Integrative analysis of multi-omics data improves model predictions: an application to lung cancer

Erica Ponzi[1], Magne Thoresen[1], Therese Haugdahl Nøst[2], Kajsa Møllersen[2]

1. Oslo Center for Biostatistics and Epidemiology, UiO, University of Oslo, Norway
2. Department of Community Medicine, UiT, The Arctic University of Norway, Tromsø, Norway

Abstract

Background:
Cancer genomic studies often include data collected from several omics platforms. Each omics data source contributes to the understanding of the underlying biological process via source specific ("individual") patterns of variability. At the same time, statistical associations and potential interactions among the different data sources can reveal signals from common biological processes that might not be identified by single source analyses. These common patterns of variability are referred to as "shared" or "joint". To capture both contributions of variance, integrative dimension reduction techniques are needed. Integrated PCA is a model based generalization of principal components analysis that separates shared and source specific variance by iteratively estimating covariance structures from a matrix normal distribution. Angle based JIVE is a matrix factorization method that decomposes joint and individual variation by permutation of row subspaces. We apply these techniques to identify joint and individual contributions of DNA methylation, miRNA and mRNA expression collected from blood samples in a lung cancer case control study nested within the Norwegian Woman and Cancer (NOWAC) cohort study.
Results:
In this work, we show how an integrative analysis that preserves both

components of variation is more appropriate than analyses considering uniquely individual or joint components. Our results show how both joint and individual components contribute to a better quality of model predictions, and facilitate the interpretation of the underlying biological processes.
Conclusion:
When compared to a non integrative analysis of the three omics sources, integrative models that simultaneously include joint and individual components result in better prediction of cancer status and metastatic cancer at diagnosis.

Keywords: data integration, dimension reduction, joint and individual variance explained, multi-omics, prediction models

1 Background

Cancer studies benefit from the availability of genomic data, also known as omics. The dimensionality of omics data is extremely high, suggesting the application of dimension reduction techniques. Additionally, omics are available across multiple sources (or 'blocks') of data, collected on the same organisms or tissues, and measured on different platforms. A comprehensive understanding of the key underlying biological process relies on an integrative approach able to combine the information arising from such multi-source data. To this end, a large number of statistical methods for the simultaneous analysis of multi-omics data have recently been proposed. Multiple reviews of such methods are available, for example in Tseng et al (2015); Huang et al (2017); Rappaport and Ron (2018).

Data integration techniques are often used to identify 'joint' (also referred to as 'common' or 'shared') contributions of the data sources to the observed variation, and their simultaneous effect on the biological process under study. Such patterns of variation arise from the interaction among different omics sources, and may not be detected by a separate analysis of each single source. However, the different data sources do not only contain the joint information, but also independent contributions. The separate analysis of each data source has so far been the most common approach used in the omics context, and knowledge about the individual contributions of each omics source is relevant to the understanding of the biological processes of interest. As a consequence, considering only the joint patterns might also prove insufficient,

as it overlooks the heterogeneity among single data sources, and their individual signals from the underlying relevant biological process. An example of this can be seen in genomic studies collecting DNA methylation and gene expression data. It is known that methylation regulates gene expression and that this can cause a non-negligible joint structure across the different data sources. On the other hand, methylation and gene expression correlate to the clinical outcome also through signals that are specific to each omics data source and biologically relevant independently from each other. Therefore, dimension reduction methods that take both joint and individual patterns into account are necessary.

Two of the most common approaches to dimension reduction are Principal Component Analysis (PCA) and matrix factorization, and these have recently been expanded to the case of multi source data. For example, consensus PCA (Westerhuis et al, 1998) consists of PCA on the normalized concatenated data, and distributed PCA (Fan et al, 2019) performs local PCA on the individual data sources and then uses these principal components to estimate a global covariance structure. Other dimension reduction methods have been extended to the case of multi source data, as for example canonical correlation analysis (CCA) (Hotelling, 1936) or partial least squares (PLS) analysis, which has been further generalized to O2PLS (Trygg and Wold, 2003). A similar method that allows for the presence of multiple data sources is the multiple CCA (Witten and Tibshirani, 2009), but it mainly focuses on the common variation among the components, and seems to neglect the individual contributions of the data sources.

Although the extensions of PCA mentioned above account for multiple data sources, they focus either on their joint or individual contributions but do not identify both of them simultaneously. Integrated PCA (iPCA) is a model based generalization of PCA that decomposes variance into shared and source specific variation (Tang and Allen, 2018). It is based on the assumption of a matrix variate normal distribution of the data, whose covariance structure is given as the Kronecker product of a shared and an individual, data block specific, covariance matrix. Integrated PCA estimates these two matrices via an iterative algorithm, and consequently extracts principal components for both.

Solutions based on matrix factorization also preserve both joint and individual structures in the data. In this framework, each data block is decomposed into three matrices modeling different types of variation, specifically joint variation across the blocks, individual variation for each data block, and

residual noise. JIVE stands for Joint and Individual Variance Explained, it was formulated by Lock et al (2013) and, also thanks to its implementation available in R (O'Connell and Lock, 2016), has been used in various medical applications, including clustering of cancer genomics data (Hellton and Thoresen, 2016), multisource omics data (Kuligowski et al, 2015; Kaplan and Lock, 2017) and imaging and behavioral data (Yu et al, 2017). Although JIVE solves the issue of maintaining joint and individual structures, it uses an iterative algorithm and is computationally very intensive. In Feng et al (2018), Angle Based JIVE (aJIVE) was formulated to improve this aspect. It computes the matrix decomposition by using perturbations of the row spaces to identify the joint and individual variation, and results in a much faster implementation than the original JIVE. Other similar approaches to identify both kinds of variation have been proposed, as for example DISCO (Schouteden et al, 2013) and OnPLS (Lofsted et al, 2012). An illustration of these methods and a comparison with JIVE was provided in Måge et al (2019).

In this work, we demonstrate how an integrative analysis of the data is more appropriate than a uniquely source-specific analysis, which would not reveal relations between the sources of biological information. Integrative analyses are also preferable to uniquely joint analysis, which can be redundant and use the available information to repeatedly measure the same shared components. We show how both joint and individual components contribute to a better quality of model predictions. The combination of joint and individual components can also facilitate the biological interpretation of the underlying process, although this might still fail as the dimension reduction itself bears the risk of obscuring some relevant information.

