

## Integrative Exploratory Analysis of Two or More Genomic Datasets

Chen Meng and Aedin Culhane

### Abstract

Exploratory analysis is an essential step in the analysis of high throughput data. Multivariate approaches such as correspondence analysis (CA), principal component analysis, and multidimensional scaling are widely used in the exploratory analysis of single dataset. Modern biological studies often assay multiple types of biological molecules (e.g., mRNA, protein, phosphoproteins) on a same set of biological samples, thereby creating multiple different types of omics data or multiassay data. Integrative exploratory analysis of these multiple omics data is required to leverage the potential of multiple omics studies. In this chapter, we describe the application of co-inertia analysis (CIA; for analyzing two datasets) and multiple co-inertia analysis (MCIA; for three or more datasets) to address this problem. These methods are powerful yet simple multivariate approaches that represent samples using a lower number of variables, allowing a more easily identification of the correlated structure in and between multiple high dimensional datasets. Graphical representations can be employed to this purpose. In addition, the methods simultaneously project samples and variables (genes, proteins) onto the same lower dimensional space, so the most variant variables from each dataset can be selected and associated with samples, which can be further used to facilitate biological interpretation and pathway analysis. We applied CIA to explore the concordance between mRNA and protein expression in a panel of 60 tumor cell lines from the National Cancer Institute. In the same 60 cell lines, we used MCIA to perform a cross-platform comparison of mRNA gene expression profiles obtained on four different microarray platforms. Last, as an example of integrative analysis of multiassay or multi-omics data we analyzed transcriptomic, proteomic, and phosphoproteomic data from pluripotent (iPS) and embryonic stem (ES) cell lines.

**Key words** Multivariate, Dimension reduction, Multi-omics, Multiassay, Data integration

---

### 1 Introduction

High throughput technologies including microarray, sequencing, mass spectrometry based proteomics which assay biological molecules have developed rapidly in the past decades. These technologies generate vast amounts of data that describe biological samples at genomic scale and are often called omics data. The capacity and performance of these technologies have improved concurrently with dramatic decreases in cost, and therefore modern

omics studies frequently apply multiple omics techniques to describe the same set of biological observations, such studies include The Cancer Genome Atlas (TCGA), Cancer Cell Line Encyclopedia (CCLE), and ENCyclopedia of DNA Elements (ENCODE). These projects systematically profile large numbers of biological samples resulting in multiple levels of qualitative or quantitative omics data. Whilst the systematically measuring large number of biological molecules (genes, proteins) can reveal novel knowledge that cannot be discovered by traditional methods, the accumulation of multiple omics data presents new challenges for data integration and interpretation.

Several exploratory data analysis (EDA) methods including correspondence analysis (CA), principal component analysis (PCA) have been widely applied to study single omics data [1, 2]. These EDA methods are frequently performed in the early stage of analysis for quality control, detecting batch effect or exploring basic cluster structure in a dataset. In the analysis of multiple omics data, EDA also needs to identify correlations and associations between each of the high dimensional datasets. In this chapter, we describe the following EDA methods that enable researchers to identify relationships between two or more high dimensional datasets:

- Co-inertia analysis (CIA) can be used to explore relationships between two datasets [3];
- Multiple co-inertia analysis (MCIA) can be applied to analyze multiple datasets [4].

Both methods project observations (samples) and variables onto a lower dimensional space, but constrain the dimension reduction such that the new variables represent covariant structure among datasets. Variables from each dataset are transformed onto the same scale. The association between variables and samples can be visualized in this new space, which greatly facilitates the detection of global variance structure and identification of the most informative variables across datasets.

---

## 2 Methods

### 2.1 Analysis of Two Datasets Using Co-inertia Analysis

#### 2.1.1 Co-inertia Analysis

Co-inertia analysis is a multivariate exploratory approach used to identify the covariance between two datasets that have the same set of observations [5]. In the field of omics data analysis, CIA was first introduced to cross-platform comparison of microarray data [3]. With increasing availability of other omics data, it has also been applied to integration of different types of omics data [6].

Two omics datasets may be represented by two matrices,  $X$  and  $Y$ . In this chapter, we assume the rows of a matrix are samples (observations) and columns are variables, such as genes, proteins, other small molecules, etc. Similar to PCA, CIA is a dimension

reduction technique but it considers two datasets simultaneously. For the  $i$ th dimension, CIA finds a pair of new variables, designated as co-inertia components or dimensions, using a linear combination of the original variables in  $X$  and  $Y$ , so as to maximize the squared covariance between them.

$$\operatorname{argmax}_{u^i, v^i} \operatorname{cov}^2(Xu^i, Yv^i) \quad (1)$$

$Xu^i$  and  $Yv^i$  are the co-inertia components for matrix  $X$  and  $Y$ , respectively;  $v^i$  and  $u^i$  are the linear combination coefficients, which is comparable to the loading vectors in the PCA. Due to the optimized criteria, the co-inertia components capture the most important covariance structure between the two datasets. The co-structure between the two datasets may be visualized by the co-inertia components in a lower dimensional space.

## 2.2 Case Study I: Integration of NCI-60 Cell Line Transcriptomic and Proteomic Data

The NCI-60 panel is a collection of 60 cancer cell lines from nine different tissues of origin. It includes leukemia, melanoma, ovarian, renal, breast, prostate, colon, lung, and central nervous system (CNS). These cell lines are widely used for in vitro screening of anti-cancer compounds. In attempts to discover gene–drug interactions, several genome wide data profiling approaches have been applied to these cell lines, including DNA copy number variation, DNA mutation, gene expression, protein expression, drug sensitivity, etc. In this case study, we will examine the mRNA expression measure by Agilent GE 4x44K microarray platform (downloaded from [7]) and the proteome data (mass spectrometry based proteomics) [8]. We will use CIA to explore the similarity between datasets and cell lines.

