Searching for best lower dimensional visualization angles for high dimensional RNA-Seq data

Wanli Zhang Corresp., [1] , **Yanming Di** [1]

[1] Statistics, Oregon State University, Corvallis, Oregon, United States

Corresponding Author: Wanli Zhang
Email address: zhangwa@stat.oregonstate.edu

The accumulation of RNA-Seq gene expression data in recent years has resulted in large and complex data sets of high dimensions. Exploratory analysis, including data mining and visualization, reveals hidden patterns and potential outliers in such data, but is often challenged by the high dimensional nature of the data. The scatterplot matrix is a commonly used tool for visualizing multivariate data, and allows us to view multiple bivariate relationships simultaneously. However, the scatterplot matrix becomes less effective for high dimensional data because the number of bivariate displays increases quadratically with data dimensionality. In this study, we introduce a selection criterion for each bivariate scatterplot and design/implement an algorithm that automatically scan and rank all possible scatterplots, with the goal of identifying the plots in which separation between two pre-defined groups is maximized. By applying our method to a multi-experiment *Arabidopsis* RNA-Seq data set, we were able to successfully pinpoint the visualization angles where genes from two biological pathways are the most separated, as well as identify potential outliers.

Searching for best lower dimensional visualization angles for high dimensional RNA-Seq data

Wanli Zhang[1¶] and Yanming Di[1¶]

[1]Department of Statistics, Oregon State University,
Corvallis, Oregon, United States of America

¶ These authors contributed equally to this work.

Abstract

The accumulation of RNA-Seq gene expression data in recent years has resulted in large and complex data sets of high dimensions. Exploratory analysis, including data mining and visualization, reveals hidden patterns and potential outliers in such data, but is often challenged by the high dimensional nature of the data. The scatterplot matrix is a commonly used tool for visualizing multivariate data, and allows us to view multiple bivariate relationships simultaneously. However, the scatterplot matrix becomes less effective for high dimensional data because the number of bivariate displays increases quadratically with data dimensionality. In this study, we introduce a selection criterion for each bivariate scatterplot and design/implement an algorithm that automatically scan and rank all possible scatterplots, with the goal of identifying the plots in which separation between two pre-defined groups is maximized. By applying our method to a multi-experiment *Arabidopsis* RNA-Seq data set, we were able to successfully pinpoint the visualization angles where genes from two biological pathways are the most separated, as well as identify potential outliers.

1. INTRODUCTION

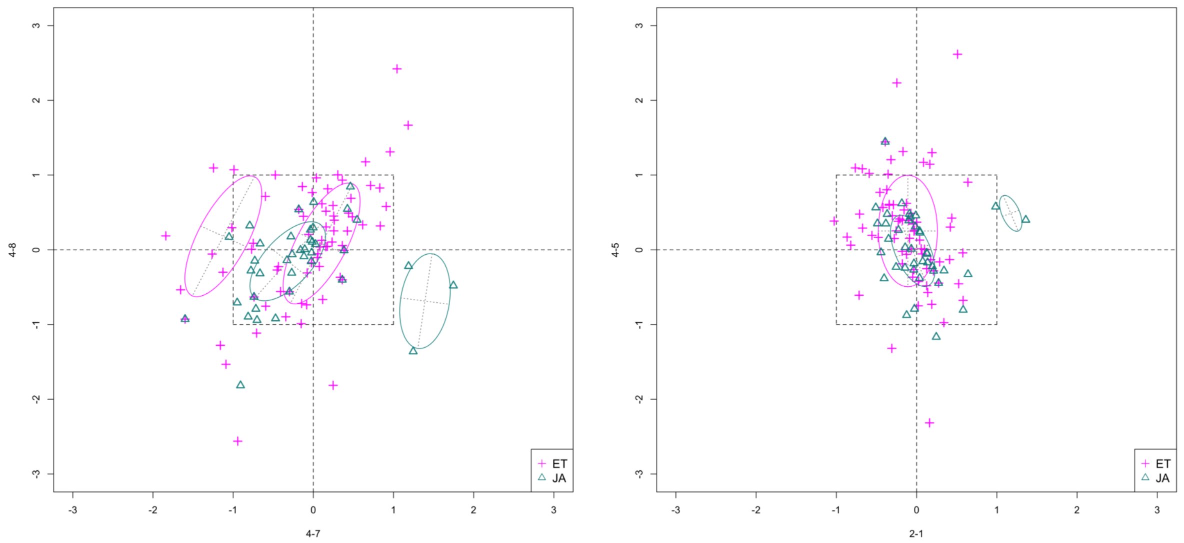
High throughput RNA sequencing (RNA-Seq) has been widely adopted for quantifying relative gene expression in comparative transcriptome analysis. In recent years, the increasing number of RNA-seq studies on the model plant *Arabidopsis thaliana* have resulted in an ever-accumulating amount of data from multiple RNA-Seq experiments. In this article, we will develop tools for the exploration and visualization of such multi-experiment data.

For examining treatment effects of individual genes under multiple conditions and across multiple experiments, a vector summarizing the differential expression (DE) results under different treatment conditions seems adequate. To visualize the DE profile under different treatments, a line plot can be used. However, since genes work interactively in all biological processes, it is of interest to examine expression patterns of groups of genes, through which the genes' biological context can be better understood. In light of this, researchers often would like to both identify the general trend and pinpoint individual aberrations in the expression profile of genes belonging to the same biological pathway, as well as compare the profiles between multiple pathways.

When multiple genes are being examined together, the line plots are less effective for visualizing DE or expression profiles: The lines often cross each other, making it difficult to identify the grouping and understand the behavior of individual genes. One common alternative visualization method is the scatterplot, which shows expression level under two treatment conditions at a time. Scatterplots are effective in showing clustering patterns and outliers, greatly assisting with data exploration (Elmqvist et al., 2008). For high dimensional data, one has the option of using the scatterplot matrix, in which each panel is the scatterplot for the corresponding pairs of features. However, manual scanning of all possible pairwise scatterplots can be arduous or even fruitless at times, because the number of possible visualization angles increases quadratically with respect to data dimensionality (p choose 2 possible angles).

In this paper, we propose to automatically search for the best low dimensional visualization angles (2-, 3-, or 4-dimensional) based on a context-sensitive, numeric measure of importance, thereby reducing the amount of effort invested in scatterplot scanning. In our study, we hope to explore the patterns and differences in gene expression profile between two phytohormone signaling pathways, and therefore, we would like the top ranked scatterplots to contain as much information as possible on pathway classification. We thus define such an importance measure for the dimension subsets such that the scatterplots will show the largest separation between different pre-defined groups in the data set.

For this study, we will look for feature subsets upon which the pathways ethylene (ET) and jasmonate (JA) are the most separated, and quantify the between-group separation by calculating the repeated cross-validation (RCV) error of misclassification using MclustDA (Fraley and Raftery, 2002), a model-based classification method. In Figure 1, we show one of the top ranked 2-subset feature combinations that give the greatest separation between two pathways, as well as subset giving the smallest separation. Comparing the two scatterplots, we can observe that the two pathway groups in Figure 1(a) are more visually distinguishable than those in Figure 1.



(a) Group Separation = 0.7082

(b) Group Separation = 0.5355

Figure 1: Scatterplots of 2-dim feature subsets reflecting maximum and minimum group separations. Dashed-line square marks ±1 range from the origin. Different classes distinguished with color. Ellipses correspond to component mean and covariance fitted by MclustDA. Treatment i-j represents the jth treatment in experiment i.

The rest of the paper is formatted as follows: Section 2 outlines the collection and processing of the data and information on the experiments and biological pathways. The statistical methods are described in Section 3. In Section 4, we list the results obtained by applying our method to the collected data. Finally, we state our conclusion and discuss limitations and possibilities for future work in Section 5. Additional proofs and graphs are included in the Appendix.

2. DATA DESCRIPTION AND PROCESSING

2.1 Collecting experimental data

In this study, we use the data collected and processed by Bin Zhuo (Zhuo et al., 2016). All 5 datasets were acquired from the National Center for Biotechnology Information (NCBI) website `www.ncbi.nlm.nih.gov` and processed through a customized assembly pipeline to obtain a matrix of counts for genes in samples. All datasets originate from RNA-Seq experiments on the model plant *Arabidopsis thaliana*, with treatment conditions (including genetic variants) varying between experiments. All experiments were conducted on the leaf tissue. The number of treatments/factor levels also vary among the experiments. The GEO (Gene Expression Omnibus) accession numbers (which can be used to directly search for the experiment/dataset information) are available as part of the meta-data, and the assembly pipeline is described in detail in Zhuo et al. (2016). We have included the basic information on the experiments in Table 1.

ID	**GEO accession #**	**Title**
1	GSE36626	Dynamic Deposition of the Histone H3.3 Variant Accompanies Developmental Remodeling of Arabidopsis Transcriptome (mRNA-Seq)
2	GSE39463	Time-course RNA-seq analysis of the barley MLA1 immune receptor-mediated response to barley powdery mildew fungus Bgh in Arabidopsis thaliana
3	GSE48235	Four distinct types of dehydration stress memory genes in Arabidopsis thaliana
4	GSE51304	Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis
5	GSE54677	Transcriptional gene silencing by Arabidopsis Microrchidia homologues involves the formation of heteromers

Table 1: Experiment information

For each of the 5 experiments, a negative binomial regression model is fitted to the normalized counts, where the normalization factors are computed using the 104 genes shared by top 1000 most stably expressed genes in three tissue groups (Zhuo et al., 2016). After removing the columns corresponding to the baseline expression levels of the control groups, the resulting matrix summarizes the log (base 2) fold changes under different treatments across the 5 experiments: Each column represents the log fold changes of gene expression between one treatment group (or a gene knockout mutant) and the control group (or wildtype) in one of the experiments.

2.2 Finding pathway genes

For this study, we focus our attention on the signaling pathways of two phytohormones: ethylene (ET) and jasmonic acid (JA). As a plant hormone, ethylene is commercially important due to its regulation on fruit ripening (Lin et al., 2009). JA acts as a key cellular signal involved in the activation of immune responses to most insect herbivores and necrotrophic microorganisms (Ballaré, 2010).

For each pathway, we first use AmiGO 2 (`http://amigo.geneontology.org/amigo/landing`) to search for the list of genes involved, and then identify the subset of genes in our data set that belong to the pathway through cross-reference. Genes with a fold change of < 2 under all treatment conditions are filtered out. The name, GO accession number, and the number of genes in each pathway are listed in Table 2.

ID	Pathway name	GO accession #	# Genes
ET	Ethylene-activated signaling pathway	GO: 0009873	86
JA	Jasmonic acid mediated signaling pathway	GO: 0009867	48

Table 2: Pathway information

In Figure 2, we display the expression profile of genes that belong to each pathway group. Under certain individual treatment-control contrasts (e.g. 2-3, 4-3, 5-1), there exist observable similarities between the distribution of expression levels, while it is more difficult to tell under other treatments.