We use both a matrix factorization and a PCA based method to identify individual and joint components in a real data set on lung cancer. We chose to use aJIVE among the matrix factorization methods, because it inherits a good subspace recovery in comparison to other methods (Måge et al, 2019), as well as the robustness to model misspecification from JIVE, but it also solves the issue of correlated individual subspaces (Feng et al, 2018), and provides a much faster implementation. Furthermore, McCabe et al (2020) show that aJIVE performs best in terms of consistency and lack of overfitting when compared to other integrative methods. We validate the aJIVE results with a PCA based method, which has the advantage of being independent from the initial ranks selection. We use iPCA because it is the only method in this framework that can identify both individual and joint patterns, and because it has been shown to perform well in prediction models (Tang and

Allen, 2018).

The data we use stem from a lung cancer case control study nested within the Norwegian Woman and Cancer (NOWAC) cohort study (Lund et al, 2008). The associations among three levels of omics data analyzed in blood samples, specifically DNA methylation, mRNA and miRNA expression, are investigated and their joint and individual contributions are used to predict future cancer cases, and the characterization of future cancers as metastatic or non-metastatic at diagnosis. We show that both kinds of components contain information that reveal properties about biological processes and that using joint components improves model predictions. Specifically, we show that joint components increase the AUC of prediction models, when compared to models with solely individual components, to models uniquely based on clinical, patient-level covariates, and most importantly to non-integrative models, i. e. based on independent analyses of data from each source.

2 Methods

2.1 Data integration setup

Throughout the manuscript, we will denote each data block with $\boldsymbol{X}_k$, where $k = 1, ..., K$ and K is the number of data sources used in the study. Each block is a matrix with n columns, where n is the number of study subjects. The kth matrix $\boldsymbol{X}_k$ has p_k rows, corresponding to the variables in data source k. The overall dimensionality is denoted as $p = p_1 + ... + p_K$. The low-rank decomposition we want to obtain from the different methods is:

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{J}_1 + \boldsymbol{I}_1 + \boldsymbol{\epsilon}_1 \\ &\vdots \\ \boldsymbol{X}_K &= \boldsymbol{J}_K + \boldsymbol{I}_K + \boldsymbol{\epsilon}_K \end{aligned} \tag{1}$$

where $\boldsymbol{I}_k$ is the individual component for data block k, $\boldsymbol{\epsilon}_k$ is its residual component and

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 \\ ... \\ ... \\ \boldsymbol{J}_K \end{bmatrix} \tag{2}$$

is the joint structure matrix, where each $\boldsymbol{J}_k$ is the submatrix of the joint structure $\boldsymbol{J}$ associated with $\boldsymbol{X}_k$. The methods we use in our application obtain such decomposition with very different algorithms, but both focus on the distinction between the $\boldsymbol{J}$ and $\boldsymbol{I}_k$ components, and on the calculation of their loadings and scores.

2.2 Angle Based JIVE

Angle based Joint and Individual Variation Explained (aJIVE) is a variant of the JIVE method, based on perturbation of row subspaces. JIVE aims to minimize the squared residual components ϵ_1, ..., ϵ_K, using an iterative algorithm that alternatively estimates the joint and individual components by singular value decomposition (SVD). AJIVE builds on this method but constructs the algorithm in a more efficient and computationally feasible way. Besides resulting in a faster implementation of the algorithm, aJIVE provides a more intuitive interpretation of the decomposition, especially in the case of high correlations among individual components (Feng et al, 2018). The aJIVE algorithm is structured in three phases: First the low-rank approximation of each data block $\boldsymbol{X}_k$ is obtained by SVD. Secondly, the joint structure between the obtained low-rank approximations is extracted computing the SVD of the stacked row basis matrices, by using basic principles of Principal Angle Analysis. Finally, the joint components $\boldsymbol{J}_k$ are obtained by projection of each data block onto the joint basis, while the individual components $\boldsymbol{I}_k$ are calculated by orthonormal basis subtraction.

The first step is based on the choice of the initial ranks, which are used as a threshold value in the first SVD decomposition of the data blocks. This choice is rather subjective and involves taking into account some bias variance trade-off in the joint signals representation. Although Feng et al (2018) provide guidelines on how to determine the initial ranks, the recommended choice is based on the observation of scree plots, which remains highly subjective. As an alternative, Zhu and Ghodsi (2006) present a choice of initial ranks based on the profile likelihood of the single data blocks.

From the aJIVE decomposition, it is possible to obtain the full matrix representation of the original features, as well as the block specific decompositions of each data source and the common normalized scores. The algorithm requires much lower computation time than the original JIVE, and its implementation is available in `Matlab` (Jiang, 2018) and `R` (Carmichael, 2019).

2.3 Integrated PCA

Integrated Principal Component Analysis (iPCA) is a generalization of principal component analysis to multiple data sources. The core idea of iPCA is the assumption that each data source $\boldsymbol{X}_k$ follows a matrix variate normal distribution:

$$\boldsymbol{X}_k \sim \mathsf{N}_{(n,p)}\left(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}_k\right), \tag{3}$$

where $\boldsymbol{M}$ is the mean matrix, $\boldsymbol{\Delta}_k$ is the column covariance specific to $\boldsymbol{X}_k$ and $\boldsymbol{\Sigma}$ is the row covariance shared by all data matrices $\boldsymbol{X}_k$. The assumption of normality is very common for methylation data, and for gene expression after applying a log transformation. Moreover, Tang and Allen (2018) show that the iPCA method is robust to deviations from normality.

Integrated PCA performes PCA on the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}_k$ and denotes with $\hat{\mathbf{U}}$ the eigenvectors of $\boldsymbol{\Sigma}$, which represent the joint components $\boldsymbol{J}$. The eigenvectors of $\boldsymbol{\Delta}_k$ are denoted as $\hat{\mathbf{V}}_k$ and represent the individual components $\boldsymbol{I}_k$ specific to data set k. To obtain these quantities, iPCA uses the "Flip-Flop algorithm" to calculate maximum likelihood estimators (MLEs) of the covariance matrices. Because of the complexity of the model and of the non-existence of MLEs for a large number of cases, iPCA uses a penalized maximum likelihood, with a penalty term defined as:

$$P_q(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Delta}_1^{-1}, ..., \boldsymbol{\Delta}_k^{-1}) = \lambda_\Sigma \|\boldsymbol{\Sigma}^{-1}\|_q + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\Delta}_k^{-1}\|_q$$

with norm q. Different possibilities for q are illustrated in Tang and Allen (2018): L_1 or multiplicative Frobenius or additive Frobenius, and suggest using the multiplicative Frobenius to ensure convergence. The choice of the tuning parameters λ_Σ and λ_1, ..., λ_k is based on a missing data imputation framework. This step of the algorithm is computationally very intensive and requires the input of initial values, which need to be chosen carefully by the user.