We load the required package and data using:

```
library(omicade4)
library(made4)
load("../data/NCI60_rnaprotein.RDA")
```

NCI60\_rnaprotein is an object of class *list*, which consists of two numerical matrices, mRNA and protein. These two matrices have the following dimensions:

```
summary(NCI60_rnaprotein)

##           Length Class  Mode
## mRNA      640958 -none- numeric
## protein  431288 -none- numeric

sapply(NCI60_rnaprotein, dim)

##           mRNA protein
## [1,] 11051      7436
## [2,]   58        58
```

Each of the matrices has 58 cell lines in columns. Due to the problem of data quality, we removed two cell lines, resulting in 58 cell lines included in this analysis. CIA requires that the columns in the matrices are correctly matched, to verify this:

```
identical(colnames(NCI60_rnaprotein$mRNA), colnames(NCI60_rnaprotein$protein))

## [1] TRUE
```

However, the number of rows in different matrices may be different. To facilitate the visualization later, we first create some auxiliary variables to indicate the names of cell lines, tissues of origin of cell lines, and the color for each.

```
names <- strsplit(colnames(NCI60_rnaprotein$mRNA), "\\.")
tumorType <- sapply(names, "[", 1)
cellline <- sapply(names, "[", 2)

# color vector
tumorColor <- as.factor(tumorType)
levels(tumorColor) <- c("red", "green", "blue", "cyan", "pink",
                        "brown", "gray25", "orange", "gray75")
tumorColor <- as.character(tumorColor)

phenoData<-cbind(tumorType=tumorType, cellline=cellline, tumorColor=tumorColor)
rownames(phenoData) = colnames(NCI60_rnaprotein$mRNA)
phenoData[1:4,]

##           tumorType cellline      tumorColor
## BR.MCF7          "BR"      "MCF7"          "red"
## BR.MDA_MB_231    "BR"      "MDA_MB_231"      "red"
## BR.HS578T        "BR"      "HS578T"          "red"
## BR.BT_549        "BR"      "BT_549"          "red"
```

Note that Bioconductor is developing a “multi-assay” data object class <https://github.com/vjcitn/biocMultiAssay> which should be helpful and will be recommended when analyzing multi-assay data.

### 2.2.1 PCA of Individual Datasets

Before performing the integrative analysis, we first perform basic exploratory analysis and PCA on each individual dataset. For example, exploring the distribution of datasets:

```
layout(matrix(1:2, 1, 2))
boxplot(NCI60_rnaprotein$mRNA, main="mRNA", col=tumorColor)
boxplot(NCI60_rnaprotein$protein, main="Protein", col=tumorColor)
```

The plot is not shown here. But the lower boundary of the boxes in proteomic data reaches 0. This is because the missing values in the proteomics data are replaced with 0. Before the integrative analysis of two datasets, we analyze each of single datasets using PCA:

```
pca_mrna <- prcomp(t(NCI60_rnaprotein$mRNA))
pca_protein <- prcomp(t(NCI60_rnaprotein$protein))
```

The variance of principal components (PCs) and cell lines in the first two PCs may be visualized by:

```
layout(matrix(1:4, 2, 2))
par(mar=c(3, 3, 1.5, 0.5))
plot(pca_mrna, main="mRNA")
legend("topright", col = unique(tumorColor), pch=20, legend = unique(tumorType))
plot(pca_mrna$x[, 1:2], col=tumorColor, pch=20)
abline(v=0, h=0)
plot(pca_protein, main="Protein")
plot(pca_protein$x[, 1:2], col=tumorColor, pch=20)
abline(v=0, h=0)
```

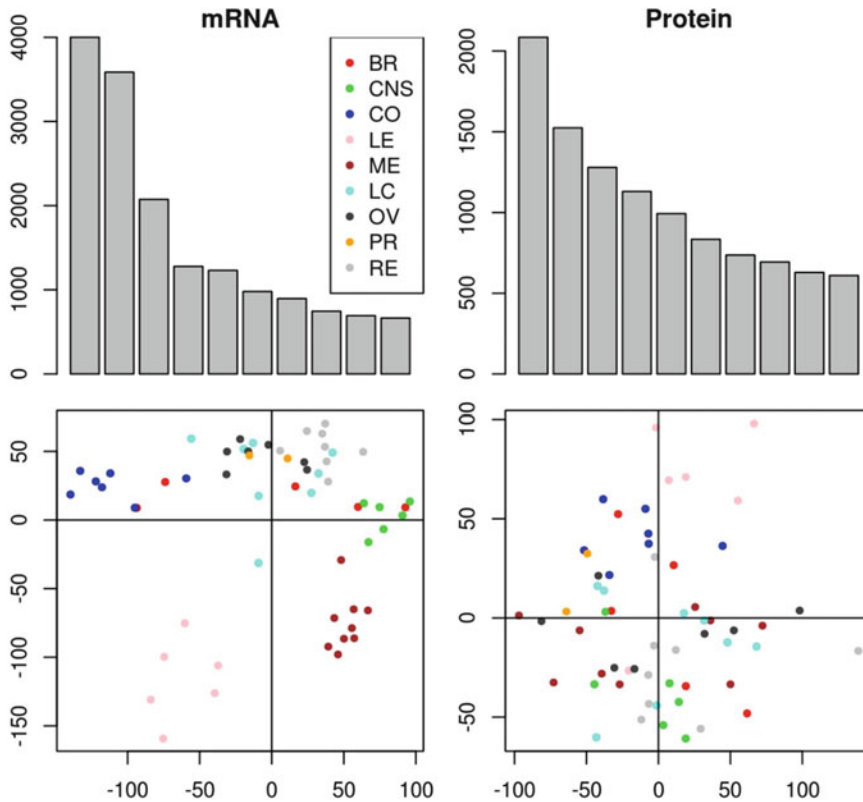
The output is shown in Fig. 1. We observe that the first two PCs in transcriptomic data explain a larger proportion of variance than those in proteomic data. We see cell lines with different anatomical tissue of origin are better separated in the transcriptomic data. But this analysis does not evaluate the co-structure between the two datasets.