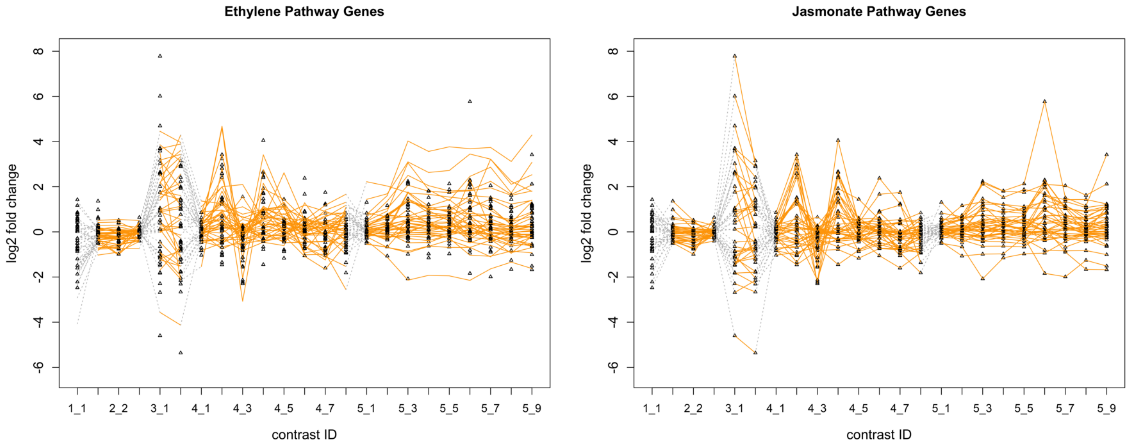


Figure 2: Gene expression profile plot for pathways ET and JA. Treatments from the same experiment are joined by orange lines. Different experiments are joined by grey dashed lines. Feature i-j represents the jth treatment-control contrast in experiment i.

3. METHOD

3.1 Mixture discriminant analysis via MclustDA

In this section, we will start by introducing a classification method named MclustDA, and then define a measure for group separation using cross-validation results with MclustDA. Finally, we lay out our strategy for reducing data dimensionality with the ultimate goal of simplifying navigation of scatterplots.

MclustDA model In discriminant analysis (DA), known classifications of some observations are used to classify others. The number of classes, G, is assumed to be known. For probabilistic DA methods, it is assumed that observations in class k follow a class specific probability distribution $f_k(\cdot)$. Let τ_k represent the proportion of observations in class k. According to Bayes's theorem, it follows that

$$P(\boldsymbol{y} \in \text{class } j) = \frac{\tau_j f_j(\boldsymbol{y})}{\sum_{k=1}^{G} \tau_k f_k(\boldsymbol{y})},$$

where observation $\boldsymbol{y}$ is assigned to the most probable class.

Commonly used DA methods, including Fisher's linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), assume a multivariate normal density for each class:

$$f_k(\boldsymbol{y}) = \phi(\boldsymbol{y}|\mu_k, \Sigma_k).$$

The method is called LDA if the covariance matrices for all classes coincide ($\Sigma_k = \Sigma$ for $k = 1, ..., G$), and is called QDA if the class covariances are allowed to vary.

MclustDA (Fraley and Raftery, 2002), an extension and generalization to LDA and QDA, models each class density as a mixture of multivariate normals. The density for class j is as follows:

$$f_j(\boldsymbol{y}|\theta_k) = \sum_{k=1}^{G_j} \tau_{jk}\phi(\boldsymbol{y}|\mu_{jk}, \Sigma_{jk}),$$

where G_j is the number of components for class j, $\{\tau_{jk}\}$ are mixing proportions for components in class j, and θ_k is the vector of parameters for the normal mixture. Component covariances Σ_{jk} are allowed to vary both within and between classes.

Parameters within each class are separately estimated by maximum likelihood via the EM algorithm (Dempster et al., 1977), which is equivalent to fitting a Mclust (Fraley and Raftery, 2002) model for each class. And just like Mclust, MclustDA performs model selection within each class for the number of mixture components as well as covariance matrix parameterizations with Bayesian information criterion (Schwarz, 1978).

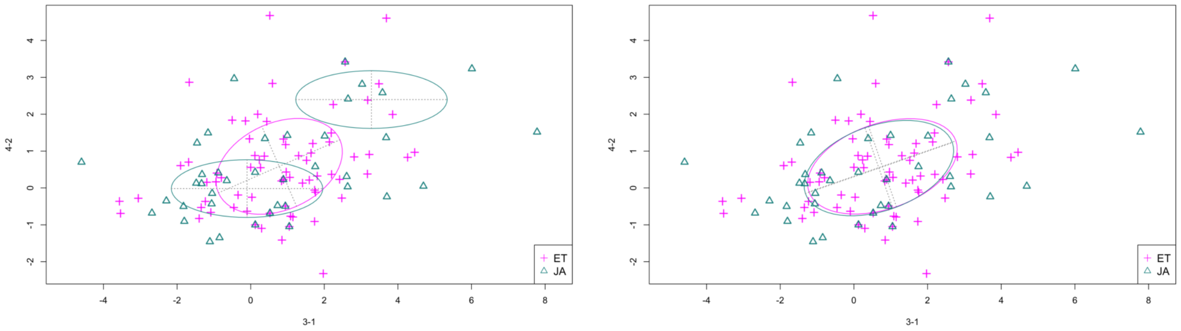
Comparison with LDA In our study, MclustDA is chosen over LDA/QDA as the classifier due to its greater flexibility in describing the data. In RNA-Seq analysis, we typically assume that the majority of genes are not differentially expressed, and therefore we expect to see a cluster of points around the origin. Since MclustDA proposes to fit more than one normal component to each class, it readily captures the cluster of non-DE genes as well as any abnormalities that might be of interest.

In Figure 3, we fitted a MclustDA model and a LDA model on dimensions [3-1, 4-2] of our data, separately. In MclustDA fit, each class is described with a mixture of two bivariate normal components, with the ellipses representing fitted covariance matrix estimates. For details in how the ellipses are constructed, see Appendix A.

Class JA is fitted with a component centered near the origin, representing genes with low expression levels under both treatments, as well as a component centered at

$(2.276, 1.663)$ that encompasses relatively active genes. Class ET is represented by a single normal component centered at $(0.537, 0.406)$.

In comparison, due to model assumptions, LDA fitted a bivariate normal density to each class with covariances being equal, and in this case, the estimated centers almost coincide with each other. The fitted normal densities are only able to capture the general shape and orientation of each class, while MclustDA provides us with a more detailed anatomy of geometric and distributional properties in each class.



(a) MclustDA fit with 1 and 2 components in each class

(b) LDA model assuming equal covariance matrix for each class

Figure 3: Comparison of MclustDA and LDA fit of the same data. Fitted components and points from different classes are distinguished with color. Ellipses correspond to component covariances.

3.2 Quantification of group separation

Our definition of group separation measure is motivated by the relationship between visualized separation and misclassification probability (from a model-based classifier).

Suppose we wish to separate two populations π_1 and π_2. Let $X = [X_1, ..., X_p]$ denote the p-dimensional measurement vector of an observation. We assume that densities $f_1(x)$ and $f_2(x)$ describe the variability of the two populations. Let p_1 and p_2 denote prior probability of each population. Define $c(1|2)$ and $c(2|1)$ as costs of misclassifying an object from class 2(1) as class 1(2). Here we let $c(1|2) = c(2|1) = 1$ to simplify the formulation. Let Ω denote the entire sample space, and $\Omega = R_1 \cup R_2$, where R_1 is the set of values of x for which we classify objects into π_1, and $R_2 = \Omega - R_1$.

The probability of misclassifying an object from π_1 as π_2 is:

$$P(2|1) = P(X \in R_2|\pi_1) = \int_{R_2} f_1(x)\mathrm{d}x,$$

and similarly, we have

$$P(1|2) = P(X \in R_1|\pi_2) = \int_{R_1} f_2(x)\mathrm{d}x.$$

By definition, we can calculate the probability of misclassifying any object:

$$P(\text{misclassified as } \pi_1) = P(X \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2,$$
$$P(\text{misclassified as } \pi_2) = P(X \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1.$$

The Total Probability of Misclassification (TPM) is defined as the probability of either misclassifying a π_1 object or misclassifying a π_2 object, i.e.

$$\text{TPM} = p_1P(2|1) + p_2P(1|2). \tag{1}$$

Suppose $Y = \{Y_1, ..., Y_{N_1}\} \sim \pi_1$ and $Z = \{Z_1, ..., Z_{N_2}\} \sim \pi_2$ are two i.i.d samples from the two populations. Assume that a classification system has been trained and tested on this data set, and results in the following confusion matrix in Table 3:

		Predicted Class		
		π_1	π_2	**total**
Actual Class	π_1	$n_{1\mid 1}$	$n_{2\mid 1}$	N_1
	π_2	$n_{1\mid 2}$	$n_{2\mid 2}$	N_2
	total	N'_1	N'_2	

Table 3: Confusion matrix

Then the misclassification error rate (MER), i.e. probability of misclassifying any object, is given by:

$$\text{MER} = \frac{n_{1|2} + n_{2|1}}{N_1 + N_2} = \frac{n_{1|2}}{N_2} \cdot \frac{N_2}{N_1 + N_2} + \frac{n_{2|1}}{N_1} \cdot \frac{N_1}{N_1 + N_2}. \tag{2}$$

Under the assumption that each object is independently classified, the number of misclassified π_1 objects, $N_{2|1}$, follows a Binomial distribution with parameters $(N_1, P(2|1))$. Likewise, the number of misclassified π_2 objects, $N_{1|2}$, follows a Binomial distribution with parameters $(N_2, P(1|2))$. The maximum likelihood (ML) estimators for $P(2|1)$ and $P(1|2)$ can be easily computed:

$$\widehat{P(2|1)} = \frac{n_{2|1}}{N_1}; \quad \widehat{P(1|2)} = \frac{n_{1|2}}{N_2}.$$

Now, if we set $p_1 = N_1/(N_1 + N_2)$ and $p_2 = N_2/(N_1 + N_2)$ as prior probabilities for π_1 and π_2, then under independence assumption, it follows that

$$\text{MER} = p_1\widehat{P(2|1)} + p_2\widehat{P(1|2)},$$

that is, MER is a maximum likelihood, and hence consistent, estimate of TPM.

In practice, however, the MER tends to underestimate TPM because the same data has been used for both training and testing. In this study, we use cross-validation to address this issue.

Repeated stratified cross-validation One of the most commonly used method to estimate the expected error rate is cross-validation (CV). For a K-fold CV, the original data is randomly split into K equally sized subsamples, of which $K-1$ (training set) are used to train a classifier and the remaining one (validation set) is used to test the trained classifier. For a binary classification problem, the misclassification error rate (MER), as defined in (2), is typically computed using the validation set as a performance measure for the classifier. The training-validation process is iterated over all K folds, each time using a different subsample as validation set, and the resulting K MER values are averaged. In stratified cross-validation, the folds are selected so that they contain approximately the same proportion of classes as the original data. It has been shown in previous studies that stratified CV tends to perform uniformly better than CV, in terms of both bias and variance (Kohavi, 1995).

Due to the randomness in partitioning the sample into K folds, we have introduced variation into the K-fold CV estimator. One way to reduce this variation is to repeat the whole cross-validation process multiple times using different pseudorandom allocations of instances to training and validation folds for each repetition (Kim, 2009), and report the average of CV estimators across all repetitions. This method is often referred to as the repeated cross-validation (RCV). For improved repeatability of results, common seeding has been recommended in earlier studies (Powers and Atyabi, 2012). In our implementation, we set a fixed random number seed for each repetition of CV.