Integrated PCA has been implemented in `R` and its code is available at Tang (2018). Note that iPCA, unlike aJIVE, does not require the specification of the initial ranks, although it requires initial values for the tuning parameters. While aJIVE is built on an optimization procedure, focused on minimizing the approximation error, iPCA is a model-based approach for the estimation of the underlying subspace. This ensures the convergence to

a global solution for iPCA, and a better recovery of the subspace, but on the other hand it results in less robustness to model misspecification, which does not affect aJIVE, and less robustness to errors. Both aJIVE and iPCA do not rely on any independence assumptions, and correlations among individual components do not represent an issue in any of the two methods. An important difference between the two methods, as we highlighted above, is the computation time of the two algorithms, which is significantly lower in aJIVE. The interpretability of the results is comparable in aJIVE and iPCA.

2.4 Application to the NOWAC data

The dataset

The dataset used in the following analyses stems from blood samples in a lung cancer case control study nested within the Norwegian Women and Cancer Study (NOWAC) (Lund et al, 2008). All participating subjects are women who did not have a cancer diagnosis at recruitment (1991–2006) or at time of blood sampling (2003–2006). The time from blood sampling to cancer diagnosis ranges from 0.3 to 7.9 years, with a median time equal to 4.2 years. For each case, one control was matched on time since blood sampling and birth year. All participants gave written informed consent and the study was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Inspectorate. Three levels of omics data are available for $n = 230$ individuals, with numbers of variables respectively equal to $p_1 = 485512$ CpG methylation, $p_2 = 11610$ mRNA expression and $p_3 = 199$ miRNA expression. Information about individual covariates, including age, body mass index (BMI), dietary and smoking habits was also collected for all participants. Outcomes of interest are the classification of case versus control, as well as the characterization of cancers as metastatic or non-metastatic at diagnosis.

Both methylation and gene expression have been shown to associate with the occurrence and characteristics of lung cancer (Yanaihara et al, 2006; Hu and Chen, 2015; Zhang et al, 2016; Baglietto et al, 2017) but it is also known that the different data sources might contribute together and jointly relate to these biological outcomes (Heller et al, 2012; Sandanger et al, 2018). We apply aJIVE and iPCA to estimate such joint and individual contributions, and use these components in prediction models for the occurrence of lung

cancer and classification of tumor types.

Filtering and preprocessing

Laboratory processing and microarray analyses for mRNA expression and DNA methylation are described in Sandanger et al (2018). For miRNA, laboratory processing included miRNA isolation and purification from 100 μl plasma using the Qiagen miRNeasy Serum/Plasma Kit. Small RNA sequencing libraries were prepared using the NEXTflex small RNA-seq kit v3 (Bioo Scientific, Austin, TX, USA) and sequencing of fragments was performed using a Illumina HiSeq4000 flowcell, according to the manufacturer's instructions (Illumina, Inc., San Diego, CA, USA), at 50 bp SE, resulting in approximately $7 - 9$M reads per sample.

The filtering of miRNA expressions was based on the counts per million, that is the total read counts of a miRNA divided by the total read counts of the sample and multiplied by 10^6, and signals having less than one count per million were excluded. Additionally, signals with null reads on more than 5 patients were excluded.

Because of the high computational requirements, we reduced the number of mRNA expressions to $p_2 = 5000$, by selecting the variables with higher variance. We then reduced the number of CpGs methylation sites by selecting the CpGs located on the same genes as the filtered mRNAs. Among these, we excluded CpGs with more than 40% missing data, as well as CpGs with extreme M-values ($|M| > 3$, see Zhang et al (2012); Ma et al (2014)). This resulted in $p_1 = 18545$. All $p_3 = 199$ available miRNAs were analysed. We used log2 transformed expressions for both mRNA and miRNA, and M-values for methylation (Du et al, 2010). We accounted for missing values in the data by using `SVDmiss`, as suggested in Lock et al (2013), and centered the data for easier interpretation and comparison of the two methods.

aJIVE and iPCA

We performed aJIVE on the three levels of omics data. The initial ranks were selected by maximizing the profile likelihood (Zhu and Ghodsi, 2006), and set respectively to 43, 11 and 9. Different choices of initial ranks were also explored. Additionally, we ran aJIVE on pairs of omics sources, with the same ranks as above for each two way association.

We performed iPCA on a smaller subset of variables, due to the high

computational requirements. We selected variables with the same procedure as reported above, but starting with $p_2 = 500$ mRNAs, and resulting in $p_1 = 1588$ CpGs. The $p_3 = 199$ miRNAs were all included. We also repeated aJIVE on this smaller subset. We set the initial penalty terms to 1, 0.01 and 0.001, as suggested in Tang and Allen (2018), and used the multiplicative Frobenius norm.

Using joint components for prediction

Joint and individual components were used in prediction models. The outcome of interests were the occurrence of lung cancer (yes/no) and metastasis (yes/no).

We fitted logistic models on each outcome using joint and individual components as explanatory variables, in addition to age, BMI and smoking. These models were compared in terms of AUC with the respective models with only age, BMI and smoking as covariates. To assess the performance of the models, we measured the average AUC in a 10 fold cross validation. To identify the amount of information added by the joint components, we compared these models to models fitted on the individual components from aJIVE and the clinical covariates as predictors. Finally, we compared these results with a non-integrative analysis, obtained by performing PCA separately on each single data source. We fitted a model on the first principal components (PCs) of each data source, and on the same clinical covariates. We chose to include five PCs for each data source, to have a comparison with the number of joint and individual components, and based on the variance explained by the first PCs and on the analysis of screeplots. Models were repeated with a higher number of PCs and results did not change substantially.

In addition, we used a random forest of 500 trees to predict case vs control on the basis of joint and individual components, and patient covariates as above (age, BMI, smoking). We extracted the AUCs and the out of the bag (OOB) classification errors from the random forest, and we ranked all the variables in importance on the basis of their mean decrease in gini index (Jiang et al, 2009).

3 Results

We will focus on aJIVE as this is where we have the lowest computational requirements and can include a higher number of variables from each data source. A comparison with iPCA will be given in Section 3.3.

3.1 aJIVE

Using initial ranks obtained with the profile likelihood method resulted in a joint rank equal to 5, and individual ranks respectively equal to 38, 7 and 7. The decomposition is illustrated in the heatmap in Figure 1 and shows that the joint component dominates the individual and residual components for all three data sources. Figure 2 reports the proportions of variance explained that are due to the joint, individual and residual components and also shows that the joint components explain most of the total variance in the data.

Repeating the algorithm with lower initial ranks resulted in lower joint rank estimates, while higher ranks gave estimates of joint ranks that were stable and equal to 5, suggesting that overestimating the initial ranks should not affect the results, while an underestimation of the ranks can fail to detect some joint components in the data. The pairwise analysis starting with the same initial ranks resulted in different joint ranks, suggesting that most joint components are shared between methylation and mRNA. The aJIVE joint and individual ranks for pairwise analyses are reported in Table 1.

