, C. Caramanis and S. Sanghavi, ”Robust PCA via Outlier Pursuit”, *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3047-3064, 2012.

[25] N. Shahid, V. Kalofolias, X. Bresson, M. Bronstein and P. Vandergheynst, ”Robust Principal Component Analysis on Graphs”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2812-2820.

[26] B. Wang et al., ”Similarity network fusion for aggregating data types on a genomic scale”, *Nature Methods*, vol. 11, no. 3, pp. 333-337, 2014.

[27] B. Jiang, C. Ding, B. Luo and J. Tang, ”Graph-Laplacian PCA: Closed-Form Solution and Robustness”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3492-3498.

[28] I. Joliffe, *Principal Component Analysis and factor analysis*. 1986.

[29] N. Shahid, N. Perraudin, V. Kalofolias, G. Puy and P. Vandergheynst, ”Fast Robust PCA on Graphs”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 740-756, 2016.

[30] F. Shang, L. Jiao, and F. Wang, Graph dual regularization non-negative matrix factorization for co-clustering, *Pattern Recognition*, vol. 45, no. 6, pp. 22372250, Jun. 2012.

[31] J. Chen, G. Wang, Y. Shen, and G. B. Giannakis, ”Canonical correlation analysis of datasets with a common source graph,” *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 43984408, 2018.

[32] N. Yu, M. J. Wu, J. X. Liu, C. H. Zheng, and Y. Xu ”Correntropy-Based Hypergraph Regularized NMF for Clustering and Feature Selection on Multi-Cancer Integrated Data,” *IEEE Transactions on Cybernetics*, pp. 112, Jun. 2020.

[33] O. Shetta and M. Niranjan, ”Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality”, *Royal Society Open Science*, vol. 7, no. 2, 2020. Available: 10.1098/rsos.190714.

[34] J. Chen, G. Wang and G. Giannakis, ”Graph Multiview Canonical Correlation Analysis”, *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 2826-2838, 2019.

[35] Y. Yuan and Q. Sun, ”Graph regularized multiset canonical correlations with applications to joint feature extraction”, Pattern Recognition, vol. 47, no. 12, pp. 3907-3919, 2014.

[36] ”UCSC Xena”, Xenabrowser.net, 2020. [Online]. Available: https://xenabrowser.net/. [Accessed: 09- Jan- 2020].

[37] J. Rhee et al., ”Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer”, *Nucleic Acids Research*, vol. 41, no. 18, pp. 8464-8474, 2013.

[38] M.J. Wu, Y.L. Gao, J.X. Liu, C.H. Zheng, and J. Wang ”Integrative Hypergraph Regularization Principal Component Analysis for Sample Clustering and Co-Expression Genes Network Analysis on Multi-Omics Data,” *IEEE journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1823-1834, Jun. 2020.

[39] Y.L. Gao, M.X. Hou, J.X. Liu, and X.Z. Kong ”An Integrated Graph Regularized Non-Negative Matrix Factorization Model for Gene Co-Expression Network Analysis”, *IEEE Access*, vol. 7, Sep. 2019.

[40] R. Shen, A. B. Olshen, M. Ladanyi ”Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis”, *Bioinformatics*, vol. 25, no. 22, pp. 29062912, Sep. 2009.

[41] S. Mitra, S. Saha, M. Hasanuzzaman ”Multi-View Clustering for Multi-Omics Data Using Unified Embedding,” *Scientific Reports*, vol, 10, no. 1, pp 1-16, Aug. 2020.

[42] L. Hubert, P. Arabie ”Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp 193-218, 1985.

[43] T. Cover and J. Thomas, ”*Elements of Information Theory*”. New York: Wiley, 1991.

[44] C. Chen, B. He, Y. Ye and X. Yuan, ”The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent”, *Mathematical Programming*, vol. 155, no. 1-2, pp. 57-79, 2014.

[45] M. Piles et al., ”Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs”, Genetics Selection Evolution, vol. 51, no. 1, pp. 1-15, 2019.

Omar Shetta received the BEng degree in Electrical and Electronics Engineering (2016) and PhD degree in Machine Learning (2021) from the University of Southampton, Southampton, UK. His research interest is in structured low rank approximations to various machine learning problems including computational biology. He is currently a post doctoral researcher in The Electronics and Computer Science (ECS) at Southampton University his research focus is on deep learning for image segmentation.

Mahesan Niranjan received the B.Sc. degree from the University of Peradeniya, Sri Lanka, in 1982, the M.E.E. degree from Eindhoven University of Technology, The Netherlands, in 1985, both in electronic engineering, and the Ph.D. degree from the University of Cambridge, Cambridge, UK, in 1990. He is currently Professor of Electronics and Computer Science at the University of Southampton, Southampton, UK, where he was Head of the Information: Signals, Images and Systems (ISIS) research group. Prior to this appointment in February 2008, he has held a professorship in the University of Sheffield (1999-2008) and a lectureship in t he University of Cambridge (1990-1998). At the University of Sheffield, he served as Head of Computer Science (2002-2004) and Dean of the Faculty of Engineering (2006-2008). His research interests are in the algorithmic and applied aspects of machine learning, and he has authored or coauthored about 100 papers in peer reviewed journals and conferences. He has been Program Chair of several international workshops and has acted as a co-organizer of a six month program on neural networks and machine learning at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK. His current research has a strong focus on computational biology and biomedical signal processing.

Srinandan Dasmahapatra received the BSc (Calcutta) and PhD (Stony Brook) degrees in physics. After a couple of post-docs in physics, working at the interface between statistical mechanics and quantum field theory, he has worked in various aspects of artificial intelligence, including speech recognition and knowledge representation. Currently, he mostly focuses on tackling problems in biology and physics based on approaches to inference and representation in machine learning