Let $C \times K$-CV denote a K-fold CV with C repetitions. There has been much discussion on the optimal choice of C and K (Kohavi, 1995; Kim, 2009; Powers and Atyabi, 2012). Increasing C tends to decrease the variance of the RCV estimator, but at the same time increases the computational time. The choice of K takes into account the tradeoff between bias and variance of the CV estimator (of the expected error rate). For small K, less data is used to train the classifier and therefore the error estimate tends to be biased. For large K, the estimator becomes less biased due to more data being used in training, but its variance is inflated due to higher correlation between different training folds. Kohavi (1995) recommends using a stratified 10-fold CV with multiple runs, and we chose $C = 10$ considering the amount of computation required as well as the specs of our hardware.

Quantify group separation We define the group separation index (GSI) as

$$\mathrm{GSI} = 1 - \hat{\epsilon}_{\mathrm{rcv}}, \tag{3}$$

where $\hat{\epsilon}_{\mathrm{rcv}}$ denotes the repeated stratified CV estimator of the total misclassification probability using MclustDA as the classifier.

Intuitively, for a chosen feature subset, a small CV error indicates that the data can be more easily classified when projected onto these dimensions, which, in our expectation, can be reflected in the graphical representation of the data by showing that different classes can be more easily distinguished through simple visualization.

3.3 Feature subset selection via GSI ranking

In this section, we describe the data in each pathway with a low dimensional representation for easier interpretation by selecting a parsimonious subset of features (treatments) that contain as much information on pathway classification/separation as possible. In other words, we hope to find the dimensions to project the data onto such that the separation between two pathways is as large as possible. We use GSI, as defined in (3), to measure the separation between two pathway groups.

In order to find the optimal feature subset in terms of group separation, we designed and implemented the following algorithm:

Step 1: Determine the number of features M to keep. Choose M from $\{2, 3, 4\}$.

Step 2: List all M-subsets of features exhaustively. Call this collection of subsets $\mathcal{F}_M$.

Step 3: For each member of $\mathcal{F}_M$, subset the data accordingly. Calculate and record a 10×10 stratified CV error rate (and equivalently, GSI) with MclustDA as classifier on each subsetted data. For each fold of CV, use misclassification error rate as measure of fit.

- CV model fitting: First fit a MclustDA model to the entire subsetted data, setting maximum number of components as 2. Then use the same fitted model (number of components, parameterization) for every fold of CV.

Step 4: Rank the feature subsets in $\mathcal{F}_M$ according to their GSI values. Feature subsets with higher GSI values are ranked higher.

Step 5: Repeat above steps for other values of M.

For the purpose of finding "good" angles for data visualization, we will examine the scatterplots and scatterplot matrices generated by top-ranked feature subsets. The results will be discussed in Section 4.

Random number seed To ensure reproducibility of our results, for each of 2-, 3- and 4-subset selection process, we followed the following protocol to set random number seeds:

Step 1: Choose a list of 50 random number seeds. Partition the list into 5 batches of 10 seeds.

Step 2: For each feature subset, run 10-fold stratified CV for 50 times, each time using a different seed from the list.

Step 3: Average results within each of 5 batches of 10 random seeds to obtain 10×10 stratified CV result. For instance, average of seeds 1∼10 results serves as first run of 10×10 RCV; average of seeds 11∼20 serves as second run, etc.

3.4 Dimension reduction via PCA

Principal component analysis (PCA) maps the data onto a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. As a dimension reduction technique, usually only the first few principal components (PCs) are used. Despite its popularity in the field of data visualization, the formulation of PCA does not involve any class information in the data, which implies that the projected directions corresponding to the largest variance may not contain the best separability information.

To verify this observation, using the expression data from all 5 experiments, we calculated its principal components, and treat them as the new (projected) features. Then for the first 2, 3 and 4 PCs, respectively, we calculated the group separation index for each case using 10×10-CV with MclustDA and compare the results with ours.

4. RESULTS

4.1 Repeated cross-validation with MclustDA

With the secondary purpose of testing the stability of repeated CV, we executed multiple runs for each of the 2-, 3-, and 4-subset feature selection procedures. The top ranked feature subsets as well as their corresponding GSI values are presented in Tables 4∼6.

Table 4: Top ranked 2-subsets from multiple runs of 10×10 RCV. Ties are marked with asterisk ($*$). Combinations appearing in all 5 runs are highlighted with distinguishing colors.

(a) Run 1

Rank	Subset	GSI
1	[4-7, 4-8]	0.698
2	[3-1, 3-2]	0.694
3	[4-7, 5-2]	0.688
4	[3-1, 5-2]	0.681
5	[2-1, 4-7]	0.680

(b) Run 2

Rank	Subset	GSI
1	[4-7, 4-8]	0.703
2	[3-1, 3-2]	0.696
3	[4-7, 5-2]	0.690
4	[3-1, 5-8]	0.682
5	[3-1, 5-2]	0.681

(c) Run 3

Rank	Subset	GSI
1$*$	[3-1, 3-2]	0.697
2$*$	[4-7, 5-2]	0.697
3	[4-7, 4-8]	0.689
4	[3-1, 5-2]	0.685
5	[2-1, 4-7]	0.675

(d) Run 4

Rank	Subset	GSI
1	[3-1, 3-2]	0.704
2	[4-7, 4-8]	0.695
3	[3-1, 5-2]	0.687
4	[3-1, 5-8]	0.684
5	[4-7, 5-2]	0.681

(e) Run 5

Rank	Subset	GSI
1	[4-7, 4-8]	0.708
2	[3-1, 3-2]	0.702
3	[4-7, 5-2]	0.689
4	[4-4, 5-2]	0.688
5	[3-1, 5-2]	0.683

Table 5: Top ranked 3-subsets from multiple runs of 10 × 10 RCV. Ties are marked with asterisk (∗). Combinations appearing in all 5 runs are highlighted with distinguishing colors.

(a) Run 1

Rank	Subset	GSI
1	[3-2, 4-7, 5-2]	0.724
2	[3-2, 5-2, 5-6]	0.720
3	[1-1, 3-1, 5-2]	0.718
4	[3-1, 5-2, 5-6]	0.710
5	[2-2, 2-3, 3-1]	0.707

(b) Run 2

Rank	Subset	GSI
1∗	[1-1, 3-1, 5-2]	0.717
2∗	[3-2, 4-7, 5-2]	0.717
3	[3-2, 5-2, 5-6]	0.713
4	[1-1, 2-2, 3-1]	0.708
5	[2-2, 2-3, 3-1]	0.708

(c) Run 3

Rank	Subset	GSI
1	[3-2, 5-2, 5-6]	0.725
2	[1-1, 3-1, 5-2]	0.717
3∗	[3-1, 5-2, 5-9]	0.715
4∗	[3-2, 4-7, 5-2]	0.715
5	[3-1, 4-7, 5-2]	0.714

(d) Run 4

Rank	Subset	GSI
1	[3-2, 5-2, 5-6]	0.736
2	[1-1, 3-1, 5-2]	0.728
3	[2-2, 2-3, 3-1]	0.713
4∗	[3-1, 4-7, 5-2]	0.712
5∗	[3-2, 4-7, 5-2]	0.712

(e) Run 5

Rank	Subset	GSI
1	[1-1, 3-1, 5-2]	0.726
2	[3-2, 4-7, 5-2]	0.724
3	[3-2, 5-2, 5-6]	0.720
4	[2-3, 3-1, 5-2]	0.715
5	[2-2, 2-3, 3-1]	0.710

Table 6: Top ranked 4-subsets from multiple runs of 10 × 10 RCV. Ties are marked with asterisk (∗). Combinations appearing in all 5 runs are highlighted with distinguishing colors.

(a) Run 1

Rank	Subset	GSI
1	[1-1, 4-2, 4-4, 5-2]	0.730
2	[4-4, 4-5, 4-7, 5-8]	0.728
3	[3-1, 4-4, 4-5, 5-2]	0.727
4	[4-1, 4-7, 4-8, 5-1]	0.725
5	[3-2, 4-5, 4-7, 5-7]	0.724

(b) Run 2

Rank	Subset	GSI
1	[1-1, 4-2, 4-4, 5-2]	0.745
2	[3-1, 4-4, 4-5, 5-2]	0.741
3	[2-3, 3-1, 4-4, 4-5]	0.734
4	[2-3, 3-1, 4-7, 5-2]	0.725
5	[3-2, 4-5, 4-7, 5-7]	0.724

(c) Run 3

Rank	Subset	GSI
1	[1-1, 4-2, 4-4, 5-2]	0.735
2	[4-4, 4-5, 4-7, 5-8]	0.731
3∗	[3-2, 4-5, 4-7, 5-7]	0.726
4∗	[4-1, 4-7, 4-8, 5-1]	0.726
5	[3-1, 4-4, 4-5, 5-2]	0.725

(d) Run 4

Rank	Subset	GSI
1	[1-1, 4-2, 4-4, 5-2]	0.739
2∗	[3-1, 4-4, 4-5, 5-2]	0.731
3∗	[3-2, 4-5, 4-7, 5-7]	0.731
4	[4-4, 4-5, 4-7, 5-8]	0.723
5	[3-1, 3-2, 4-7, 5-2]	0.722

(e) Run 5

Rank	Subset	GSI
1	[1-1, 4-2, 4-4, 5-2]	0.740
2	[3-1, 4-4, 4-5, 5-2]	0.735
3∗	[3-2, 4-5, 4-7, 5-7]	0.730
4∗	[2-2, 2-3, 3-1, 4-7]	0.730
5	[3-2, 4-6, 4-7, 5-6]	0.723

4.1.1 Stability of RCV model selection results

Although the top ranked feature subsets sometimes differ between multiple RCV runs, we are still able to observe high degree of overlap between the results:

For 4-subset (Table 6), [1-1, 4-2, 4-4, 5-2], [3-1, 4-4, 4-5, 5-2] and [3-2, 4-5, 4-7, 5-7] are among top ranked feature combinations in all 5 runs.

For 3-subset (Table 5), feature combinations [3-2, 4-7, 5-2], [3-2, 5-2, 5-6] and [1-1, 3-1, 5-2] are ranked top for all 5 runs.

For 2-subset (Table 4), [4-7, 4-8], [3-1, 3-2], [4-7, 5-2] and [3-1, 5-2] are among top ranked feature combinations for all runs.

4.1.2 Top Ranked Scatterplot: Same Experiment

In Figure 4, we show the scatterplot of the data projected onto dimensions [4-7, 4-8], one of the top ranked 2-subset feature combinations. These two features originate from the same experiment.