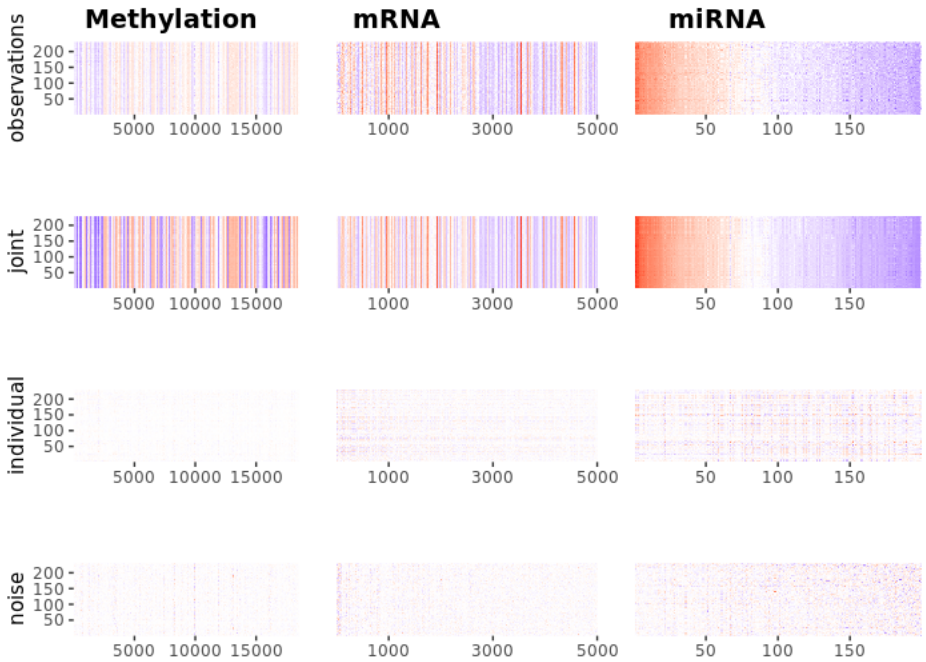


Figure 1: aJIVE decomposition of the three omics sources. The joint component dominates the individual and residual components for all three datasets.

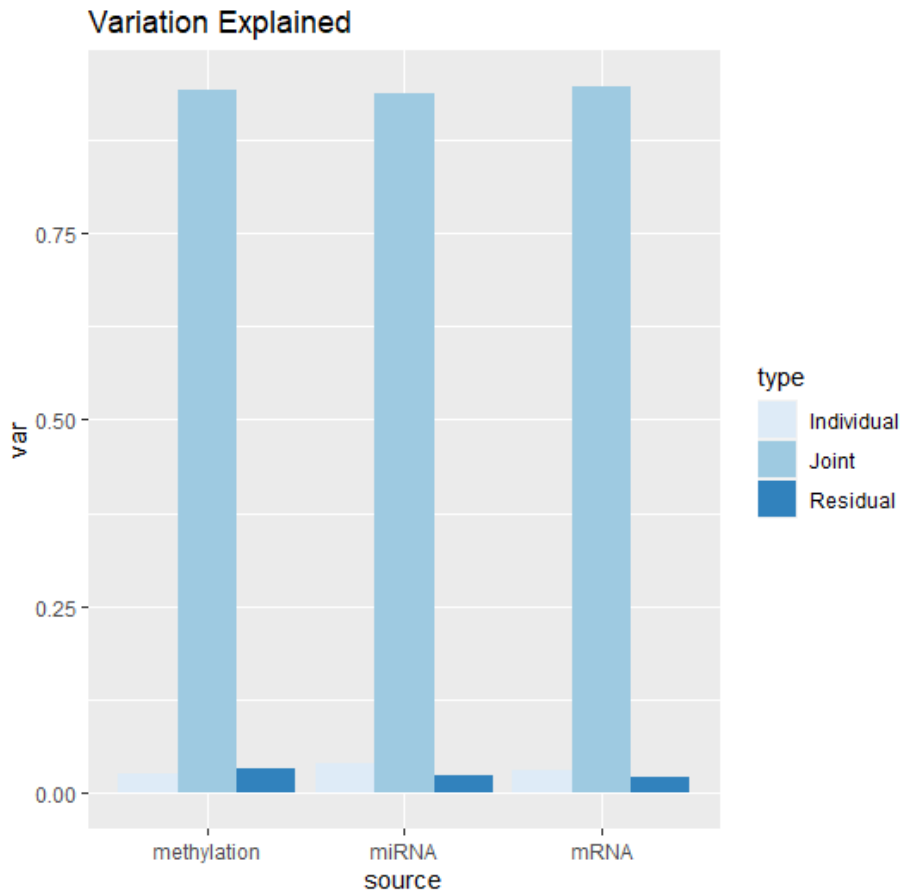


Figure 2: Joint and individual proportions of variance explained. The joint component is prevalent on the individual and residual components for all three datasets.

Comparison	Joint Rank	Ind Rank Methylation	Ind Rank mRNA	Ind Rank miRNA
methyl-mRNA	5	38	6	-
methyl-miRNA	2	41	-	7
mRNA-miRNA	3	-	8	6

Table 1: Joint and individual ranks obtained by pairwise aJIVE.

3.2 Prediction models

Figure 3 reports the ROC curves relative to the logistic models fitted on the full dataset. The model with only patient covariates (age, BMI and smoking) as explanatory variables is compared in terms of AUC to three "integrative" models, reported in Table 2: the model with aJIVE joint components and patient covariates as explanatory variables, the model with aJIVE individual components and patient covariates, and finally the "full" model with patient covariates, aJIVE joint components and the first five aJIVE individual components for each data source as explanatory variables. These are further compared to "non-integrative" models, using the first five individual PCs obtained for each dataset separately.

These results were validated by 10 fold cross validation for each outcome. In the ROC studies from cross validation, the model with all components seems to improve the prediction for both case-control and metastasis status. The mean AUCs for the "full" models are 0.76 and 0.74, for case-control and metastasis status respectively. The mean AUCs for the models using only joint components and clinical covariates are 0.76 and 0.71, while the mean AUCs for the models using only clinical covariates are 0.69 and 0.69. We see that including the individual components does not always result in better models, which seems to confirm that the major role is played by the joint components identified by aJIVE rather than the source specific components. When including only individual components from aJIVE, in addition to clinical covariates, the mean AUC is 0.66 for the prediction of case vs control and 0.68 for the prediction of metastasis. The mean AUC of the non-integrative model, based on the single data PCAs and the clinical covariates, is respectively 0.68 and 0.65, lower than the AUCs obtained in the models using the aJIVE components.