### 2.2.2 CIA of Both Datasets

To visualize the correlated structure between the datasets, we perform CIA using R function `cia`,

```
mRNA <- NCI60_rnaprotein$mRNA
protein <- NCI60_rnaprotein$protein
coin <- cia(mRNA, protein, cia.nf = 5)
```

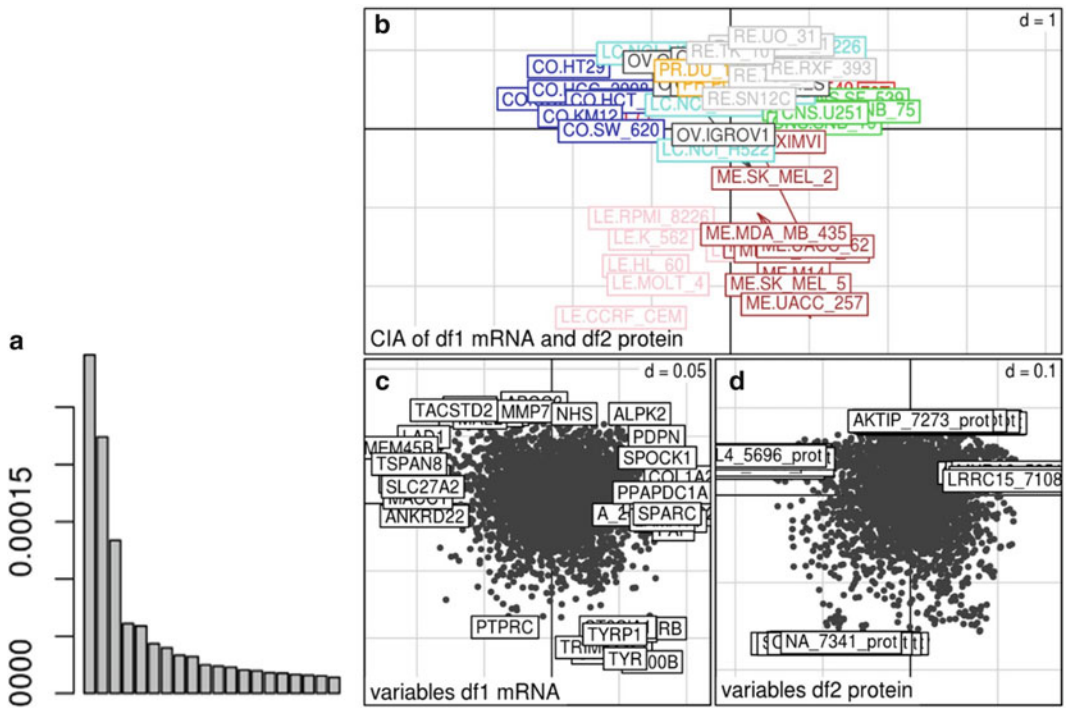
The output of the `cia` function is an object of class `cia`, which can be easily visualized using the `plot` function



**Fig. 1** PCA of individual datasets of mRNA gene expression and protein expression profiles of NCI60 cell lines. The *upper panels* show the variance associated with each principal component of the PCA. The *lower panels* show a plot of the first (horizontal axis) and second (vertical axis) PC for the mRNA and proteins data, respectively

```
barplot(coin$coinertia$eig[1:20])
plot(coin, col=tumorColor)
```

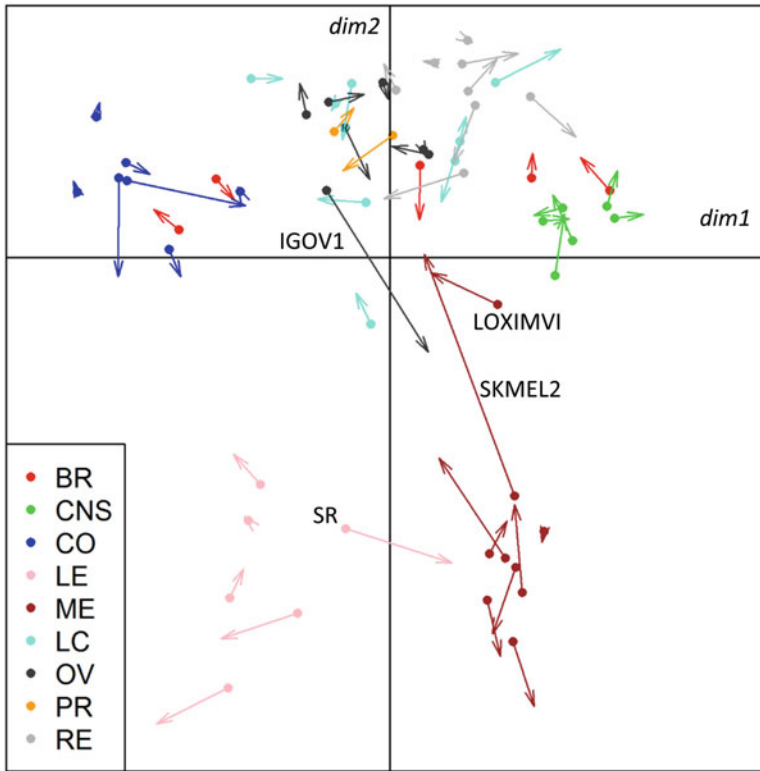
These commands generate Fig. 2. The scree plot in panel a shows the variance associated with each of the co-inertia components, which may be interpreted similarly to variance of PCs in PCA. Fig. 2a shows that the first three components have significantly higher variance than the others. Therefore, these three components should be carefully interpreted and compared with biological or batch factors. The interpretation of components often involves the visualization of samples and variables. Panel b shows the first two co-inertia components. In this plot, samples from the transcriptomic data are shown as the head of arrows. The corresponding cell lines in proteomic data are the ends of arrows. Therefore, this plot is also denoted as sample space. We can see that the leukemia cell lines and melanoma cell lines are weighted on the negative end of the second component, indicating these