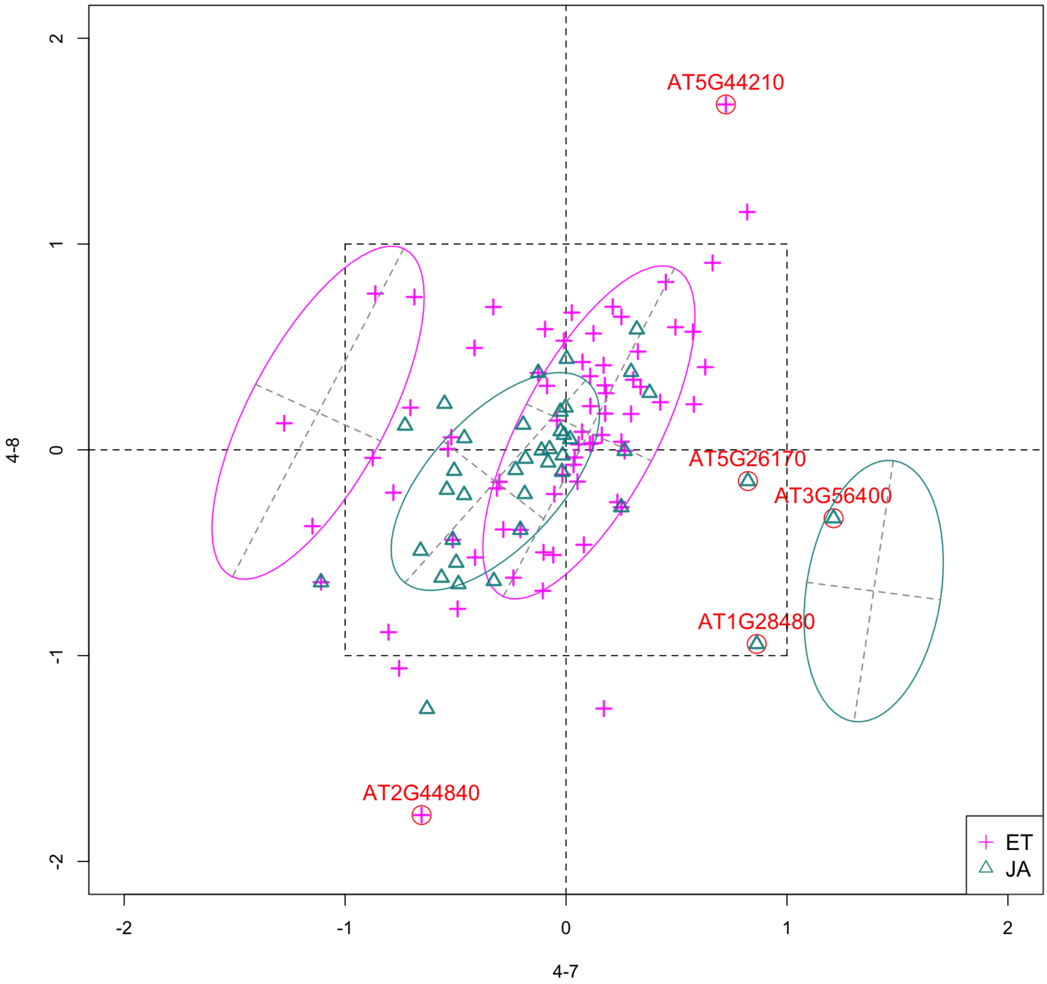


Figure 4: Scatterplot of data projected on dimensions 4-7 and 4-8. Pathways are distinguished with color. Ellipses represent estimated covariances fitted by MclustDA. Potential outliers highlighted and labeled with their names. Dashed-line square is $\pm\log(2)$ range from the origin.

Experiment 4 Since both features originate from the same experiment, we will focus on the context of this experiment and first present some background information. The purpose of Experiment 4 is to characterize non-CG methylation and its interaction with histone methylation in *Arabidopsis thaliana*(Stroud et al., 2014). Non-CG methylation is a category of DNA methylation, where methyl groups are added to the DNA molecule, altering its chemical structure and thereby changing its activity. DNA methylation is usually catalyzed by DNA methyltransferases (MTases), which transfer and covalently bind methyl groups to DNA. In *Arabidopsis*, the principal DNA MTases include chromomethylase (CMT) and domains rearranged MTase (DRM) proteins, in particular CMT3 and DRM2. Expression of DRM1 is scarcely detected, while the function of CMT2 has not been studied as well as that of CMT3.

Histone methylation is a process by which methyl groups are transferred to amino acids of histone proteins. Histone methylation can either increase or decrease gene transcription, depending on which amino acids are methylated and the degree of methylation. The methylation process is most commonly observed on lysine residues (K) of histone tails H3 and H4, among which H3K9 (lysine residue at 9th position on H3) serves as a common site for gene inactivation. Lysine methylation requires a specific MTase, usually containing an evolutionarily conserved SET domain. In *Arabidopsis*, Su(var)3-9 homologue 4 (SUVH 4), SUVH 5 and SUVH 6 are the major H3K9 MTases.

Feature 4-7 corresponds to the *drm1 drm2 cmt2 cmt3* quadruple gene knockout mutant, created by crossing *cmt2* to *cmt3* and *drm1 drm2* double mutants. It was found that non-CG methylation was eliminated in such mutants, indicating that DRM1, DRM2, CMT2 and CMT3 proteins are collectively responsible for all non-CG methylation in *Arabidopsis*. Feature 4-8 corresponds to the *suvh4 suvh5 suvh6* triple mutant. The control group of this experiment corresponds to wildtype *Arabidopsis*. Table 7 summarizes the above information.

Feature ID	**Sample GEO accession #**	**Description**
4-0 (control)	GSM1242374, GSM1242375	Wildtype
4-7	GSM1242388, GSM1242389	*drm1 drm2 cmt2 cmt3* quadruple mutant
4-8	GSM1242390, GSM1242391	*suvh4 suvh5 suvh6* triple mutant

Table 7: Feature information

Outliers Potential outliers from JA pathway, as highlighted and labeled in the scatterplot, fall into the fourth quadrant, which implicates that these genes are up-regulated under 4-7 (DNA methylation) but down-regulated under 4-8 (histone methylation). Information on these genes is collected from TAIR (Berardini et al., 2015) and displayed in Table 8. One interesting discovery we made was that one of the outliers, **AT3G56400**, functions as a repressor of JA-regulated genes. In other words, its gene product inhibits the expression of other genes related to JA regulation.

Gene name	Description
AT5G44210	encodes a member of the ERF (ethylene response factor) subfamily B-1 of ERF/AP2 transcription factor family (ATERF-9)
AT2G44840	Same function as AT5G44210; Cell-to-cell mobile mRNA
AT5G26170	WRKY Transcription Factor, Group II-c; Involved in jasmonic acid inducible defense responses.

AT3G56400	WRKY Transcription Factor, Group III;
	Repressor of JA-regulated genes;
	Activator of SA-dependent defense genes.

AT1G28400	GATA zinc finger protein

Table 8: Outlier information

Pattern differences The first thing we can observe from the scatterplot is that a majority of genes are expressed at a low level (with fold change < 2) under both treatment conditions, as demonstrated by the clustered points inside ± 1 square. Although most genes are expressed at a relatively low level, we are still able to identify the difference between the two pathways. If a differential expression (DE) analysis is performed and only DE genes are included in our model, it will be less likely for us to spot the same structural difference as before because we would lose much group level information by filtering out non-DE genes.

Secondly, not considering the outliers, genes belonging to the JA pathway are mostly concentrated around the origin as well as in quadrant III, meaning that most JA genes are down-regulated under both treatments. The expression pattern of ET pathway genes, however, is more diverse than that of JA genes. These genes populate all four quadrants of the coordinate system, with the highest density in quadrant I followed by quadrant II and III. That is, a majority of ET genes are up-regulated under both treatments, while most of the others are down-regulated under 4-7.

4.1.3 Top Ranked Scatterplot: Different Experiments

In Figure 5, we show the scatterplot of another top ranked feature combination, [3-1, 5-2], which come from two different experiments.

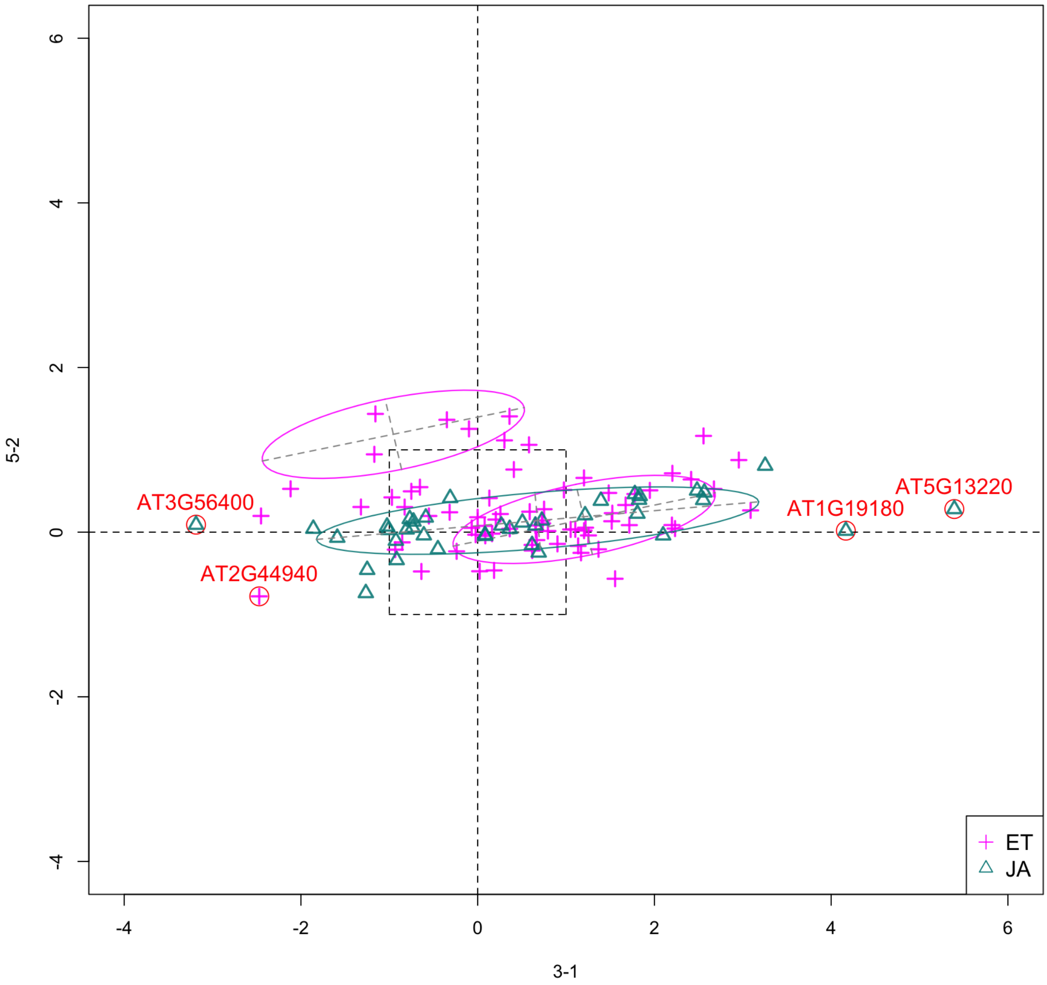


Figure 5: Scatterplot of data projected on dimensions 3-1 and 5-2. Pathways are distinguished with color. Ellipses represent estimated covariances fitted by MclustDA. Potential outliers highlighted and labeled with their names. Dashed-line square is $\pm\log(2)$ range from the origin.

