Yashita Jain[1] / Shanshan Ding[1,2] / Jing Qiu[1,2]

Sliced inverse regression for integrative multi-omics data analysis

[1] Center for Bioinformatics and Computational Biology, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA, E-mail: sding@udel.edu. https://orcid.org/0000-0003-1935-694X.
[2] Department of Applied Economics and Statistics, University of Delaware, 531 S College Ave., Newark, DE 19711, USA, E-mail: sding@udel.edu. https://orcid.org/0000-0003-1935-694X.

Abstract:
Advancement in next-generation sequencing, transcriptomics, proteomics and other high-throughput technologies has enabled simultaneous measurement of multiple types of genomic data for cancer samples. These data together may reveal new biological insights as compared to analyzing one single genome type data. This study proposes a novel use of supervised dimension reduction method, called sliced inverse regression, to multi-omics data analysis to improve prediction over a single data type analysis. The study further proposes an integrative sliced inverse regression method (integrative SIR) for simultaneous analysis of multiple omics data types of cancer samples, including MiRNA, MRNA and proteomics, to achieve integrative dimension reduction and to further improve prediction performance. Numerical results show that integrative analysis of multi-omics data is beneficial as compared to single data source analysis, and more importantly, that supervised dimension reduction methods possess advantages in integrative data analysis in terms of classification and prediction as compared to unsupervised dimension reduction methods.

Keywords: Integrative genomic analysis, sliced inverse regression, sufficient dimension reduction
DOI: 10.1515/sagmb-2018-0028

1 Introduction

With the advent of high-throughput technologies such as micro-arrays for genome wide assays, it has been possible to measure a broad range of genomic data extensively used in biomedical, in particular cancer studies (Liu, Shen & Pan, 2016). These genomic data can be of different types such as structural genomics, functional genomics, epigenomics and metagenomics. Functional genomics focuses on dynamic aspect of gene transcription, translation and protein-protein interactions.

The functional genomics such as MiRNA, MRNA and proteomic are responsible to play either oncogenic or tumor suppressive roles present in the cancer samples. The critical changes in these gene expression from the cancer cells enable tumors to initiate and progress in different tissues or aid in suppressing the different tumors (Bhattacharjee et al., 2001; Bichsel et al., 2001; Nishizuka et al., 2003; Reis-Filho & Pusztai, 2011; Wei et al., 2014; Peng & Croce, 2016; Oliveto et al., 2017). These functional genomic studies not only identify significant genes related to cancer, but can also help in identifying different cancer types. As these gene expression profiles are informative, revealing the developmental lineage and differentiation state of the tumors, they have high potential in cancer classification and diagnosis (Lu et al., 2005; Xu et al., 2016).

To reserve and store these genomic data altogether, several repositories such as The Cancer Genome Atlas (TCGA), NCI-60 and The International Cancer Genome Consortium (ICGC) have profiled thousands of cancer genome samples, generating a broad range of genomic expression profiles (Liu et al., 2010; Gholami et al., 2013) to encourage researchers for integrative analysis. Integrative genomics is based on the idea that any biological system is made up of many multiple molecular phenomena, and only by understanding the interaction between different layers of genomic structures, its phenotypic traits can be explored (Kristensen et al. 2014). As shown from the above instances, while the functional genomics are responsible for playing alternation in cancer expression, it is important to study the different gene expression simultaneously. Integrative analysis of these functional genomic data from multiple sources can potentially provide additional biological insights (Rhodes & Chinnaiyan, 2005; Nie et al., 2007). An integrative-omics approach can identify novel genes, markers, vital networks and pathways (Iliopoulos et al., 2008; De Cubas et al., 2013). It can also classify disease progression and different cancer types by analyzing different gene expression profiles simultaneously (Shen, Olshen & Ladanyi, 2009; Nibbe, Koyutürk & Chance, 2010).

Shanshan Ding is the corresponding author.

Although different omics data might have different functional information, they all possess high dimensionality features. Such high dimensional data often contain redundant information and bring extraneous variation to the goal of study. One of the commonly used solutions to handle high dimension data is to extract important features from a low dimensional projection, such that the original variables can be transformed into a set of new variables with lower/much lower dimensions. This is called dimension reduction or dimensionality reduction (James et al. (2013)). Representative methods include but not limited to principal component analysis (PCA), partial least squares (PLS), sliced inverse regression (SIR), and their extensive extensions. In recent years, several integrative dimension reduction methods have been developed and demonstrated greater power than separate analysis of each data type. For example, the integrative PCA related methods such as iCluster (Shen, Olshen & Ladanyi, 2009; Shen et al., 2012), sparse iCluster (Shen, Wang & Mo, 2013), and irPCA (Liu, Shen & Pan, 2016), and the integrative PLS related methods including sparse PLS (Lê Cao et al. 2008), integrOmics (Lê Cao, González & Déjean, 2009), sMBPLS (Li et al. 2012), among others. In particular, Shen, Olshen, and Ladanyi (2009) developed a joint probabilistic PCA model to achieve integrative clustering. Liu, Shen, and Pan (2016) proposed an integrative and regularized PCA methods using an elastic net penalty to achieve more efficient computational and numerical performance. Lê Cao et al. (2008) and Li et al. (2012) studied sparse PLS approaches for integrating two and multiple data types, respectively.

Despite of the recent development of the integrative dimension reduction techniques, there are data settings that existing approaches might not well address. For example, the integrative PCA methods are unsupervised, meaning that dimension reduction is conducted marginally on the predictors X without considering the relationship with the response variable Y. There is no reason, however, in principle that the marginal reduction can provide fully useful information about the response (Cox 1968). Thus, these type of methods might not well serve for regression or prediction problems. Though the integrative PLS methods are supervised and take care of the response variable when reducing the predictors, they inherit certain limitations from the standard PLS approach. For example, they only retain useful features for conditional mean function but might lose relevant information necessary for prediction and full regression problems (Li, Cook & Tsai, 2007). In addition, they mainly focus on data with continuous responses and is less applied to classification problems.

In this article, we propose to use sufficient dimension reduction (SDR) methods for integrative analysis of multi-omics data to improve prediction performance over unsupervised dimension reduction methods. We develop an integrative sufficient dimension reduction approach to achieve simultaneous dimension reduction of multiple data types with sharing structures while preserving full information for regression or classification. Sufficient dimension reduction (SDR) (Cook 1994; 1996) is a type of supervised dimension reduction methods that is important in both theory and practice. It serves to reduce the dimension of the predictors X by replacing them with a minimal set of linear combinations, without loss of information in modeling the relationship with Y. The proposed method extends the classical sufficient dimension reduction approach called slice inverse regression (SIR) (Li 1991) to the multiple source of data and integration setting. We referred to it as integrative slice inverse regression, or Integrative SIR (ISIR). The new ISIR method resolves the aforementioned issues in existing integrative procedures and integrates multiple omits data for simultaneous dimension reduction. We develop new algorithms for ISIR and demonstrate the advantages of sufficient dimension reduction methods in integrative omics data analysis via numerical studies, which extend the preliminary study in Jain and Ding (2017).