**Fig. 2** A plot showing results of a CIA of NCI60 transcriptomic and proteomics data (the same data as in Fig. 1). (a): the variance associated with each co-inertia components; (b): sample space; (c): variables space for mRNA data; (d): variable space for protein data

two cell lines are most different with others. The lengths of the arrows describe the similarity between the samples from the two different datasets. Highly correlated pairs of samples will be projected close to each other and, therefore, are linked by a short arrow. In practice, we often need to extract the co-inertia components and customize the plot. As an example, we remove some labels of cell lines:

```
ax1=1
ax2=2
par(mar=c(0.1, 0.1, 0.1, 0.1))
plot(coin$coinertia$mX[, c(ax1, ax2)], col=tumorColor, pch=20,
      xlim=c(-2.5, 2.5), axes=FALSE, frame.plot = TRUE)
abline(v=0, h=0)
arrows(coin$coinertia$mX[, ax1], coin$coinertia$mX[, ax2],
       coin$coinertia$mY[, ax1], coin$coinertia$mY[, ax2],
       angle = 15, length = 0.1, col=tumorColor)
legend("bottomleft", col = unique(tumorColor), pch=20, legend = unique(tumorType))
# text(coin$coinertia$mX[, c(ax1, ax2)], labels = cellline, cex=0.7)
```



**Fig. 3** A customized plot of the sample space shows the project of NCI60 cells when a CIA is performed on NCI60 transcriptomic and proteomic data analysis. Each cell line is colored by its anatomical tissue of origin

These commands generate a plot similar to Fig. 3. In the plot, we see most of the arrows are short in length, which indicates high overall similarity or considerable correlated structure between the datasets. This can be confirmed using the RV coefficient, which is also included in the `cia` object and may be extracted by

```
coin$coinertia$RV
## [1] 0.7464801
```

The RV coefficient is a multivariate extension of the Pearson correlation coefficient to measure the overall similarity between two matrices. It ranges from 0 to 1. A high RV coefficient indicates high degree of similarity. In this case, the RV coefficients of 0.75 suggest that a relative high correlation exists between the two datasets.



Despite a good overall similarity, some cell lines have lower correlation between their mRNA and protein profiles. These include lung cancer cell line IGOV1, leukemia cell line SR, and melanoma cell line SKMEL2. For example, whilst the SKMEL2 is projected close to other melanoma cell lines in transcriptomic data, in the proteomics data this cell line is closer to the plot origin and is far from most other melanoma cell lines. Similarly, the proteome of leukemia cell line SR is closer to the melanoma cell lines in comparison with other leukemia cell lines. This discrepancy may reflect biological variance, a batch effect, or a technical artifact, such as sample mis-labeling.

### 2.2.3 CIA: Exploring the Variables

In CIA, the projection of each sample is determined by its variable measurements. The variables from both datasets are transformed onto the same scale and projected into the same space, thereby enabling exploration of relationships between variables, and between samples and variables. The loadings of the mRNA and protein variables are shown in panel c and d of Fig. 2, which are also called gene space, or more generally, variable space. In these variable plots, the variables with highest weights (i.e., on the negative or positive ends of components) are the most influential variables and these define the co-inertia components. Variables and samples that are projected in the same direction from the origin have a strong association (i.e. the variables are increased or upregulated in those samples), whereas variables projected at the opposite direction to a sample are frequently have low values in those samples. Therefore, the variables with highest weights in each of components can be extract and these facilitate the biological interpretation of components. We will extract the highest weighted variables using the function `topVar`:

```
topVar(coin, axis = 1, end = "neg", topN = 10)
```

```
##      ax1_df1_negative  ax1_df2_negative
## 1      TMEM45B  CALML4_5696_prot
## 2      OVOL1  UGT1A10_3875_prot
## 3      TSPAN8  AZGP1_3143_prot
## 4      POF1B  DLG1_2351_prot
## 5      MACC1  AKR7A3_4110_prot
## 6      DDC  SDCBP2_4458_prot
## 7      SLC27A2  FABP1_844_prot
## 8      LAD1  LGALS4_779_prot
## 9      FBP1  TMEM62_5329_prot
## 10     ANKRD22  MUC13_920_prot
```

```
topVar(coin, axis = 1, end = "pos", topN = 10)
```

```
##      ax1_df1_positive  ax1_df2_positive
## 1      COL1A2  NCOA7_5477_prot
## 2      FAP  COL6A2_3829_prot
## 3      CNRIP1  HSPB7_4887_prot
## 4      SPARC  COL5A1_5880_prot
## 5      PPAPDC1A  MXRA8_5851_prot
## 6      LAMA4  PCOLCE_4369_prot
## 7      IGFBP7  COPZ2_724_prot
## 8      SPOCK1  LRRC15_7108_prot
## 9      A_24_P554882  COL3A1_1636_prot
## 10     PDPN  BMP1_704_prot
```

```

topVar(coin, axis = 2, end = "neg", topN = 10)

##      ax2_df1_negative  ax2_df2_negative
## 1          S100B  SH2D1A_2507_prot
## 2           TYR  PTPRCAP_1858_prot
## 3       C14orf34   CD1C_5757_prot
## 4       TRIM63   CD5_2006_prot
## 5       TYRP1   CD3E_1036_prot
## 6       MLANA   GRAP_1115_prot
## 7       EDNRB   FLI1_4768_prot
## 8       BCL2A1   RHOH_1471_prot
## 9       ST8SIA1   NA_7341_prot
## 10      PTPRC   TRAT1_698_prot

topVar(coin, axis = 2, end = "pos", topN = 10)

##      ax2_df1_positive  ax2_df2_positive
## 1          ABCC3   CLCN1_7161_prot
## 2          MAL2   AKTIP_7273_prot
## 3       TACSTD2  TM4SF18_2667_prot
## 4          MMP7  PI4K2A_7076_prot
## 5          ELF3  SHISA2_3899_prot
## 6          ALPK2  PAPLN_6608_prot
## 7          NHS   HLA.A_5614_prot
## 8          CST6  DHGPSL_6037_prot
## 9       KRT8P20  DNTTIP2_3974_prot
## 10         MALL  HAVCR1_2844_prot

```

The results suggest that the positive end of the first component captures several collagens, including COL1A1, COL6A2, which are the major components of the extracellular matrix and connective tissues. In tumors, collagens are associated with cancer cell metastasis and poor prognosis in patients. Therefore, we can infer that the first dimension reflects the mechanism related to different potential of metastasis of the cell lines. The negative end of the second component is associated with leukemia and melanoma cell lines. Accordingly, genes (see “ax2\_df1\_negative”) from transcriptomic data with high weights in this component include several melanogenesis genes, such as S100B and TRY, explaining why melanoma cell line LOXIMVI, a cell line that lacks melanin is projected closer to the origin in the proteome sample space. By contrast these proteins are absent in the proteomics data, which

highlighted several immune cell markers, such as CD1C, CD5, and CD3E, which are highly expressed in the leukemia cell lines. Therefore, the projection of leukemia cell lines is more influenced by the proteomic data whereas the weights of melanoma cell lines are determined by the transcriptomic data. But both are separated from other cell lines.