Table 2 reports accuracy and OOB classification error for the random forests, as well as the mean AUCs. For case-control status, both models with all components and with only joint components improve the predictions. The non-integrative models perform similarly to the model including only covariates, and have very low accuracy. This difference from the logistic models with cross validation can be due to the instability of the random forest, and to the limited sample size. We do not report the random forests results for metastasis because they are highly unstable and the accuracy is very low, most likely due to the even more limited sample size, that is only 125 (only cases) for the metastasis.

Figure 4 shows the first ten variables ranked by variable importance in the full model for case-control status. Three of the five joint components appear among the first five variables when ranked for variable importance in the random forest prediction.

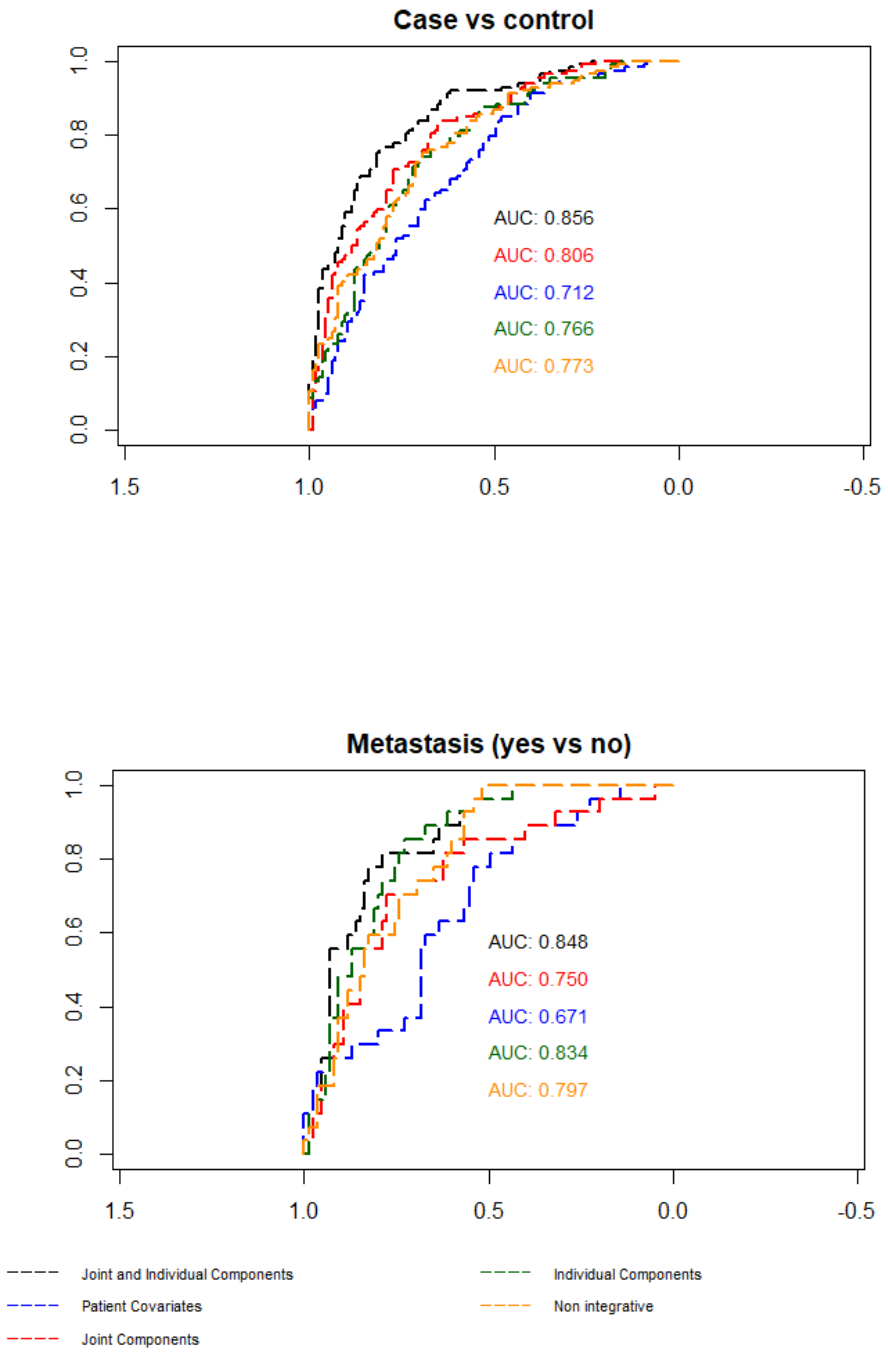


Figure 3: ROC curves from 10 fold cross validation

Model	Accuracy	OOB Classification Error	mean AUC
Joint and Individual (with covs)	0.81	26.53%	0.77
Joint (with covs)	0.68	33.67%	0.74
Patient covariates	0.65	38.27%	0.61
Individual (with covs)	0.77	30.61%	0.71
Non-integrative analysis (with covs)	0.50	38.05%	0.65