The rest of the paper is organized as follows. In Section 2, we first briefly review SDR with a focus on SIR, and then propose the Integrative SIR method and new computational algorithms. Section 3 examines the numerical performance of SIR and ISIR with both simulation studies and real data analysis and compares them with some unsupervised dimension reduction methods. It includes data description, prescreening, dimension reduction, and prediction with different methods. This section also discusses the robustness of Integrative SIR to different initial values and classification methods. Section 4 concludes the paper.

2 Method

2.1 Brief review of sufficient dimension reduction (SDR)

SDR is a powerful tool for data reduction and visualization. It is a supervised dimension reduction method that seeks to replace the predictor vector by its projection onto a lower dimensional subspace of the original predictor space without the loss of information on the response variable (Cook 1994; 1996; 2004). More specifically, consider a p-dimensional predictor vector X and a response variable Y. SDR seeks the projection of the p-dimensional X onto a d-dimensional subspace that seizes all the information we need to know about the outcome Y. Here the dimension d is usually much smaller than p and the smallest subspace (smallest d) is of

interest to achieve maximum reduction. Mathematically speaking, SDR tries to find the smallest subspace δ of the predictor space $\mathbb{R}^p$ such that

$$Y \perp\!\!\!\perp X|P_\delta X, \tag{1}$$

where $\perp\!\!\!\perp$ indicates independence, and $P_{(.)}$ stands for the projection operator. The above equation means that Y is independent of X given the reduction $P_\delta X$. The smallest subspace δ that satisfies (1) is called the central subspace (CS) for $Y \mid X$ (Cook 1994; 1996; 1998a), often denoted by $\delta_{Y|X}$. Let $d = dim(\delta)$ and let $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_d) \in \mathbb{R}^{p\times d}$ be a basis matrix of δ. By (1), the predictors X can be replaced by the linear combinations $\gamma_1^T X, \dots, \gamma_d^T X$, such that the d transformed predictors retain full information on modeling Y. In other words, $Y \perp\!\!\!\perp X|\gamma_1^T X, \dots, \gamma_d^T X$. The goal of SDR is to find the central subspace, or a basis of the central subspace and then use the basis to form the reduced predictors (linear combinations). There are a number of methods available to estimate the central subspace, such as sliced inverse regression (SIR) (Li 1991), sliced average variance estimation (SAVE) (Cook and Weisberg 1991), partial SIR (Chiaromonte, Cook & Li, 2002), directional regression (DR) (Li and Wang 2007), among many others. As SIR is one of the most widely used SDR method, we mainly focus on extending SIR to integration of multiple omics data although the proposed idea can be similarly applied to other scenarios. The next subsection reviews the SIR method and its algorithm to find the estimation of CS.

2.2 Sliced inverse regression

Introduced by Li (1991), sliced inverse regression (SIR) is a classical and popular non-parametric method for sufficient dimension reduction (SDR). It utilizes the inverse conditional mean $E(X \mid Y)$ to estimate the central subspace $\delta_{Y|X}$. To facilitate the description, let $\Sigma = cov(X)$ be the covariance matrix of X and $Z = \Sigma^{-1/2}(X - E(X))$ be the standardized predictor vector. Under a linearity condition proposed in Li (1991), it can be shown that the Z-scale inverse conditional mean $E(Z \mid Y)$ is contained in $\delta_{Y|Z}$, the central subspace for $Y \mid Z$. Hence one can use the sample version of the conditional mean under different outcomes (or slices) of Y to cover and estimate $\delta_{Y|Z}$. For more details, see Li (1991). Once $\delta_{Y|Z}$ is estimated, by the invariant property in Cook (1998b), we have $\delta_{Y|Z} = \Sigma^{1/2}\delta_{Y|X}$. The estimator of $\delta_{Y|X}$ can thus be easily obtained by estimating Σ with its sample covariance matrix.

Let $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{R}^{p\times d}$ be a semi-orthogonal basis of the central subspace $\delta_{Y|Z}$. Cook (2004) proposed a least square framework for formulating SIR. In particular, for each value y of Y, since $E(Z|Y) \in \delta_{Y|Z}$, there exists a coordinate vector $C_y \in \mathbb{R}^d$ such that

$$E(Z|Y = y) = \beta C_y.$$

This relationship can also be formulated into an inverse regression model:

$$Z_y = \beta C_y + \varepsilon, \tag{2}$$

where Z_y represents $Z \mid Y = y$ and ε is a random error and is independent of Y. Therefore, for an i.i.d. random sample (Y_i, X_i), $i = 1, \dots, n$, the estimation of β and C_y can be achieved by minimizing the least squares loss function

$$L_d(\beta, C_y) = \sum_{y=1}^{h} \| \hat{Z}_y - \beta C_y \|^2 \tag{3}$$

subject to $\beta^T\beta = I_d$, where I_d is the d by d identity matrix, h is the number of classes of Y if Y is categorical, and is the number of slices that partition the range of Y into intervals (slices) if Y is continuous, $\hat{Z}_y = \hat{E}(Z|Y = y)$ is the sample conditional mean (or sample slice mean), and $\| \cdot \|$ stands for the Euclidian norm. When Y is continuous, slicing here ensures enough observations for a good estimate $\hat{Z}_y$. In this case, the symbol y in $\hat{Z}_y$ represents a particular slice but not a particular value. For simplicity, we keep the same notation.

The following algorithm shows how to estimate β and the target central subspace $\delta_{Y|X}$ for the SIR method.

Algorithm 1 The SIR Algorithm

Input: an $n \times p$ standardized data matrix $(Z_1, \dots, Z_n)^T$ with responses $Y_1, \dots, Y_n$, where Z_i is the standardized X_i, $i = 1, \dots, n$.

Output: $p \times d$ estimated basis matrix $\hat{\beta}$ for $\delta_{Y|Z}$ and $p \times d$ estimated basis matrix for $\delta_{Y|X}$ through the following steps.

1. Assuming β is fixed, estimating C_y by minimizing the objective function (3) with $\hat{C}_y = \beta^T \hat{Z}_y$.
2. Plug the estimate of C_y back to (3). Let $\hat{\beta}_1, \ldots, \hat{\beta}_d$ be the leading d eigenvectors of

$$\hat{M} = \frac{1}{n} \sum_{y=1}^{h} n_y \hat{Z}_y \hat{Z}_y^T, \tag{4}$$

the sample covariance matrix of $E(Z|Y)$, where n_y is the sample size for each category or slice of Y. Then $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$.

3. Consequently, $\hat{\Sigma}^{-1/2}\hat{\beta}$ is an estimated basis for $\delta_{Y|X}$, where $\hat{\Sigma}$ is an estimate of Σ, e.g. sample covariance matrix of X.

The SIR algorithm provides a closed form solution for the sufficient reduction. There is no iteration needed to find the estimated basis for $\delta_{Y|Z}$ or $\delta_{Y|X}$.

To apply SIR in multiple omics data analysis, one way is to simply apply SIR to each omics data type separately, and then integrate/combine all the reduction results from multiple data sources together for prediction or classification. As shown in Section 3, using such sufficient dimension reduction methods over unsupervised dimension reduction methods as well as using integrative analysis over a single data type analysis can greatly improve prediction performance. Although it is beneficial to apply SIR to integrative analysis, the above strategy might not borrow information among different data types and might not capture sharing features across data types. Therefore, it is desirable to develop an integrative SIR method and to perform SIR on multiple data types simultaneously while taking care of sharing features among data sources and preserving full information for regression and prediction. The proposed method called Integrative SIR (ISIR) is discussed in the following subsection.