Experiment 3 The focus of this study is the response of *Arabidopsis* to multiple consecutive dehydration stresses (Ding et al., 2013). Based on the observation that pre-exposure to abiotic stresses (including dehydration) may alter plants subsequent responses by improving resistance to future exposures, the researchers hypothesized the existence of "memory genes": genes that provide altered response to subsequent stresses (Ding et al., 2012).

A RNA-Seq study is performed to determine the transcriptional responses of *Arabidopsis* plants that have experienced multiple exposures to dehydration stress and compare them with the transcriptional behavior of plants encountering the stress for the first time. The dehydration treatments are applied in the following fashion:

(1) Plants were removed from soil and air-dried for 2h. Call this exposure Stress 1 (S1).

(2) Plants were then rehydrated for 22h by being placed in humid chambers with their roots in a few drops of water. Call this step Recovery 1 (R1).

(3) Air-dry R1 plants for 2h. This is Stress 2 (S2), followed by R2, which is the same as R1.

(4) Air-dry R2 plants for 2h. This is Stress 3 (S3).

RNA-Seq analyses were then performed on leave tissues from pre-stressed/watered plants (control), S1 plants and S3 plants. For each treatment group, plants from two independent biological samples were used. In our data, feature 3-1 corresponds to S1, or first drought stress. See Table 9 for a summary.

Experiment 5 In this study, the researchers examine the functional relationship between members of the *Arabidopsis* microrchidia (AtMORC) ATPase family (Moissiard et al., 2014), which have been shown to be involved in transposon repression and gene silencing. Three of seven MORC homologs were examined: AtMORC1, AtMORC2 and AtMORC6. RNA-Seq experiment using single and double mutants indicates that AtMORC1 and AtMORC2 act redundantly in gene silencing. Wildtype *Arabidopsis* was used as control group. Treatment groups include both single and double mutant lines: *atmorc2-1*, *atmorc2-4*, *atmorc1-2*, *atmorc1-5*, and *atmorc1-2 atmorc2-1*, in which two individual alleles were used for *atmorc1* and *atmorc2*. In our data, feature 5-2 corresponds to the single mutant line *atmorc2-1*. Table 9 includes summary information on this experiment.

Feature ID	Sample GEO accession #	Description
3-0 (control)	GSM1173202, GSM1173203	Watered condition
3-1	GSM1173204, GSM1173205	First drought stress
5-0 (control)	GSM1321694, GSM1321704	Wildtype
5-2	GSM1321696, GSM1321706	*atmorc2-1* mutant

Table 9: Feature information for experiments 3 and 5

Outliers In Figure 5, we highlighted a few observations considered as outlying, and as before, looked up their information using TAIR. A brief description for each outlier is included in Table 10. Gene **AT3G56400** is again identified as an outlier, mainly because of its highly negative expression level under treatment 3-1, while the near-zero expression level under 5-2 indicates its inactivity under this treatment. Gene **AT5G13220** has the highest expression level under 3-1 among all JA genes, and at the same time not as active under 5-2. This gene is interesting because it functions as a repressor of JA signaling, and its high expression level could be an implication for repression of JA singaling for *Arabidopsis* plants going through first drought stress (3-1).

Gene name	Description

AT3G56400	WRKY Transcription Factor, Group III;
	Repressor of JA-regulated genes;
	Activator of SA-dependent defense genes.

AT1G19180	a.k.a. JAZ1 Nuclear-localized protein involved in JA signaling; JAZ1 transcript levels rise in response to a jasmonate stimulus.
AT5G13220	a.k.a. JAS1, JAZ10 Repressor of JA signaling
AT2G44940	Integrase-type DNA-binding superfamily protein

Table 10: Outlier information for 3-1 and 5-2

Pattern differences From the scatterplot, the first thing we can observe is that quite a few genes from both pathways are up- or down-regulated under treatment 3-1, while genes are expressed at an overall low level under 5-2. Nevertheless, a few genes from ET group show overexpression pattern under 5-2. JA pathway genes populate quadrants I, II and III, while ET pathway genes are mainly located in quadrants I, II and IV. Overall, under 5-2, ET genes tends to be more active than JA genes.

In the previous two sections we singled out two of the top ranked scatterplots for discussion. Interested readers are directed to the appendix for additional scatterplots for top ranked feature subsets (Figures 6∼13).

4.2 GSI for PC transformed data

In Table 11, we report the GSI for PC transformed data, as well the maximum GSI achieved by subsets of the original data. The proportion of total variation explained is 66.5% for first 2 PCs, 78.2% for first 3, and 85.6% for first 4. Through comparison, we observe that using PCs as new features does not necessarily maximize the separation between the distinct groups in the data, therefore confirming our statement in Section 3.4.

Table 11: Separation index for PC transformed data and maximum GSI for original data

# of PCs	GSI	Features	max GSI achieved
2	0.638	2	0.708
3	0.642	3	0.736
4	0.639	4	0.745

5. CONCLUSION

Conclusion In this article, we defined a numeric measure for the separation between different groups of data, and used said measure to perform low dimensional feature subset selection in order to find the most interesting angles to visualize high dimensional data. By applying our method to a multi-experiment RNA-Seq data on *Arabidopsis* leave tissues, we found that the top ranked feature subsets did demonstrate some interesting differences in expression patterns between two biological pathways, which shows that our method can be a potentially powerful tool in the exploratory analysis of such high dimensional integrated/assembled data from various sources.

Significance Firstly, our method yields well documented results. We enumerated the group separation index for every low dimensional feature subset, and constructed the scatterplots/scatterplot matrices for each case. If scientists know beforehand which features are of interest, they will be able to directly access the corresponding entry in our result. Secondly, through the application of mixture discriminant analysis, we were able to summarize the expression pattern of groups of genes using a mixture of only a handful of normal components. Furthermore, using the fitted MclustDA ellipses as visual aid, we were able to clearly show the geometric structure of each group and make comparisons. Finally, as seen in Figure 4, through visualization of the unfiltered data, we are able to identify difference in expression patterns of non-DE genes between two biological pathways.

Limitations & **Future Work** A limitation of our method is the difficulty of scaling our feature selection method to data of higher dimensions. The first concern is the heavy computational burden required for RCV. In our implementation, although we used parallel computing to speed up computation as much as possible, the actual running times for 3- and 4-dimensional subset are not quite satisfactory (Table 12), mainly due to the large number of possible subsets. However, in practice, the 2-subset results are usually more interpretable and visually appealing than its higher dimensional counterparts. Therefore, we recommend doing only 2-dim feature subset selection for exploratory purposes.

Table 12: Average running time for 10-fold cross-validation for all feature subsets, averaged over 50 runs with different random number seeds.

Subset dim.	**# of subsets**	**Avg. runtime (s)**
2	253	65.04
3	1771	512.61
4	8855	2241.43

Another reason is that the scatterplot matrix becomes less informative when the number of displayed dimensions exceeds 4. Even in our study, scatterplot matrices of dimensions 3 and 4 cannot fully reflect geometric properties of the data. For 3-d and 4-d angles, the scatterplot matrix only shows projections to all axial dimensions, which doesn't precisely convey the amount of separation between two classes, computed using all 3 or 4 dimensions. It is difficult to visualize the geometric and topological differences by only looking at individual panels of scatterplots. To more effectively visualize higher dimensional feature subsets, we can consider using interactive visualization tools, such as GGobi Swayne et al. (2003) and R Shiny Chang et al. (2017). Both tools allow users to identify the same point in all panels of a scatterplot matrix, significantly increasing its visual expressiveness.

Error rate definition In our definition of TPM in (1), we made the assumption that the cost of misclassifying an object from either class is the same, i.e. $c(1|2) = c(2|1)$. We can adjust the cost values if we are more concerned about correctly classifying a certain class of observations.

Evaluating reproducibility of experiments Currently, a typical differential expression analysis is conducted in a gene-wise manner, i.e. genes are treated as observations and the treatment conditions as features. In our study, we took the same approach because our goal was to differentiate expression pattern between two groups of genes. However, with the increase in the availability of RNA-Seq data thanks to advances in information technology, we can also study the comparability and reproducibility of RNA-Seq experiments. In this sense, we will be exploring the relationship between treatment conditions or experiments, with genes acting as features/variables. Evaluation of experiment reproducibility is usually accomplished by performing the same experiment using the same setting, which is, unfortunately, not a common practice in RNA-Seq studies. In light of this, one of our long-term goal is the quantification of similarity between RNA-Seq experiments, which not only accounts for differences in experimental designs and parameter settings, but also utilize the information hidden in the expression of genes.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge Jeff Chang, Sarah Emerson and Duo Jiang for their valuable insight and comments, and Bin Zhuo for providing the data.

REFERENCES

Carlos L. Ballaré. Jasmonate-induced defenses: a tale of intelligence, collaborators and rascals. *Trends in Plant Science*, 16(5), 2010.

Tanya Z. Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait, and Eva Huala. The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome. *genesis*, 2015. doi: 10.1002/dvg.22877.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.0.3.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Yong Ding, Michael Fromm, and Zoya Avramova. Multiple exposures to drought train transcriptional responses in *Arabidopsis. Nature Communications*, 2012. doi: 10.1038/ncomms1732.

Yong Ding, Ning Liu, Laetitia Virlouvet, Jean-Jack Riethoven, Michael Fromm, and Zoya Avramova. Four distinct types of dehydration stress memory genes in *Arabidopsis thaliana. BMC Plant Biology*, 13(229):1–38, 2013.

Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1141–1148, 2008.

Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53(11):3735–3745, 2009.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2):1137–1145, 1995.

Zhefeng Lin, Silin Zhong, and Don Grierson. Recent advances in ethylene research. *Journal of Experimental Botany*, 60(12):3311–3336, 2009.

Guillaume Moissiard, Sylvain Bischof, Dylan Husmann, William A. Pastor, Christopher J. Hale, Linda Yen, Hume Stroud, Ashot Papikian, Ajay A. Vashisht, James A. Wohlschlegel, and Steven E. Jacobsen. Transcriptional gene silencing by *Arabidopsis* microrchidia homologues involves the formation of heteromers. *Proceedings of the National Academy of Sciences*, 111(20):7474–7479, 2014.

David M. W. Powers and Adham Atyabi. The problem of cross-validation: Averaging and bias, repetition and significance. *Engineering and Technology (S-CET)*, May 2012.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6, 1978.

Hume Stroud, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson, Dinshao J. Patel, and Steven E. Jacobsen. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nature Structural & Molecular Biology*, 21(1): 64–72, 2014.

Deborah F. Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis - Data Visualization*, 43:423–444, 2003.

Bin Zhuo, Sarah Emerson, Jeffrey H. Chang, and Yanming Di. Identifying stably expressed genes from multiple RNA-Seq data sets. *PeerJ*, 4(e2791), 2016.

A. CONSTRUCTION OF COVARIANCE ELLIPSES FOR NORMAL COMPONENTS

In this section, we introduce how the covariance ellipses are constructed by MclustDA when a scatterplot or a scatterplot matrix is graphed.