### 2.3 Exploration of Three or More Datasets

#### 2.3.1 Multiple Co-inertia Analysis

CIA can be used to explore the concordance and discrepancy between two datasets. MCIA is a generalization of CIA to analyze more than two datasets [4]. In MCIA, the multiple omics data are represented by  $K$  blocks of matrices ( $X_1, X_2, \dots, X_K$ ). Similar with CIA, for the  $i$ th dimension, MCIA defines a set of block components using the linear combination of variables in each of the matrices. The goal of MCIA is to find a synthetic component,  $s^i$ , so as to maximize the sum of squared covariance between the block components and the synthetic components, that is

$$\operatorname{argmax}_{u_k^i, s^i} \sum_{k=1}^K \operatorname{cov}^2(X_k u_k^i, s^i) \quad (2)$$

where  $X_k u_k^i$  are the set of block component and the  $s^i$  is the synthetic component;  $u_k^i$  are the loading for the variables in the  $k$ th matrix. In PCA, the principal components are the optimal lower rank approximation of a high dimensional dataset, whereas the block components in MCIA are sub-optimal in terms of representing the individual matrices, but they represent the best covariant structures across multiple datasets. Similar to CIA, the block components and synthetic components can be visualized in a two dimensional space to facilitate the interpretation of multiple high dimensional datasets.

#### 2.3.2 Case Study 2: Cross Comparison of Gene Expression Data Obtained on Four Different Microarray Platforms

Cross-platform comparison is often used in EDA, such as meta-analysis or as part of cross-validation of findings. In this example, we explore the consensus in data from four transcriptomic studies of NCI60 cell line using different microarray platform (Affymetrix HG U95, U133, U133 plus2.0 and Agilent; downloaded from [7]). The goal of this analysis is to explore the concordances and discrepancies in mRNA expression measurements obtained using different platforms for each cell lines.

First, we load the data

```
load("../data/NCI60_4arrays.RDA")
```

and get an overview of dimensions of datasets using the following functions:

```
summary(NCI60)

##           Length Class  Mode
## agilent   640958 -none- numeric
## hgu95     510574 -none- numeric
## hgu133    524552 -none- numeric
## hgu133p2  602156 -none- numeric

sapply(NCI60, dim)

##      agilent hgu95 hgu133 hgu133p2
## [1,]   11051   8803   9044   10382
## [2,]     58    58    58      58

names(NCI60)

## [1] "agilent" "hgu95" "hgu133" "hgu133p2"

tumorType <- sapply(strsplit(colnames(NCI60$agilent), "\\."), "[", 1)
tumorColor <- as.factor(tumorType)
levels(tumorColor) <- c("red", "green", "blue", "cyan", "pink",
                        "brown", "gray25", "orange", "gray75")
tumorColor <- as.character(tumorColor)
```

We draw a boxplot to explore the distribution of datasets.

```
layout(matrix(1:4, 2, 2))
boxplot(NCI60$agilent, main="Agilent", col=tumorColor)
boxplot(NCI60$hgu95, main="Affy HG U95", col=tumorColor)
boxplot(NCI60$hgu133, main="Affy HG U133", col=tumorColor)
boxplot(NCI60$hgu133p2, main="Affy HG U133 plus2.0", col=tumorColor)
```

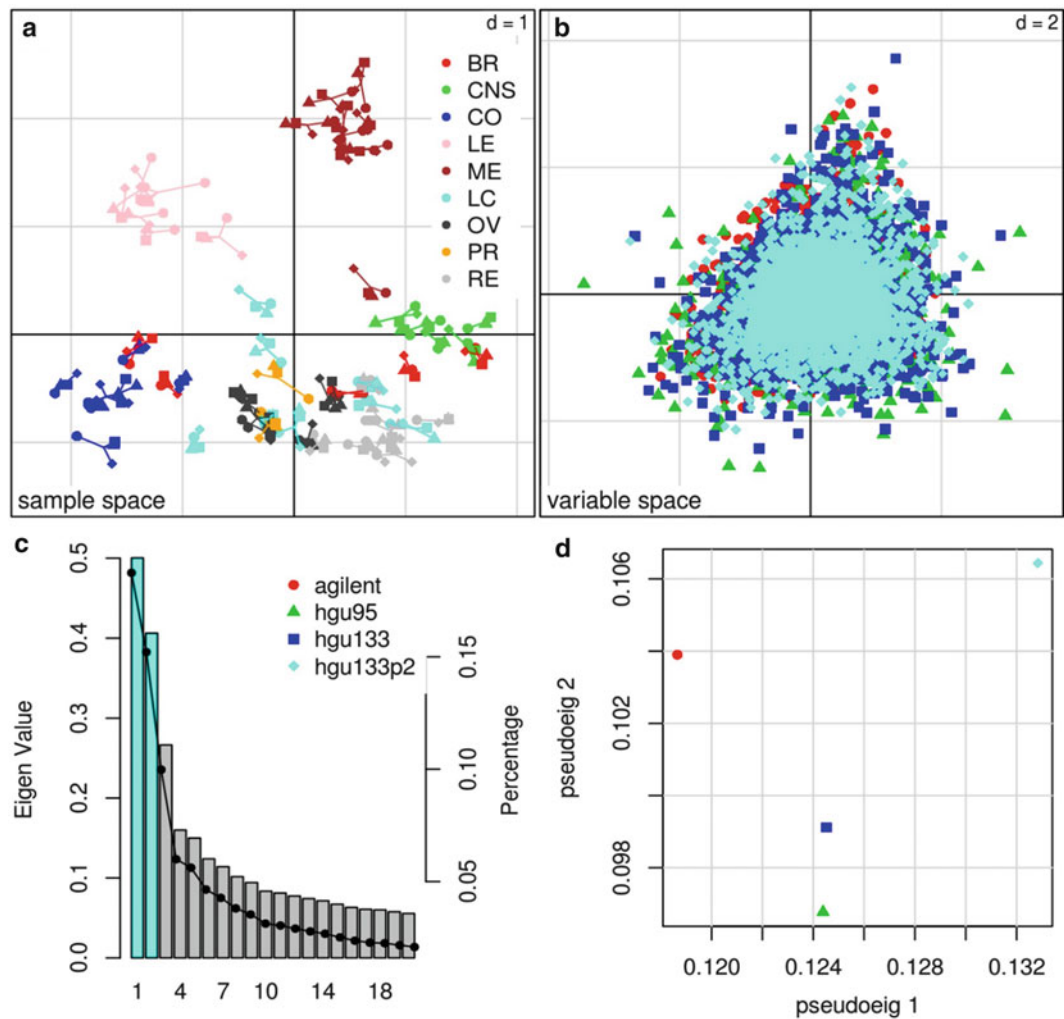
MCIA is performed using the function `mcia` in *omicade4* package and is visualized using the function `plot`, for example

```
mcoin <- mcia(NCI60)

## 'svd' fail to convergence, 'eigen' used to perform singular value decomposition

plot(mcoin, df.color = 2:5, sample.color=tumorColor,
     sample.legend = FALSE,
     sample.lab=FALSE)

legend("bottomright", col = unique(tumorColor), pch=20, legend = unique(tumorType))
```