2.3 Integrative SIR

The idea of Integrative SIR is motivated by the unsupervised integrative dimension reduction method "iCluster" (Shen, Olshen & Ladanyi, 2009). Shen, Olshen, and Ladanyi (2009) pointed out that in a latent model such as (2), the coordinate part C_y's can actually indicate a latent clustering structure among the different classes (or slices) of Y. For example, if the response variable represents tumor types, then C_y, $y = 1, \ldots, h$, can represent latent clustering structure among the different tumor types. Suppose there are totally s omics data types. Let $X^{(j)} \in \mathbb{R}^{p^{(j)}}$, $j = 1, \ldots, s$, denote the $p^{(j)}$-dimensional predictor vector in jth omics data source. For each subject, we observe the response variable Y and the predictor information $X^{(1)}, \ldots, X^{(s)}$. To integrate multiple sources of data information, the goal of Integrative SIR is to take into account all the multi-omics data information simultaneously while finding a sufficient reduction for each data type by sharing common latent clustering information. It is useful for handling multiple sources of data with similarities (Jain and Ding 2017).

Let $\beta^{(j)} \in \mathbb{R}^{p^{(j)} \times d^{(j)}}$ be a basis for the central subspace of $Y \mid Z^{(j)}$, where $Z^{(j)}$ is the standardized $X^{(j)}$. We require that $Y \perp\!\!\!\perp Z^{(j)} | \beta^{(j)^T} Z^{(j)}$, the marginal conditional independence of the outcome with each $Z^{(j)}$ (or $X^{(j)}$). Then the basic idea of Integrative SIR is to estimate $\beta^{(j)}$, $j = 1, \ldots, s$, and the latent coordinate C_y simultaneously across multi-omics data sources. Once $\beta^{(j)}$ are estimated, by the invariant property, the basis of the central subspace for $Y \mid X^{(j)}$ is obtainable. Under a similar linearity condition as used in conventional SIR (Li 1991), the mathematical form of the Integrative SIR model can be given by

$$\begin{aligned} Z_y^{(1)} &= \beta^{(1)} C_y + \varepsilon^{(1)}, \\ Z_y^{(2)} &= \beta^{(2)} C_y + \varepsilon^{(2)}, \\ &\vdots \\ Z_y^{(s)} &= \beta^{(s)} C_y + \varepsilon^{(s)} \end{aligned} \tag{5}$$

subject to $\beta^{(j)^T} \beta^{(j)} = I_{d^{(j)}}$, $j = 1, \ldots, s$, where $Z_y^{(j)}$, $j = 1, \ldots, s$, represent $Z^{(j)} \mid Y = y$, and $\varepsilon^{(j)}$ are i.i.d. random errors and independent of Y. By accommodating the latent sharing clustering structure across data types, we assume that the structural dimension $d^{(j)} = d$ to be the same for all omics data types, which is satisfied, for example, when every individual data type has a single index model structure for modeling Y. Similar to the rationale in Shen,

Olshen, and Ladanyi (2009), the latent term C_y connects the multiple data sources and reveals dependencies across data types.

Let $(Y_i, X_i^{(1)}, \ldots, X_i^{(s)})$, $i = 1, \ldots, n$, be an i.i.d. random sample, and let $\bar{X}^{(j)}$ be the sample mean and $\hat{\Sigma}^{(j)}$ be an estimate of the covariance matrix of $X^{(j)}$. To estimate the ISIR parameters, we first standardize the predictors for each data type as $\hat{Z}_i^{(j)} = \hat{\Sigma}^{(j)^{-1/2}}(X_i^{(j)} - \bar{X}^{(j)})$, for $i = 1, \ldots, n, j = 1, \ldots, s$. To obtain the estimated covariance matrix $\hat{\Sigma}^{(j)}$, we apply a shrinkage covariance method given by Schäfer and Strimmer (2005). The shrinkage method provides analytic calculation of optimal shrinkage intensity. It can well handle high dimensional genomic data when $p^{(j)} > n$. It guarantees to return a positive definite and well-conditioned covariance matrix estimator.

Next let $\hat{Z}_y^{(j)} = \hat{E}(Z^{(j)}|Y = y)$ be the sample conditional mean (or sample slice mean) for the jth type of data. Then the Integrative SIR parameters can be estimated by minimizing an integrative square loss function:

$$L_d(\beta^{(1)}, \ldots, \beta^{(s)}, C_y) = \sum_{j=1}^{s} \sum_{y=1}^{h} \| \hat{Z}_y^{(j)} - \beta^{(j)} C_y \|^2 \tag{6}$$

subject to $\beta^{(j)^T}\beta^{(j)} = I_d$, where similar to SIR, h is the number of classes for categorical Y, or is the number of slices for continuous Y. Due to the common latent term C_y, the objective function (6) has no closed form solution. We develop a fast iterative algorithm to minimize (6) and to obtain the estimates of $\beta^{(j)}$ and C_y.

Algorithm 2 The Integrative SIR Algorithm

Input: an $n \times (p^{(1)} + \ldots + p^{(s)})$ standardized predictor matrix and n responses.

Output: $p^{(j)} \times d$ estimated basis matrices $\hat{\beta}^{(j)}$ for $\delta_{Y|Z^{(j)}}$ and $p^{(j)} \times d$ estimated basis matrices for $\delta_{Y|X^{(j)}}$, $j = 1, \ldots, s$, through the following steps.

1. Initialize $\hat{\beta}^{(j)}, j = 1, \ldots, s$, with random values.
2. Given $\hat{\beta}^{(j)}, j = 1, \ldots, s$, estimate C_y by minimizing the objective function (6) with

$$\hat{C}_y = \frac{1}{s} \sum_{j=1}^{s} \hat{\beta}^{(j)T} \hat{Z}_y^{(j)}. \tag{7}$$

   Plug $\hat{C}_y$ back to (6) and apply standard singular value decomposition (SVD) to $\sum_{y=1}^{h} \hat{C}_y Z_y^{(j)^T}$ that gives $U^{(j)} D^{(j)} V^{(j)^T}$ for each data type j.

3. Estimate $\beta^{(j)}$ as $\hat{\beta}^{(j)} = V^{(j)} U^{(j)^T}$ for the jth data type.
4. Iterate *Steps 2 and 3* until the objective function (6) converges. Then the final updated $\hat{\beta}^{(j)}$ is an estimated basis for $\delta_{Y|Z^{(j)}}, j = 1, \ldots, s$. Consequently, $\hat{\Sigma}^{(j)^{-1/2}} \hat{\beta}^{(j)}$ gives an estimated basis matrix for $\delta_{Y|X^{(j)}}, j = 1, \ldots, s$.

The algorithm takes random initialization for $\beta^{(j)}$ for each data type. For example, in application, one can choose the initial values of the elements of $\beta^{(j)}$ to be random numbers generated from Uniform (0,1). As shown in Section 3, the proposed algorithm is robust to different choices of initial values. In terms of computational speed, since each step of the iterative algorithm has a closed form solution, the algorithm converges fast to the optimal solution.