Table 2: Random forest diagnostics

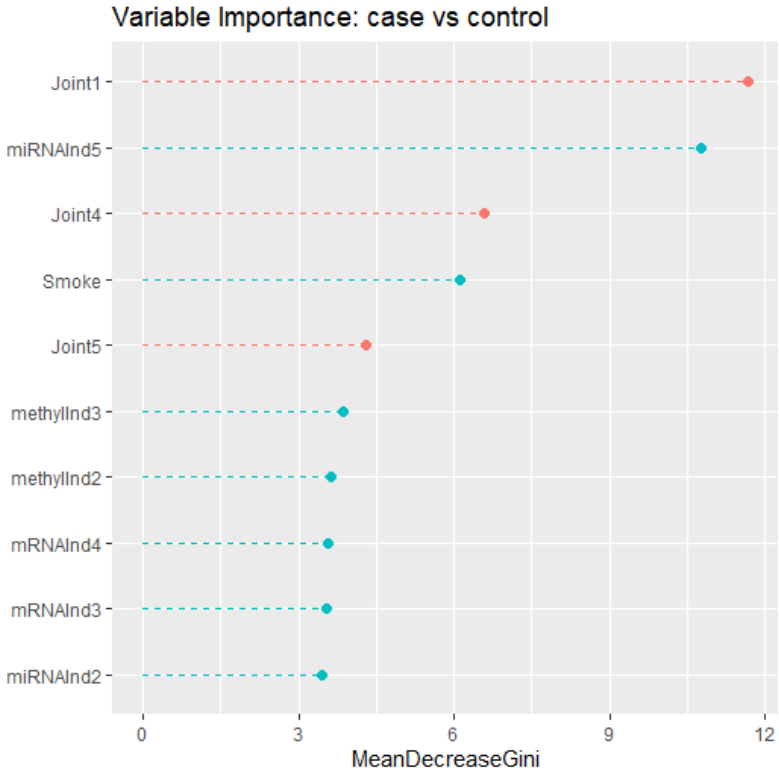
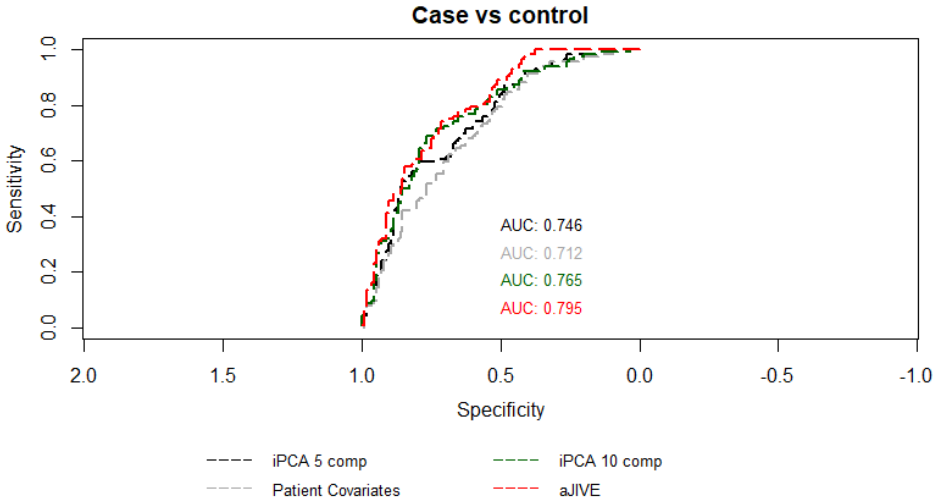


Figure 4: Variable importance plot from random forest on case vs control

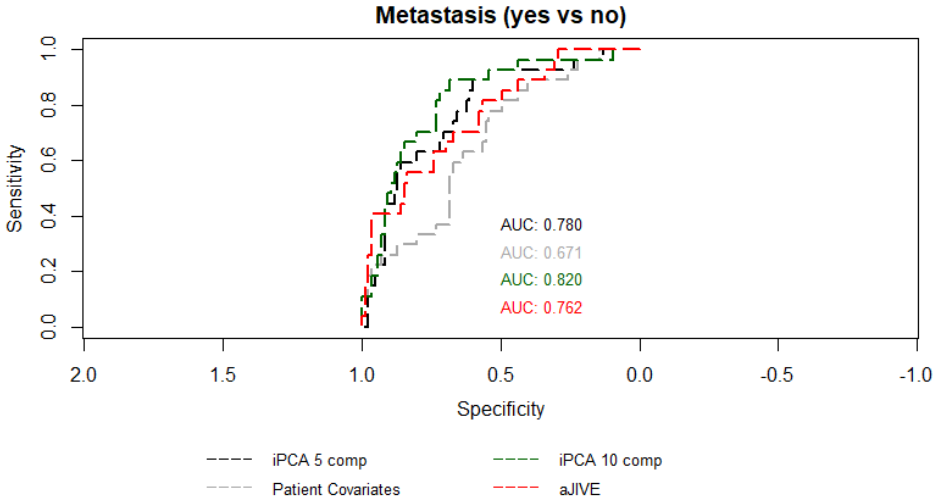
3.3 iPCA

Joint components obtained from iPCA were used in prediction models of case-control and metastasis status. Joint components were also obtained by aJIVE on the same subset of data and prediction models were compared. The initial ranks were selected for aJIVE via profile likelihood and set respectively to 23, 18 and 9. AJIVE selected a joint rank equal to 6 and individual ranks equal to 18, 13 and 6.

We fitted prediction models on case-control and metastasis status using the joint scores from iPCA and aJIVE as predictors, together with the patient covariates. Figure 5 shows ROC curves for both methods for each outcome. In both cases, using joint scores from either aJIVE or iPCA improved the prediction of the outcome to an equivalent extent. In the prediction of case-control status, aJIVE resulted in slightly higher AUCs than those based on iPCA, both when using five and 10 joint components from iPCA. When predicting metastasis, aJIVE performed better than iPCA with five joint components, but not with 10 joint components. However, the results were very similar. Results were validated via 10-fold cross validation and the AUC are comparable between iPCA and aJIVE based models. The mean AUC for the models including five iPCA components and patient covariates is 0.72 for case-control status and 0.70 for metastasis status, while including 10 iPCA components gives mean AUC equal to 0.71 and 0.73 respectively. The mean AUC of the model including aJIVE components and patient covariates is 0.76 for case-control status and 0.73 for metastasis status.



(a) Case vs control



(b) Metastasis (yes vs no)

Figure 5: Prediction models ROC from iPCA and aJIVE

4 Discussion

We use two data integration methods to identify both joint and individual components in a lung cancer study, where multiple omics data sources are available. While the individual contribution of each data source is known to be relevant and has been widely studied in this context, different data sources are also expected to jointly associate with the clinical outcomes. We show how including both joint and individual components in prediction models improves the quality of prediction for the occurrence of lung cancer, as well as its classification into metastatic or non-metastatic cancer. Models that include both components lead to better predictions when compared to non-integrative models, or to models based on clinical covariates.

Prediction models are validated in a 10 fold cross validation framework, and such results are further confirmed by random forests. From the cross validation study, we see that for case-control status, the joint components provide better prediction than non-integrative analysis. For metastasis, the improvement is evident only for the "full" integrative model with both joint and individual components. Still, the AUCs are comparable and small differences can be due to the limited sample size, since we are restricted to cases only ($n = 125$), and to the unbalance between groups in the analysis of metastasis. This limitation is more obvious in the random forests, where the small sample size leads to low accuracy and high instability. Therefore, we only show random forests for the classification of case vs controls, although the same procedure can be repeated for the prediction of metastasis.

A possible explanation of generally low AUCs is that prediction models might also be affected by the time between blood sampling and cancer diagnosis, and we expect the quality of predictions to be higher in subjects with a shorter time to diagnosis. We stratified cases into two subgroups based on time to diagnosis (higher vs lower than the median time) and obtained higher in-sample AUCs for the classification of case vs control in subjects with a closer time to diagnosis for most models. For the classification of metastasis, the sample size in the two time to diagnosis classes is not enough to draw conclusions.