**Fig. 4** A plot showing results of an MCIA which integrated and compared four different microarray gene expression datasets. (a): the sample space; (b): the variable space, variables from four different platforms are shown as different colors; (c): the scree plot shows the variance associated with each dimension. (d): Dataset space

The function `mcia` returns an object of class `mcia`. A typical visualization of the plot is shown in Fig. 2.4. Similar to CIA, the plot consists of the sample space (Fig. 4a) and variables spaces (Fig. 4b). In the sample space, samples from different datasets are shown as point with different shapes and also the same cell lines in each datasets are linked to the synthetic components. The short lines in this plot indicate a good correlation for the four datasets. The pairwise RV coefficient may be extracted by

```
round(mcoin$mcoa$RV, 3)

##           agilent hgu95 hgu133 hgu133p2
## agilent      1.000 0.953  0.955   0.955
## hgu95        0.953 1.000  0.988   0.965
## hgu133       0.955 0.988  1.000   0.969
## hgu133p2     0.955 0.965  0.969   1.000
```

All the RV coefficients are higher than 0.95, indicating a good correlation between datasets generated by different platforms.

Figure 4b shows the variable loadings, variables from different datasets are shown with different colors. The variables and samples projected on the same direction are highly associated with each other. Figure 4c shows the variance associated with each dimension. In MCIA, multiple datasets contribute to the variance of components. Therefore, the global variance can be decomposed into the contributions of each individual datasets. This information is shown in Fig. 4d, the pseudo-eigenvalue for each dimension indicates the decomposed variance of a single dataset, so this panel may also be called “dataset space.” It shows that all the datasets contribute roughly equal to the first and second components as indicated by the small range of  $x$  and  $y$  axes in Fig. 4d. Strictly, data from HGU133plus2.0 contributes slightly higher than others to both first and second components, whereas Agilent data have a lower contribution to first component, but a higher contribution to the second dimension in comparison with HGU95 and HGU133 data.

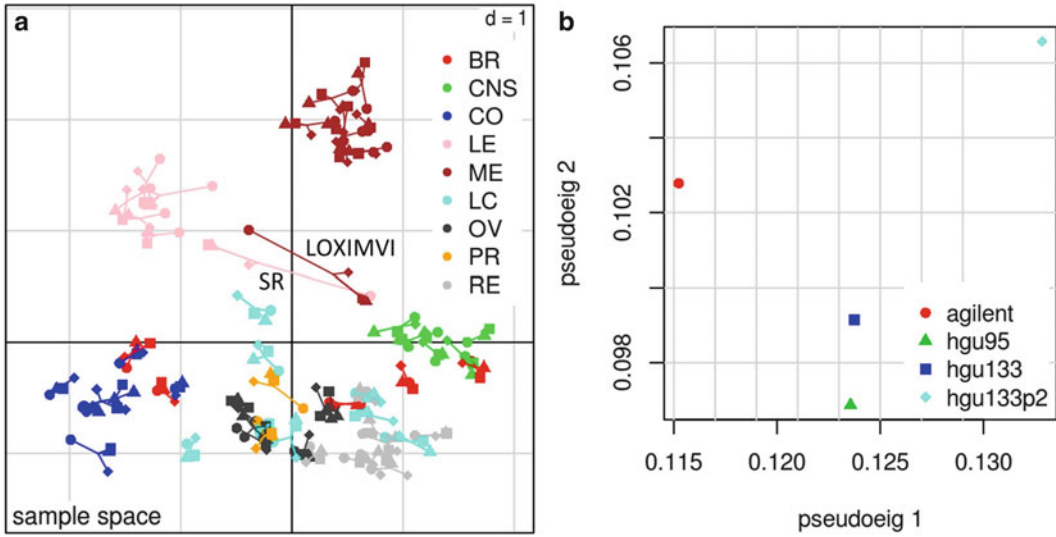
Next, we will show how to use MCIA to detect outlier or “abnormal” samples using MCIA cross-platform comparison. To do so, we swap the names of two samples in the Agilent data to simulate a mis-labeling problem. We exchange the leukemia cell line SR and melanoma cell line LOXIMVI in Agilent data and run MCIA on the datasets

```
NCI60_rand <- NCI60
NCI60_rand$agilent[, c("ME.LOXIMVI", "LE.SR")] <-
  NCI60_rand$agilent[, c("LE.SR", "ME.LOXIMVI")]
```

```
mcoin_rand <- mcia(NCI60_rand)

## 'svd' fail to convergence, 'eigen' used to perform singular value decomposition

plot(mcoin_rand, df.color = 2:5, sample.color=tumorColor,
      sample.legend = FALSE,
      sample.lab=FALSE)
```