To select the structural dimension d, we used leave-one-out cross validation to choose the best d from $d = 0$ to $d = h - 1$ that gives the overall minimum prediction error.

3 Numerical analysis and results

3.1 Simulation studies

In this section, we demonstrate the advantages of sufficient dimension reduction methods in integrative analysis over unsupervised dimension reduction methods and compare integrative SIR with existing methods. We simulated several cases of $n = 50$,100 and 200 samples and in each case simulated predictor vectors from three different data types with $p^{(j)} = 500$, $j = 1, 2, 3$. Each $n \times p^{(j)}$ data type was generated from the inverse regression model with the formula, $X^{(j)} = \beta^{(j)} C_y + \varepsilon^{(j)}, j = 1, \ldots, s$, where the parameters $\beta^{(j)}$, C_y and the covariance matrix

of ε were obtained from the real data estimation as discussed in the following subsection with $d = 2$. For example, the estimated $\beta^{(j)}$, $j = 1, \ldots, s$, from each data type in the real data example were chosen to be the true $\beta^{(j)}$ in the simulation study. The C_y, $y = 1,2,3$, were the coordinates of each outcome having same latent structure for all the data types. The random errors $\varepsilon^{(j)}$, $j = 1, \ldots, s$, were generated from multivariate normal distribution.

We performed both SIR and ISIR estimation and evaluated the estimation accuracy by $\| P_{\beta^{(j)}} - P_{\hat{\beta}^{(j)}} \|_F$, the Frobenious norm of the differences between $P_{\beta^{(j)}}$ and $P_{\hat{\beta}^{(j)}}$, where $P_{\beta^{(j)}}$ and $P_{\hat{\beta}^{(j)}}$ are the projection matrices onto the subspaces spanned by $\beta^{(j)}$ and $\hat{\beta}^{(j)}$, respectively, $j = 1, \ldots, s$, for both methods. Table 1 shows the comparison results on the Frobenious norms for SIR and ISIR. It can be seen that ISIR provides more accurate estimation at different sample sizes.

Table 1: Averaged Frobenious norm for the difference between $P_{\beta^{(j)}}$ and $P_{\hat{\beta}^{(j)}}$ for SIR and Integrative SIR over 50 simulations.

n	p_1, p_2, p_3	Method	Data type 1	Data type 2	Data type 3
50	500	SIR	1.086	1.084	1.081
		ISIR	0.974	0.973	0.971
100	500	SIR	0.867	0.865	0.870
		ISIR	0.773	0.773	0.775
200	500	SIR	0.660	0.659	0.665
		ISIR	0.594	0.595	0.595

We also intended to find classification errors based on the results obtained from SIR and Integrative SIR and compare the prediction performance of the sufficient dimension reduction approaches with unsupervised dimension reduction methods such as PCA and irPCA for comparison. The PCA is the classical principle component analysis, and the irPCA is the integrated and regularized PCA method proposed in Liu, Shen, and Pan (2016). We used leave-one-out cross validation to select tuning parameters and to obtain prediction results for all the methods. Each classification error represents the misclassification rate that was calculated by applying random forest classifier to dimension reduced data in test sets. We then evaluated the prediction performance for each method based on individual data and combined data. For example, for SIR or ISIR, we used the reduction ${\beta^{(j)}}^T X$ of the single data type j, $j = 1, \ldots, 3$, to predict Y separately, and also used the combined data ${\beta^{(1)}}^T X$, ${\beta^{(2)}}^T X$, and ${\beta^{(3)}}^T X$ to predict Y integrally.

Table 2 demonstrates the comparison results among different dimension reduction methods at different sample sizes for both single data prediction and combined data prediction. We see that integrative analysis is beneficial compared to single data source prediction as integrating multiple sources of omics data can potentially reduce classification errors. In addition, the supervised dimension reduction methods improve over unsupervised dimension reduction methods in terms of prediction performance. For example, at sample size 50, SIR reduces the prediction error for the integrative analysis from 0.197 to 0.021 as compared to PCA, and ISIR reduces the error from 0.050 to 0.014 as compared to irPCA. The sufficient dimension reduction methods lead to significant improvement. The results also signal that Integrative SIR is competitive to SIR and can potentially improve over SIR by capturing common information across data types.

Table 2: Averaged classification errors for different dimension reduction methods over 50 simulations.

n	p_1, p_2, p_3	Method	Combined	Data type 1	Data type 2	Data type 3
50	500	PCA	0.197	0.362	0.244	0.270
		irPCA	0.050	0.140	0.132	0.122
		SIR	0.021	0.071	0.068	0.076
		ISIR	0.014	0.092	0.015	0.077
100	500	PCA	0.167	0.363	0.233	0.230
		irPCA	0.020	0.085	0.086	0.087
		SIR	0.019	0.069	0.069	0.062
		ISIR	0.011	0.050	0.040	0.037
200	500	PCA	0.146	0.419	0.238	0.243
		irPCA	0.013	0.072	0.074	0.072
		SIR	0.007	0.057	0.058	0.056
		ISIR	0.005	0.049	0.047	0.051

3.2 Real data analysis

To demonstrate the integration of multiple types of data sets, SIR and Integrative SIR were applied to analyze MRNA, MiRNA and proteomics expression profile of a subset of a melanoma, leukemia and CNS cell lines from the NCI-60 panel. The data is taken from Meng et al. (2016).

The NCI-60 Human Tumor Cell Lines Screen (Shoemaker 2006) has been a part of the global cancer research community for more than 20 years. The NCI-60 subset data set have 21 observations each having MiRNA expressed for 537 genes, MRNA expression for 12895 genes and proteomics expression for 7016 genes with known outcome for each observation. There are three outcome tumor type namely CNS, leukemia and melanoma. As most of the gene expression overlaps and not all the genes have significant information, selecting a handful of genes would be a crucial step.

Gene selection

There are a number of ways to perform a gene selection to get the most informative genes from the entire sample. Most gene selection approaches in class prediction problems combine ranking genes with different test models (Lee et al., 2005; Yeung, Bumgarner & Raftery, 2005). Another approach is applying same classifier progressively on smaller sets of genes until a satisfactory solution is achieved (Van't Veer et al., 2002; Roepman et al., 2005). Frequently an arbitrary decision as to the number of genes to retain is made (for example, keep the 500 top ranked genes) according to different statistical measures like fold-change, variance, etc. (Li, Zhang & Ogihara, 2004). There are also several other methods which can take care of multi-class prediction problems in different samples (Pavlidis, 2003; Chen et al., 2005; Díaz-Uriarte & De Andres, 2006).

In this study, two different gene selection approaches were used to select those genes which are important to be included in this study. The first approach used to select the genes of interest is by calculating p-values of each gene using ANOVA test. Analysis of variance is a statistical method to find the variability in the gene expression partitioned into various sources (Pavlidis 2003). ANOVA examines whether this variability due to a particular factor, or a combination of factors, is statistically significant compared to the measured variability due to multiple sources. Therefore, ANOVA can be used to examine differences between classes.