For 2D data, suppose the mean and covariance estimates for component k of class j are $\hat{\mu}_{jk}$ and $\hat{\Sigma}_{jk}$, respectively. Also suppose that $\hat{\Sigma}_{jk}$ has eigenvalues $\lambda_1 \geqslant \lambda_2$ and their corresponding eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. Then MclustDA computes the major and minor axes of the ellipse centered at $\hat{\mu}_{jk}$ the following way:

$$\text{major axis} = \hat{\mu}_{jk} \pm \sqrt{\lambda_1}\boldsymbol{e}_1, \quad \text{minor axis} = \hat{\mu}_{jk} \pm \sqrt{\lambda_2}\boldsymbol{e}_2,$$

and the resulting ellipse has coverage probability of approximately 0.393.

In the case of higher dimensional data, MclustDA constructs the scatterplot and graphs the ellipses two dimensions at a time. Suppose $\hat{\mu}_{jk}$ and $\hat{\Sigma}_{jk}$ are defined the same way as above, and consider data dimensions p and q for visualization via scatterplot. Let $\Sigma^{(p,q)} = [\hat{\Sigma}_{jk}]_{(p,q)}$ be the covariance submatrix corresponding to the two dimensions, and $\mu^{(p,q)} = [\hat{\mu}_{jk}]_{(p,q)}$ be the corresponding mean vector. Now, suppose $\Sigma^{(p,q)}$ has eigenvalue/eigenvector pairs $\{\lambda_1^{(p,q)}, \boldsymbol{e}_1^{(p,q)}\}$ and $\{\lambda_2^{(p,q)}, \boldsymbol{e}_2^{(p,q)}\}$ with $\lambda_1^{(p,q)} \geqslant \lambda_2^{(p,q)}$. Then the ellipse plotted by MclustDA has major and minor axes as follows:

$$\text{major axis} = \mu^{(p,q)} \pm \sqrt{\lambda_1^{(p,q)}}\boldsymbol{e}_1^{(p,q)}, \quad \text{minor axis} = \mu^{(p,q)} \pm \sqrt{\lambda_2^{(p,q)}}\boldsymbol{e}_2^{(p,q)},$$

where the ellipse has the same coverage probability as the case above.