**Fig. 5** Demonstration showing the application of MCIA to detecting outliers or “problem” samples. Here, the same analysis (MCIA) is performed as in Fig. 4. However, the names of the melanoma cell lines LOXIMVI and the leukemia cell lines SR are swapped. (a): the sample space; (b): Dataset space

Figure 5 shows the corresponding sample space and dataset space. In the sample space, it is clearly shown that the Agilent data for SR and LOXIMVI are projected away from their counterpart in other datasets. The relative long line in this plot suggests a mis-labeling problem. In addition, the exchange of the labels of cell lines in Agilent data results in that the covariate structure in both dimensions are less data, which can be seen from the decreased pseudo-eigenvalue for Agilent data in Fig. 5b (compare with Fig. 4d).

### 2.3.3 Example 2: Integrative Analysis of Transcriptome, Proteome, and Phosphoproteome of Stem Cell Lines

In this example, we use the data generated by Phanstiel et al. [9]. The goal of this research was to compare protein expression and phosphorylation between embryonic stem cell (ES) and induced pluripotent stem (iPS) cell lines. In the study, 4 ES and 4 iPS cell lines were selected and their transcriptome, proteome, and phosphoproteome data were measured in triplicates. Here, we only analyze one of the replicates as the aim to illustrate how to use MCIA to integrative analysis of multiple omics data of different levels.

First, load the data

```
load("../data/iPSES8ples.RDA")
```

and summarize the data



```
summary(ipSES)

##           Length Class  Mode
## mrna      124648 -none- numeric
## protein   30928 -none- numeric
## phospho    62272 -none- numeric

sapply(ipSES, dim)

##           mrna protein phospho
## [1,] 15581      3866      7784
## [2,]      8        8        8
```

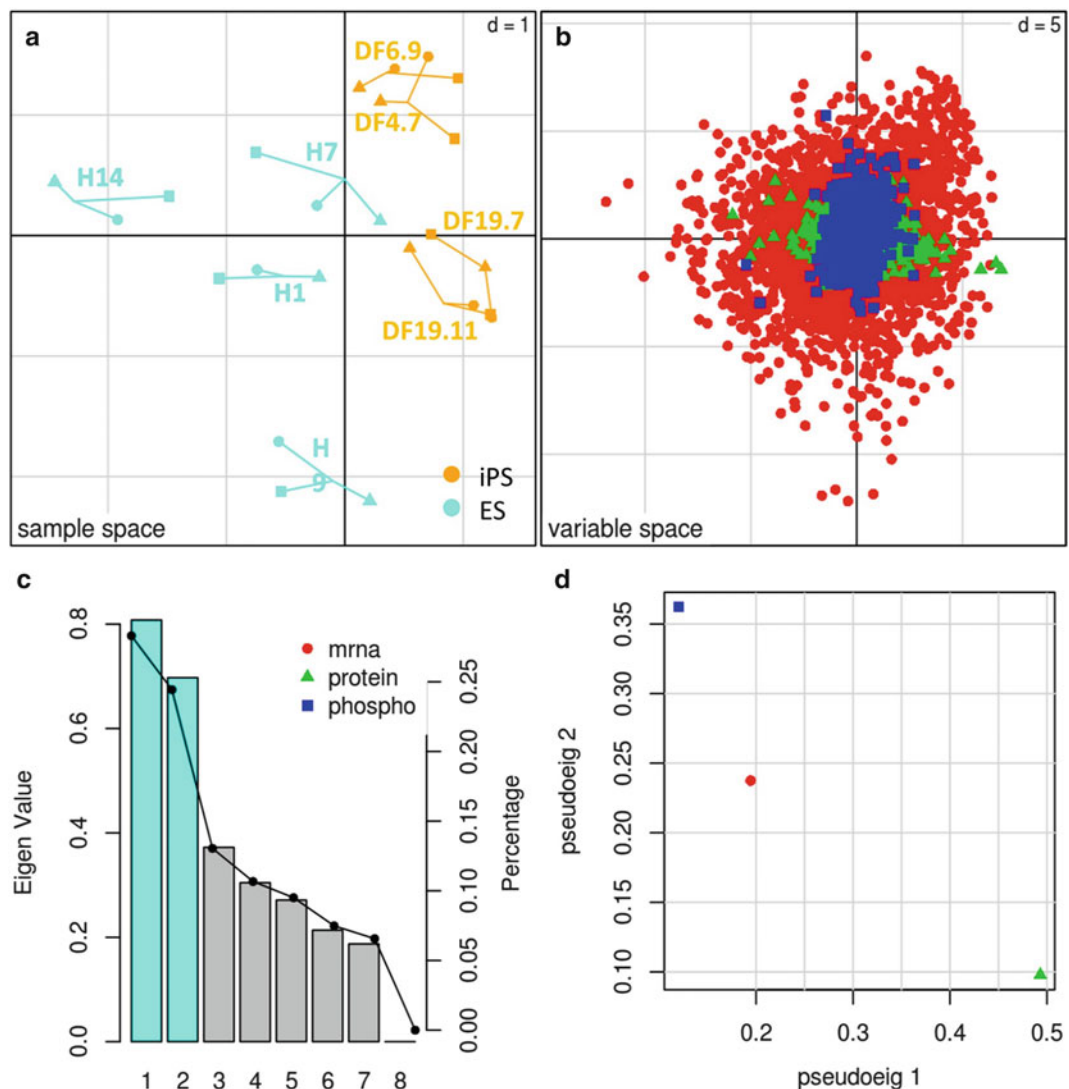
perform the MCIA and plot the result

```
mcoin <- mcia(ipSES)
plot(mcoin, df.color = 2:4, sample.color=rep(c("cyan", "orange"), each=4),
     sample.legend = FALSE, sample.lab=FALSE)
```