ANOVA test was carried out to each gene to find out the p-values. As the p-values were calculated, adjusted p-values were computed using false discovery rate (FDR). This step was done because as the gene being statistically tested independently, the risk of false negative increases. To prevent these errors, false discovery rate method can be used to conceptualize the type I error (Benjamini and Hochberg 1995). There are several control procedures such as Benjamini–Hochberg procedure, Benjamini–Hochberg–Yekutieli procedure, and others. Out of all, Benjamini–Hochberg procedure or BH procedure is used in this analysis to control the FDR.

We also used variance as a criterion of choosing the top ranked genes for the whole sample. It means that the more the variances across tumor types, the more likely the genes to be selected. For both criteria, the top 500 ranked genes were selected from each data type for the analysis.

Sufficient dimension reduction

The conventional SIR was firstly applied to the top 500 genes selected by the above two criteria (variance and ANOVA test), respectively. The optimal dimension d was selected to be 2 by cross validation. After performing SIR, the estimated $\beta^{(j)}$s were used to get the reduced predictors from the original data. The data dimension was thus reduced from 21×500 to 21×2 for each data type. The two variables received after dimension reduction thus can be used to visualize how well they classify the outcomes. Figure 1 and Figure 2 show the separation of the three tumor types namely CNS, leukemia and melanoma, by the dimension reduced predictors for each data type.

We also applied Integrative SIR to the analysis of the multi-omics data and to achieve sufficient dimension reduction simultaneously. After performing Integrative SIR with genes prescreened from the aforementioned two criteria, the data were reduced from 21×500 to 21×2 for each data type. Here the optimal dimension d was selected to be 2 by cross validation. Again, the two variables obtained after dimension reduction for each data type can be plotted against each other to visualize how well these variables separate the outcomes. Figure 3 and Figure 4 show the separation results.

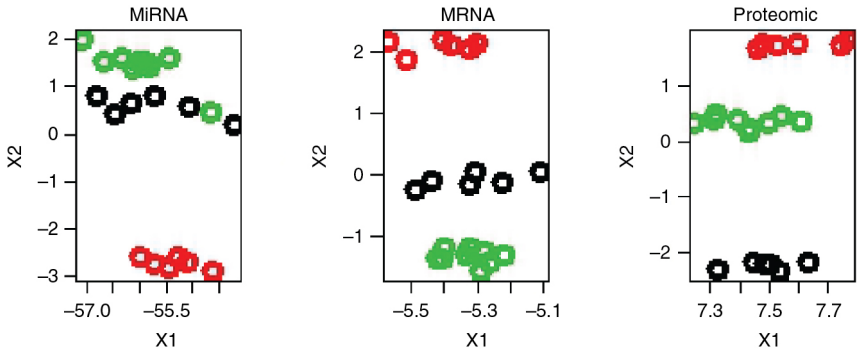

Figure 1: Conventional SIR applied to genes selected by the variance criterion: 2D plots with estimated sufficient predictors obtained by conventional SIR. The red circles represent CNS, the green circles represent Leukemia and the black circles represent Melanoma tumor types.

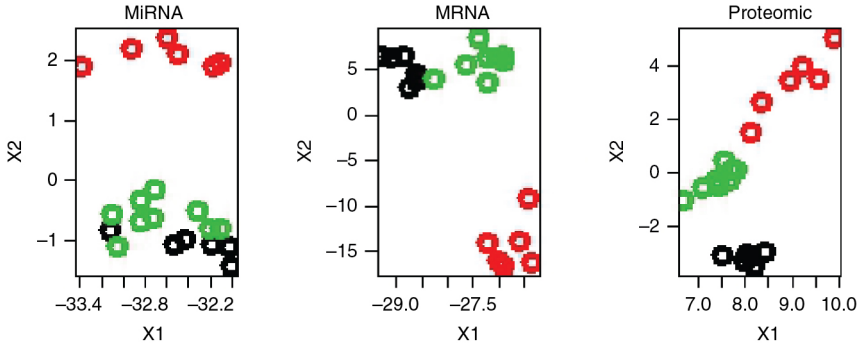

Figure 2: Conventional SIR applied to genes selected using ANOVA test: 2D plots with estimated sufficient predictors obtained by conventional SIR . The red circles represent CNS, the green circles represent Leukemia and the black circles represent Melanoma tumor types.

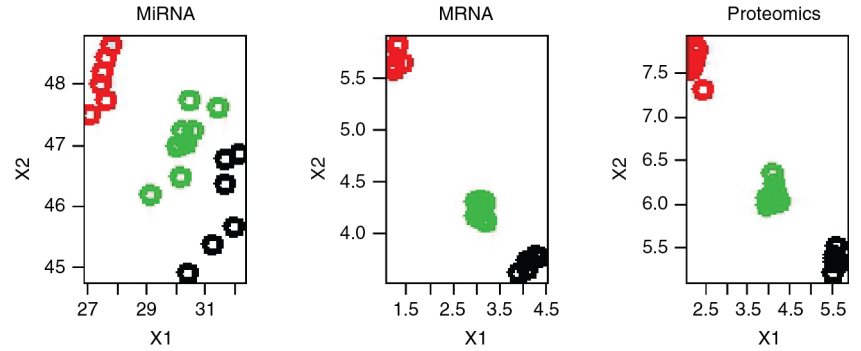

Figure 3: Integrative SIR applied to top 500 genes selected by the variance criterion: 2D plots with estimated sufficient predictors obtained by integrated SIR. The red circles represent CNS, the green circles represent Leukemia and the black circles represent Melanoma tumor types.

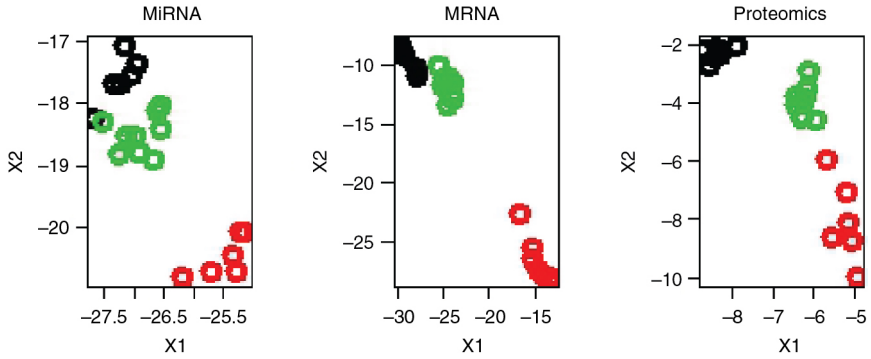

Figure 4: Integrative SIR applied to genes selected using ANOVA test: 2D plots with estimated sufficient predictors obtained by Integrative SIR. The red circles represent CNS, the green circles represent leukemia and the black circles represent melanoma tumor types.

As shown in Figure 1, for MiRNA data type, the cluster of leukemia and melanoma are overlapping with each other, indicating that SIR might not successfully classify the two outcomes. Exploring further at the other two data types, MRNA and proteomics, all the three clusters are reasonably close to each other. In comparison, Figure 3 shows better separation of the different outcomes by the ISIR method.

Similarly, Figure 2 and Figure 4 show how the reduced variables from SIR and ISIR separate outcomes with genes prescreened by ANOVA test. In Figure 2 the clusters for MiRNA and MRNA are overlapping with each other and in Figure 4 the clusters for data type MiRNA is overlapping, while the clusters for other data types are well separated and the points within clusters are compact. These two figures also demonstrate that Integrative SIR gives relatively better separation of the outcomes as compared to SIR.