B. SCATTERPLOTS AND SCATTERPLOT MATRICES FOR SELECT TOP RANKED FEATURE COMBINATIONS

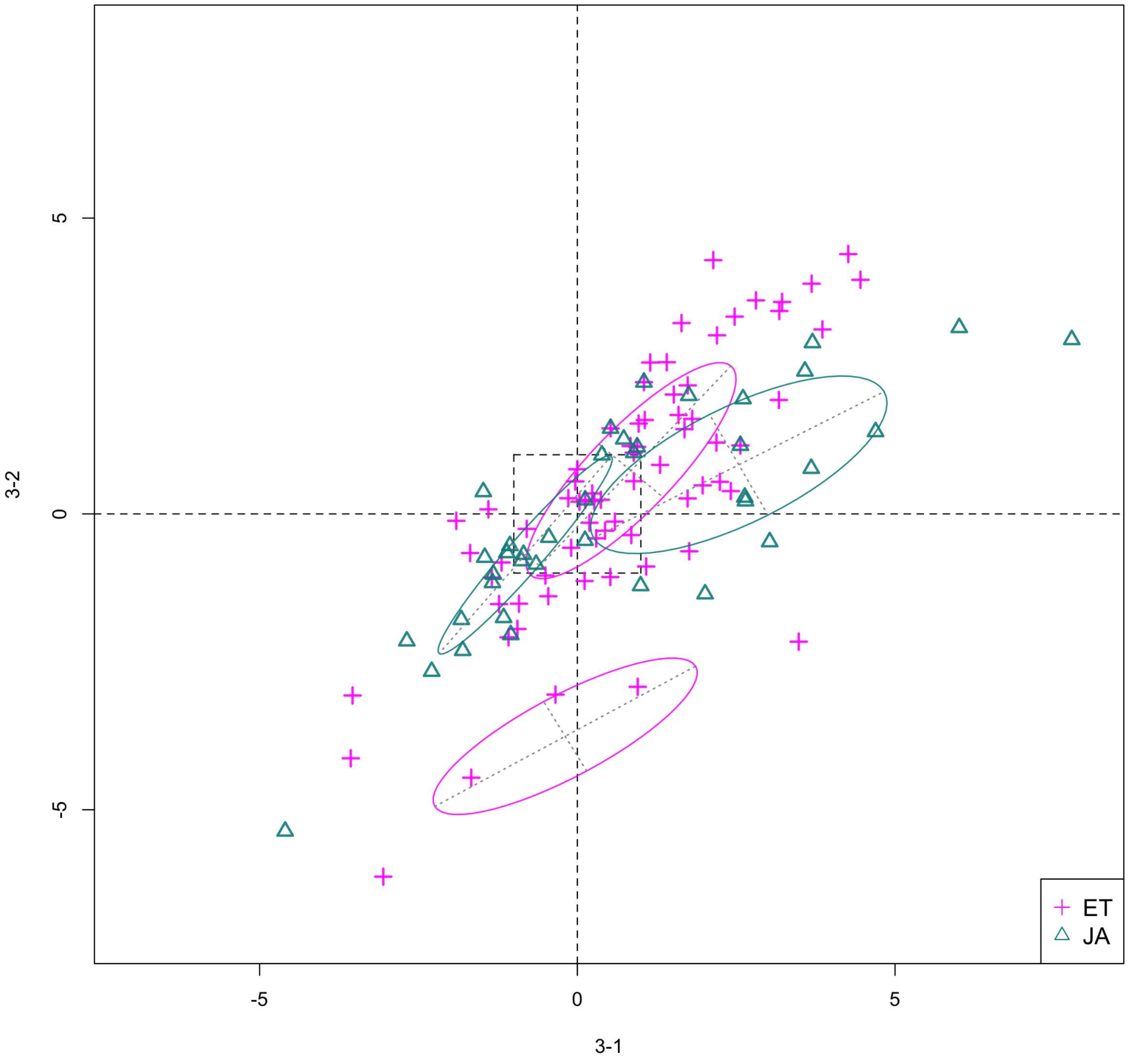


Figure 6: Scatterplot for 3-1 and 3-2

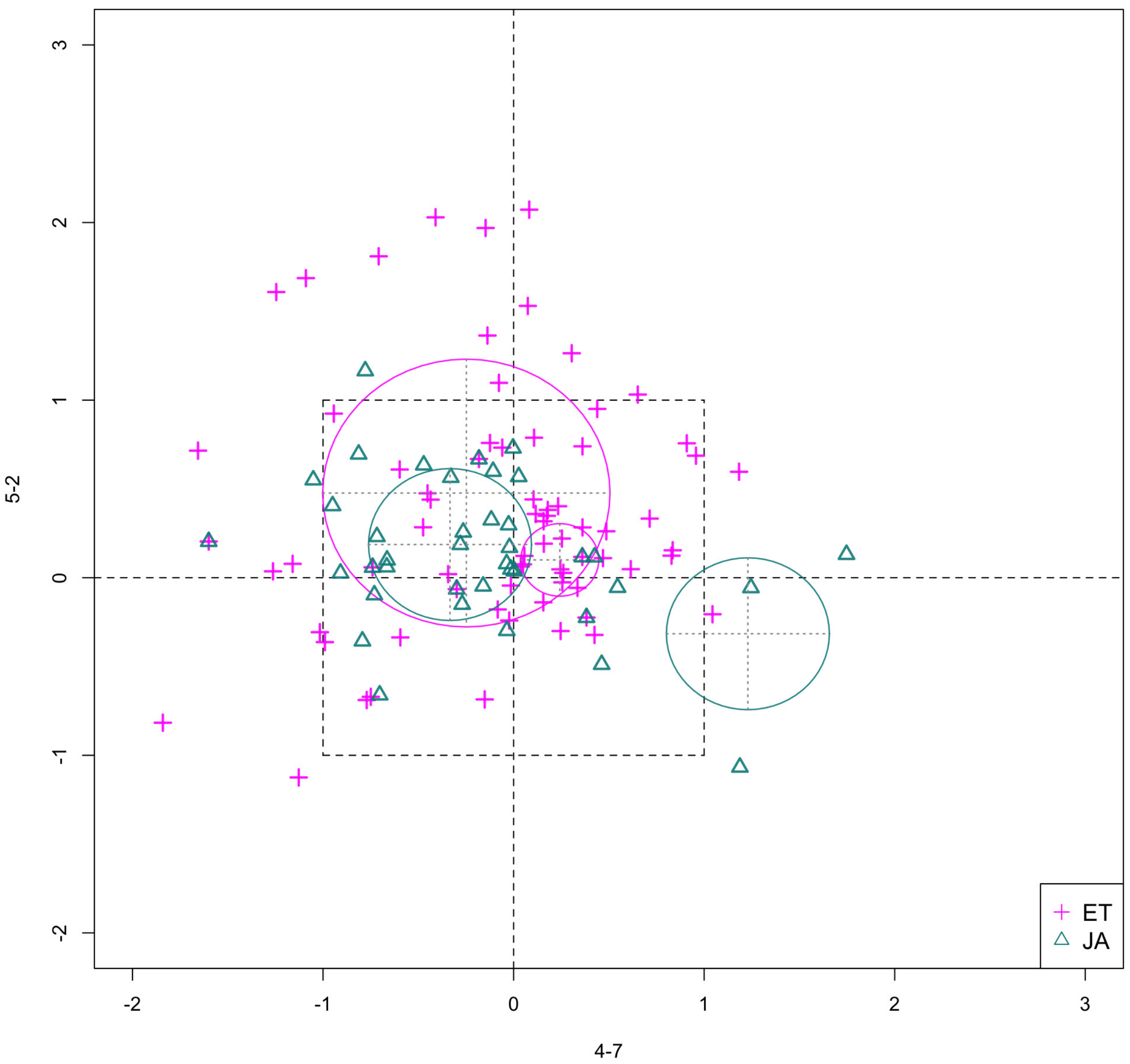


Figure 7: Scatterplot for 4-7 and 5-2

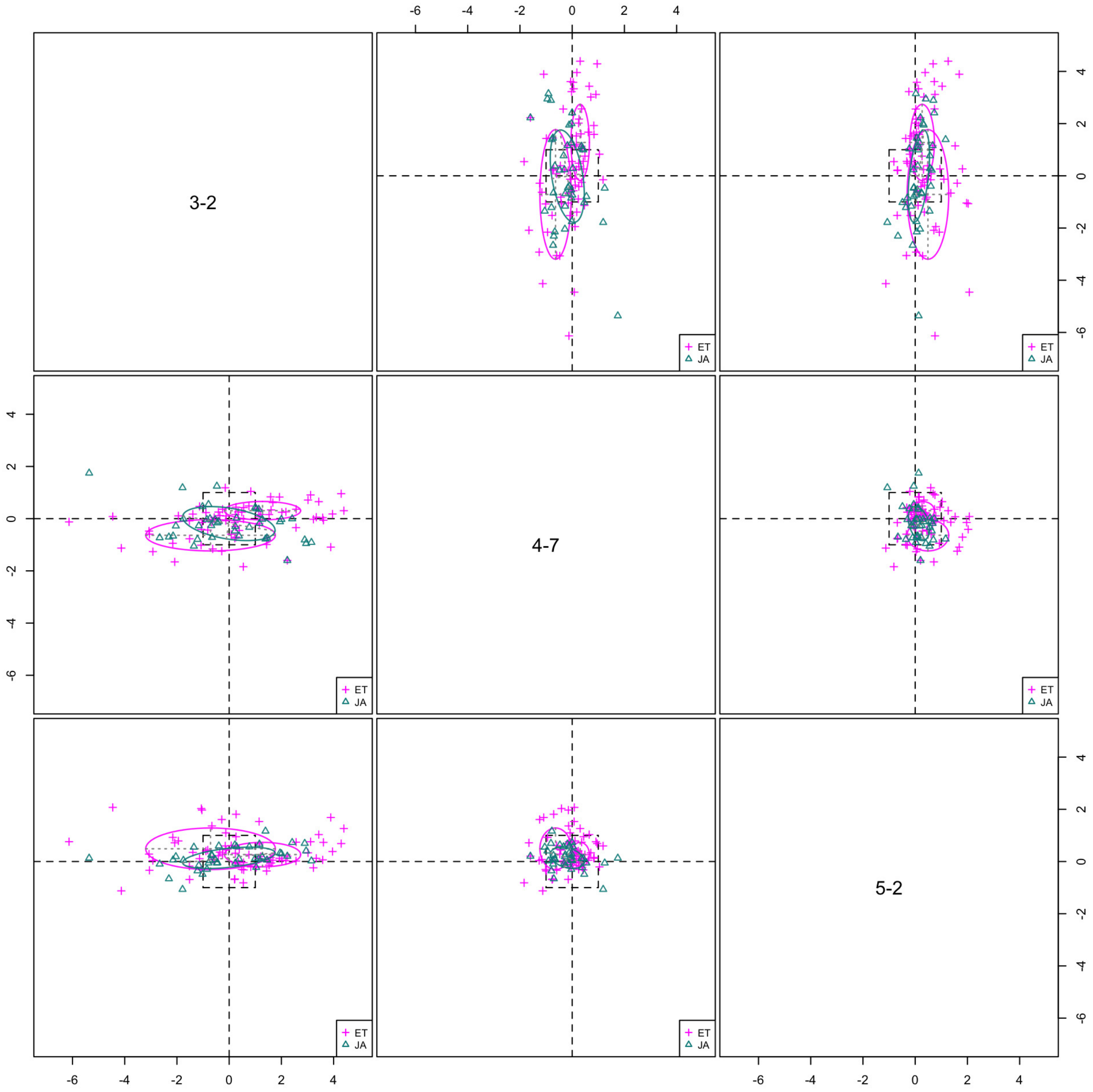


Figure 8: Scatterplot matrix for [3-2, 4-7, 5-2]

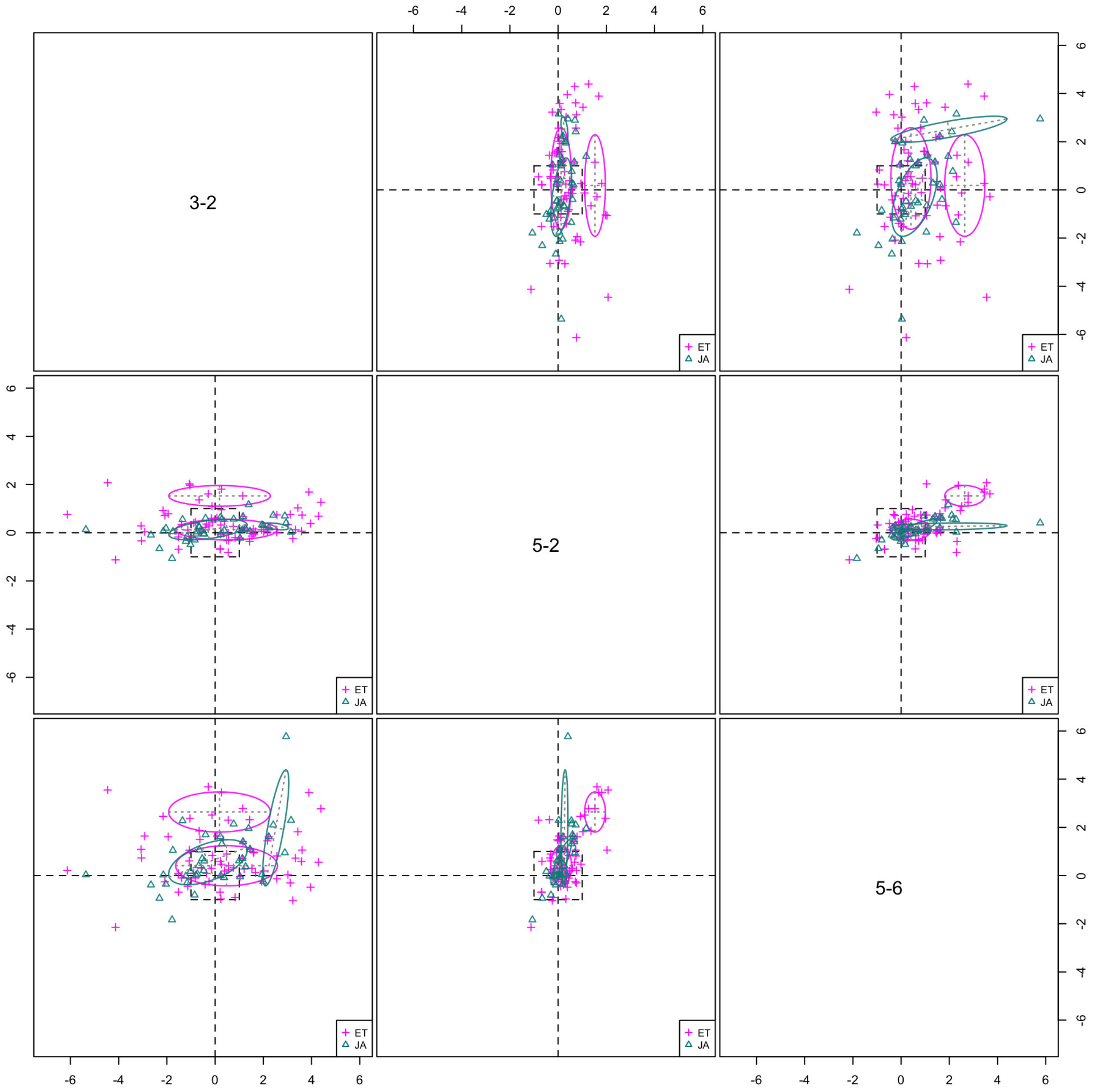


Figure 9: Scatterplot matrix for [3-2, 5-2, 5-6]

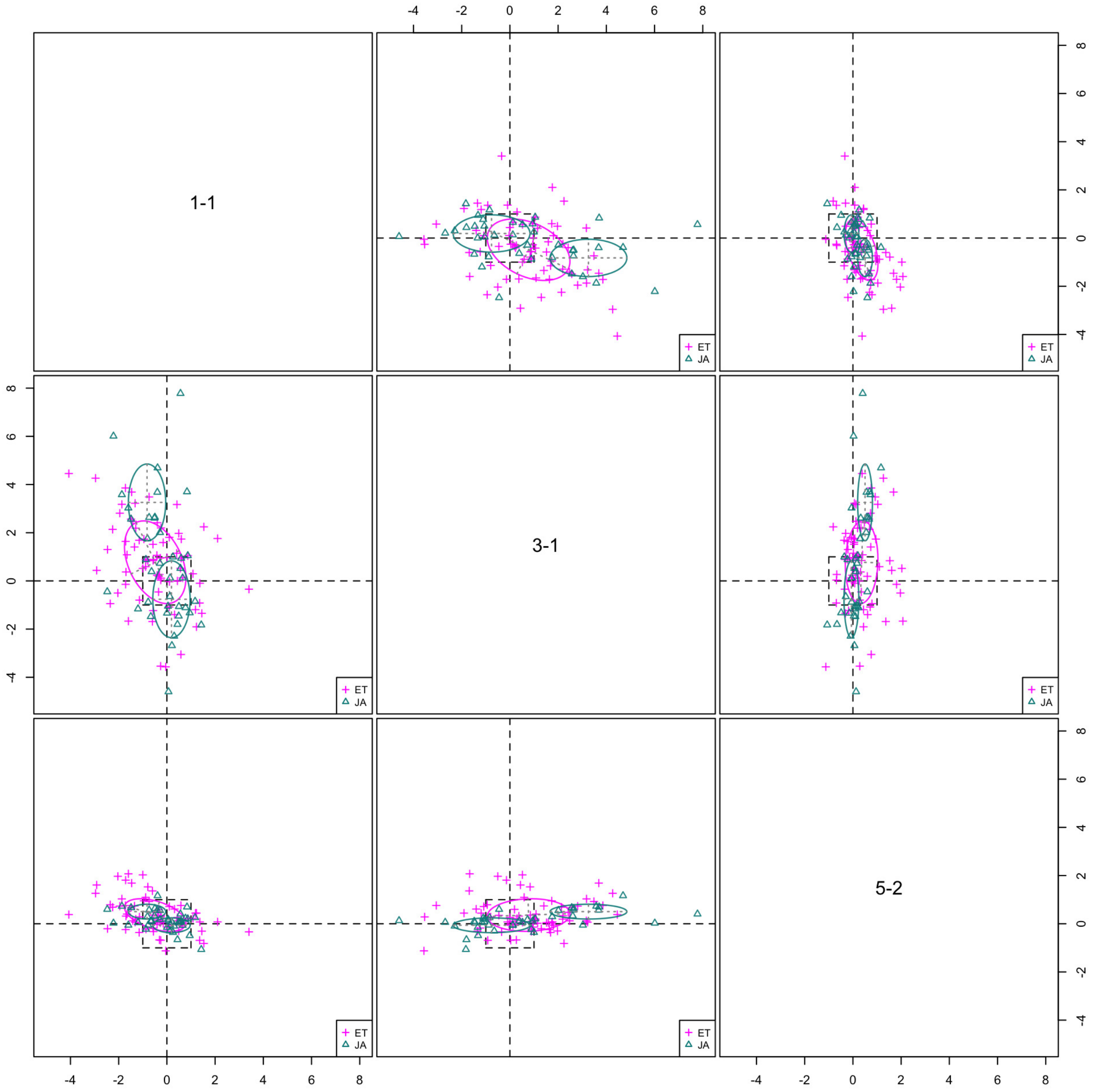


Figure 10: Scatterplot matrix for [1-1, 3-1, 5-2]