It is interesting to observe that three genomic components identified from aJIVE rank above smoking in importance for case-control classification (Figure 4). Smoking is known to be the one, major risk factor for lung cancer. We speculate that one reason might be a weak mediation of the smoking effect through the genomic components Fasanelli et al (2015). While this work

provides a preliminary evidence of the importance of an integrative analysis of the omics sources, a more thorough investigation of the joint and individual components could help identify relevant biological patterns for future research. An example can be given by the underlying biological processes involving smoking and lung cancer: the omics signals that are dominating the components could be important risk factors for lung cancer, in addition to information on active or past smoking, and their interaction could shed light on the relevant underlying biological processes. Although a functional interpretation of such processes and of their link to the clinical outcomes is not straightforward, an investigation of the aJIVE components could provide further information that would not be identified by a non integrative analysis of the separate omics sources.

The chosen approach for variable filtering is based on variance for mRNA, and on genomic location for methylation. Specifically, the top 5000 most variable mRNAs are selected and CpGs are then selected based on their location on genes, by including CpGs located on the same genes as filtered mRNAs.The joint contributions are therefore expected to be most relevant, given the natural association between signals on the same gene location, and this aspect needs to be carefully considered in the interpretation of the results. Also the filtering of the miRNAs needs to be taken into account, where less restrictive criteria might result in the estimation of different joint and individual components. Other choices could be made in this phase, for example applying the variance criterion independently on each data source, which could yield different joint and individual components. Alternative criteria are the interquantile range (IQR), or the association with the clinical outcome of interest, estimated by an appropriate regression model. Another choice we made in the preprocessing and filtering of the data is the use of M-values for methylation. This choice is motivated by Du et al (2010), but β values could also be employed for the purpose of this work.

To identify joint and individual components, we use one matrix factorization based method, aJIVE, and one PCA based method, iPCA. Although iPCA requires higher computation time when compared to aJIVE, it solves one of the main issues in aJIVE, which is the selection of initial ranks. The most common method for the choice of the initial ranks in aJIVE is the visualization of screeplots, which is subjective and highly sensitive to noise in the data. The profile likelihood idea suggested by Zhu and Ghodsi (2006) partly addresses the problem, but it still lacks some objectivity and automation. Nevertheless, the correct choice of ranks is fundamental for aJIVE, and

ranks misspecification can lead to incorrect results (Feng et al, 2018). For this reasons, iPCA should be preferred. On the other hand, the implementation and computation of aJIVE is much more efficient and fast, and allows for inclusion of more variables in the analysis. Our results also show that aJIVE gives slightly better predictions than iPCA. Therefore, a sensible strategy could be to use aJIVE when the choice of initial ranks is non-controversial and quite straightforward from the data, and to opt for iPCA when such choice is uncertain. In any case, it is always recommended to fit the models with more than one method and compare and validate results.

The high dimensionality of the data also motivates the use of sparse methods, which reduce the number of variables included in the model and provide an easier interpretation of the results. A sparse version of the aJIVE method could be used for this purpose, by introducing a penalty term in the decomposition to induce variable sparsity. This has not been specifically implemented for aJIVE, but Lock et al (2013) discuss and provide an implementation of a sparse version of the JIVE method. It has been shown in Tang and Allen (2018) that sparsity can be achieved in the iPCA method by using an additive L1 norm in the estimation of the covariance matrices.

One aspect that is not accounted for by either methods is the presence of partially shared components. When joint components are only shared by, for example, two out of the three data sources, they will not be identified by aJIVE or iPCA. This is a limitation of most data integration methods, and we expect partially shared components to result in even better prediction models. A way to investigate partially shared patterns is provided in the SLIDE method by Gayananova and Li (2019), and is a potential starting point for further work in this direction.

5 Conclusion

Our study shows how integrative models that include both joint and individual contribution of multiple datasets lead to more accurate model predictions, and facilitate the interpretation of the underlying biological processes. We use iPCA and aJIVE to identify joint and individual contributions of DNA methylation, miRNA and mRNA expression in a lung cancer case control study. We include both joint and individual components in prediction models for the occurrence of lung cancer, and for its classification into metastatic

or non-metastatic cancer. Our results show that models that include joint and individual components lead to better predictions when compared to non-integrative models, or to models based on clinical covariates only.

Ethics approval and consent to participate

All participants gave written informed consent and the study was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Inspectorate. More information is available in Lund et al (2008).

Availability of data and materials

The code for the statistical analysis is available at https://github.com/ericaponzi.

Data cannot be shared publicly because of local and national ethical and security policies. Data access for researchers will be conditional on adherence to both the data access procedures of the Norwegian Women and Cancer Cohort and the UiT –The Arctic University of Norway (contact via Tonje Braaten tonje.braaten@uit.no and Arne Bastian Wiik arne.b.wiik@uit.no) in addition to an approval from the local ethical committee.

Competing interests

The authors declare that they have no competing interests.

Funding

Norwegian Research Council - grant number 248804: National training initiative to make better use of biobanks and health registry data.
Norwegian Research Council - FRIMEDBIO grant number 262111: Identifying biomarkers of metastatic lung cancer using gene expression, DNA methylation and microRNAs in blood prior to clinical diagnosis (Id-Lung).
The miRNA and mRNA analyses were provided by the Genomics Core Facility (GCF), Norwegian University of Science and Technology (NTNU). GCF

is funded by the Faculty of Medicine and Health Sciences at NTNU and Central Norway Regional Health Authority.

Author's contributions

K. M., M. T. and E. P. conceived the research idea. E.P. conducted the statistical analyses. T. H. N. was responsible for the acquisition of data and the biological interpretation of the results. E.P. K. M. and M.T. wrote the manuscript, with inputs from T. H. N. All authors gave final approval.

References

Baglietto L, Ponzi E, Haycock P, Hodge A, Assumma MB, Jung CH, Chung J, Fasanelli F, Guida F, Campanella G, Chadeau-Hyam M, Grankvist K, Johansson M, Ala U, Provero P, Wong EM, Joo J, English DR, Kazmi N, Lund E, Faltus C, Kaaks R, Risch A, Barrdahl M, Sandanger TM, Southey MC, Giles GG, Johansson M, Vineis P, Polidoro S, Relton CL, Severi G (2017) DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. International Journal of Cancer 140:50–61

Carmichael I (2019) ajive: Angle based joint and individual variation explained. `https://github.com/idc9/r_jive`, URL `https://github.com/MeileiJiang/AJIVE_Project`

Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11:587

Fan J, Wang D, Wang K, Zhu Z (2019) Distributed estimation of principal eigenspaces. Annals of Statistics 47:3009–3031

Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, Chadeau-Hyam M, Ala U, F, Kaaks R, Risch A, De Stavola B, Hodge A, Giles GG, Southey MC, Relton CL, Haycock PC, Lund E, Polidoro S, Sandanger TM, Severi G, Vineis P (2015) Hypomethylation of smoking-related genes is associated

with future lung cancer in four prospective cohorts. Nature Communications 6:10,192

Feng Q, Jiang M, Hannig J, Marron J (2018) Angle-based joint and individual variation explained. Journal of Multivariate Analysis 166:241 – 265, DOI https://doi.org/10.1016/j.jmva.2018.03.008, URL http://www.sciencedirect.com/science/article/pii/S0047259X1730204X

Gayananova I, Li G (2019) Structural learning and integrative decomposition of multi-view data. Biometrics 75:1121–1132

Heller G, Weinzierl M, Noll C, Babinsky V, Ziegler B, Altenberger C, Minichsdorfer C, Lang G, Döme B, End-Pfützenreuter A, Arns BM, Grin Y, Klepetko W, Zielinski CC, Zöchbauer-Müller S (2012) Genome-wide mirna expression profiling identifies mir-9-3 and mir-193a as targets for dna methylation in non–small cell lung cancers. Clinical Cancer Research 18(6):1619–1629, DOI 10.1158/1078-0432.CCR-11-2450, URL https://clincancerres.aacrjournals.org/content/18/6/1619, https://clincancerres.aacrjournals.org/content/18/6/1619.full.pdf

Hellton KH, Thoresen M (2016) Integrative clustering of high-dimensional data with joint and individual clusters. Biostatistics 17(3):537–548, DOI 10.1093/biostatistics/kxw005, URL https://doi.org/10.1093/biostatistics/kxw005, https://academic.oup.com/biostatistics/article-pdf/17/3/537/17741839/kxw005.pdf

Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377

Hu Y, Chen G (2015) Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation. Oncology Letters 10:1350–1370

Huang S, Chaudhary K, Garmire LX (2017) More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics 8

Jiang M (2018) AJIVE project. https://github.com/MeileiJiang/AJIVE_Project, URL https://github.com/MeileiJiang/AJIVE_Project

Jiang R, Tang W, Wu X, Wenhui F (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10:S65

Kaplan A, Lock EF (2017) Prediction with dimension reduction of multiple molecular data sources for patient survival. Cancer Inform 16:1–11

Kuligowski J, Perez-Guaita D, Sanchez-Illana Z A ad Leon-Gonzalez, de la Guardia M, Vento M, Lock EF, Quintas G (2015) Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE). Analyst 13:4521–4529

Lock EF, Hoadley KA, Marron JS, Nobel AB (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Annals of Applied Statistics 7:523–542

Lofsted T, Hoffman D, Trygg J (2012) Global, local and unique decomposition in OnPLS for multiblock data analysis. Analytica Chimica Acta 791:13–24

Lund E, Dumeaux V, Braaten T, Hjartåker A, Engeset D, Skeie G, Kumle M (2008) Cohort profile: the Norwegian Women and Cancer study: NOWAC - kvinner og kreft. International Journal of Epidemiology 37:36 – 4

Ma B, Wilker EH, Willis-Owen SAG, Byun HM, Wong KCC, Motta V, Baccarelli AA, Schwartz J, Cookson WOCM, Khabbaz K, Mittleman MA, Moffatt MF, Liang L (2014) Predicting DNA methylation level across human tissues. Nucleic Acids Research 42(6):3515–3528, DOI 10.1093/nar/gkt1380, URL `https://doi.org/10.1093/nar/gkt1380`

Måge I, Smilde AK, van der Kloet FM (2019) Performance of methods that separate common and distinct variation in multiple data blocks. Journal of Chemometrics 33:e3085

McCabe SD, Lin DY, Love MI (2020) Consistency and overfitting of multi-omics methods on experimental data. Briefings in Bioinformatics 21:1277–1284

O'Connell MJ, Lock EF (2016) R.JIVE for exploration of multi-source molecular data. Bioinformatics 32(18):2877–2879, DOI 10.1093/bioinformatics/btw324, URL `https://doi.org/10.1093/bioinformatics/btw324`,

https://academic.oup.com/bioinformatics/article-pdf/32/18/2877/25416720/btw324.pdf

Rappaport N, Ron S (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research 42:10,546–10,562

Sandanger TM, Haugdahl Nøst T, Guida F, Rylander C, Campanella G, Muller DC, van Dongen J, Boomsma DI, Johansson M, Vineis P, Vermeulen R, Lund E, Chadeau-Hyam M (2018) DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the norwegian women and cancer cohort. Scientific Reports 8:16,714

Schouteden M, Van Deun TF, Wilderjans T, Van Mechelen I (2013) Performing DISCO-SCA to search for distinctive and common information in linked data. Behavior Research Methods 46:576–587

Tang TM (2018) Integrated principal components analysis (iPCA). https://github.com/DataSlingers/iPCA, URL https://github.com/MeileiJiang/AJIVE_Project

Tang TM, Allen GI (2018) Integrated principal components analysis. 1810.00832

Trygg J, Wold H (2003) O2-PLS, a two-block (x-y) latent variable regression (LVR) method with an integral OSC filter. Journal of Chemometrics 17:53–64

Tseng G, Ghosh D, Zhou XJ (2015) Integrating omics data. Cambridge University Press

Westerhuis JA, Kourti T, MacGregor JF (1998) Analysis of multiblock and hierarchical PCA and PLS models. Journal of Chemometrics 12:301–321

Witten D, Tibshirani R (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8:Article 28

Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM, Harris CC (2006) Unique microrna molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9(3):189 – 198, DOI https://doi.org/10.

1016/j.ccr.2006.01.025, URL http://www.sciencedirect.com/science/article/pii/S153561080600033X

Yu Q, Risk BB, Zhang K, Marron J (2017) JIVE integration of imaging and behavioral data. NeuroImage 152:38 – 49, DOI https://doi.org/10.1016/j.neuroimage.2017.02.072, URL http://www.sciencedirect.com/science/article/pii/S1053811917301830

Zhang Y, Breitling LP, Balavarca Y, Holleczek B, Schöttker B, Brenner H (2016) Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. International Journal of Cancer 139(11):2482–2492, DOI 10.1002/ijc.30374, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.30374, https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.30374

Zhang Z, Liu J, Kaur M, Krantz ID (2012) Characterization of DNA methylation and its association with other biological systems in lymphoblastoid cell lines. Genomics 99(4):209 – 219, DOI https://doi.org/10.1016/j.ygeno.2012.01.002, URL http://www.sciencedirect.com/science/article/pii/S0888754312000031

Zhu M, Ghodsi A (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. Computational Statistics and Data Analysis 51:918–930