Classification and prediction

In this section, we evaluate the prediction performance of SIR and Integrative SIR and compare them with unsupervised dimension reduction methods. Similar to the simulation study, to find classification errors for each method, random forest classifier was performed on the dimension reduced data with leave-one-out cross-validation.

Figure 5 and Figure 6 show the three plots of classification errors using random forest method to single data types after applying SIR and Integrative SIR, respectively. Leave-one-out cross-validation was conducted to evaluate the prediction performance as the data had only 21 observations. Each plot in the two figures reveals the classification results for all 21 observations.

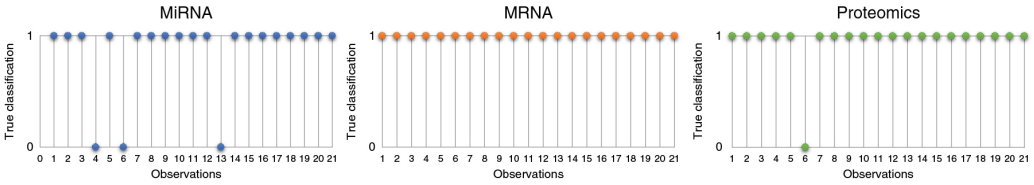

Figure 5: SIR applied to top 500 genes selected by the variance criterion: Classification results of SIR for each data type. 1 represents the correct classification and 0 represents the incorrect classification.

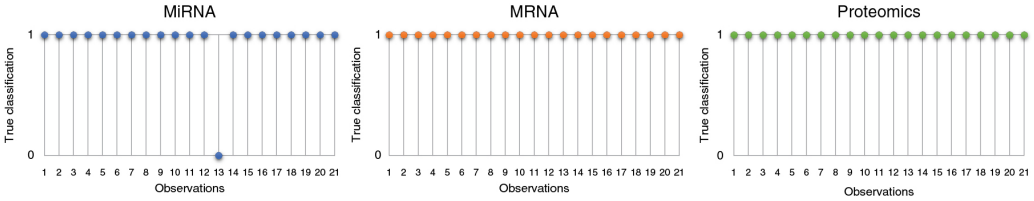

Figure 6: Integrative SIR applied to top 500 genes selected by the variance criterion: Classification results of Integrative SIR for each data type. 1 represents the correct classification and 0 represents the incorrect classification.

Similarly, classification errors were produced when SIR and Integrative SIR were applied to the set of genes selected using ANOVA test. These errors were also generated using random forest classification with leave-one-out cross validation. The results were similarly plotted in Figure 7 and Figure 8. Figure 5–Figure 6 show that ISIR provides more accurate prediction results than SIR does in the real example.

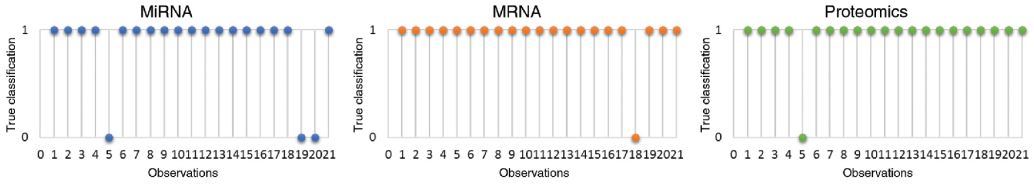

Figure 7: SIR applied to genes selected by ANOVA test: Classification results of SIR for each data type. 1 represents the correct classification and 0 represents the incorrect classification.

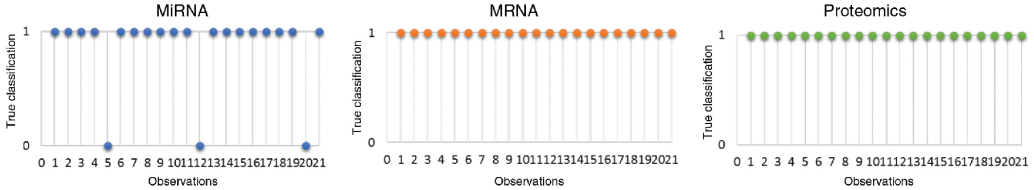

Figure 8: Integrative SIR applied to genes selected by ANOVA test: Classification results of Integrative SIR for each data type. 1 represents the correct classification and 0 represents the incorrect classification.

Table 3 further compares the classification errors for supervised and unsupervised dimension reduction methods including PCA, irPCA, SIR and Integrative SIR. Similar to the simulation studies, we compared the classification errors in terms of single data prediction as well as in terms of integrative prediction. In addition, we considered two cases: In the first case, we used top 500 genes selected by ANOVA test for each data type. In the second case, we kept all 537 genes for the MiRNA data, and chose top 1000 genes for MRNA and proteomics, respectively, by ANOVA test. For both cases, the optimal dimensions selected for SIR and ISIR are 2, the optimal dimensions selected for PCA and irPCA are 3 and 4, respectively, all by leave-one-out cross validation.

Table 3: Classification error for each data type for different methods.

n	p_1, p_2, p_3	Method	Combined	Data type 1	Data type 2	Data type 3
21	500,500,500	PCA	0.095	0.190	0.095	0.190
		irPCA	0.000	0.143	0.048	0.095
		SIR	0.000	0.048	0.000	0.048
		ISIR	0.000	0.000	0.000	0.000
21	537,1000,1000	PCA	0.095	0.238	0.095	0.095
		irPCA	0.095	0.095	0.095	0.095
		SIR	0.000	0.190	0.000	0.095
		ISIR	0.000	0.048	0.095	0.000

Comparing all the dimension reduction methods listed in Table 3, the supervised dimension reduction methods SIR and ISIR outperform the unsupervised PCA type of methods in terms of prediction performance and the improvement is significant. In addition, predicting using combined data overall does better than predicting using single data source, especially for the supervised dimension reduction (SDR) methods. This makes sense as adding more data sources, one can gain additional information that has been identified to be useful by SDR for modeling and predicting the outcome variable. Therefore, integrative analysis of multi-omics data can be beneficial as compared to single data type analysis. The proposed integrative SIR method can potentially further improve over SIR.

3.3 Robustness analysis

In this section, we conduct robustness analysis to show the stability of the propose algorithm and the prediction results.

Using different initial values of βs

In the beginning of Integrative SIR method algorithm, the initial values of βs were randomly chosen from the uniform distribution between 0 and 1. To show that the Integrative SIR method is robust to the initial values, we simply change the random seed number initialized in the start of the program. Table 4 shows the classification error for each method with different seed numbers when top 500 genes were prescreened for each data type by ANOVA test. The results obtained from Integrative SIR with different initial values are identical. We also performed massive numerical work to estimate the central subspaces with many other initial values including those obtained from other dimension reduction methods such as standard SIR, the algorithm performed quite stably. The results similarly hold when more genes were prescreened. Though the numerical work shows robust performance, in application, to avoid possible local minimizers, one might use preliminary estimates from SIR, PCA or factor analysis as initial values to potentially improve convergence.

Table 4: Classification error (in %) for each data type for different methods with different initial values.