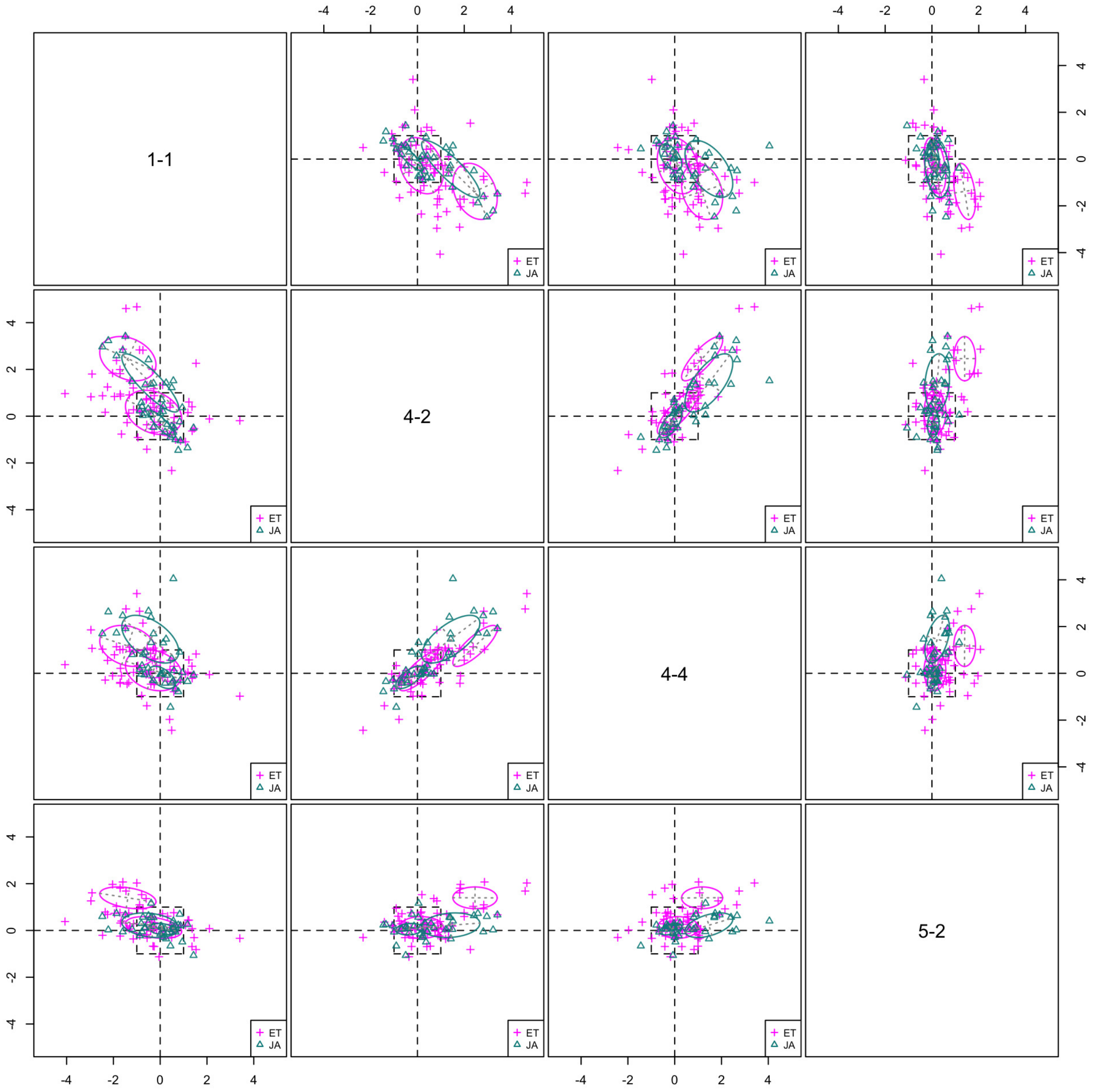


Figure 11: Scatterplot matrix for [1-1, 4-2, 4-4, 5-2]

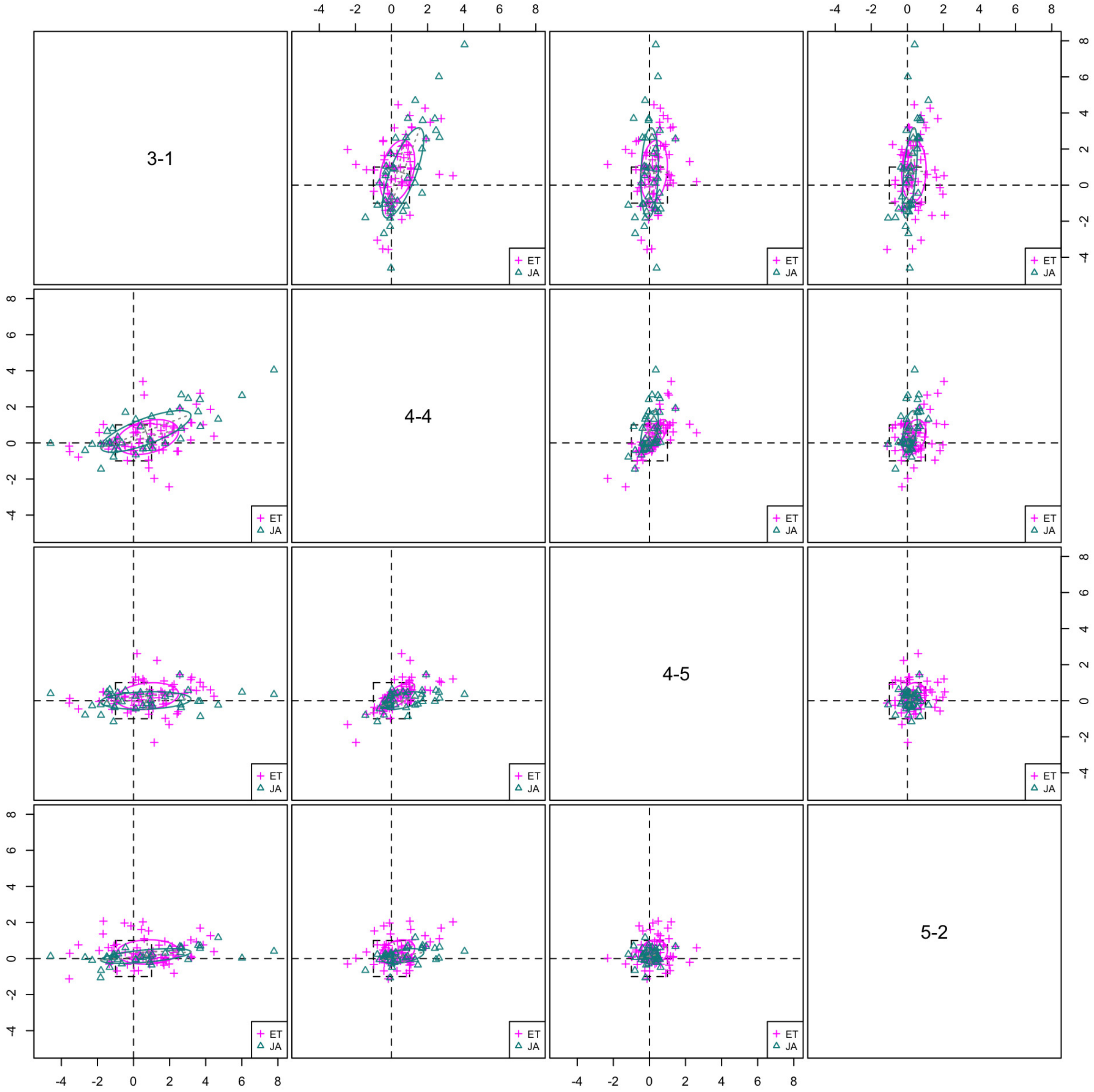


Figure 12: Scatterplot matrix for [3-1, 4-4, 4-5, 5-2]

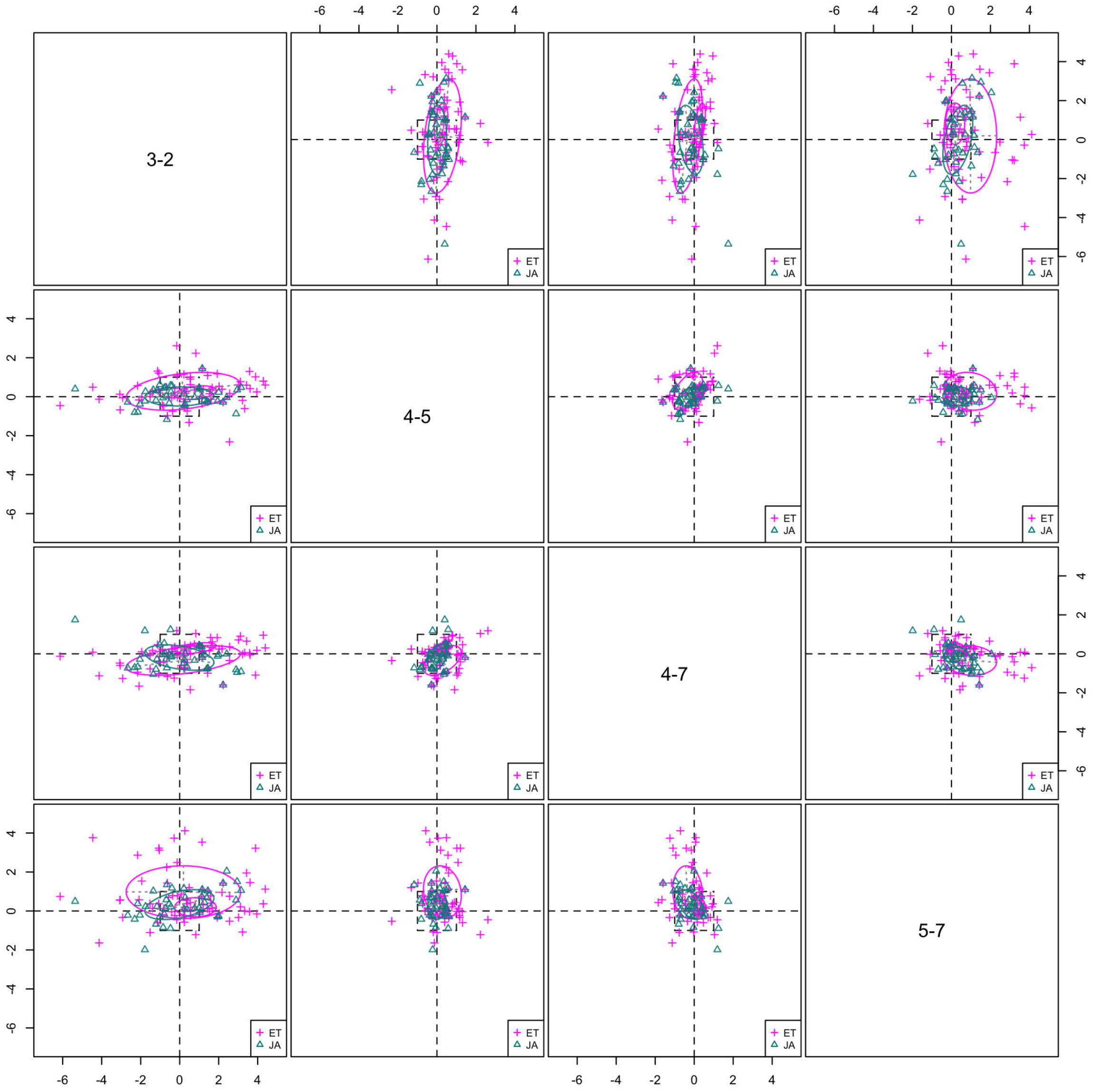


Figure 13: Scatterplot matrix for [3-2, 4-5, 4-7, 5-7]