Figure 6a–d shows the sample space, variable space, component variance, and dataset space, respectively. Panel c suggests that the first two dimensions have significantly higher variance than others. The corresponding sample space suggests that the iPS and ES cell lines could be distinguished by the first dimension. Particularly, iPS cell lines DF6.9 and DF4.7, DF19.7 and DF19.11 are highly correlated. The ES cell lines are more dispersely projected onto the space. H14 is on the negative end of the first component and H9 is highly weighted on the negative end of the second component. The relative short lines between samples indicate a good correlation between the datasets. However, the dataset space (panel d) suggests that the protein data have more variance on the first dimension, whereas the protein phosphorylation data contribute more variance on the second one. This finding is inconsistent with the variable space, where the proteomic data are more spread on the first component and the phosphoproteomic data have a wider range on the second component. The RV coefficients between datasets are

```
round(mcoin$mcoa$RV, 3)

##           mrna protein phospho
## mrna      1.000    0.705    0.845
## protein  0.705    1.000    0.457
## phospho  0.845    0.457    1.000
```



**Fig. 6** Plot showing results of an MCIA of iPS and ES data. Three datasets were integrated in this analysis, including mRNA expression data (RNA sequencing), protein expression and phosphorylation data (Mass spectrometry based proteomic) (a): the sample space; (b): the variable space, variables from four different platforms are shown as different colors; (c): the scree plot shows the variance associated with each dimension. (d): Dataset space

Unexpectedly we find that the phosphorylation data have a better correlation with the mRNA data rather than the protein data. In addition, similarly to our previous analyses, we can select the variables with the greatest weights on each of the dimensions using the function `topVar`

```

topVar(mcoin, axis = 1, end = "neg", topN = 5)

##   ax1_mrna_negative ax1_protein_negative ax1_phospho_negative
## 1             ZIM2      IPI00946792.1   IPI00022628.5_s387
## 2             MMP1      IPI00555956.2   IPI00657687.1_s182
## 3             CYP4F11    IPI00012989.2   IPI00969114.1_s55
## 4             PEG3      IPI00219774.3   IPI00742682.2_s2155
## 5             LGALS4    IPI00873459.3   IPI00186139.8_s181

topVar(mcoin, axis = 1, end = "pos", topN = 5)

##   ax1_mrna_positive ax1_protein_positive      ax1_phospho_positive
## 1             C17orf50    IPI00026993.1      IPI00066543.2_s40.t44
## 2             C12orf39    IPI00306642.3      IPI00798034.2_s280.s281
## 3             IAPP       IPI00026219.4      IPI00304935.6_s6
## 4             IL4I1      IPI00915008.1      IPI00008422.5_s239.s242
## 5             GTF2H2D    IPI00607808.2 IPI00008422.5_s239.s242.s245

topVar(mcoin, axis = 2, end = "neg", topN = 5)

##   ax2_mrna_negative ax2_protein_negative ax2_phospho_negative
## 1             OLIG1      IPI00442171.4   IPI00292975.4_s1012
## 2             HLA.DQA1    IPI00444452.3   IPI00871890.1_s16
## 3             OR7D2      IPI00843802.2   IPI00217467.3_s104
## 4             OLIG2      IPI00218271.5   IPI00657687.1_s182
## 5             PSTPIP1    IPI00828098.2   IPI00011913.1_s188

topVar(mcoin, axis = 2, end = "pos", topN = 5)

##   ax2_mrna_positive ax2_protein_positive ax2_phospho_positive
## 1             ZNF560      IPI00306406.4   IPI00010800.2_s352
## 2             VWA5B1      IPI00472164.2   IPI00292059.2_s240
## 3             CSH1       IPI00549381.5   IPI00640088.1_s228
## 4             CYP2E1      IPI00386731.3   IPI00337315.1_s873
## 5             FGA        IPI00183002.6   IPI00005978.8_y23

```

To reveal the biological meaning of different components, additional methods, including gene set enrichment analysis (GSEA), may be applied to further analyze the selected genes on each component.

### 3 Session Info

```
toLatex(sessionInfo())
```

- R version 3.2.2 (2015-08-14), x86\_64-pc-linux-gnu
- Locale: LC\_CTYPE=en\_US.UTF-8, LC\_NUMERIC=C, LC\_TIME=en\_US.UTF-8, LC\_COLLATE=en\_US.UTF-8, LC\_MONETARY=en\_US.UTF-8, LC\_MESSAGES=en\_US.UTF-8, LC\_PAPER=en\_US.UTF-8, LC\_NAME=C, LC\_ADDRESS=C, LC\_TELEPHONE=C, LC\_MEASUREMENT=en\_US.UTF-8, LC\_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: ade4 1.7-2, gplots 2.17.0, knitr 1.11, made4 1.44.0, omicade4 1.10.0, RColorBrewer 1.1-2, scatterplot3d 0.3-36
- Loaded via a namespace (and not attached): BiocStyle 1.8.0, bitops 1.0-6, caTools 1.17.1, codetools 0.2-14, digest 0.6.8, evaluate 0.8, formatR 1.2.1, gdata 2.17.0, gtools 3.5.0, highr 0.5.1, KernSmooth 2.23-15, magrittr 1.5, stringi 1.0-1, stringr 1.0.0, tcltk 3.2.2, tools 3.2.2

### References

1. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M (2001) Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA* 98:10781–10786
2. Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. In: *Pacific Symposium on Bio-computing*, pp 455–466
3. Culhane AC, Perriere G, Higgins DG (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* 21(4):59
4. Meng C, Kuster B, Culhane A, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 29(15):162
5. Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshw Biol* 31:277–294
6. Culhane AC, Fagan A, Higgins DG (2007) A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* 7:2162–2171
7. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y (2012) Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 72(14):3499–511
8. Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* 4:609–620
9. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V, Stewart R, Thomson JA, Coon JJ (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* 8:821–827

Statistical Genomics

Methods and Protocols

Mathé, E.; Davis, S. (Eds.)

2016, XI, 418 p. 113 illus., 85 illus. in color., Hardcover

ISBN: 978-1-4939-3576-5

A product of Humana Press