Seed number	Methods	Data type 1	Data type 2	Data type 3′
1	SIR	0.048	0.000	0.048
	ISIR	0.000	0.000	0.000
34	SIR	0.048	0.000	0.048

	ISIR	0.000	0.000	0.000
1234	SIR	0.048	0.000	0.048
	ISIR	0.000	0.000	0.000

Using different classifiers

Different classifier may give different results. For the simulation and real data analysis, we have utilized the random forest classifier to predict the different classes of the outcome variable. We next used another popular classifier called support vector machine (SVM) to evaluate the methods. Support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression.

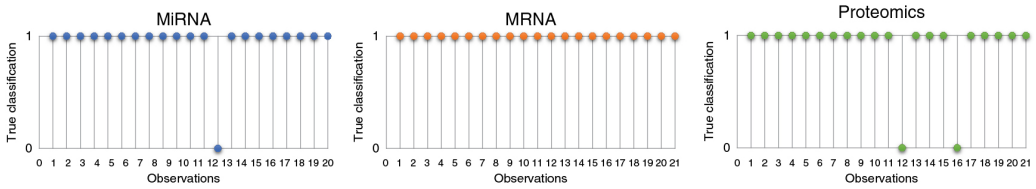

Figure 9: Support vector machine classification with conventional SIR, 1 represents the correct classification and 0 represents the incorrect classification.

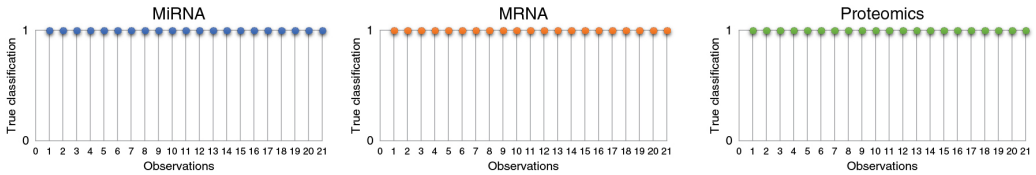

Figure 10: Support vector machine classification with Integrative SIR, 1 represents the correct classification and 0 represents the incorrect classification.

SVM classifier was applied with leave-one-out cross validation. Figure 9 and Figure 10 show the classification error plots for SIR and Integrative SIR. Similar to the results in the previous section, Integrative SIR showed competitive classification performance as compared to SIR.

4 Conclusion

We propose a novel use of sufficient dimension reduction methods in integrative multi-omics data analysis and demonstrate the improvement over unsupervised dimension reduction methods. In particular, we introduced sliced inverse regression (SIR) and proposed an Integrative SIR method for multi-omics data analysis. The proposed method reduces the dimensions of multiple omics data simultaneously while taking into account latent sharing information across data types without loss of information in regression and prediction. By capturing the relationship between response and predictors while performing dimension reduction, the sufficient dimension reduction methods possess advantages in terms of classification and prediction as compared to PCA types of methods. In addition, by considering common information across data sources, the Integrative SIR method can potentially improve over SIR. The performance of the proposed method is robust to the choice of initial values in the algorithm and to the classifiers used to evaluate the classification errors.

As genomic data have become increasingly affordable, it is important to study and perform integrative analysis of multi-omics data for comprehensive understanding of underlying data structures. Future directions can be thought of extending integrative settings to sparse SDR methods (Qian, Ding & Cook, 2018) to explore new biological insights, prediction and accuracy of the model. The proposed framework can also be extended to likelihood-based dimension reduction methods such as principal fitted components analysis (Cook & Forzani, 2008; Ding & Cook, 2014). Moreover, by simultaneous sufficient dimension reduction and statistical modeling, applying envelope models (Cook, Li & Chiaromonte, 2010; Su & Cook, 2011; Cook, Helland & Su, 2013; Cook & Zhang, 2015; Su et al., 2016; Ding & Cook, 2018) to integrative data analysis might potentially gain further efficiency in comparison to classical SDR methods. Furthermore, integrative SDR can be extended to matrix or tensor frameworks (Li, Kim & Altman, 2010; Ding & Cook 2015a; 2015b) with temporal information incorporated. These directions are worthy of future investigation.

Funding: This work was supported by the DE-CTR ACCEL, Grant Number: 18A00364.

References

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Stat. Soc. Series B Stat. Methodol., 57, 289–300.

Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno and M. Gillette (2001): "Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses," Proc. Natl. Acad. Sci. U. S. A., 98, 13790–13795.

Bichsel, V. E., J. M. Worth, V. V. Prabhu, J. S. Gutkind, L. A. Liotta, P. J. Munson III, E. F. Petricoin and D. B. Krizman (2001): "Proteomic profiling of the cancer microenvironment by antibody arrays," Proteomics, 1, 1271–1278.

Chen, D., Z. Liu, X. Ma and D. Hua (2005): "Selecting genes by test statistics," Biomed Res. Int., 2005, 132–138.

Chiaromonte, F., R. D. Cook and B. Li (2002): "Sufficient dimension reduction in regressions with categorical predictors," Ann. Stat., 30, 475–497.

Cook, R. D. (1994): "On the interpretation of regression plots," J. Am. Stat. Assoc., 89, 177–189.

Cook, R. D. (1996): "Graphics for regressions with a binary response," J. Am. Stat. Assoc., 91, 983–992.

Cook, R. D. (1998a): "Principal hessian directions revisited," J. Am. Stat. Assoc., 93, 84–94.

Cook, R. D. (1998b): *Regression Graphics*. New York, NY: John Wiley & Sons.

Cook, R. D. (2004): "Testing predictor contributions in sufficient dimension reduction," Ann. Stat., 32, 1062–1092.

Cook, R. D. and L. Forzani (2008): "Principal fitted components for dimension reduction in regression," Stat. Sci., 23, 485–501.

Cook, R. D. and S. Weisberg (1991): "Discussion of 'sliced inverse regression for dimension reduction"'. J. Am. Stat. Assoc., 86, 328–332.

Cook, R. D. and X. Zhang (2015): "Foundations for envelope models and methods," J. Am. Stat. Assoc., 110, 599–611.

Cook, R. D., B. Li and F. Chiaromonte (2010): "Envelope models for parsimonious and efficient multivariate linear regression," Stat. Sin., 20, 927–960.

Cook, R., I. Helland and Z. Su (2013): "Envelopes and partial least squares regression," J. R. Stat. Soc. Series B Stat. Methodol., 75, 851–877.

Cox, D. R. (1968): "Notes on some aspects of regression analysis," J. R. Stat. Soc. Ser. A Stat. Soc., 131, 265–279.

De Cubas, A. A., L. J. Leandro-García, F. Schiavi, V. Mancikova, I. Comino-Méndez, L. Inglada-Perez, M. Perez-Martinez, N. Ibarz, P. Ximénez-Embún and E. López-Jiménez (2013): "Integrative analysis of mirna and mrna expression profiles in pheochromocytoma and paraganglioma identifies genotype-specific markers and potentially regulated pathways," Endocr. Relat. Cancer, 20, 477–493.

Díaz-Uriarte, R. and S. A. De Andres (2006): "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, 7, 3.

Ding, S. and R. D. Cook (2014): "Dimension folding PCA and PFC for matrix-valued predictors," Stat. Sin., 24, 463–492.

Ding, S. and R. D. Cook (2015a): "Higher-order sliced inverse regressions," Wiley Interdiscip Rev. Comput. Stat., 7, 249–257.

Ding, S. and R. D. Cook (2015b): "Tensor sliced inverse regression," J. Multivar. Anal., 133, 216–231.

Ding, S. and R. D. Cook (2018): "Matrix variate regressions and envelope models," J. R. Stat. Soc. Series B Stat. Methodol., 80, 387–408.

Gholami, A. M., H. Hahne, Z. Wu, F. J. Auer, C. Meng, M. Wilhelm and B. Kuster (2013): "Global proteome analysis of the nci-60 cell line panel," Cell Rep., 4, 609–620.

Iliopoulos, D., K. N. Malizos, P. Oikonomou and A. Tsezou (2008): "Integrative microrna and proteomic approaches identify novel osteoarthritis genes and their collaborative metabolic and inflammatory networks," PloS One, 3, e3740.

Jain, Y. and S. Ding (2017): "Integrative sufficient dimension reduction methods for multi-omics data analysis." In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, p. 616. ACM.

James, G., D. Witten, T. Hastie and R. Tibshirani (2013): *An introduction to statistical learning*. New York: Springer.

Kristensen, V. N., O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi and A.-L. Børresen-Dale (2014): "Principles and methods of integrative genomic analyses in cancer," Nat. Rev. Cancer, 14, 299.

Lê Cao, K.-A., D. Rossouw, C. Robert-Granié and P. Besse (2008): "A sparse pls for variable selection when integrating omics data," Stat. Appl. Genet. Mol. Biol., 7, Article 35.

Lê Cao, K.-A., I. González and S. Déjean (2009): "Integromics: an r package to unravel relationships between two omics datasets," Bioinformatics, 25, 2855–2856.

Lee, J. W., J. B. Lee, M. Park and S. H. Song (2005): "An extensive comparison of recent classification tools applied to microarray data," Comput. Stat. Data Anal., 48, 869–885.

Li, K.-C. (1991): "Sliced inverse regression for dimension reduction," J. Am. Stat. Assoc., 86, 316–327.

Li, B. and S. Wang (2007): "On directional regression for dimension reduction," J. Am. Stat. Assoc., 102, 997–1008.

Li, T., C. Zhang and M. Ogihara (2004): "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," Bioinformatics, 20, 2429–2437.

Li, L., R. D. Cook and C.-L. Tsai (2007): "Partial inverse regression," Biometrika, 94, 615–625.

Li, B., M. K. Kim and N. Altman (2010): "On dimension folding of matrix-or array-valued statistical objects," Ann. Stat., 38, 1094–1121.

Li, W., S. Zhang, C.-C. Liu and X. J. Zhou (2012): "Identifying multi-layer gene regulatory modules from multi-dimensional genomic data," Bioinformatics, 28, 2458–2466.

Liu, H., P. D'Andrade, S. Fulmer-Smentek, P. Lorenzi, K. W. Kohn, J. N. Weinstein, Y. Pommier and W. C. Reinhold (2010): "mrna and microrna expression profiles of the nci-60 integrated with drug activities," Mol. Cancer Ther., 9, 1080–1091.

Liu, B., X. Shen and W. Pan (2016): "Integrative and regularized principal component analysis of multiple sources of data," Stat. Med., 35, 2235–2250.

Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub (2005): "Microrna expression profiles classify human cancers," Nature, 435, 834–838.
Meng, C., O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami and A. C. Culhane (2016): "Dimension reduction techniques for the integrative analysis of multi-omics data," Brief. Bioinform., 17, 628–641.
Nibbe, R. K., M. Koyutürk and M. R. Chance (2010): "An integrative-omics approach to identify functional sub-networks in human colorectal cancer," PLoS Comput. Biol., 6, e1000639.
Nie, L., G. Wu, D. E. Culley, J. C. Scholten and W. Zhang (2007): "Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications," Crit. Rev. Biotechnol., 27, 63–75.
Nishizuka, S., L. Charboneau, L. Young, S. Major, W. C. Reinhold, M. Waltham, H. Kouros-Mehr, K. J. Bussey, J. K. Lee and V. Espina (2003): "Proteomic profiling of the nci-60 cancer cell lines using new high-density reverse-phase lysate microarrays," Proc. Natl. Acad. Sci. U.S.A., 100, 14229–14234.
Oliveto, S., M. Mancino, N. Manfrini and S. Biffo (2017): "Role of micrornas in translation regulation and cancer," World J. Biol. Chem., 8, 45.
Pavlidis, P. (2003): "Using anova for gene selection from microarray studies of the nervous system," Methods, 31, 282–289.
Peng, Y. and C. M. Croce (2016): "The role of micrornas in human cancer," Signal. Transduct. Target. Ther., 1, 15004.
Qian, W., S. Ding and R. D. Cook (2018): "Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension," J. Am. Stat. Assoc., 1–48.
Reis-Filho, J. S. and L. Pusztai (2011): "Gene expression profiling in breast cancer: classification, prognostication, and prediction," Lancet., 378, 1812–1823.
Rhodes, D. R. and A. M. Chinnaiyan (2005): "Integrative analysis of the cancer transcriptome," Nat. Genet., 37, S31–S37.
Roepman, P., L. F. Wessels, N. Kettelarij, P. Kemmeren, A. J. Miles, P. Lijnzaad, M. G. Tilanus, R. Koole, G.-J. Hordijk and P. C. van der Vliet (2005): "An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas," Nat. Genet., 37, 182.
Schäfer, J. and K. Strimmer (2005): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," Stat. Appl. Genet. Mol. Biol., 4.
Shen, R., A. B. Olshen and M. Ladanyi (2009): "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," Bioinformatics, 25, 2906–2912.
Shen, R., Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi and C. Sander (2012): "Integrative subtype discovery in glioblastoma using icluster," PloS One, 7, e35236.
Shen, R., S. Wang and Q. Mo (2013): "Sparse integrative clustering of multiple omics data sets," Ann. Appl. Stat., 7, 269.
Shoemaker, R. H. (2006): "The nci60 human tumour cell line anticancer drug screen," Nat. Rev. Cancer, 6, 813–823.
Su, Z. and R. D. Cook (2011): "Partial envelopes for efficient estimation in multivariate linear regression," Biometrika, 98, 133–146.
Su, Z., G. Zhu, X. Chen and Y. Yang (2016): "Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression," Biometrika, 103, 579–593.
Van't Veer, L. J., H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton and A. T. Witteveen (2002): "Gene expression profiling predicts clinical outcome of breast cancer," Nature, 415, 530–536.
Wei, X., J. Li, H. Xie, Q. Ling, J. Wang, D. Lu, L. Zhou, X. Xu, S. Zheng (2014): "Proteomics-based identification of the tumor suppressor role of aminoacylase 1 in hepatocellular carcinoma," Cancer Lett., 351, 117–125.
Xu, T., T. D. Le, L. Liu, R. Wang, B. Sun and J. Li (2016): "Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data," PloS One, 11, e0152792.
Yeung, K. Y., R. E. Bumgarner and A. E. Raftery (2005): "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," Bioinformatics, 21, 2394–2402.